

Academic Press is an imprint of Elsevier  
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK  
225 Wyman Street, Waltham, MA 02451, USA  
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2011

Copyright © 2011 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK:  
phone: (+44) (0) 1865 843830; fax: (+44) (0) 1865 853333;  
email: [permissions@elsevier.com](mailto:permissions@elsevier.com).

Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting, *Obtaining permission to use Elsevier material*.

#### Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

ISBN: 978-0-12-386485-7

ISSN: 1876-1623

For information on all Academic Press publications  
visit our website at [www.elsevierdirect.com](http://www.elsevierdirect.com)

Printed and bound in USA

11 12 13 14 10 9 8 7 6 5 4 3 2 1

Working together to grow  
libraries in developing countries

[www.elsevier.com](http://www.elsevier.com) | [www.bookaid.org](http://www.bookaid.org) | [www.sabre.org](http://www.sabre.org)

ELSEVIER

BOOK AID  
International

Sabre Foundation

# APPLICATION OF COMPUTATIONAL METHODS TO THE DESIGN OF FATTY ACID AMIDE HYDROLASE (FAAH) INHIBITORS BASED ON A CARBAMIC TEMPLATE STRUCTURE

By ALESSIO LODOLA, SILVIA RIVARA, AND MARCO MOR

Dipartimento Farmaceutico, Università degli Studi di Parma,  
Parco Area delle Scienze 27/A, Parma, Italy

I. Introduction .....	2
II. Ligand-Based Drug Design .....	5
III. Structure-Based Drug Design .....	11
A. QM/MM Mechanistic Modeling .....	12
B. LIE Calculations .....	15
IV. Recent Advances .....	21
References.....	22

## ABSTRACT

Computer-aided approaches are widely used in modern medicinal chemistry to improve the efficiency of the discovery phase. Fatty acid amide hydrolase (FAAH) is a key component of the endocannabinoid system and a potential drug target for several therapeutic applications. During the past decade, different chemical classes of inhibitors, with different mechanisms of action, had been developed. Among them, alkyl carbamic acid biphenyl-3-yl esters represent a prototypical class of active site-directed inhibitors, which allowed detailed pharmacological characterization of FAAH inhibition. Both ligand- and structure-based drug design approaches have been applied to rationalize structure–activity relationships and to drive the optimization of the inhibitory potency for this class of compounds.

In this chapter, we review our contribution to the discovery and optimization of therapeutically promising FAAH inhibitors, based on a carbamic template structure, which block FAAH in an irreversible manner exerting analgesic, anti-inflammatory and anxiolytic effects in animal models. The peculiar catalytic mechanism of FAAH, and the covalent interaction with carbamate-based inhibitors, prompted the application of different computer-aided tools, ranging from ligand-based approaches to docking

procedures and quantum mechanics/molecular mechanics (QM/MM) hybrid techniques. Latest advancements in the field are also reported.

## I. INTRODUCTION

Fatty acid amide hydrolase (FAAH) is a mammalian membrane protein responsible for the hydrolysis and inactivation of biologically active amides (Piomelli, 2003), including the endocannabinoid anandamide and agonists of the peroxisome proliferator-activated receptors, such as oleylethanolamide and palmitoylethanolamide (Muccioli, 2010).

The catalytic mechanism of FAAH is unique among mammalian enzymes in that it involves a catalytic triad consisting of two serine residues (Ser217 and Ser241) and one lysine residue (Lys142), rather than the more common serine–histidine–aspartate triad found in classical serine hydrolases (McKinney and Cravatt, 2005). It has been proposed that Lys142 might serve as a key acid and base in distinct steps of the catalytic cycle (Fig. 1). As a base, it would activate the Ser241 nucleophile for attack on the substrate carbonyl. As an acid, Lys142 would protonate the substrate leaving group, leading to its expulsion. The effect of Lys142 on Ser241 nucleophile strength and on leaving group protonation occurs indirectly, via the bridging Ser217 of the triad which acts as a “proton shuttle” (Lodola et al., 2005; McKinney and Cravatt, 2005).

Genetic or pharmacological inactivation of FAAH enzyme leads to analgesic, anti-inflammatory, anxiolytic, and antidepressant effects in animal models (Bambico et al., 2009), without producing the undesirable side

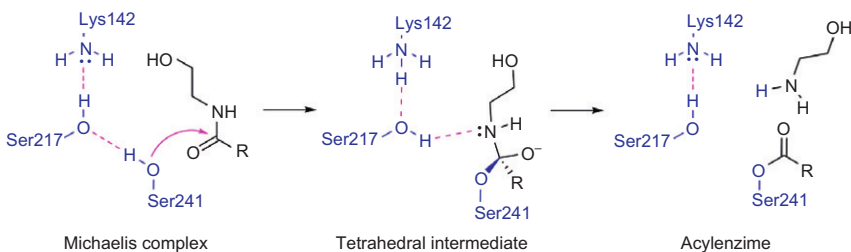


FIG. 1. Proposed catalytic mechanism of FAAH in presence of fatty acid ethanolamides. R represents the lipophilic chain of the substrate. Hydrogen bonds are displayed with pink dotted lines.

effects observed with cannabinoid receptor agonists (Piomelli, 2005). FAAH represents therefore an attractive therapeutic target for the treatment of several central nervous system disorders (Petrosino and Di Marzo, 2010).

FAAH enzyme activity is blocked by a variety of classical serine hydrolase inhibitors such as sulfonyl fluorides, fluorophosphonates,  $\alpha$ -ketoesters,  $\alpha$ -ketoamides, trifluoromethylketones, and acyl-heterocycles (Seierstad and Breitenbucher, 2008). Other classes of inhibitors, characterized by an improved drug-like profile, have also been reported (Minkkilä et al., 2010). These include piperazinyl-(pyridinyl)urea- and carbamate-based compounds (Mor and Lodola, 2009) which have been shown to inhibit FAAH by covalently modifying the enzyme's active site, that is, through carbamylation of the nucleophile Ser241 (Alexander and Cravatt, 2005; Ahn et al., 2007).

Among these carbamoylating agents, *N*-alkylcarbamic acid aryl esters emerged as the first promising class of compounds capable to inhibit FAAH *in vivo*, gaining considerable interest for the treatment of anxiety, inflammation, and pain (Kathuria et al., 2003; Piomelli et al., 2006; Sit et al., 2007). More recently, other classes of carbamate derivatives and related compounds (Gattinoni et al., 2010) have been developed by academic and industrial groups. For more detailed information, the reader is referred to reviews dedicated to FAAH inhibitors (Seierstad and Breitenbucher, 2008; Minkkilä et al., 2010).

The design of *N*-alkylcarbamic acid aryl esters as FAAH inhibitors has been widely supported by the application of computer-aided drug design (CADD) techniques (Marshall and Beusen, 2003). By definition, CADD uses computational methods to discover and improve biologically active compounds. This was also the case for FAAH, as both ligand-based drug design (LBDD) and structure-based drug design (SBDD) have been applied to rationalize structure–activity relationships (SARs), helping the design of novel FAAH inhibitors.

The LBDD approach is usually applied when structural information on the target macromolecule is missing (Marshall and Beusen, 2003). LBDD relies on the hypothesis that compounds with comparable physicochemical properties behave similarly in biological systems. Pharmacophore models as well as quantitative SARs (QSARs) can therefore be developed based on the analysis of known ligands. The QSAR approach is based on the search for a mathematical relationship between the biological activity of a series of compounds and their structural descriptors, usually encoding

a chemical or physicochemical information (e.g., lipophilicity, electronic properties, steric hindrance, etc.) (Hansch and Leo, 1995). Classical QSAR variables usually account for the magnitude of a structural property, but they do not provide information about their spatial distribution in the molecular surroundings (Selassie, 2003). Thanks to computer graphics, vector descriptors have been developed, allowing the rationalization of structure–activity data within a three-dimensional (3D) setting. The possibility to represent molecular properties in a 3D space is evocative of the supposed ligand–receptor interaction process and makes intuitive the meaning of the QSAR models (Favia, 2011). The most popular 3D-QSAR methodologies are comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) (Tropsha, 2003). These methods, correlating differences in biological activity with changes in shape and in the intensity of noncovalent interaction fields “around” (CoMFA) or “on” (CoMSIA) the molecules, have been successfully applied in numerous drug-discovery projects, both in retrospective analysis and in supporting the design of new compounds (Tropsha, 2003; Mor et al., 2005).

The SBDD approach is based on availability of the 3D structure of the biological target, usually obtained by X-ray crystallography or NMR studies (Hardy et al., 2003). If an experimental structure of the target is not available, homology models can be developed based on the experimental structure of a related protein (Fiser et al., 2002). Given the 3D-structure of the target, ligands can be (i) designed directly into the target binding site using interactive graphic tools (Marshall and Beusen, 2003) or (ii) built and placed within the binding site using a molecular docking approach (Kitchen et al., 2004). Molecular docking attempts to predict the preferred conformation and orientation of a compound into a specific cavity (i.e., the binding site) of the target molecule, assigning a “score” to all the identified binding modes (Kroemer, 2007). The reliability of a docking strategy mainly relies on the quality of the scoring function (Leach et al., 2006). In the past decades, several approaches have been developed to estimate the free energy of binding, with different levels of accuracy. The most rapid and less computationally demanding methods are the empirical or knowledge-based scoring approaches, which are based either on simple energy functions or on the frequency of occurrence of different atom–atom contact pairs in complexes of known structure (Klebe, 2006). The minimalism of the energy function together with the lack of conformational sampling make these approaches

extremely fast, but rather inaccurate (Michel and Essex, 2010). However, the most rigorous and accurate methods, which involve slow gradual transformations between the states of interest, by using molecular dynamics (MD) simulations, are extremely time-consuming (Deng and Roux, 2009). In this respect, computational approaches based on enhanced sampling methods (Branduardi et al., 2007; Colizzi et al., 2010; Woods et al., 2011) seem quite promising, as they have the potential to make accurate predictions at reasonable computational costs.

One of the most important aspects when trying to predict the binding mode of an active compound along with the potencies of a set of similar ligands is the time required for calculating their affinity. While screening of virtual libraries demands a high throughput of ligands, and thus the time spent on evaluating a single compound needs to be short, when the binding mode of a “lead” compound is relatively certain it may be desirable to perform time-consuming calculations, to improve the accuracy of the prediction (Jorgensen, 2009). In spite of the theoretical aspects behind the “scoring problem,” various lead identification (Villoutreix et al., 2009) and optimization (Andricopulo et al., 2009; Carmi et al., 2010; Solorzano et al., 2010) projects have been successfully carried out by applying SBDD techniques, indicating that theoretical approaches can give a practical and valuable contribution to the design of bioactive compounds.

This review focuses on the application of computational methods to the design and development of FAAH inhibitors belonging to the class of *N*-alkylcarbamic acid aryl esters. Early investigations, when the 3D structure of FAAH was still unknown, were based on LBDD techniques, including QSAR and 3D-QSAR methods, while more recent advancements were obtained applying SBDD approaches. These included (i) molecular docking, (ii) combined quantum mechanics/molecular mechanics (QM/MM) simulations, and (iii) linear interaction energy (LIE) calculations.

## II. LIGAND-BASED DRUG DESIGN

QSAR and 3D-QSAR methods have been successfully applied to the design of *N*-alkylcarbamic acid aryl esters as FAAH inhibitors (Tarzia et al., 2003; Mor et al., 2004; Minkkila et al., 2010), suggesting that for covalent ligands of similar reactivity, the recognition phase plays a pivotal role in explaining differences in the inhibitory potency (Tarzia et al., 2006).

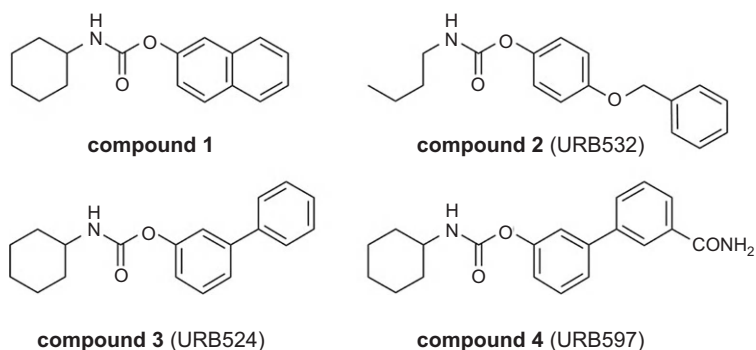


FIG. 2. Representative FAAH inhibitors synthesized during the discovery phase.

Carbamates **1** and **2** reported in Fig. 2 are representative of the most active compounds developed in the early phase of our FAAH project, having  $IC_{50}$  values of 324 and 396 nM, respectively. Analysis of their molecular structures allowed to get a first insight into the shape requirements for the aromatic substituent. Conformational analysis of the benzyloxyphenyl fragment of **2** revealed two families of accessible conformations, differing in the torsion angle around the O—CH<sub>2</sub> bond, with the two phenyl rings in *anti* or in *gauche* conformation (Tarzia et al., 2003). The *gauche* conformation of **2** more closely resembled the shape of the naphthyl derivative **1** when the compounds were superimposed via their common carbamate group (Fig. 3A). This led us to hypothesize that a bent shape of the carbamate *O*-substituent could favor enzyme inhibition, possibly by allowing a better steric complementarity between the inhibitor and the FAAH active site. To test this hypothesis, we conducted a systematic exploration of the steric requirements of the aromatic substituent by preparing a series of carbamate derivatives where the shape of the *O*-group was modified. Compounds with lipophilic *O*-substituents characterized either by a straight (e.g., 6-ethylnaphthalen-2-yl, (*E*)-4-styrylphenyl, biphenyl-4-yl) or by a bent shape (e.g., 8-bromonaphthalen-2-yl, (*Z*)-4-styrylphenyl, biphenyl-3-yl) were prepared. As a result, greater inhibitory potencies were obtained for those compounds characterized by a bent shape. In particular, we observed the strongest FAAH inhibition for the *m*-biphenyl derivative URB524 (compound **3**, Fig. 2), whose  $IC_{50}$  value (63 nM) indicates a 36-fold greater potency than the isomeric *p*-biphenyl derivative ( $IC_{50}$  = 2297 nM).

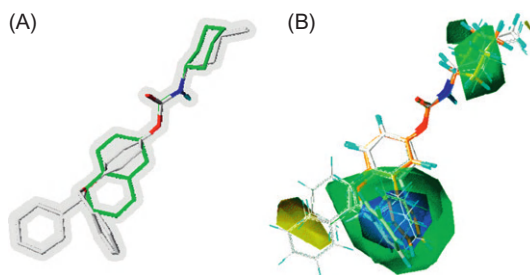


FIG. 3. (A) Superposition of compounds **1** (green carbon atoms) and **2** (white carbon atoms) in its *gauche* and *anti* conformations. (B) CoMSIA contour plots for a set of carbamate FAAH inhibitors. Compounds are represented with lines, with the exception of **2** (white carbons) and **3** (orange carbons) represented with capped sticks. The surfaces highlight regions of space where the influence of the steric potential on  $pIC_{50}$  is more significant. The color codes are: blue, very positive; green, positive; yellow, negative.

The comparison between 4-styrylphenyl isomers and between the differently substituted 2-naphthyl derivatives was suggestive of a similar trend. This prompted us to calculate a 3D-QSAR model, trying to correlate steric descriptors with inhibitory potency (Tarzia et al., 2003). The inhibitors were mutually superposed via their common carbamate group and a CoMSIA model was obtained, correlating inhibitor potency, expressed on a  $-\log$  scale ( $pIC_{50}$ ), with the molecular shape. A partial least squares (PLS) model with two latent variables provided good descriptive and predictive power ( $R^2=0.82$ ,  $s=0.32$ ,  $q^2_{LOO}=0.54$ ) for the 14-compound set of *O*-aryl *N*-alkylcarbamic acid esters (Tarzia et al., 2003). The coefficients of the steric field are depicted in Fig. 3B as isopotential surfaces. A large and deep favorable region was observed for the aryl substituent, as illustrated by the green and blue volumes at the bottom of Fig. 3B respectively, indicating the positive effect on inhibitory potency exerted by the presence of a substituent in this region of space. This region encompasses the second ring of the  $\beta$ -naphthyl substituent and the distal phenyl of the styryl substituent in its (*Z*)-configuration. It is reasonable to assume the proximity of this region to the binding site surface of FAAH, which would result in an improvement of steric interactions between the enzyme and the inhibitor. Thus, the *O*-aromatic moiety, which is hypothesized to serve as a leaving group in the reaction leading to enzyme



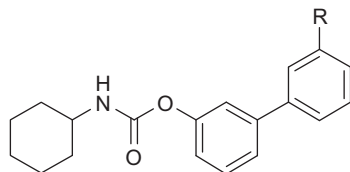
carbamoylation, would exert its effect on inhibitory potency at an early recognition stage of the process. A small region with moderately negative coefficients is represented by the yellow surface at the bottom left of Fig. 3B, opposite to the point of attachment of the phenyl *O*-substituent on the carbamate group. It indicates that straight substituents can be accommodated at the binding pocket in a less efficient manner than the folded ones. As mentioned earlier, the most relevant example is represented by the *p*-biphenyl derivative, whose potency is much lower than that of the *m*-biphenyl isomer. The CoMSIA coefficients suggest the existence of a large cavity with a curved shape in the active site of the enzyme, where suitable *O*-substituents can be accommodated, favoring the interaction of their carbonyl group with the active serine.

The most promising compound of this series, the biphenyl-3-yl derivative URB524 (compound **3**, Fig. 2), was selected as the lead structure for potency optimization. A two levels experimental design, based on positive and negative levels for lipophilicity ( $\pi$ ) and for an electronic descriptor ( $\sigma$ ), was performed, introducing four substituents (methyl, trifluoromethyl, amino, and carbamoyl) in *meta* and in *para* position of the distal phenyl ring (Mor et al., 2004). The 3'-methyl and 3'-amino derivatives resulted as potent as the parent compound (Table I), while the 3'-carbamoyl derivative (compound **4**, URB597, Fig. 2) was more potent than URB524. Substitution in the *para* position was not favorable, as all the *para*-derivatives were less active than URB524 (Mor et al., 2004). This limited exploration led to the identification of the best inhibitor of the carbamate series, the 3'-carbamoyl derivative URB597 endowed with an  $IC_{50}$  of 4 nM (Mor et al., 2004) which has become a standard reference in the field of FAAH inhibition.

The significant increase in potency of URB597, compared to the parent compound URB524, suggests that the 3'-carbamoyl group could undertake polar interactions at the binding site, supporting the idea that weak forces might have a pivotal role in controlling biological processes that involve the formation and break of covalent bonds.

To search for a statistical relationship between physicochemical properties and inhibitor potency, additional substituents were inserted at the 3' position of the biphenyl-3-yl group. These substituents were selected to introduce a balanced variation of their lipophilic, steric, and electronic properties. Analysis of the  $IC_{50}$  values shows that hydrophilic groups

TABLE I  
Inhibitory Potency ( $\text{pIC}_{50}$ ) on FAAH and Physicochemical Descriptors for a Series of Cyclohexylcarbamic Acid 3'-Substituted Biphenyl-3-yl esters



Compounds	R	$\text{pIC}_{50}$	$\pi^a$	$\text{MR}^b$	$\text{HB}^c$
3	—H	7.20	0.00	1.03	0
4	—C(O)NH <sub>2</sub>	8.34	-1.49	9.81	1
5	—CF <sub>3</sub>	6.84	0.88	5.02	0
6	—CH <sub>3</sub>	7.21	0.56	5.65	0
7	—NH <sub>2</sub>	7.19	-1.23	5.42	1
8	—F	7.02	0.14	0.92	0
9	—OC(O)NH $\epsilon$ -C <sub>6</sub> H <sub>11</sub>	6.44	1.06	36.13	1
10	—C <sub>6</sub> H <sub>5</sub> O	6.38	2.08	27.68	1
11	—C <sub>6</sub> H <sub>5</sub>	6.25	1.96	25.36	0
12	—CH <sub>2</sub> C <sub>6</sub> H <sub>5</sub>	5.73	2.01	30.01	0
13	— <i>n</i> -C <sub>3</sub> H <sub>7</sub>	6.96	1.55	14.96	0
14	—NO <sub>2</sub>	7.30	-0.28	7.36	1
15	—SO <sub>2</sub> NH <sub>2</sub>	7.58	-1.82	12.28	1
16	—C(O)CH <sub>3</sub>	8.04	-0.55	11.18	1
17	—CN	7.47	-0.57	6.33	1
18	—OH	8.06	-0.67	2.85	1
19	—CH <sub>2</sub> OH	8.06	-1.03	7.19	1
20	—(CH <sub>2</sub> ) <sub>2</sub> OH	7.73	-0.77	11.8	1

<sup>a</sup>Substituent lipophilicity.

<sup>b</sup>Molar refractivity.

<sup>c</sup>Hydrogen bonding capability.

(15–20, Table I) have a favorable effect on inhibitory activity. On the contrary, the introduction of large, lipophilic substituents (11–13) led to a drop in inhibitory activity. Several compounds in this set were more active than URB524, although none of them was better than the 3'-carbamoyl derivative URB597. A plot of  $\text{pIC}_{50}$  values versus  $\pi$  (Fig. 4)

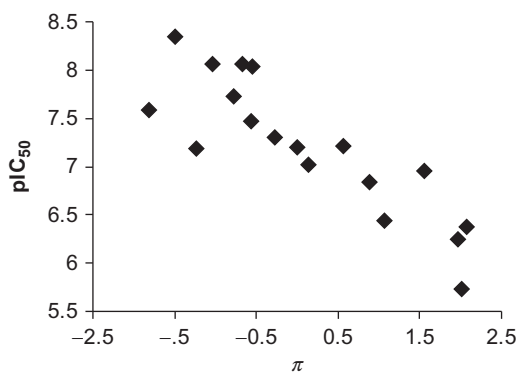


FIG. 4. Plot of FAAH inhibitory potency ( $\text{pIC}_{50}$ ) versus lipophilicity ( $\pi$ ) for compound **3** (URB524) and its *meta*-substituted derivatives (**4–20**).

shows a negative correlation between inhibitory activity and lipophilicity, also indicated by Eq. (1):

$$\text{pIC}_{50} = -0.49(\pm 0.07)\pi + 7.26(\pm 0.09) \quad (1)$$

$$n = 18 \quad r^2 = 0.74 \quad s = 0.37 \quad F = 46.0 \quad q^2 = 0.66 \quad \text{SDEP} = 0.40$$

The inclusion of an indicator variable, set to one for substituents able to undertake hydrogen bonds (HB) and to zero for lack of hydrogen bonding capability, in combination with MR provided an alternative model:

$$\text{pIC}_{50} = -0.046(\pm 0.009)\text{MR} + 0.80(\pm 0.18)\text{HB} + 7.29(\pm 0.17) \quad (2)$$

$$n = 18 \quad r^2 = 0.76 \quad s = 0.37 \quad F = 23.2 \quad q^2 = 0.67 \quad \text{SDEP} = 0.39$$

These QSAR models strongly suggest that the introduction of polar substituents at the *meta* position of the distal phenyl ring leads to a significant improvement of the  $\text{pIC}_{50}$  value, likely due to formation of polar interaction (i.e., H bonds) with hydrophilic amino acid residues within the FAAH channel.

## III. STRUCTURE-BASED DRUG DESIGN

The availability of the crystal structure of FAAH covalently bound to methyl arachidonyl phosphonate (Bracey et al., 2002) allowed us to look for a molecular rationalization of the QSAR models (reported in the previous section of this review) by performing docking simulations. Docking of URB597 within FAAH active site suggested two alternative binding orientations, both consistent with the observed SAR and with the carbamylation of the nucleophile Ser241 (Basso et al., 2004; Mor et al., 2004). In the first binding orientation (Fig. 5A), the *m*-biphenyl moiety of URB597 occupies the acyl chain binding (ACB) channel of FAAH, while in the second one (Fig. 5B), the cyclohexyl ring occupies the ACB channel and the *O*-aryl group is placed in the cytoplasmic access (CA) channel. In both orientations, residues able to undertake H bonds (Thr488 in orientation A; Gln273 in orientation B, see Fig. 4) could be found close to the 3'-position of URB597 biphenyl portion, accounting for both QSAR equations (1) and (2). To discriminate between these two binding modes, we modeled the mechanism of covalent adduct formation by URB524 in FAAH (Lodola et al., 2008) using a hybrid QM/MM approach (Mulholland, 2005), validated for FAAH catalysis (Lodola et al., 2005, 2009).

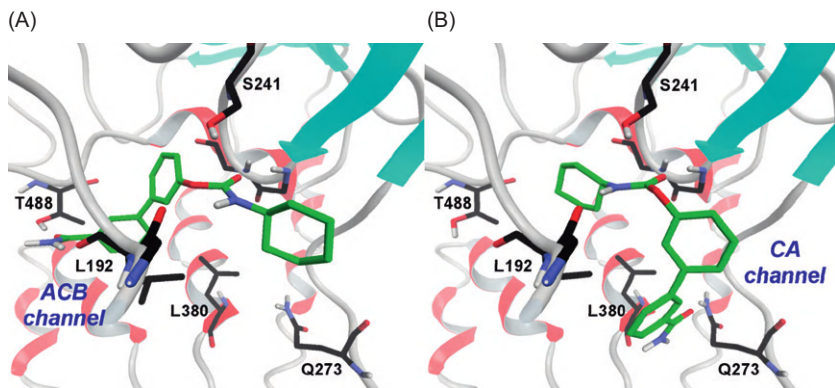


FIG. 5. Docking of URB597 into FAAH binding site, in two alternative orientations (A and B). Carbons of the inhibitor are colored in green, those of FAAH in black. The secondary structure of the enzyme is also displayed ( $\beta$  sheets are colored in cyan,  $\alpha$  helices in red, loops in gray).

At the same time, we prepared a new series of *N*-alkylcarbamic acid biphenyl-3-yl esters (Table 2). Starting from the lead compound URB524, steric and lipophilic requirements of the *N*-substituent for FAAH inhibition were explored, and the results were further analyzed applying molecular modeling techniques (Mor et al., 2008). The LIE method (Aqvist and Mareljus, 2001) was employed to estimate the binding affinity of the compounds docked in both orientations A and B. Correlative models based on LIE descriptors were built and compared.

### A. QM/MM Mechanistic Modeling

Application of hybrid QM/MM methods (Lonsdale et al., 2010) allows the simulation of enzyme-catalyzed reactions. In the QM/MM approach, the simulation system (i.e., the enzyme–substrate complex) is computationally separated into two subsets: the “core” that contains the reacting fragments and is described by a QM method (semiempirical, *ab initio*, or density-functional theory (DFT)), and the contiguous protein, represented by a classical force field (Senn and Thiel, 2009). With this approach, it is possible to treat systems composed by thousands of atoms and to describe the potential energy surfaces (PESs) relevant to enzymatic chemistry (Lonsdale et al., 2010) with an affordable computational effort.

In the case of FAAH, noncovalent complexes with URB524 were built according to orientations A and B (Fig. 5). These complexes were solvated and equilibrated by MD simulations. The geometry of the resulting FAAH-inhibitor structures was optimized applying a hybrid QM/MM potential and then used for mechanistic investigation. In the QM/MM modeling, the terminal methylamine fragment of Lys142 side chain, the side chains of Ser217 and Ser241, and the whole inhibitor were treated at the PM3 QM level, while the other atoms were treated with the CHARMM22 force field (MacKerell et al., 1998). The covalent bonds crossing the boundary between the QM and MM regions were treated by introducing three link atoms (Field et al., 1990), which are included in the QM subsystem (composed by 62 atoms in total). The adiabatic mapping approach (Lonsdale et al., 2010) was used to calculate PESs, generating models of the transition states (TSs) and intermediates along the carbamoylation pathway. To correct for possible shortcomings in the energetics due to the known limitations of the PM3, DFT energy corrections, at B3LYP/6-31+G(d) level, were also applied. The carbamoylating reaction of the

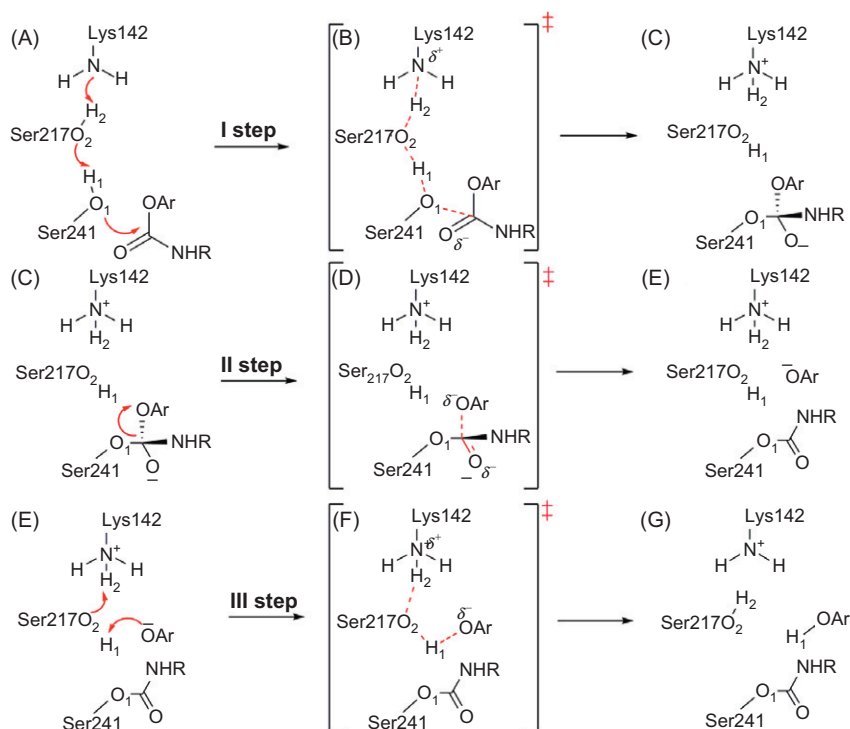


FIG. 6. Main steps of Ser241 carbamoylation in FAAH. Structures A–G are significant configurations along the reaction pathway.

nucleophile Ser241 was modeled in three main steps (Fig. 6): (i) formation of the tetrahedral intermediate (TI, C); (ii) expulsion of the *m*-biphenate with formation of *O*-carbamoylated Ser241 (E); and (iii) *m*-biphenate protonation and formation of neutral Lys142 (G). Appropriate reaction coordinates were applied (Lodola et al., 2008) to ensure the overall progress of the reaction. Energetics of Ser241 carbamoylation by URB524 in both orientations, at B3LYP/6-31+G(d)//PM3-CHARMM22, is reported in Fig. 7.

In orientation A, the first step of the carbamoylation reaction (activation of Ser241 followed by nucleophilic attack on the inhibitor carbonyl carbon forming the TI (C)) has an energy barrier of 35 kcal/mol. Although stabilized by the oxyanion hole, the TI is much less stable than the reactant

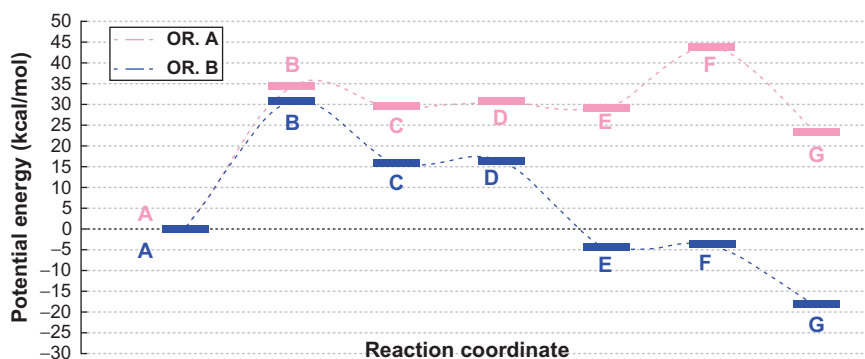


FIG. 7. B3LYP/6-31+G(d)//PM3-CHARMM22 energy profile for Ser 241 carbonylation by compound **3** (URB524), starting from binding orientations A (pink) and B (blue).

complex (by 29 kcal/mol). Expulsion of the *m*-biphenate anion produces **E** with a very low barrier, indicating that this event is effectively concerted with TI formation. During this process, the carbonyl carbon assumes a planar geometry, while the carbonyl oxygen maintains its interaction with the oxyanion hole. A double proton transfer (**E–G**) terminates the catalytic cycle. Protonation of the biphenate oxygen by Ser217 is concerted with proton transfer between Lys142 and Ser217 and represents the rate-limiting step of the whole process, with a barrier of 44 kcal/mol. Carbonylation occurs much more easily in orientation B. The barrier for the formation of the TI (**C**) (30 kcal/mol) is 5 kcal/mol lower than in orientation A. The TI (**C**) is a transient configuration and is greatly stabilized by the oxyanion hole (the energy of **C** is only 15 kcal/mol above the reactant). Breakdown of the tetrahedral intermediate takes place with a very low barrier, and so is effectively concerted with the first reaction step. Opposite to what observed for orientation A, the product of the reaction, **E**, is more stable than the starting structure **A** by 4 kcal/mol. This key difference arises from crucial interactions at the active site. Indeed, when the cyclohexyl ring is placed in the ACB channel (orientation B), it assumes a more favorable orientation, allowing the carbonyl oxygen of the carbamoylated Ser241 to better interact with the oxyanion hole. Moreover, the charged oxygen on the biphenate leaving group accepts a

short hydrogen bond from Ser217 H<sub>2</sub> and is also well positioned to “feel” the field effect of the positively charged Lys142 which at this stage of the reaction lives in its protonated form. This stabilization is weaker in orientation A, as the *m*-biphenate oxygen, residing in the ACB channel, remains further away from the catalytic triad than in orientation B.

The third step of carbamoylation (**E–G**) takes place without a significant energy barrier in orientation B, as protonation of the *m*-biphenate is favored by the proximity of Ser217, which is also well oriented to deprotonate Lys142. The resulting product **G** is very stable: it is the most stable configuration along the modeled pathway in orientation B (–18 kcal/mol), consistent with the experimentally observed irreversible inhibition of FAAH (Tarzia et al., 2003).

These calculations suggest that carbamoylation of Ser241 likely occurs starting from binding orientation B, as in orientation A the reaction has a significantly higher barrier, and leads to an unstable product. This prediction has been recently confirmed by the crystallographic structure of the FAAH-URB597 carbamoylated adduct (Mileni et al., 2010), suggesting that QM/MM-based mechanistic modeling can give a practical contribution in ongoing inhibitor design (De Vivo, 2011).

### B. *LIE Calculations*

In the case of covalent inhibitors, it is difficult to obtain an accurate estimation of the binding free energy to an enzyme target, as it depends not only from the stereoelectronic complementarity between the enzyme and the inhibitor, but also on the chemical reactivity of the inhibitor (Tarzia et al., 2006). Our investigation on the SAR of *N*-alkylcarbamic acid biphenyl-3-yl esters started from the approximation that, for compounds with similar reactivity, the inhibitory potency should be linearly related to the free energy of the recognition process. In this context, it should be possible to predict differences in pIC<sub>50</sub> for a series of inhibitors by simulating the enzyme–inhibitor recognition process (i.e., with molecular docking), and then estimating the binding affinity with a suitable and relatively accurate computational method.

The LIE approach is based on the linear response approximation, which estimates  $\Delta G$  of “noncovalent” binding of a small molecule to a target protein as a function of polar and nonpolar energy components, that are considered linearly related to electrostatic and Van der Waals interactions



between the ligand and its environment. The free energy of binding for the protein–ligand complex is calculated considering two states: the “free” ligand, in a solvent environment, and the ligand bound to the solvated protein.

The LIE method applied to FAAH inhibitors implements the formulation proposed by Jorgensen (Carlson and Jorgensen, 1995), where the Surface Generalized Born (SGB) continuum model is used for solvent representation (Ghosh et al., 1998). In the resulting SGB-LIE approach the free energy of binding is calculated as:

$$\Delta G_{\text{bind}} = \alpha(\langle U_{\text{vdw}}^{\text{b}} \rangle - \langle U_{\text{vdw}}^{\text{f}} \rangle) + \beta(\langle U_{\text{elec}}^{\text{b}} \rangle - \langle U_{\text{elec}}^{\text{f}} \rangle) + \gamma(\langle U_{\text{cav}}^{\text{b}} \rangle - \langle U_{\text{cav}}^{\text{f}} \rangle) \quad (3)$$

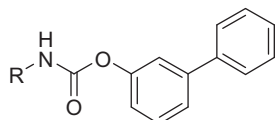
where b refers to bound state and f refers to the free state.



In Eq. (3),  $(\langle U_{\text{vdw}}^{\text{b}} \rangle - \langle U_{\text{vdw}}^{\text{f}} \rangle)$  estimates, by means of a Lennard–Jones potential, the variation of steric energy associated with ligand binding;  $(\langle U_{\text{elec}}^{\text{b}} \rangle - \langle U_{\text{elec}}^{\text{f}} \rangle)$  describes the change of electrostatic energy due to ligand desolvation and its accommodation into the protein binding site; the last term  $(\langle U_{\text{cav}}^{\text{b}} \rangle - \langle U_{\text{cav}}^{\text{f}} \rangle)$  accounts for the energy penalty due to the formation of a cavity within the solvent. The bracket notation indicates that an ensemble average of the energy terms should be taken into account for binding energy calculations. However, local sampling with energy minimization proved to be able to provide reasonable results in several cases, with limited or no reduction in the accuracy of  $\Delta G$  estimation, and this approach was applied also to our set of FAAH inhibitors.

In the SGB-LIE equation, Eq. (3), all the terms are evaluated for the interaction between ligand, both in the free and in the bound state, and its environment.  $\alpha$ ,  $\beta$ , and  $\gamma$  are free coefficients which were calculated by fitting the experimental free energies of binding for a training set of ligands with known protein affinity values. This empirical fitting can partially compensate the limits of the method, due to the neglect of conformational changes, intramolecular strain, and entropic effects.

The 22 *N*-alkylcarbamic acid biphenyl-3-yl esters, with different sizes, shapes and branching of the substituent on the nitrogen atom (Table II) were docked into the FAAH binding site. Two families of complexes, corresponding to orientations A and B, were generated, and SGB-LIE calculations were performed (Mor et al., 2008). The interaction energy terms, referring to van der Waals (vdw), electrostatic (elec), and cavity (cav) components, were calculated for free and bound inhibitors.

TABLE II  
 Inhibitory Potency ( $pIC_{50}$ ) on FAAH and SGB-LIE Components (Expressed in kcal/mol) in Binding Orientation B for a Series of *N*-Alkylcarbamic acid biphenyl-3-yl esters



Compounds	R	$pIC_{50}$	Orientation B		
			$\Delta U_{vdw}$	$\Delta U_{elec}$	$\Delta U_{cav}$
URB524		7.20	-46.75	2.91	-2.68
21		4.86	-34.97	1.43	-2.14
22		6.16	-38.04	2.13	-2.40
23		6.28	-42.66	3.92	-2.45
24	$H_3C-(CH_2)_3-$	6.95	-43.63	2.91	-2.58
25	$H_3C-(CH_2)_5-$	7.24	-42.28	2.39	-2.89
26	$H_3C-(CH_2)_7-$	7.28	-46.81	4.89	-3.18
27		7.27	-45.58	4.02	-2.50

(Continued)

TABLE II (Continued)

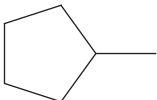
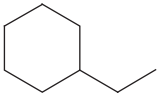
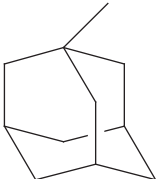
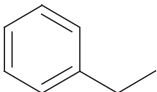
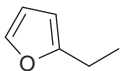
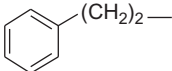
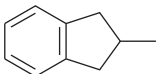
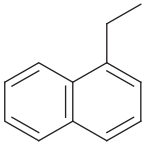
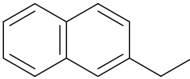
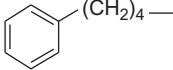
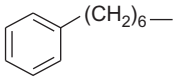
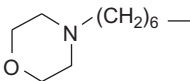
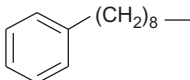
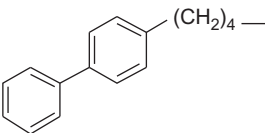
Compounds	R	pIC <sub>50</sub>	Orientation B		
			$\Delta U_{\text{vdw}}$	$\Delta U_{\text{elec}}$	$\Delta U_{\text{cav}}$
28		7.47	-45.32	4.93	-2.60
29		7.18	-43.90	2.12	-2.72
30		5.39	-42.36	7.45	-2.86
31		6.86	-47.30	9.45	-2.68
32		6.76	-45.99	3.82	-2.53
33		6.32	-52.43	12.96	-2.88
34		6.67	-52.13	9.83	-2.89

TABLE II (Continued)

Compounds	R	pIC <sub>50</sub>	Orientation B		
			$\Delta U_{\text{vdw}}$	$\Delta U_{\text{elec}}$	$\Delta U_{\text{cav}}$
35		7.23	-54.51	13.02	-2.99
36		8.27	-55.45	13.22	-3.01
37		8.03	-52.62	8.14	-3.21
38		7.89	-51.31	-0.17	-3.45
39		7.40	-50.89	5.21	-3.45
40		8.27	-60.87	9.82	-3.81
41		8.11	-60.85	10.73	-3.77

The difference between these energy values (bound minus free) was used to build LIE equations by multiple regression analysis (MRA). While for orientation A no significant model was found, an acceptable equation was obtained for orientation B, using the standard SGB-LIE approach. The resulting Eq. (4) explained 71% of pIC<sub>50</sub> variation and showed a good predictive power ( $q^2=0.61$ ).

$$\begin{aligned} \text{pIC}_{50} = & -0.187(\pm 0.046)\Delta U_{\text{vdw}} - 0.141(\pm 0.034)\Delta U_{\text{elec}} \\ & + 0.375(\pm 0.513)\Delta U_{\text{cav}} \end{aligned} \quad (4)$$

$$n = 22 \quad r^2 = 0.71 \quad s = 0.49 \quad F = 15.9 \quad q^2 = 0.61 \quad \text{SDEP} = 0.53$$

Internal correlation among  $X$  variables ( $r_{\Delta U_{\text{vdw}}, \Delta u_{\text{elec}}} = -0.71$ ;  $r_{\Delta U_{\text{vdw}}, \Delta u_{\text{cav}}} = 0.84$ ;  $r_{\Delta U_{\text{elec}}, \Delta u_{\text{cav}}} = -0.40$ ) affects the uncertainty for the cavity term coefficient ( $0.375 \pm 0.513$ ), suggesting that  $\Delta U_{\text{cav}}$  itself is a negligible term. The model indicated that vdw interactions give the most significant contribution (i.e., with the largest coefficient/standard error ratio) to binding energy. vdw energy is strongly related to the closeness of ligand and enzyme surfaces. Electrostatic interactions also showed a significant effect: because chemical modulation in this set of compounds mainly addressed size and shape, this result can be a consequence of the complementarity between inhibitors and the binding site. In fact, the carbamic group of all these inhibitors may form several hydrogen bonds (e.g., with oxyanion hole residues and with Met191 backbone carbonyl, see Fig. 8), and a high steric complementarity allows a more efficient electrostatic interaction. However, interpretation of the  $\Delta U_{\text{elec}}$  term is complicated by the fact that it also includes the contribution of the SGB solvent reaction-field energy. On the basis of these results, SGB-LIE values fairly reproduce the SAR profile for the carbamate inhibitors reported in Table II only when these compounds are placed within the FAAH active site in binding orientation B. It is nice to observe that a similar conclusion emerged from QM/MM mechanistic simulations.

The reliability of these theoretical models was tested by introducing at the 3'-position of the biphenyl nucleus of one of the most potent inhibitors of the series, compound **36**, a substituent able to form hydrogen bonds, that is, the carbamoyl group, also present in URB597. The significant gain in pIC<sub>50</sub> (from 8.27 to 9.20) displayed by URB880 (Fig. 8)

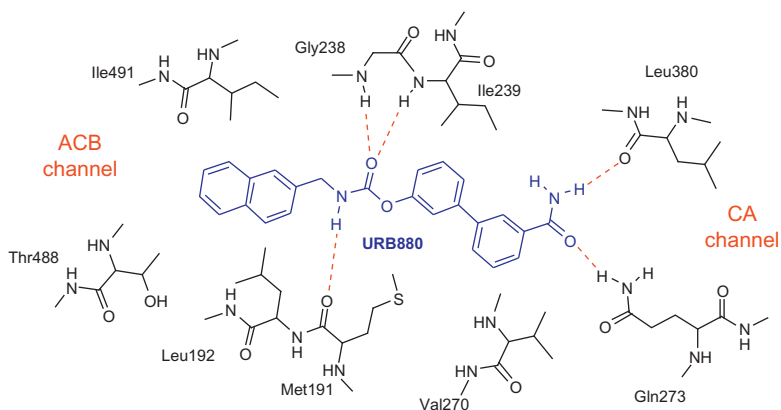


FIG. 8. Two-dimensional representation of URB880 in binding orientation B. Hydrogen bonds between the inhibitor and the enzyme are indicated with red-dotted lines.

confirmed that concurrent positioning of the lipophilic *N*-alkyl group within the ACB pocket and of the biphenyl moiety within the more hydrophilic CA cavity favors inhibitory potency (Mor et al., 2008).

#### IV. RECENT ADVANCES

Despite their relatively long history, FAAH inhibitors characterized by a carbamic structure are still a hot topic both in the field of computational medicinal chemistry and in pharmacology. Brief highlights of the most recent developments in these fields are presented in this paragraph. It has been recently shown that *N*-alkylcarbamic acid biphenyl-3-yl esters are a highly versatile class of covalent inhibitors, as their intrinsic reactivity can be easily tuned by chemical manipulation (Vacondio et al., 2009). In fact, it is possible to enhance their chemical (and metabolic) stability by simply introducing electron-donor substituents in conjugated positions of the proximal phenyl ring. This increases the electron density around the carbonyl carbon, limiting its reactivity toward nucleophiles. However, while the introduction of electron-donor groups (e.g., *p*-OH or *p*-NH<sub>2</sub>) significantly improves the stability of these carbamates versus nucleophiles,

including those present in liver and plasmatic carboxylesterases (Clapper et al., 2009), the same substitution does not affect the interaction with FAAH. This unexpected lack of correlation between reactivity and FAAH inhibitory potency might be due to the “unique” catalytic mechanism of FAAH. QM/MM mechanistic modeling of FAAH carbamoylation in presence of the cyclohexylcarbamic acid biphenyl-3-yl ester URB524 and its *p*-OH (URB694) and *p*-NH<sub>2</sub> (URB618) analogues showed that FAAH is insensitive to the intrinsic reactivity of the carbamate group, as the crucial TS of the reaction is dominated by a proton transfer and not by a nucleophilic attack (Lodola et al., 2011). This finding could help in the development of a new generation of “stabilized” carbamate inhibitors that, while retaining good *in vitro* potency for FAAH, would display longer half-life in plasma, making them significantly more potent *in vivo*, and more selective versus off-target carboxylesterases, than current inhibitors.

In this scenario, novel FAAH inhibitors with an unprecedentedly seen pharmacokinetic profile have been recently reported (Clapper et al., 2010). These new compounds markedly differed in their ability to access the central nervous system from the first generation of carbamic-based FAAH inhibitor. Among them, the *p*-hydroxyl derivative of URB597, namely URB937, suppressed FAAH activity in peripheral tissues of mice and rats but failed to affect FAAH activity in the brain. Despite the inability to access brain and spinal cord, URB937 attenuated behavioral responses indicative of persistent pain in rodent models of peripheral nerve injury and inflammation. These findings indicate that brain-impenetrant FAAH inhibitors might offer a new therapeutic option for pain treatment.

## REFERENCES

- Ahn, K., Johnson, D. S., Fitzgerald, L. R., Liimatta, M., Arendse, A., Stevenson, T., et al. (2007). Novel mechanistic class of fatty acid amide hydrolase inhibitors with remarkable selectivity. *Biochemistry* **46**, 13019–13030.
- Alexander, J. P., Cravatt, B. F. (2005). Mechanism of carbamate inactivation of FAAH: implications for the design of covalent inhibitors and *in vivo* functional probes for enzymes. *Chem. Biol.* **12**, 1179–1187.
- Andricopulo, A. D., Salum, L. B., Abraham, D. J. (2009). Structure-based drug design strategies in medicinal chemistry. *Curr. Top. Med. Chem.* **9**, 771–790.
- Aqvist, J., Marelius, J. (2001). The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High. Throughput Screen.* **4**, 613–626.

- Bambico, F. R., Duranti, A., Tontini, A., Tarzia, G., Gobbi, G. (2009). Endocannabinoids in the treatment of mood disorders: evidence from animal models. *Curr. Pharm. Des.* **15**, 1623–1646.
- Basso, E., Duranti, A., Mor, M., Piomelli, D., Tontini, A., Tarzia, G., et al. (2004). Tandem mass spectrometric data-FAAH inhibitory activity relationships of some carbamic acid *O*-aryl esters. *J. Mass Spectrom.* **39**, 1450–1455.
- Bracey, M. H., Hanson, M. A., Masuda, K. R., Stevens, R. C., Cravatt, B. F. (2002). Structural adaptation in a membrane enzyme that terminates endocannabinoid signaling. *Science* **298**, 1793–1796.
- Branduardi, D., Gervasio, F. L., Parrinello, M. (2007). From A to B in free energy space. *J. Chem. Phys.* **126**, 054103.
- Carlson, H. A., Jorgensen, W. L. (1995). An extended linear response method for determining free energies of hydration. *J. Phys. Chem.* **99**, 10667–10673.
- Carmi, C., Cavazzoni, A., Vezzosi, S., Bordi, F., Vacondio, F., Silva, C., et al. (2010). Novel irreversible epidermal growth factor receptor inhibitors by chemical modulation of the cysteine-trap portion. *J. Med. Chem.* **53**, 2038–2050.
- Clapper, J. R., Vacondio, F., King, A. R., Duranti, A., Tontini, A., Silva, C., et al. (2009). A second generation of carbamate-based fatty acid amide hydrolase inhibitors with improved activity in vivo. *ChemMedChem* **4**, 1505–1513.
- Clapper, J. R., Moreno-Sanz, G., Russo, R., Guijarro, A., Vacondio, F., Duranti, A., et al. (2010). Anandamide suppresses pain initiation through a peripheral endocannabinoid mechanism. *Nat. Neurosci.* **13**, 1265–1270.
- Colizzi, F., Perozzo, R., Scapozza, L., Recanatini, M., Cavalli, A. (2010). Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *J. Am. Chem. Soc.* **132**, 7361–7371.
- De Vivo, M. (2011). Bridging quantum mechanics and structure-based drug design. *Front. Biosci.* **16**, 1619–1633.
- Deng, Y., Roux, B. (2009). Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **113**, 2234–2346.
- Favia, A. D. (2011). Theoretical and computational approaches to ligand-based drug discovery. *Front. Biosci.* **16**, 1276–1290.
- Field, M. J., Bash, P. A., Karplus, M. (1990). A combined quantum-mechanical and molecular mechanical potential for molecular-dynamics simulations. *J. Comput. Chem.* **11**, 700–733.
- Fiser, A., Feig, M., Brooks, C. L., 3rd, Sali, A. (2002). Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* **35**, 413–421.
- Gattinoni, S., De Simone, C., Dallavalle, S., Fezza, F., Nannei, R., Amadio, D., et al. (2010). Enol carbamates as inhibitors of fatty acid amide hydrolase (FAAH) endowed with high selectivity for FAAH over the other targets of the endocannabinoid system. *ChemMedChem* **5**, 357–360.
- Ghosh, A., Sendrovc Rapp, C., Friesner, R. A. (1998). Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B* **102**, 10983–10990.
- Hansch, C., Leo, A. (1995). Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. American Chemical Society, Washington, DC.



- Hardy, L. W., Abraham, D. J., Safo, M. K. (2003). Structure based drug design. In: Burger's Medicinal Chemistry and Drug Discovery, Abraham, D. J. (Ed.), vol. 1, pp. 417–469. John Wiley & Sons, Hoboken.
- Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**, 724–733.
- Kathuria, S., Gaetani, S., Fegley, D., Valiño, F., Duranti, A., Tontini, A., et al. (2003). Modulation of anxiety through blockade of anandamide hydrolysis. *Nat. Med.* **9**, 76–81.
- Kitchen, D. B., Decornez, H., Furr, J. R., Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949.
- Klebe, G. (2006). Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **11**, 580–594.
- Kroemer, R. T. (2007). Structure-based drug design: docking and scoring. *Curr. Protein Pept. Sci.* **8**, 312–328.
- Leach, A. R., Shoichet, B. K., Peishoff, C. E. (2006). Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **49**, 5851–5855.
- Lodola, A., Mor, M., Hermann, J. C., Tarzia, G., Piomelli, D., Mulholland, A. J. (2005). QM/MM modelling of oleamide hydrolysis in fatty acid amide hydrolase (FAAH) reveals a new mechanism of nucleophile activation. *Chem. Commun.* 4399–4401.
- Lodola, A., Mor, M., Rivara, S., Christov, C., Tarzia, G., Piomelli, D., et al. (2008). Identification of productive inhibitor binding orientation in fatty acid amide hydrolase (FAAH) by QM/MM mechanistic modelling. *Chem. Commun.* 214–216.
- Lodola, A., Mor, M., Sirirak, J., Mulholland, A. J. (2009). Insights into the mechanism and inhibition of fatty acid amide hydrolase from quantum mechanics/molecular mechanics (QM/MM) modeling. *Biochem. Soc. Trans.* **37**, 363–367.
- Lodola, A., Capoferri, L., Rivara, S., Chudyk, E., Sirirak, J., Dyguda-Kazimierowicz, E., et al. (2011). Understanding the role of carbamate reactivity in fatty acid amide hydrolase inhibition by QM/MM mechanistic modelling. *Chem. Commun.* **47**, 2517–2519.
- Lonsdale, R., Ranaghan, K. E., Mulholland, A. J. (2010). Computational enzymology. *Chem. Commun.* **46**, 2354–2372.
- MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. A* **102**, 3586–3616.
- Marshall, G. R., Beusen, D. D. (2003). Molecular modelling in drug design. In: Burger's Medicinal Chemistry and Drug Discovery, Abraham, D. J. (Ed.), vol. 1, pp. 77–168. John Wiley & Sons, Hoboken.
- McKinney, M. K., Cravatt, B. F. (2005). Structure and function of fatty acid amide hydrolase. *Annu. Rev. Biochem.* **74**, 411–432.
- Michel, J., Essex, J. W. (2010). Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided. Mol. Des.* **24**, 639–658.

- Mileni, M., Kamtekar, S., Wood, D. C., Benson, T. E., Cravatt, B. F., Stevens, R. C. (2010). Crystal structure of fatty acid amide hydrolase bound to the carbamate inhibitor URB597: discovery of a deacylating water molecule and insight into enzyme inactivation. *J. Mol. Biol.* **400**, 743–754.
- Minkkilä, A., Saario, S., Nevalainen, T. (2010). Discovery and development of endocannabinoid-hydrolyzing enzyme inhibitors. *Curr. Top. Med. Chem.* **10**, 828–858.
- Mor, M., Lodola, A. (2009). Pharmacological tools in endocannabinoid neurobiology. *Curr. Top. Behav. Neurosci.* **1**, 87–110.
- Mor, M., Rivara, S., Lodola, A., Plazzi, P. V., Tarzia, G., Duranti, A., et al. (2004). Cyclohexylcarbamic acid 3'- or 4'-substituted biphenyl-3-yl esters as fatty acid amide hydrolase inhibitors: synthesis, quantitative structure-activity relationships, and molecular modelling studies. *J. Med. Chem.* **47**, 4998–5008.
- Mor, M., Rivara, S., Lodola, A., Lorenzi, S., Bordini, F., Plazzi, P. V., et al. (2005). Application of 3D-QSAR in the rational design of receptor ligands and enzyme inhibitors. *Chem. Biodivers.* **2**, 1438–1451.
- Mor, M., Lodola, A., Rivara, S., Vacondio, F., Duranti, A., Tontini, A., et al. (2008). Synthesis and structure-reactivity relationship of fatty acid amide hydrolase inhibitors: modulation at the N-portion of biphenyl-3-yl alkylcarbamates. *J. Med. Chem.* **51**, 3484–3498.
- Muccioli, G. G. (2010). Endocannabinoid biosynthesis and inactivation, from simple to complex. *Drug Discov. Today* **15**, 474–483.
- Mulholland, A. J. (2005). Modelling enzyme reaction mechanisms, specificity and catalysis. *Drug Discov. Today* **10**, 1393–1402.
- Petrosino, S., Di Marzo, V. (2010). FAAH and MAGL inhibitors: therapeutic opportunities from regulating endocannabinoid levels. *Curr. Opin. Invest. Drugs* **11**, 51–62.
- Piomelli, D. (2003). The molecular logic of endocannabinoid signalling. *Nat. Rev. Neurosci.* **4**, 873–884.
- Piomelli, D. (2005). The endocannabinoid system: a drug discovery perspective. *Curr. Opin. Invest. Drugs* **6**, 672–679.
- Piomelli, D., Tarzia, G., Duranti, A., Tontini, A., Mor, M., Compton, T. R., et al. (2006). Pharmacological profile of the selective FAAH inhibitor KDS-4103 (URB597). *CNS Drug Rev.* **12**, 21–38.
- Seierstad, M., Breitenbucher, J. G. (2008). Discovery and development of fatty acid amide hydrolase (FAAH) inhibitors. *J. Med. Chem.* **51**, 7327–7343.
- Selassie, C. D. (2003). History of quantitative structure-activity relationship. In: *Burger's Medicinal Chemistry and Drug Discovery*, Abraham, D. J. (Ed.), vol. 1, pp. 1–48. John Wiley & Sons, Hoboken.
- Senn, H. M., Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **48**, 1198–1229.
- Sit, S. Y., Conway, C., Bertekap, R., Xie, K., Bourin, C., Burris, K., et al. (2007). Novel inhibitors of fatty acid amide hydrolase. *Bioorg. Med. Chem. Lett.* **17**, 3287–3291.
- Solorzano, C., Antonietti, F., Duranti, A., Tontini, A., Rivara, S., Lodola, A., et al. (2010). Synthesis and structure-activity relationships of N-(2-oxo-3-oxetanyl)amides as N-acylethanolamine-hydrolyzing acid amidase inhibitors. *J. Med. Chem.* **53**, 5770–5781.

- Tarzia, G., Duranti, A., Tontini, A., Piersanti, G., Mor, M., Rivara, S., et al. (2003). Design, synthesis, and structure-activity relationship of alkylcarbamic acid aryl esters, a new class of fatty acid amide hydrolase inhibitors. *J. Med. Chem.* **46**, 2352–2360.
- Tarzia, G., Duranti, A., Gatti, G., Piersanti, G., Tontini, A., Rivara, S., et al. (2006). Synthesis and structure-activity relationships of FAAH inhibitors: cyclohexylcarbamic acid biphenyl esters with chemical modulation at the proximal phenyl ring. *ChemMedChem* **1**, 130–139.
- Tropsha, A. (2003). Recent trends in quantitative structure-activity relationship. In: *Burger's Medicinal Chemistry and Drug Discovery*, Abraham, D. J. (Ed.), vol. 1, pp. 49–76. John Wiley & Sons, Hoboken.
- Vacondio, F., Silva, C., A., Fioni, A., Rivara, S., Duranti, A., et al. (2009). Structure-property relationships of a class of carbamate-based fatty acid amide hydrolase (FAAH) inhibitors: chemical and biological stability. *ChemMedChem* **4**, 1495–1504.
- Villoutreix, B. O., Eudes, R., Miteva, M. A. (2009). Structure-based virtual ligand screening: recent success stories. *Comb. Chem. High Throughput Screen.* **12**, 1000–1016.
- Woods, C. J., Malaisree, M., Hannongbua, S., Mulholland, A. J. (2011). A water-swap reaction coordinate for the calculation of absolute protein-ligand binding free energies. *J. Chem. Phys.* **134**, 054114.

# RECENT THEORETICAL AND COMPUTATIONAL ADVANCES FOR MODELING PROTEIN–LIGAND BINDING AFFINITIES

By EMILIO GALLICCHIO AND RONALD M. LEVY

Department of Chemistry and Chemical Biology, BioMaPS Institute for Quantitative Biology,  
Rutgers University, Piscataway, New Jersey, USA

I. Introduction .....	28
II. Theory of Noncovalent Binding .....	30
A. Statistical Mechanics Formulation of Molecular Association Equilibria .....	30
B. Alchemical Formulation .....	32
C. Potential of Mean Force Formulation .....	34
D. Implicit Representation of the Solvent .....	35
E. Definition of the Bound State .....	40
F. Thermodynamic Decompositions .....	44
III. Computational Methods .....	53
A. Free Energy Estimators .....	54
B. Double Decoupling .....	59
C. Binding Energy Distribution Analysis Method .....	62
D. PMF Approach .....	64
E. Relative Binding Free Energies .....	65
F. RE Conformational Sampling .....	66
G. Mining Minima .....	67
H. MM/PBSA and MM/GBSA Approaches .....	69
I. Studies of Ligand and Receptor Reorganization .....	71
IV. Conclusions .....	73
References .....	74

## ABSTRACT

We review recent theoretical and algorithmic advances for the modeling of protein ligand binding free energies. We first describe a statistical mechanics theory of noncovalent association, with particular focus on deriving the fundamental formulas on which computational methods are based. The second part reviews the main computational models and algorithms in current use or development, pointing out the relations with each other and with the theory developed in the first part. Particular emphasis is given to the modeling of conformational reorganization and entropic effect. The methods reviewed are free energy perturbation, double

decoupling, the Binding Energy Distribution Analysis Method, the potential of mean force method, mining minima and MM/PBSA. These models have different features and limitations, and their ranges of applicability vary correspondingly. Yet their origins can all be traced back to a single fundamental theory.

## I. INTRODUCTION

Molecular recognition forms the basis for virtually all biological processes. Understanding the interactions between proteins and their ligands is key to rationalize molecular aspect of enzymatic processes and the mechanisms by which cellular systems integrate and respond to regulatory signals. From a medicinal perspective, there is great interest in the development of computer models capable of predicting accurately the strength of protein–ligand association (Jorgensen, 2004). Structure-based drug discovery models seek to predict receptor–ligand binding free energies from the known or presumed structure of the corresponding complex (Guvench and Mackerell, 2009; Mobley et al., 2010). Within this class of methods, docking and empirical scoring approaches (Brooijmans and Kuntz, 2003; McInnes, 2007), which are useful in virtual screening applications (Shoichet, 2004; Zhou et al., 2007), are now routinely employed in drug discovery programs. This review focuses on a class of computational methodologies based on the fundamental physical and chemical principles that govern molecular association equilibria (Gilson and Zhou, 2007; Shirts et al., 2007; Deng and Roux, 2009; Mobley and Dill, 2009; Chodera et al., 2011). Given a sufficiently accurate model of molecular interactions, these methods have the potential to incorporate greater detail and achieve sufficient accuracy to address aspects of drug development such as ligand optimization, and to address questions such as drug specificity and resistance.

Despite their potential, physics-based models of protein–ligand binding are not widely employed in academic and industrial research, and their effectiveness as predictive tools remains uncertain (Mobley and Dill, 2009; Mobley et al., 2010; Chodera et al., 2011). There are clearly many reasons that this is the case. Models of this kind are more computationally demanding than alternative empirical techniques and require expert training for setting them up properly. Early applications of physics-based models of binding, when molecular models, computer algorithms, and computer hardware technologies had not reached a sufficient level of maturity,

eventually yielded discouraging results, likely dissuading adoption by the current generation of researchers (Chipot and Pohorille, 2007).

In the past decade, however, a revival of the field has taken place with the development of better atomistic models and simulation algorithms, and more powerful computers. A new awareness of the limits of applicability of the technologies and the interplay between the various elements of the models have recently led to more trustworthy and realistic outcomes. As the models become more widely employed and these technical developments progress to produce more precise and reproducible results, it is also important to remain aware and deepen our understanding of the statistical mechanics theory of binding on which these models are based.

Thermodynamically, the strength of the association between a ligand molecule and its target receptor is measured by the standard free energy of binding. A statistical mechanics theory of molecular association equilibria exists which is nowadays well understood and widely accepted (Gilson et al., 1997). Various computational implementations of this theory have been proposed. Computational models cannot capture all of the complexities of molecular interactions, and all of them, implicitly or explicitly, apply approximations or simplifications. Knowledge of the relationships between the theory and its implementation helps to appreciate the meaning and limits of approximations. This knowledge can also serve as a guide in the design of more realistic computational models and can suggest approaches for the analysis of the results in ways that further our understanding of the binding process. It is only relatively recently found that subtle but potentially critical aspects of the theory have been fully appreciated and are being incorporated into computational models.

Theoretical accounts of the theory of binding are somewhat scattered in the current literature and the various descriptions are often tailored to specific numerical implementations and applications, making it often difficult to resolve commonalities. The purpose of this review is to partially fill this gap. The first part describes a statistical mechanics theory of noncovalent association, with particular focus on deriving the fundamental formulas on which computational methods are based. This section also introduces the thermodynamic quantities that often appear in the recent literature as well as their nomenclature. The second part reviews the main computational models and algorithms in current use or development, pointing out the relations with each other and with the theory developed in the first part.

## II. THEORY OF NONCOVALENT BINDING

### A. *Statistical Mechanics Formulation of Molecular Association Equilibria*

Consider an ideal solution of receptor molecules R and ligand molecules L in equilibrium with their complexes RL. The affinity between the two species can be expressed by the standard binding free energy  $\Delta G_b^\circ$  associated with the bimolecular reaction



given by

$$\Delta G_b^\circ = -kT \ln K_b, \quad (2)$$

where  $K_b$  is the dimensionless binding constant expressed as

$$K_b = \left[ \frac{[RL]/C^\circ}{([R]/C^\circ)([L]/C^\circ)} \right]_{\text{eq}}, \quad (3)$$

where [...] are concentrations,  $C^\circ$  is the standard state concentration (often set as 1 M or 1 molecule/1668 Å<sup>3</sup>), and the eq subscript states that all concentrations are evaluated at equilibrium. It should be noted that this quasi-chemical description of binding is based on the idea that the bound complex RL can be treated as a distinct chemical species. As further discussed below, this is a reasonable approach if the interaction between the ligand and the receptor is strong, yielding a thermodynamically stable complex. We make this implicit assumption in what follows, noting, however, that if the receptor–ligand interactions are weak and nonlocalized, it would be more appropriate to treat the receptor/ligand mixture as a nonideal solution of the components.

A statistical mechanics expression for the binding constant is available under these assumptions, which, when a generally small pressure–volume term is neglected, can be expressed as (Gilson et al., 1997)

$$K_b = \frac{C^\circ}{8\pi^2} \frac{Z_{N,RL} Z_N}{Z_{N,R} Z_{N,L}}, \quad (4)$$

where  $Z_N$  is the configurational partition function of the solvent bath composed of  $N$  molecules, and  $Z_{N,RL}$ ,  $Z_{N,R}$ , and  $Z_{N,L}$  are the configurational partition functions of the complex, receptor, and ligand, respectively, in solution. A critical aspect of this formulation is that each partition

function includes only the internal degrees of freedom of each species.<sup>1</sup> For example (to simplify notation here and elsewhere, we omit Jacobian factors for curvilinear coordinates)

$$Z_{N,L} = \int d\mathbf{x}_L d\mathbf{r}_s e^{-\beta U(\mathbf{x}_L, \mathbf{r}_s)} \quad (5)$$

is the configurational partition function of the ligand placed in an arbitrary position and orientation in solution integrated over the  $3n_L - 6$  internal degrees of freedom of the ligand  $\mathbf{x}_L$ , where  $n_L$  is the number of atoms of the ligand,  $\mathbf{r}_s$  denotes the degrees of freedom of the solvent, and  $U(\mathbf{x}_L, \mathbf{r}_s)$  is the potential energy of solvent + ligand system. The six external degrees of freedom of the ligand  $\zeta_L$  (three translations and three rotations) correspond to as many additional internal degrees of freedom of the complex specifying the position and orientation of the ligand relative to the receptor (Boresch et al., 2003). The configurational partition function of the complex is then written as

$$Z_{N,RL} = \int_{\text{bound}} d\mathbf{x}_R d\mathbf{x}_L d\zeta_L d\mathbf{r}_s e^{-\beta U(\mathbf{x}_R, \mathbf{x}_L, \zeta_L, \mathbf{r}_s)}, \quad (6)$$

where the integral runs over all conformations of the complex that are deemed bound, for example, those in which the ligand is within a specified binding site. A convenient choice is to use the the external coordinates of the ligand relative to the receptor to define this state (Gilson et al., 1997; Boresch et al., 2003). An indicator function  $I(\zeta_L)$  is introduced set to 1 for values of  $\zeta_L$  corresponding to positions and orientations of the ligand which are considered bound to the receptor and zero otherwise. Note that, in this formalism, the value of the binding constant depends on this arbitrary definition of the complex, raising the question of how to choose it appropriately. This is a more general issue which is further discussed below. The integral of  $I(\zeta_L)$  measures the extent of the defined bound state

$$\int d\zeta_L I(\zeta_L) = V_{\text{site}} \Omega_{\text{site}}, \quad (7)$$

where  $V_{\text{site}}$  is the integral over translational coordinates and  $\Omega_{\text{site}}$  the integral over the orientational coordinates.  $V_{\text{site}}$  represents the physical volume of the binding site, while  $\Omega_{\text{site}}$  measures the allowed range of orientations of the

<sup>1</sup>The separation of the overall translations is exact, while the separation of rotational degrees of freedom neglects vibrational-rotational couplings. The latter is generally a valid approximation at physiological temperature.



ligand in the complex. If  $I(\zeta_L)$  is independent of the orientational coordinates (such that is the definition of the complex is based only on the position of the ligand relative to the receptor), then  $\Omega_{\text{site}} = 8\pi^2$ .

### B. Alchemical Formulation

In order to make Eq. (4) amenable to computation, it is convenient to express it in terms of combinations of ensemble averages. To do so, we need to express ratios of partition functions in Eq. (4) such that numerators and denominators have the same number and types of degrees of freedom. This is achieved by multiplying and dividing Eq. (4) by Eq. (7) times the configurational partition function of the ligand in vacuum

$$Z_L = \int d\mathbf{x}_L e^{-\beta U(\mathbf{x}_L)}, \quad (8)$$

yielding the following equivalent expression for  $K_b$

$$K_b = \frac{V_{\text{site}} \Omega_{\text{site}}}{V^\circ} e^{-\beta(\Delta G_2 - \Delta G_1)}, \quad (9)$$

where  $V^\circ = 1/C^\circ$ . In Eq. (9),  $\Delta G_2$ , defined by

$$\begin{aligned} e^{-\beta \Delta G_2} &= \frac{\int d\mathbf{x}_R d\mathbf{x}_L d\zeta_L dr_s I(\zeta_L) e^{-\beta U(\mathbf{x}_R, r_s)} e^{-\beta U(\mathbf{x}_L)} e^{-\beta u(\mathbf{x}_L, \zeta_L, \mathbf{x}_R, r_s)}}{\int d\mathbf{x}_R d\mathbf{x}_L d\zeta_L dr_s I(\zeta_L) e^{-\beta U(\mathbf{x}_R, r_s)} e^{-\beta U(\mathbf{x}_L)}} \\ &= \left\langle e^{-\beta u(\mathbf{x}_L, \zeta_L, \mathbf{x}_R, r_s)} \right\rangle_{\text{R}_{\text{slv}} + \text{L}_{\text{gas}}}, \end{aligned} \quad (10)$$

is the free energy for establishing receptor–ligand and solvent–ligand interactions, while the ligand is in the receptor binding site (where  $I(\zeta_L)$  is nonzero). The quantity

$$u(\mathbf{x}_L, \zeta_L, \mathbf{x}_R, r_s) = U(\mathbf{x}_R, \mathbf{x}_R, \zeta_L, r_s) - U(\mathbf{x}_R, r_s) - U(\mathbf{x}_L) \quad (11)$$

is the *binding energy* between the ligand and the receptor plus solvent environment;  $U(\mathbf{x}_R, r_s)$  is the potential energy of the receptor–solvent system in absence of the ligand, and  $U(\mathbf{x}_L)$  is the internal potential energy of the ligand. Similarly,  $\Delta G_1$ , defined by

$$\begin{aligned} e^{-\beta \Delta G_1} &= \frac{\int d\mathbf{x}_L d\zeta_L dr_s I(\zeta_L) e^{-\beta U(r_s)} e^{-\beta U(\mathbf{x}_L)} e^{-\beta u(\mathbf{x}_L, \zeta_L, r_s)}}{\int d\mathbf{x}_L d\zeta_L dr_s I(\zeta_L) e^{-\beta U(r_s)} e^{-\beta U(\mathbf{x}_L)}} \\ &= \left\langle e^{-\beta u(\mathbf{x}_L, \zeta_L, r_s)} \right\rangle_{\text{slv} + \text{L}_{\text{gas}}}, \end{aligned} \quad (12)$$

is the free energy for establishing ligand–solvent interactions (the same as the *solvation free energy* of the ligand).

As specified in Eqs. (10) and (12), the free energy changes  $\Delta G_2$  and  $\Delta G_1$  are expressed as averages over the ensembles corresponding to, respectively, the free solvated receptor with the ligand in the gas phase ( $\mathbf{R}_{\text{solv}} + \mathbf{L}_{\text{gas}}$ ), and the pure solvent with the ligand in the gas phase ( $\text{solv} + \mathbf{L}_{\text{gas}}$ ). In either case, the ligand is located in the binding site, as specified by the indicator function  $I(\zeta_L)$ , but not interacting with the receptor and the solvent. We will therefore refer to these states as *decoupled*.<sup>2</sup>

By inserting Eq. (9) in Eq. (2), we finally obtain an expression for the standard binding free energy

$$\Delta G_b^\circ = \Delta G_t^\circ + \Delta G_r + \Delta G_2 - \Delta G_1, \quad (13)$$

where

$$\Delta G_r = -kT \ln \frac{\Omega_{\text{site}}}{8\pi^2} \quad (14)$$

is a free energy penalty ( $\Omega_{\text{site}}$  is smaller than  $8\pi^2$ ) for restricting the isotropic distribution of ligand orientations in solution to the those allowed in the complex, and

$$\Delta G_t^\circ = -kT \ln \frac{V_{\text{site}}}{V^\circ} \quad (15)$$

is the free energy for transferring the ligand from a solution at concentration  $C^\circ$  to a volume of size  $V_{\text{site}}$ . For later use, we define here the quantity  $\Delta G_r$ , as the concentration-independent component of the standard free energy of binding,

$$\Delta G_1 = \Delta G_2 - \Delta G_r, \quad (16)$$

which will be referred to as the *interaction free energy* of binding. As the other terms in Eq. (13) can be evaluated analytically, it is the computation of the interaction free energy which is the main goal of computer simulations of binding.

The alchemical thermodynamic path underlying Eq. (13) is illustrated in Fig. 1. The overall binding process (upper horizontal equilibrium) is

<sup>2</sup>However, note that integration over the external degrees of the freedom  $\zeta_L$  for the solvation free energy calculation (Eq. (12)) is unnecessary and has been explicitly indicated only for consistency with the thermodynamic cycle indicated below; both the solution and gas phases are homogeneous and isotropic, and therefore, integration over the translational and rotational degrees of freedom  $\zeta_L$  yields a canceling factor of  $V_{\text{site}}\Omega_{\text{site}}$  in both the numerator and the denominator of Eq. (12).

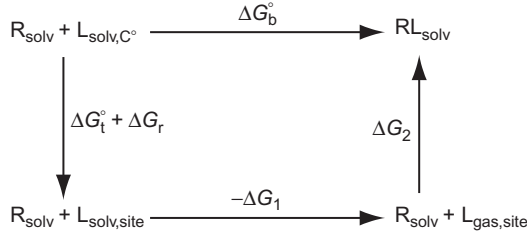


FIG. 1. Thermodynamic cycle illustrating the decomposition of the standard binding free energy [Eq. (13)].  $R_{\text{solv}}$  is the solvated receptor,  $L_{\text{solv},C^{\circ}}$  (upper left) is the ligand in solution at concentration  $C^{\circ}$ ,  $L_{\text{solv},\text{site}}$  (lower left) is the ligand solvated sequestered in the binding site,  $L_{\text{gas},\text{site}}$  (lower right) is the ligand in the gas phase in a volume equal to the binding site volume, and  $RL_{\text{solv}}$  is the solvated complex.

decomposed into a thermodynamic cycle with three distinct processes. The ligand is first transferred from the bulk solution at concentration  $C^{\circ}$  to a volume in the bulk solution identical to the binding site volume (left downward process) including any imposed orientational restraints. The free energy associated with this first step is  $\Delta G_{\text{t}}^{\circ} + \Delta G_{\text{r}}$ , given by Eqs. (15) and (14). In the second step (bottom horizontal process), the ligand is transferred from this volume in solution to an equivalent volume in the gas phase; as noted above, the free energy change for this step is the negative of the solvation free energy of the ligand. Finally (right upward process), the interactions of the ligand with the receptor and the solvent are turned on while the ligand is confined within the receptor binding site. This decomposition of the binding free energy forms the basis of the double-decoupling class (Deng and Roux, 2009; Mobley and Dill, 2009) of computational methods that will be discussed later in this review.

### C. Potential of Mean Force Formulation

An equivalent statistical mechanics formulation for the binding constant follows from the direct binding process corresponding to the upper horizontal process in Fig. 1. The binding constant effectively measures the probability of occurrence of configurations of the system in which the ligand is found within the binding site, that is conformations in which  $I(\zeta_{\text{L}})$  is nonzero, relative to the unbound conformations where  $I(\zeta_{\text{L}}) = 0$ . It should be therefore possible to compute the binding constant by means of a suitable

direct thermodynamic path connecting these two conformational states without resorting to intermediate gas phase thermodynamic states. To derive such a formalism, note that the product of partition functions in the numerator of Eq. (4) can be written as  $Z_{N,RL}Z_N = Z_{2N,RL}$ , where  $Z_{2N,RL}$  is the configurational partition function of the complex in a solution with twice as many solvent molecules. Similarly, the denominator can be written as  $Z_{2N,R+L}$ , the partition function of the unbound state when the receptor and the ligand are at infinite separation in a solution with  $2N$  solvent molecules. For sufficiently large  $N$  so that finite size effects are negligible, the ratio between  $Z_{2N,RL}$  and  $Z_{2N,R+L}$  is independent of  $N$  and can be written as  $Z_{N,RL}/Z_{N,R+L}$ . The expression for the binding constant then becomes

$$K_b = \frac{C^\circ}{8\pi^2} \frac{\int d\mathbf{x}_R d\mathbf{x}_L d\zeta_L dr_s I(\zeta_L) e^{-\beta U(\mathbf{x}_R, \mathbf{x}_L, \zeta_L, r_s)}}{\int d\mathbf{x}_R d\mathbf{x}_L d\mathbf{r}_s e^{-\beta U(\mathbf{x}_R, r_s)} e^{-\beta U(\mathbf{x}_L, \zeta_L^*, r_s)}}, \quad (17)$$

where  $\zeta_L^*$  specifies an arbitrary position of the ligand in the solvent bulk sufficiently removed from the receptor so that it does not interact with it. Equation (17) can be rewritten as (Jorgensen, 1989; Luo and Sharp, 2002)

$$K_b = \frac{C^\circ}{8\pi^2} \int d\zeta_L I(\zeta_L) e^{-\beta \Delta F(\zeta_L)}, \quad (18)$$

where  $\Delta F(\zeta_L)$  is the potential of mean force (PMF) along the  $\zeta_L$  coordinates, that is the free energy of the system when the position and orientation of the ligand are fixed at  $\zeta_L$  relative to the receptor. From Eq. (17), we see that  $\Delta F(\zeta_L)$  is defined as

$$e^{-\beta \Delta F(\zeta_L)} = \frac{\int d\mathbf{x}_R d\mathbf{x}_L d\mathbf{r}_s e^{-\beta U(\mathbf{x}_R, \mathbf{x}_L, \zeta_L, r_s)}}{\int d\mathbf{x}_R d\mathbf{x}_L d\mathbf{r}_s e^{-\beta U(\mathbf{x}_R, \mathbf{x}_L, \zeta_L^*, r_s)}}, \quad (19)$$

which explicitly sets to zero the PMF at  $\zeta_L^*$ . In practice, the binding PMF is computed along only one of the dimensions of  $\zeta_L$  (a receptor–ligand distance  $d$ , typically), while the other five coordinates are averaged or kept fixed (Woo and Roux, 2005; Lee and Olson, 2006).

#### D. *Implicit Representation of the Solvent*

More concise expressions for the binding constant are obtained by removing explicit integration over the solvent degrees of freedom by introducing the solvent PMF. Starting, for example, from Eq. (4), we

multiply and divide by  $Z_N^2$  and divide each partition function by  $Z_N$ . The solvent partition function yields a factor of 1. The  $Z_{N,R}/Z_N$  ratio can be expressed as

$$\frac{Z_{N,R}}{Z_N} = \frac{\int d\mathbf{x}_R d\mathbf{r}_s e^{-\beta U(\mathbf{x}_R)} e^{-u(\mathbf{x}_R, \mathbf{r}_s)} e^{-\beta U(\mathbf{r}_s)}}{\int d\mathbf{r}_s e^{-\beta U(\mathbf{r}_s)}} = \int d\mathbf{x}_R e^{-\beta U(\mathbf{x}_R)} e^{-\beta W(\mathbf{x}_R)}, \quad (20)$$

where  $U(\mathbf{x}_R)$  is the intramolecular potential energy of the receptor,  $u(\mathbf{x}_R, \mathbf{r}_s)$  denotes the receptor–solvent interaction energy,  $U(\mathbf{r}_s)$  is the solvent–solvent potential energy, and  $W(\mathbf{x}_R)$  is the solvent PMF for the  $\mathbf{x}_R$  conformation of the receptor defined by (Roux and Simonson, 1999)

$$e^{-\beta W(\mathbf{x}_R)} = \frac{\int d\mathbf{r}_s e^{-\beta u(\mathbf{x}_R, \mathbf{r}_s)} e^{-\beta U(\mathbf{r}_s)}}{\int d\mathbf{r}_s e^{-\beta U(\mathbf{r}_s)}} = \left\langle e^{-\beta u(\mathbf{x}_R, \mathbf{r}_s)} \right\rangle_{\text{solv.}} \quad (21)$$

Based on Eq. (21), the solvent PMF is interpreted as the solvation free energy of the receptor when this is fixed in conformation  $\mathbf{x}_R$ . The other ratios of partition functions can be treated similarly to define the solvent potentials of mean force,  $W(\mathbf{x}_L)$  and  $W(\mathbf{x}_R, \mathbf{x}_L, \zeta_L)$ , for the ligand and the complex. Finally, by a similar derivation that yielded Eq. (9), we can write (Gilson et al., 1997)

$$K_b = \frac{V_{\text{site}}}{V^\circ} \frac{\Omega_{\text{site}}}{8\pi^2} \frac{Z_{\text{RL}}}{Z_{\text{R+L}}} = \frac{V_{\text{site}}}{V^\circ} \frac{\Omega_{\text{site}}}{8\pi^2} e^{-\beta \Delta G_{\text{T}}}, \quad (22)$$

where  $Z_{\text{RL}}$  and  $Z_{\text{R+L}}$  are the configurational partition functions of the complex in the bound and uncoupled states, respectively, and the interaction free energy  $\Delta G_{\text{T}}$  is defined by their ratio as

$$\begin{aligned} e^{-\beta \Delta G_{\text{T}}} &= \frac{\int d\mathbf{x}_R d\mathbf{x}_L d\zeta_L I(\zeta_L) e^{-\beta[U(\mathbf{x}_R)+W(\mathbf{x}_R)]} e^{-\beta[U(\mathbf{x}_L)+W(\mathbf{x}_L)]} e^{-\beta u(\mathbf{x}_L, \zeta_L, \mathbf{x}_R)}}{\int d\mathbf{x}_R d\mathbf{x}_L d\zeta_L I(\zeta_L) e^{-\beta[U(\mathbf{x}_R)+W(\mathbf{x}_R)]} e^{-\beta[U(\mathbf{x}_L)+W(\mathbf{x}_L)]}} \\ &= \left\langle e^{-\beta u(\mathbf{x}_L, \zeta_L, \mathbf{x}_R)} \right\rangle_{\text{R+L}}, \end{aligned} \quad (23)$$

which is formally equivalent to Eq. (10) with potential energies  $U$  replaced by *effective* potential energies  $U_{\text{eff}} = U + W$ . The effective binding energy  $u$  in Eq. (23) has the same form as in Eq. (11) expressed in terms of differences of effective potential energies

$$u(\mathbf{x}_L, \zeta_L, \mathbf{x}_R) = U_{\text{eff}}(\mathbf{x}_R, \mathbf{x}_L, \zeta_L) - U_{\text{eff}}(\mathbf{x}_R) - U_{\text{eff}}(\mathbf{x}_L). \quad (24)$$

It is straightforward to show, from the definition of the solvent PMF (Eq. (21)), that the effective binding energy is the interaction free energy

with explicit solvation (Eq. (16)) for a fixed conformation ( $x_L$ ,  $\zeta_L$ ,  $x_R$ ) of the complex. Eq. (23) then expresses a combination rule to obtain the total interaction free energy for binding by averaging over the ensemble of the conformations of the uncoupled state of the complex.

Note that the meaning of the average  $\langle \rangle_{R+L}$  in Eq. (23) is different than in Eq. (10). In both averages, the ligand is sequestered in the binding site region; however, in Eq. (10), the ligand is considered as not interacting with either the receptor or the solvent, whereas in Eq. (23), the average is over the conformations of the receptor and the ligand while both of these interact with the solvent continuum in absence of the binding partner (note the absence of the binding energy term in the denominator of Eq. (23)). The standard binding free energy can then be written as

$$\Delta G_b^\circ = \Delta G_t^\circ + \Delta G_1 + \Delta G_2, \quad (25)$$

where  $\Delta G_t^\circ$  and  $\Delta G_1^\circ$  have the same meaning as in Eq. (13), and  $\Delta G_1$  is defined by Eq. (23). The PMF  $\Delta F(\zeta_L)$  in Eq. (19) can be similarly expressed in terms of the solvent PMF and the effective potential energy.

From a computational point of view, the most noticeable difference between the expression for the binding free energy in explicit solvent (Eq. (13)) and that in implicit solvent (Eq. (25)) is that the latter involves only one free energy calculation ( $\Delta G_1$ ), whereas the former is based on the difference between two free energy calculations (one for the transfer of the ligand in solution, yielding  $\Delta G_1$ , and another for its transfer to the complex,  $\Delta G_2$ ).

### 1. Connection with Potential Distribution Theory

A useful representation for the standard binding free energy  $\Delta G_b^\circ$  in the implicit solvent representation is obtained by writing the average  $\langle \exp(-\beta u) \rangle_{R+L}$  in Eq. (23) in terms of a probability distribution density of the effective binding energy (Gallicchio et al., 2010):

$$e^{-\beta \Delta G_1} = \langle \exp(-\beta u) \rangle_{R+L} = \int du p_0(u) e^{-\beta u}, \quad (26)$$

where  $p_0(u)$ , formally defined as

$$p_0(u) = \langle \delta[u(x_L, \zeta_L, x_R) - u] \rangle_{R+L}, \quad (27)$$

is the probability distribution for the effective binding energy over the ensemble of conformations in the uncoupled state (see above) that is the

state in which the ligand is in the binding site of the receptor, but both interact only with the solvent continuum. Note that, as discussed above, Eq. (26), although derived in the implicit solvent representation, is valid in general. In the explicit solvent representation,  $p_0(u)$  is interpreted as the distribution of binding free energies for fixed conformations of the complex drawn from the ensemble of conformations obtained when the ligand and the receptor are not interacting.

The larger the value of the integral in Eq. (26), the more favorable is the binding free energy. An example of a  $p_0(u)$  distribution is illustrated in Fig. 2. As further discussed in Section III.C, the magnitude of the  $p_0(u)$  distribution at positive, unfavorable, values of the binding energy  $u$  measures the entropic thermodynamic driving force which opposes binding, whereas the tail at negative, favorable, binding energies measures the energetic gain for binding due to the formation of ligand–receptor interactions. The interplay between these two opposing forces ultimately determines the strength of binding.

Equation (26) has the same form as the fundamental equation of the potential distribution theorem (PDT) (Widom, 1982; Beck et al., 2006), of

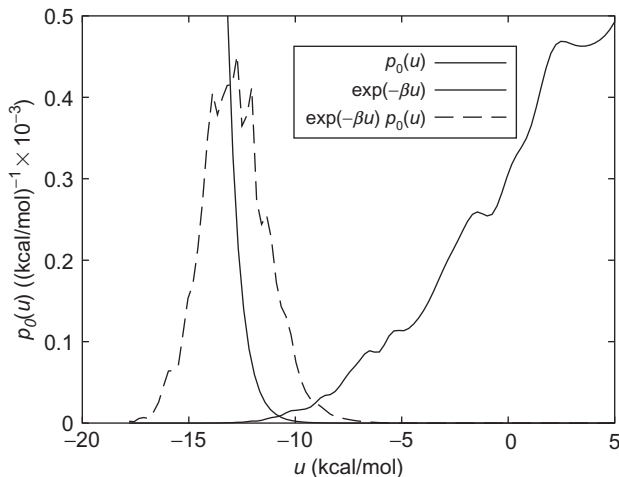


FIG. 2. Example of a calculated binding energy distribution  $p_0(u)$  from reference (Gallicchio et al., 2010). The curves to the left correspond to the  $\exp(-\beta u)$  and  $h(u) \sim \exp(-\beta u) p_0(u)$  functions (rescaled to fit within the plotting area). The integral of the latter is proportional to the binding constant (Eq. (26)).

which the particle insertion method of solvation thermodynamics (Pohorille and Pratt, 1990) is a particular realization (Widom, 1963). In particle insertion, the standard chemical potential of the solute,  $\mu$ , is written in terms of the probability distribution  $p_0(v)$  of solute–solvent interaction energies,  $v$ , corresponding to the ensemble in which the solute is not interacting with the solvent:

$$e^{-\beta\mu} = \int dv p_0(v) e^{-\beta v}. \quad (28)$$

This expression is equivalent to Eq. (26) with the solute–solvent interaction energy  $v$  replaced by the protein–ligand binding energy  $u$ . It follows that the formalism described above for the binding free energy can be regarded as a *ligand insertion* theory for protein–ligand binding, where the protein atoms and the solvent continuum play the same role as the solvent molecules in particle insertion.

A known result of PDT is a relationship between  $p_0(v)$ , the probability distribution of solute–solvent interaction energies in the absence of solute–solvent interactions, and  $p_1(v)$ , the corresponding probability distribution in the presence of solute–solvent interactions (Lu et al., 2003). In the present notation, we have

$$p_1(v) = e^{\beta\mu} e^{-\beta v} p_0(v), \quad (29)$$

where  $\mu$  is the chemical potential. The corresponding expression linking  $p_0(u)$ , the probability distribution of ligand–protein binding energies for the uncoupled (R+L) reference state, and  $p_1(u)$ , the probability distribution for the bound state RL, is

$$p_1(u) = e^{\beta\Delta G_1} e^{-\beta u} p_0(u), \quad (30)$$

where  $\Delta G_1$  is defined by Eq. (26). It follows that  $p_1(u)$  is proportional to the integrand in Eq. (26) for the interaction free energy. Note, however, that this does not imply that the interaction free energy can be computed by integration of  $p_1(u)$ , as obtained, for example, from a conventional simulation of the complex in the presence of ligand–receptor interactions. The integral of the normalized probability distribution  $p_1(u)$ , which is by definition unitary, does not contain any information about the interaction free energy. As expressed by Eq. (30), the proportionality constant between  $p_1(u)$  and the integrand of Eq. (26) is related to the interaction free energy, which is exactly the quantity we are seeking to compute.



The  $p_1(u)$  distribution is nevertheless a useful quantity for the analysis of the relative contributions to the binding free energy. Using Eq. (26), we can write Eq. (22) as

$$K_b = \int du k(u), \quad (31)$$

where, based on Eq. (30),

$$k(u) = \frac{V_{\text{site}} \Omega_{\text{site}}}{V^\circ 8\pi^2} e^{-\beta u} p_0(u) \quad (32)$$

can be interpreted as a measure of the contribution of the conformations of the complex with binding energy  $u$  to the binding constant. We thus call the function  $k(u)$  the *binding affinity density* (Gallicchio et al., 2010) (see Fig. 2). The binding affinity density  $k(u)$  is proportional to  $p_1(u)$ , the binding energy probability distribution in the bound state. (The critical distinction between the two is that the integral of the latter is equal to 1, whereas the integral of the binding affinity density is equal to the binding constant.) It thus follows that the relative contributions to the binding constant of two macrostates, one with binding energy  $u_1$  and another with binding energy  $u_2$ , are simply given by their relative populations in the ligand-bound state when the interactions between the ligand and the receptor are fully turned on.

### E. Definition of the Bound State

The expressions for the standard binding free energy presented above depend on the definition of the bound state through the indicator function  $I(\zeta_L)$ . This function can be chosen, for example, so as to as much as possible include only conformations that lack receptor–ligand clashes, or it can be defined at a coarser level by specifying, for example, an enveloping sphere containing the binding site of interest. As the choice of  $I(\zeta_L)$  is to some level arbitrary, there is a question as to which definition is appropriate. This issue has been reviewed in a number of studies (Gilson et al., 1997; Luo and Sharp, 2002; Mihailescu and Gilson, 2004). The main conclusion is that if the binding is strong and specific (as formally defined below), the specific choice for the definition of the bound state is for the most part irrelevant as long as it covers all important conformations of the complex. The conditions of strong and localized binding are the same

conditions at the basis of the quasi-chemical description of the noncovalent binding equilibrium embodied in Eq. (3).

Consider, for example, Eq. (18). The largest contributions to the integral come from regions where the binding PMF  $\Delta F(\zeta_L)$  is large and favorable and  $\exp[-\beta\Delta F(\zeta_L)]$  is large compared to 1, the value obtained in regions where the receptor and the ligand are not significantly interacting. If the minima of  $\Delta F(\zeta_L)$  are deep and localized, that is binding is strong and specific, the choice of the domain of integration has a small effect on the value of the integral as long as it covers all the regions where  $\Delta F(\zeta_L)$  is deep.

This analysis has been confirmed in at least one recent molecular simulation study (Gallicchio et al., 2010), in which the binding constant of a T4-Lysozyme complex was computed using Eq. (22) by varying the extent of the definition of the binding site region (Fig. 3). The results showed that, provided that it contains the main binding site, the binding site volume has a small effect on the computed binding constant. The variations at small binding site volumes in Fig. 3 are due to the fact that in this regime, the binding site definition misses some important conformations of the complex. The nearly constant behavior at larger binding site volumes are found to be due to a cancellation between the increasing  $V_{\text{site}}$  term in Eq. (22) and the linear decrease of the  $\exp[-\beta\Delta G_{\text{I}}]$  term with increasing binding site volume definition. Enlarging the binding site definition beyond the space that can be physically occupied by the ligand does not appreciably change the value of the integral in the numerator of Eq. (23) because the additional volume contains only points  $\zeta_L$  that cause ligand-receptor overlaps, where  $u(x_L, \zeta_L, x_R)$  is large and  $\exp[-\beta u(x_L, \zeta_L, x_R)]$  is small. However, the integral at the denominator, which does not contain the  $u(x_L, \zeta_L, x_R)$  energy term, increases linearly with increasing binding site volume definition, thereby canceling the  $V_{\text{site}}$  term at the numerator of Eq. (22). The result is a nearly invariant value of the binding constant. This example also shows that the values of  $\Delta G_{\text{I}}^{\circ}$ ,  $\Delta G_{\text{R}}$  and  $\Delta G_{\text{I}}$  in Eqs. (13), (16), and (25) are not unique. An increase in the chosen binding site volume, for instance, lowers the values of  $\Delta G_{\text{I}}^{\circ}$  and  $\Delta G_{\text{R}}$  at the expense of  $\Delta G_{\text{I}}$  that becomes less favorable so that their sum remains nearly constant. Therefore, it is important in binding free energy calculations of this kind to include the appropriate standard state terms to obtain answers that are not as affected by arbitrary model parameters.

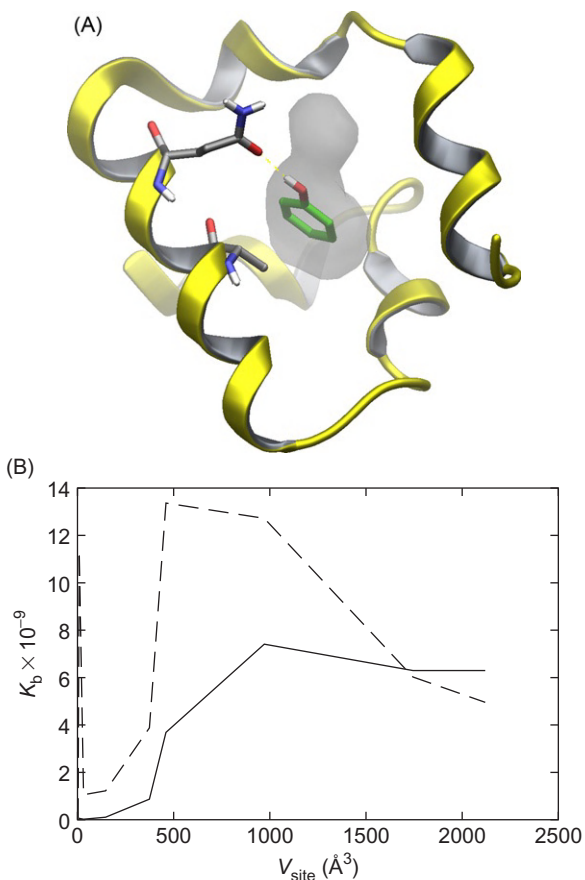


FIG. 3. The complex between phenol and the L99A/M102Q T4 mutant of lysozyme (PDB ID 1LI2, A). The ligand is highlighted in green. The surface surrounding the ligand represents the binding site which is buried and completely surrounded by protein atoms. The computed binding constant for this complex as a function of the binding site volume (B), using Eq. (22) with (full line) and without (dashed line) the inclusion of the  $V_{\text{site}}/V^p$  term (in this calculation,  $\Omega_{\text{site}}/8\pi^2=1$ ). The binding constant (full line) is fairly constant around  $K_b=6 \times 10^9$  for  $V_{\text{site}} > 500 \text{\AA}^3$ , whereas  $\exp[-\beta\Delta G_I]$  (dashed line) decreases linearly in this region. The two curves meet fortuitously at  $V_{\text{site}}=1668 \text{\AA}^3$ , where  $V_{\text{site}}/V^p=1$ . These calculations were conducted with a distance-dependent model (Gallicchio et al., 2010), which underestimates desolvation effects and overestimates affinity. The dependence on site is, however, representative of systems of this kind.

The example above involved a buried binding site. For calculations involving surface sites (as well as buried sites for binding site volumes large enough to extend into the solvent), however, the binding constant is expected to vary linearly with the volume of the binding site for large enough binding sites. Which value of the binding site volume is then appropriate? One simple answer is that in practical terms, as discussed above, if the binding is strong and localized, most reasonable choices for the binding site will yield reasonably accurate results. For example, doubling  $V_{\text{site}}$  would decrease the binding constant by a factor of 2 and increase the binding free energy by only  $\sim 0.4$  kcal/mol at room temperature, a relatively small change compared to typical strong protein–ligand binding affinities of the order of  $-10$  kcal/mol. This occurs because the slow logarithmic dependence of the binding free energy on  $V_{\text{site}}$  is not as significant compared to the larger effect due to strong ligand–receptor interactions.

For weak and less localized binding, however, the dependence on  $V_{\text{site}}$  would be more noticeable. In addition, from a theoretical perspective, we would like to understand the paradox that, even though Eq. (4) depends on an arbitrary definition of the complex, the binding constant is a measurable quantity. This has led to the conclusion that, apparently, “Nature knows how to define the complex, even if we do not” (Groot, 1992). Mihailescu and Gilson (2004) have reviewed this issue and concluded that, first of all, the theoretical expression for the binding constant depends on the experimental technique used. Only methods based on spectroscopic reporting (such as fluorescence quenching) (Barbieri et al., 2007) can be shown to be modeled by the quasi-chemical theory considered here. (Equilibrium dialysis techniques, e.g., follow a different but related law (Mihailescu and Gilson, 2004), which does not require a definition of the binding site volume.) Moreover, Mihailescu and Gilson conclude that the definition of the binding site volume most appropriate to reproduce measurements based on spectroscopic reporting is the *exclusion zone* of the complex, generally defined as the region that includes the binding minimum and the source of the spectroscopic signal, and extends up to a point where there would be enough space to allow a second ligand to interact more strongly with the receptor (Mihailescu and Gilson, 2004).

### F. Thermodynamic Decompositions

The free energy of binding is the result of a delicate balance between opposing thermodynamic forces. The main driving force toward binding is the formation of receptor–ligand interactions. However, these occur at the expense of solvent interactions producing desolvation effects that often oppose binding. Intuitively, binding is necessarily accompanied by the loss of translational freedom, and therefore, entropic forces tend to disrupt complex formation. In addition, both the ligand and the receptor lose free energy to adapt their conformations to match those compatible for binding. Given the complexity of the process, it is very difficult to predict variations of the binding equilibrium. To rationalize binding affinities, it is therefore often beneficial to consider contributions to the binding free energy each easier to rationalize than the total. We summarize below three relevant decompositions.

#### 1. Enthalpy/Entropy Decomposition

A decomposition of the binding free energy into entropic and enthalpic contributions seeks to separate energetic factors from factors related to the loss of conformational freedom (Zhou and Gilson, 2009). Obvious candidates in this role are the entropy and enthalpy of binding, which reflect changes in standard thermodynamic potentials. The standard binding entropy is by definition given by the temperature derivative of the standard binding free energy. From Eq. (13):

$$\Delta S_b^\circ = -\frac{\partial \Delta G_b^\circ}{\partial T} = k \ln \frac{\Omega_{\text{site}} V_{\text{site}}}{8\pi^2 V^\circ} - \frac{\Delta G_2 - \Delta G_1}{T} + \frac{\Delta U_2 - \Delta U_1}{T}, \quad (33)$$

where

$$\Delta U_2 = \langle U \rangle_{\text{RL-slv}} - \langle U \rangle_{\text{R-slv+L-gas}} \quad (34)$$

is the change in average potential energy for establishing receptor–ligand and solvent–ligand interactions, and

$$\Delta U_1 = \langle U \rangle_{\text{L-slv}} - \langle U \rangle_{\text{slv+L-gas}} \quad (35)$$

the change in average potential energy for establishing solvent–ligand interactions. The standard binding enthalpy is given by:

$$\Delta H_b^\circ = \Delta H_b = \Delta G_b^\circ + T\Delta S_b^\circ = \Delta U_2 - \Delta U_1. \quad (36)$$

From these expressions, we immediately see that only the entropy of binding depends on the standard concentration  $C^\circ = 1/V^\circ$  through the first term on the r.h.s. of Eq. (33) which corresponds to the work  $\Delta G_t^\circ + \Delta G_r$  for imposing translational and orientational constraints. We will refer to this term as the translational entropy of binding

$$\Delta S_t^\circ = k \ln \frac{\Omega_{\text{site}} V_{\text{site}}}{8\pi^2 V^\circ}, \quad (37)$$

whereas we will use the term interaction entropy to refer to the concentration-independent remainder  $\Delta S_I$  defined from Eq. (16) by

$$\Delta S_I = -\frac{\partial \Delta G_I}{\partial T} = -\frac{\Delta G_2 - \Delta G_1}{T} + \frac{\Delta U_2 - \Delta U_1}{T}. \quad (38)$$

The standard entropies and enthalpies of binding are measurable quantities. They are often obtained directly by isothermal calorimetry or by measuring variations of binding constant with temperature (Serdyuk et al., 2007). Although they yield quantities directly comparable to experimental measurements, Eqs. (33) and (36) are rarely used in computational studies with explicit solvation because of the difficulties of converging the changes in total average potential energies  $\Delta U_2$  and  $\Delta U_1$ , which are given by the difference of two large values (each average in Eqs. (34) and (35) scales as  $O(N)$ , where  $N$  is the size of the system, whereas their difference, which is local to the binding site, is  $O(1)$ ). Estimating  $\Delta S_b^\circ$  by evaluating  $\Delta G_b^\circ$  over a range of temperatures and evaluating the derivative by finite differences (Levy and Gallicchio, 1998) is also problematic because using a small temperature range causes amplification of statistical errors, whereas using a large temperature range may introduce systematic bias.

Equation (33) is not valid with implicit solvation because in this case, unlike the potential energy  $U(x)$ , the effective potential energy  $U_{\text{eff}}(x)$  is temperature dependent. From Eq. (23), we have (Chang et al., 2007)

$$\Delta S_I = -\frac{\partial \Delta G_I}{\partial T} = -\frac{\Delta G_I}{T} + \frac{\Delta U_{\text{eff}}}{T} - \Delta \left( \frac{\partial W}{\partial T} \right), \quad (39)$$

where

$$\Delta U_{\text{eff}} = \langle U_{\text{eff}} \rangle_{\text{RL}} - \langle U_{\text{eff}} \rangle_{\text{R+L}} \quad (40)$$

is the change in total effective potential energy upon turning on receptor–ligand interactions and

$$\Delta \left( \frac{\partial W}{\partial T} \right) = \left\langle \frac{\partial W}{\partial T} \right\rangle_{\text{RL}} - \left\langle \frac{\partial W}{\partial T} \right\rangle_{\text{R+L}}. \quad (41)$$

is the corresponding change in the average temperature derivative of the solvent PMF. The binding enthalpy is again given by  $\Delta G_{\text{b}}^{\circ} + T\Delta S_{\text{b}}^{\circ}$  or

$$\Delta H_{\text{b}} = \Delta U_{\text{eff}} - T\Delta \left( \frac{\partial W}{\partial T} \right). \quad (42)$$

The sum of the first two terms in the r.h.s. of Eq. (39) is usually referred to as the *configurational entropy* of binding (Zhou and Gilson, 2009)

$$\Delta S_{\text{conf}} = -\frac{\Delta G_{\text{I}}}{T} + \frac{\Delta U_{\text{eff}}}{T}, \quad (43)$$

whereas the last term, which would be zero for a temperature-independent potential, corresponds to the change in solvent entropy. Similarly, the last term in the r.h.s. of Eq. (42) is the solvent contribution to the binding enthalpy.

It can be shown that (Zhou and Gilson, 2009) Eq. (43) is equivalent to taking the difference of the entropies of the bound and uncoupled states each evaluated using the fundamental equation

$$S = -k \int dx \rho(x) \ln \rho(x), \quad (44)$$

where  $\rho(x) = \exp[-\beta U(x)]/Z$  is the configurational distribution function.<sup>3</sup>

One interesting result from Eqs. (39) and (42) is that the  $\partial W/\partial T$  terms cancel out when evaluating the interaction free energy as  $\Delta G_{\text{I}} = \Delta H_{\text{b}} - T\Delta S_{\text{I}}$ , yielding

$$\Delta G_{\text{I}} = \Delta U_{\text{eff}} - T\Delta S_{\text{conf}}. \quad (45)$$

Consequently, the configurational entropy and the effective enthalpy of binding form a valid decomposition in that their sum, together with the appropriate concentration-dependent terms in Eq. (25), and without approximation, gives the standard binding free energy. On the other

<sup>3</sup>In principle, Eq. (44) should include an additional constant term corresponding to the multiplicative factor necessary to make the classical partition function dimensionless. This term, which cancels the dimensions of the distribution function within the logarithm in Eq. (44), is omitted here for brevity because it cancels out when taking differences between the quantities corresponding to the unbound and bound states.

hand,  $\Delta U_{\text{eff}}$  and  $\Delta S_{\text{conf}}$ , lacking proper solvent contributions, do not directly reflect the measurable entropies and enthalpies of binding. Conversely,  $\Delta U_{\text{eff}}$  and  $\Delta S_{\text{conf}}$  are not directly measurable thermodynamic quantities. Nevertheless, the effective enthalpy/configurational entropy decomposition can yield valuable insights on the driving forces in favor and against association. Moreover, because they are evaluated with implicit solvation, these quantities are also more amenable to computation relative to the full binding entropies and enthalpies. Indeed, as discussed below, some computational methods with implicit solvation, such as molecular mechanics/Poisson–Boltzmann plus surface area (MM/PBSA), are based on Eq. (45) and independent estimates of  $\Delta U_{\text{eff}}$  and  $\Delta S_{\text{conf}}$ .

## 2. The Reorganization Free Energy

Working within the implicit solvent representation, we can think of the binding process as occurring in two separate steps. First, the ligand and the receptor reorganize their conformational ensembles to match those of the bound complex, and then receptor–ligand interactions are established. As there is no change in the configurational distributions of the binding partners, from Eq. (44), we see that the entropy change for the second step is zero. Moreover, the enthalpy change for the second step is limited to the establishment of the receptor–ligand interaction energy  $\langle u \rangle_{\text{RL}}$ , where  $u$  is the binding energy defined by Eq. (24) and the RL subscript denotes averaging over the bound conformations of the complex. The remainder,  $\Delta G_{\text{reorg}}$ , defined by the identity

$$\Delta G_{\text{I}} = \Delta G_{\text{reorg}} + \langle u \rangle_{\text{RL}} \quad (46)$$

is then the free energy for the reorganization step.

By adding and subtracting  $\langle U_{\text{eff}}(\mathbf{x}_{\text{R}}) + U_{\text{eff}}(\mathbf{x}_{\text{L}}) \rangle_{\text{R+L}}$  from Eq. (46) and using Eqs. (24), (40), and (43), we can rewrite the reorganization free energy as

$$\Delta G_{\text{reorg}} = \Delta U_{\text{reorg}} - T\Delta S_{\text{conf}}, \quad (47)$$

where  $\Delta S_{\text{conf}}$  is the configurational entropy defined above, and

$$\Delta U_{\text{reorg}} = \langle U_{\text{eff}}(\mathbf{x}_{\text{R}}) + U_{\text{eff}}(\mathbf{x}_{\text{L}}) \rangle_{\text{RL}} - \langle U_{\text{eff}}(\mathbf{x}_{\text{R}}) + U_{\text{eff}}(\mathbf{x}_{\text{L}}) \rangle_{\text{R+L}} \quad (48)$$

is the *reorganization energy* defined as the change in the average internal potential energies of the receptor and the ligand in going from the unbound state to the bound state while they are not interacting.



Equation (47) confirms that the configurational entropy corresponds to the entropic cost of reorganizing the conformational ensembles of the binding partners to form the complex.

The reorganization free energy is necessarily positive because without mutual interactions, the ligand and the receptor would spontaneously relax to their conformational ensembles at a lower free energy. Therefore based on Eq. (46), we conclude that the average binding energy  $\langle u \rangle_{\text{RL}}$  is the only term that can be favorable to binding, while reorganization always opposes it.

In some applications, other definitions of the reorganization free energy appear in which the intermediate state is one in which the receptor and the ligand conformational ensembles by construction do not match exactly those of the complex (Mobley et al., 2007a). Consider, for example, Fig. 4 in which the binding free energy (here, the ligand is assumed to be already placed in the binding site) is decomposed into the free energy  $\Delta G_{\text{reorg}}^*$  of restraining the ensembles of conformations of the receptor and the ligand in solution to chosen macrostates  $\text{R}^*$  and  $\text{L}^*$  (for instance, an application is described below in which the  $\text{R}^*$  macrostate is defined with respect to a side-chain conformation). The free energy for this process is related to the population  $P_{\text{R}+\text{L}}^*$ , defined as the probability of finding a conformation belonging to the macrostate, in the absence of restraints:

$$\Delta G_{\text{reorg}}^* = -kT \ln P_{\text{R}+\text{L}}^*. \quad (49)$$

Following this step, we consider the binding free energy,  $\Delta G_{\text{I}}^*$ , between the  $\text{R}^*$  and  $\text{L}^*$  species, that is, the binding free energy when the receptor

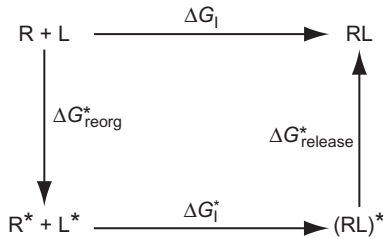


FIG. 4. Thermodynamic cycle illustrating the restrain-and-release decomposition of the interaction free energy (Eq. (51)). Although not indicated, the ligand here is assumed to be always sequestered in the binding site.  $\text{R}$  and  $\text{L}$  represent the free receptor and ligand,  $\text{R}^*$  and  $\text{L}^*$  represent the receptor and ligand restrained within a conformational macrostate,  $(\text{RL})^*$  represents the complex in which receptor and ligand are restrained within their macrostates, and  $\text{RL}$  represents the free complex.

and the ligand are limited to the chosen macrostates.  $\Delta G_{\text{I}}^*$  is defined, for example, as in Eq. (23) where in addition to the binding site indicator function  $I(\zeta_{\text{L}})$ , indicator functions  $I(x_{\text{R}})$  and  $I(x_{\text{L}})$  are present which limit the range of the receptor and ligand internal degrees of freedom. In general, the resulting state of the complex, denoted by  $(\text{RL})^*$  in Fig. 4, does not match the full complexed state RL because in the former, the receptor and the ligand are limited to their respective macrostates. If the chosen macrostate encompasses most of the conformational ensemble of the complex, the  $(\text{RL})^*$  and RL species are virtually equivalent. Otherwise, we need to consider the free energy difference,  $\Delta G_{\text{release}}^*$ , of releasing the macrostate restraints in the complexed state, given by

$$\Delta G_{\text{release}}^* = kT \ln P_{\text{RL}}^* \quad (50)$$

where  $P_{\text{RL}}^*$  is the population of the macrostate when the ligand and the receptor are interacting. Putting all together, we finally obtain

$$\Delta G_{\text{I}} = \Delta G_{\text{I}}^* + kT \ln \frac{P_{\text{RL}}^*}{P_{\text{R+L}}^*}, \quad (51)$$

which expresses  $\Delta G_{\text{I}}$  as the sum of a term,  $\Delta G_{\text{I}}^*$  corresponding to the binding free energy of a macrostate of the complex plus a free energy term corresponding to the preparation and release of this macrostate.

The result in Eq. (51) also very clearly shows that to accurately estimate the binding free energy, it is sufficient to sample only those macrostates whose population is affected by the binding reaction. From Eq. (51), we see that  $\Delta G_{\text{I}} = \Delta G_{\text{I}}^*$  as long as  $P_{\text{R+L}}^* = P_{\text{RL}}^*$ , that is, the binding free energy computed within a chosen macrostate is an accurate estimate of the binding free energy if the population of the macrostate is approximately the same in the unbound and bound states. So, for example, it is not strictly necessary to thoroughly sample regions of a protein receptor far away from the binding site as these are often not substantially affected by the binding of the ligand. Arguably, it is precisely for this reason that computer simulations, which necessarily sample a very small fraction of conformational space, can be applied to the computation of binding free energies. Equation (51) is also the basis for the ‘‘restrain-and-release’’ double-decoupling method discussed below which is useful in cases when it is convenient to conduct the binding free energy calculation within a limited portion of conformational space.

### 3. Conformational Decomposition

We showed in Section II.D.1 that the binding affinity density measures the contribution of the conformations with a particular binding energy to the overall binding constant. In this section, we generalize this result in the conformational dimension. Often, the affinity between a receptor and a ligand is the result of not one but multiple binding modes differing, for example, in the orientation of the ligand in the binding site. We would then like to estimate the contribution of each mode to the total binding free energy. As discussed later, this question has computational relevance in that if we have a way to combine the binding free energies of multiple modes into a single overall binding free energy, then it would be possible to simplify the calculation by treating each mode separately. As we show, in this section, a conformational decomposition of this kind is possible.

Let us work in the implicit solvent representation using the binding energy distribution formalism presented in Section II.D.1. Given a set of macrostates  $i=1, \dots, n$  of the complex, we consider the joint probability distribution  $p_0(u, i)$ , expressing the probability of observing the binding energy  $u$  while the complex is in macrostate  $i$ . Assuming that the set of macrostates collectively covers all possible conformations of the complex (which is always possible by including a ‘‘catch-all’’ macrostate), we can express  $p_0(u)$  as a marginal of  $p_0(u, i)$ :

$$p_0(u) = \sum_i p_0(u, i) = \sum_i P_0(i) p_0(u|i), \quad (52)$$

where we have introduced the conditional distribution  $p_0(u|i)$  and the population  $P_0(i)$  of macrostate  $i$  in the uncoupled reference state and used the relationship  $p_0(u, i) = P_0(i) p_0(u|i)$  between the joint and conditional distributions. By inserting Eq. (52) into Eq. (32), we have

$$k(u) = \sum_i P_0(i) k_i(u), \quad (53)$$

where

$$k_i(u) = \frac{V_{\text{site}} \Omega_{\text{site}}}{V^\circ 8\pi^2} p_0(u|i) e^{-\beta u} \quad (54)$$

represents the binding affinity density for macrostate  $i$ . In analogy with Eq. (31), we define a macrostate-specific binding constant

$$K_b(i) = e^{-\beta\Delta G_b^\circ(i)} = \int du k_i(u) = \frac{V_{\text{site}} \Omega_{\text{site}}}{V^\circ 8\pi^2} \langle e^{-\beta u} \rangle_{R+L,i}, \quad (55)$$

where  $\langle \dots \rangle_{R+L,i}$  represents an ensemble average in the unbound state of the complex limited to macrostate  $i$ . The macrostate-specific binding constant  $K_b(i)$  represents therefore the binding constant that would be measured if the conformations of the complex were limited to macrostate  $i$ . From Eqs. (55) and (53), the sum of the macrostate-specific binding constants weighted by the macrostate populations  $P_0(i)$  is the total binding constant:

$$K_b = \sum_i P_0(i) K_b(i). \quad (56)$$

Equation (56) expresses the fact that each conformational macrostate contributes to the total binding constant proportionally to its macrostate-specific binding constant  $K_b(i)$  weighted by the population,  $P_0(i)$ , of the macrostate in the unbound state (Jayachandran et al., 2006). Using Eq. (2), the composition formula for the binding free energy corresponding to Eq. (56) is

$$\Delta G_b^\circ = -kT \ln \sum_i P_0(i) e^{-\beta\Delta G_b^\circ(i)}, \quad (57)$$

where  $\Delta G_b^\circ(i)$  is the standard binding free energy for macrostate  $i$ .

Although Eqs. (56) and (57) have been derived in the implicit solvation representation, it can be shown that they are valid in general. In the explicit solvent representation, the macrostate  $i$  refers to the solvated state for the receptor and for the gas phase for the ligand, and it is assumed that the same definition of macrostate  $i$  is used for both legs of the double-decoupling process (Eqs. (10) and (12)). Equation (57) also forms the basis of *integration over parts* approaches (Jayachandran et al., 2006; Mobley et al., 2006; Boyce et al., 2009) to the calculation of binding free energies. The idea is that the binding free energy can be obtained by the appropriate combination of the binding free energies of a series of binding modes. These methods are attractive because it is easier to localize the calculation to a macrostate than achieving equilibration between distinct binding modes. The challenge is to identify the collection of modes that contribute the most to the total binding free energy.

Misidentification of the highest contributing mode can introduce major errors, while neglecting secondary modes generally has a smaller effect on accuracy (Moblely et al., 2006; Gallicchio et al., 2010).

The ratio  $P_0(i)K_b(i)/K_b$  measures the relative contribution of macrostate  $i$  to the overall binding constant. We can see that a large macrostate-specific binding constant  $K_b(i)$  is not a sufficient condition for a large contribution to the overall affinity. It must be also the case that the macrostate has a significant population  $P_0(i)$  in the unbound state. This result can be interpreted as a generalization of the reorganization free energy concepts developed in Section II.F.2.  $\Delta G_{\text{reorg}} = kT \ln P_0(i)$  measures the reorganization free energy penalty for restraining the system into macrostate  $i$  in the unbound state, whereas  $\Delta G_b^\circ(i)$  measures the association free energy in that macrostate. For a macrostate to contribute significantly to the binding affinity, the reorganization penalty and the association gain must combine so as to be favorable overall to binding.

It is straightforward to show from Eqs. (55) and (30) that (Gallicchio et al., 2010)

$$\frac{P_0(i)K_b(i)}{K_b} = P_1(i), \quad (58)$$

where

$$P_1(i) = \int du p_1(u, i) \quad (59)$$

is the population of macrostate  $i$  in the bound state. In other words, this analysis shows that the relative contribution of macrostate  $i$  to the binding constant is equal to the physical population of that macrostate of the complex. If a particular binding mode of the complex can be observed, by, for example, X-ray crystallography, it can be concluded therefore that its population is high and that it likely contributes significantly to the binding affinity.

It is also of interest to estimate the effect of having missed a particular binding mode in a binding free energy calculation. An expression for the binding constant,  $K_b(-j)$ , when macrostate  $j$ , say, has been missed can be derived by removing the corresponding term in the sum in Eq. (56) and, in addition, by renormalizing the macrostate populations so that they add to one. The result is

$$K_b(-j) = \frac{K_b - P_0(j)K_b(j)}{1 - P_0(j)}. \quad (60)$$

From this result, we can see that, as expected, missing macrostate  $j$  has a large effect in the computed binding constant if this macrostate provides a large contribution to the overall binding constant (the  $P_0(j) K_b(j)$  term in Eq. (60)). It also shows, however, that the binding constant can also be severely overestimated if the  $j$  macrostate is highly populated in solution (the  $1 - P_0(j)$  term at the denominator is small). In other words, large errors in binding free energy calculations are expected either if important macrostates of the bound complex are missed or if important macrostates of the unbound states are missed. The latter occurs because the calculation would underestimate the free energy required to reorganize the binding partners into their bound ensembles.

### III. COMPUTATIONAL METHODS

The development of a statistical mechanics theory of noncovalent association is only the first step in the development of computational models and methods for the calculation of binding affinities. To begin with, the expressions for the free energy of binding presented above depend on the definition of a potential energy function  $U(x)$ . We also require some prescription to generate ensembles, or set of conformations  $x$  of the system, compatible with the thermodynamic state of the system and the potential energy model. In this review, we focus on all-atom *classical force fields* (Cornell et al., 1995; Jorgensen et al., 1996; MacKerell et al., 1998; Schuler et al., 2001) energy models, and on molecular dynamics (MD) or Monte Carlo (MC)-based conformational sampling methods, which are most commonly applied models for protein–ligand binding free energy estimation. Atomistic force field models are not reviewed further here except to say that they are parametrized functions of the Cartesian coordinates of the atoms of the system, describing electrostatic, dispersion, and steric noncovalent interactions as well as covalent interactions between atoms. Force fields are used with explicit representations of solvent molecules (water in the applications described below), as well as in conjunction with implicit models of hydration (Lazaridis and Karplus, 1999; Bashford and Case, 2000; Wagoner and Baker, 2006; Chen et al., 2008; Gallicchio et al., 2009).

A very active and rich area of research is focused on the development of computer algorithms for the evaluation of free energies (Chipot and Pohorille, 2007) given an energy model. One class of free energy methods applicable to binding free energy simulations is based on connecting the unbound and bound states by a suitable thermodynamic path. At a fundamental level, *thermodynamic path methods* are capable of computing ratios of partition functions as in Eq. (4). Another class of free energy methods, often referred to as *end point methods*, compute binding free energies by explicitly estimating the free energies of the bound and unbound states (Swanson et al., 2004).

### A. Free Energy Estimators

Equations (10) and (12), for explicit solvation, and Eq. (23), for implicit solvation, suggest a simple algorithm to the computational evaluation of binding free energies by means of exponential averaging of the binding energy in an appropriate reference ensemble. In practice, these expressions suffer from several limitations and are rarely implemented as such. Instead, suitable *free energy estimators* have been developed which are discussed in this section.

Equations (10), (12), and (23) are particular realizations of the free energy perturbation (FEP) identity (Zwanzig, 1954), which states that the free energy difference  $\Delta G$  between two states 1 and 0 is

$$\Delta G = -kT \ln \frac{Z_1}{Z_0} = -kT \ln \left\langle e^{-\beta \Delta U(x)} \right\rangle_0, \quad (61)$$

where  $Z_1$  and  $Z_0$  are the corresponding configurational partition functions and  $\Delta U(x) = U_1(x) - U_0(x)$  is the difference of potential energies between state 1 and 0 (the perturbation), and the average is over conformations  $x$  sampled from the reference state 0. In our case, state 1 is the bound state and state 0 is the uncoupled state of the complex. Because they are very difficult to converge, however, in binding free energy applications, the FEP formulas are rarely evaluated directly. To understand why, consider, for example, Eq. (26) and Fig. 2. The distribution of binding energies in the unbound state,  $p_0(u)$ , is largest for large positive values of  $u$ . This is expected since in this state the ligand is restrained in the binding site where, in the absence of receptor–ligand interactions, the ligand is more

likely to sample conformations with unfavorable clashes with receptor atoms rather than conformations with favorable interactions. The values of  $u$  in the extreme negative binding energy range correspond to the low-energy conformations of the complex, which are very rarely visited in absence of ligand–receptor interactions. On the other hand, the exponential factor,  $\exp(-\beta u)$ , amplifies the contribution of these conformations to the integral in Eq. (26), causing the average to be dominated by rare events. This results in unreliable results, requiring the accumulation of an inordinate, and practically unachievable, number of independent samples to reach convergence (Pohorille et al., 2010).

An equivalent way to assess this problem is to consider the distribution,  $p_1(u)$  of binding energies in the bound ensemble (illustrated in Fig. 2 as a dashed curve). We concluded above (Eq. (31)) that most of the contribution to binding comes from conformations where  $p_1(u)$  is large. The amount of overlaps between  $p_1(u)$  and  $p_0(u)$  is a measure of the probability that one of these conformations is generated by chance in the uncoupled ensemble. As we can see from Fig. 2, the amount of overlap is small and the binding affinity is expected to be difficult to assess by sampling only the uncoupled ensemble. This is a general result, which states that the FEP formula is applicable for the computation of free energy difference between closely related states whose distributions of the perturbation energy overlap significantly (Lu and Kofke, 2001; Chipot and Pohorille, 2007; Pohorille et al., 2010).

The technique known as *stratification* (Chipot and Pohorille, 2007) is a general way to circumvent the problem of poor overlap between energy distribution functions in FEP binding free energy calculations. The first ingredient is a  $\lambda$ -dependent hybrid potential, which at  $\lambda=0$  typically corresponds to the unbound state and at  $\lambda=1$  corresponds to the bound state. A straightforward, although not necessarily optimal, choice for the hybrid potential in binding free energy calculations is

$$U(x_R, x_L, \zeta_L | \lambda) = U(x_R) + U(x_L) + \lambda u(x_R, x_L, \zeta_L), \quad (62)$$

where  $U(x_R) + U(x_L)$  represents the energy in the unbound state and  $u$  is the binding energy. Here, we have used the notation for implicit solvation denoting for simplicity the effective potential as  $U$ . The expression for hybrid potential, Eq. (62), can easily be adapted to the solvation and binding steps (Eqs. (12) and (10)) with explicit solvation. The hybrid potential defines a thermodynamic path connecting the unbound and bound states



through an arbitrary number of unphysical intermediate states at  $0 < \lambda < 1$  in which the receptor and the ligand are only partially coupled. In addition, states with similar  $\lambda$  have similar characteristics and, in particular, similar binding energy distributions with significant overlap, allowing the application of the FEP formula for the computation of their free energy difference:

$$G(\lambda_2) - G(\lambda_1) = -kT \ln \frac{Z_{\lambda_2}}{Z_{\lambda_1}} = -kT \ln \langle e^{-\beta \Delta \lambda u} \rangle_{\lambda_1}, \quad (63)$$

where  $\Delta \lambda = \lambda_2 - \lambda_1$ . Given a set of  $n$  intermediate states at  $\lambda = \lambda_i$ , the free-energy difference can then be evaluated as the sum of the free-energy differences between intermediate states

$$\Delta G = G(\lambda = 1) - G(\lambda = 0) = -kT \sum_i \ln \langle e^{-\beta \Delta \lambda_i u} \rangle_{\lambda_i}, \quad (64)$$

where  $\Delta \lambda_i = \lambda_{i+1} - \lambda_i$ . More generally, when the expression for the hybrid potential is not linear in  $\lambda$ ,  $\Delta \lambda_i u$  in Eq. (64) is replaced by  $U(\lambda_{i+1}) - U(\lambda_i)$ .

Because it is based on the sum of well-behaved terms, the FEP stratification formula, Eq. (64), is much easier to convergence than the direct application of the FEP formula between the unbound and bound states. The procedure entails performing multiple MD or MC simulations to collect samples at each  $\lambda$ . The more intermediate states are employed; the fewer samples are needed to converge each term but more terms need to be evaluated. A number of techniques have been developed to optimize the  $\lambda$  schedule in FEP calculations and to assess the reliability of individual free energy estimates based, for example, on the analysis of neighboring distributions (Chipot and Pohorille, 2007; Pohorille et al., 2010).

The *thermodynamic integration* (TI) formula, which is sometime used in binding free energy calculations (Michel and Essex, 2010), can be considered the continuous limit of Eq. (64) for  $\Delta \lambda_i \rightarrow 0$

$$\Delta G = \int_0^1 d\lambda \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda} = \int_0^1 d\lambda \langle u \rangle_{\lambda}, \quad (65)$$

where the last equality follows from Eq. (62). The TI formula is formally derived from the identity

$$\frac{\partial G(\lambda)}{\partial \lambda} = -kT \frac{\partial \ln Z(\lambda)}{\partial \lambda} = \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{\lambda}. \quad (66)$$

Equation (64) expresses each individual free energy difference in terms of an exponential average. One limitation of the exponential average is that, as

discussed above, it works well only if conformations relevant for the target state are sampled in the reference state, or in other words, if the binding energy distribution in the reference state envelopes that of the target state. The result is that often one perturbation direction gives different results than the other (hysteresis), with the one going in the direction of decreasing entropy (for binding the one starting from the unbound state) usually being more accurate (Lu and Kofke, 2001). In some cases, however, neither direction may work well unless the  $\lambda$  spacing is made very small. In recent years, more efficient free energy estimators have been developed. The Bennet acceptance ratio (BAR) formula (Bennett, 1976; Lu et al., 2003)

$$\Delta G(\lambda) = C - kT \ln \frac{\langle f[-\beta(\Delta\lambda u - C)] \rangle_{\lambda_1}}{\langle f[\beta(\Delta\lambda u - C)] \rangle_{\lambda_2}}, \quad (67)$$

where  $f(x) = 1/[1 + \exp(x)]$  is the Fermi function and  $C$  is a constant determined iteratively, has been shown to be an optimal free energy estimator with respect to the minimization of the statistical variance. It is also symmetric with respect to the perturbation direction. The BAR formula is based on the introduction of a fictitious intermediate state whose distribution is enveloped by the distributions of both end states and peaks where they most overlap. Consequently, the BAR formula requires only that the two distributions overlap to some extent, rather than requiring that one is enveloped in the other as for the exponential averaging formula. The BAR formula has for the most part replaced the exponential averaging formula in modern FEP binding free energy calculations.

A FEP approach can also be used to compute the binding free energy using the binding PMF approach (Eqs. (18) and (19)). In this case, techniques to compute free energy changes along a thermodynamic path described by a structural order parameter can be considered. For example, the distance measure  $d(\lambda)$  of the ligand from the binding site. Samples are generated at a reference receptor–ligand distance, and the potential energy changes  $\Delta U$  resulting from displacing the ligand distance from the receptor by  $\Delta d = d(\lambda_{i+1}) - d(\lambda_i)$  are computed in the context of Eq. (64) or (67). More commonly, however, the binding PMF is expressed in terms of the probability density  $p(d)$  of the receptor–ligand distance

$$\Delta F(d) = -kT \ln \frac{p(d)}{p(d^*)}, \quad (68)$$

where  $d^*$  is some reference large distance corresponding to the solvent bulk. Because it is difficult to sample a large range of distances in one

simulation, multiple simulations are conducted each employing a different auxiliary confining potential designed to bias sampling in one limited range of distances (Woo and Roux, 2005). In this technique, generally known as *umbrella sampling*, each simulation generates a biased distribution. The data from all of the simulations are then combined and unbiased using reweighting techniques such as the weighted histogram analysis method (WHAM) (Ferrenberg and Swendsen, 1989; Kumar et al., 1992; Gallicchio et al., 2005). The WHAM equations in this case are expressed as

$$P(d_i) = \frac{n(d_i)}{\sum_{\lambda} n_{\lambda} f_{\lambda} \exp[-\beta \omega_{\lambda}(d_i)]}, \quad (69)$$

where  $P(d_i) = p(d_i) \Delta d_i$  is the unbiased probability to find the system at distance bin  $i$  of size  $\Delta d_i$  centered at  $d_i$  and  $n(d_i)$  is the number of samples collected from all simulations in this bin. The denominator is a sum over the simulations, each at a different value of  $\lambda$ .  $n_{\lambda}$  is the total number of samples collected at the simulation at  $\lambda$ ,  $\omega_{\lambda}(d_i)$  is the value of the biasing potential at  $\lambda$  corresponding to bin  $i$ , and finally,

$$f_{\lambda}^{-1} = \sum_i \exp[-\beta \omega_{\lambda}(d_i)] P(d_i) \quad (70)$$

is a normalization factor related to the free energy,  $kT \ln f_{\lambda}$ , of the system at  $\lambda$  relative to the unbiased system. Equations (69) and (70) are solved iteratively until convergence. The binding free energy is then computed by integrating the binding PMF over the binding site region (Eq. (18)).

The usefulness of WHAM as a binding free energy estimator extends to alchemical methods as well. As further described below, WHAM has been used to implement Eq. (23) by choosing the binding energy  $u$  as thermodynamic path parameter and setting as biased potential  $\omega_{\lambda}(u) = \lambda u$  (Gallicchio et al., 2010). From Eq. (62), the unbiased system at  $\lambda=0$  is the unbound state and  $\lambda=1$  corresponds to the bound system, and consequently, Eq. (70) evaluated at  $\lambda=1$  yields the interaction component of the binding free energy:

$$\Delta G_1 = kT \ln f_{\lambda=1}. \quad (71)$$

More recently, the multistate Bennett acceptance ratio (MBAR) method has been developed (Tan, 2004; Shirts and Chodera, 2008), which, in a way, unifies the BAR and WHAM free energy estimators. Like WHAM, it

combines in a statistically optimal way data from multiple values of  $\lambda$  to compute the overall binding free energy (rather than from a sum of pairwise terms as in the FEP (Eq. (64)). It also resembles WHAM in terms of formulation. In fact, it is equivalent to WHAM in the limit that bin sizes are made so small so as to contain only one sample, or none. However, MBAR reduces to the BAR estimator when only two states are considered. The MBAR free energy estimator is preferable to WHAM because it does not require the definition of a histogram grid, and it is preferable to BAR because it more efficiently utilizes the samples generated at each  $\lambda$  so that all of them contribute to free energy differences. Because, in addition, it combines the generality of both methods, the MBAR is expected to become a widely employed estimator in binding free energy calculations.

### B. Double Decoupling

The double-decoupling method (Gilson et al., 1997; Deng and Roux, 2009; Mobley and Dill, 2009) is an alchemical approach to the calculation of standard binding free energies (often referred to as *absolute* binding free energies in the literature). It implements Eq. (13), where the computations of the free energies of transfer,  $\Delta G_1$  and  $\Delta G_2$ , of the ligand from the gas phase to, respectively, the solution and receptor environments form the core of the method. The name double decoupling comes from thinking of the two opposite processes of decoupling the ligand from the solution and receptor environments. Equations (12) and (10) are implemented using either the TI (Eq. (65)) or the staged FEP/BAR (Eqs. (64) and (67)) free energy estimators.

Double decoupling has been used recently to compute the standard binding free energies of a variety of protein–ligand complexes. The L99A and L99A/M102Q mutants of T4-lysozyme (Eriksson et al., 1992; Graves et al., 2005) have been the most studied systems; the small size of the ligands, the relative simplicity of the binding sites, and the availability of high-quality structural and thermodynamic data (Morton et al., 1995; Wei et al., 2002) have made these systems particularly well suited for testing the validity of various computational protocols (Deng and Roux, 2006; Mobley et al., 2007b; Boyce et al., 2009). A number of double-decoupling studies (Jayachandran et al., 2006; Wang et al., 2006) have also targeted a series of inhibitors of the FKBP12 receptor (Holt et al., 1993). Applications to the

trypsin (Jiao et al., 2008; Jiao et al., 2009) and the ribosomal peptidyl-transferase receptors (Ge and Roux, 2010) have also been recently reported.

From a computational perspective, the three main issues in double-decoupling simulations are (i) the extent of conformational sampling (discussed in detail in Section III.F), (ii) the definition of the binding site volume by restraining potentials, and (iii) the use of soft-core hybrid potentials.

As discussed above, the definition of the complexed state and the concentration dependence of the standard state are formally introduced by a binding site indicator function  $I(\zeta_L)$ . As discussed (Gilson et al., 1997; Boresch et al., 2003),  $I(\zeta_L)$  can be defined in terms of a continuous function which interpolates from values near 1 within the binding site region to values near 0 outside it. A common choice is to set

$$I(\zeta_L) = e^{-\beta U_{\text{restr}}(\zeta_L)}, \quad (72)$$

where  $U_{\text{restr}}$  is a suitable restraining potential that depends only on the external coordinates of the ligand. This definition is computationally convenient because it is differentiable and, as we can see by inserting Eq. (72) in Eq. (10) or in Eq. (23), the indicator function can be implemented by means of restraining potentials easily included in potential energy routines of MD packages. Note that, because the restraining potential is present in both the unbound states, it does not contribute to the binding energy (Eqs. (11) and (24)). Also note that the definition above makes the definition of the complexed state temperature dependent, potentially affecting in unwanted ways the temperature dependence of binding free energies. This dependence can be removed by adjusting the strength of  $U_{\text{restr}}$  according to the simulation temperature.

Some early absolute binding free energy calculations (Jorgensen et al., 1988), as well as more recent ones (Fujitani et al., 2005), did not account properly for the standard state definition. Moreover, ligand restraints are sometime described as a convenient computational device to enhance convergence by not letting the ligand wander into the whole simulation volume when it is uncoupled from the receptor (Deng and Roux, 2009). But, as discussed above, they are in fact a necessary input of the method; they implicitly provide a definition of the complexed state without which it is not possible to define its free energy. Boresch et al. (2003) have

introduced a general framework to define the six external degrees of freedom  $\zeta_{\text{L}}$  of the ligand based on the positions (expressed in spherical polar coordinates) of three reference atoms of the ligand relative to three reference atoms of the receptor. This leads to three coordinates that specify the overall translation of the ligand (one distance and two angles) and another set of three coordinates (three angles) that determine the orientation of the ligand in the binding site. Restraining potentials can be applied only on the translational coordinates or also on the orientational coordinates. For harmonic or flat-bottom harmonic restraints, the binding site volume  $V_{\text{site}}\Omega_{\text{site}}$  in Eq. (7) can be evaluated analytically. In other circumstances, the integration of the indicator function can be obtained numerically with high accuracy, as it involves at most six coordinates. Some early studies (Miyamoto and Kollman, 1993) employed multiple distance restraints between ligand atoms and receptor atoms, which, as pointed out by Boresch et al. (2003), is incorrect based on this formalism, as it would introduce couplings between the external ligand coordinates and the internal coordinates of the receptor and the ligand.

It has been observed that a hybrid potential linear in  $\lambda$  as in Eq. (62) leads to instabilities in the calculations of free energies near  $\lambda=0$  (Steinbrecher et al., 2007; Michel and Essex, 2010), when the ligand and the receptor are nearly uncoupled. Under these conditions, conformations are generated in which receptor and ligand atoms interpenetrate each other and yielding very large values of the binding energies. These cause instabilities in Eq. (63) which are difficult to overcome unless the  $\lambda$  spacing is very fine (small  $\Delta\lambda$ ). These difficulties have led to the development of so-called *soft-core hybrid potentials* which avoid large perturbation energies near the end point of the transformation. A popular class of soft-core potential employs a  $\lambda$ -dependent modified distance function in the evaluation of Lennard-Jones and Coulombic interactions. For example,

$$u_{\text{LJ}}(r|\lambda) = 4\epsilon_{\text{LJ}} \left\{ \frac{1}{\left[ \alpha\lambda + (r/\sigma_{\text{LJ}})^6 \right]^2} - \frac{1}{\left[ \alpha\lambda + (r/\sigma_{\text{LJ}})^6 \right]} \right\} \quad (73)$$

is a soft-core version of the Lennard-Jones pair potential. Note that  $u_{\text{LJ}}(r|\lambda)$  above is finite for any nonzero value of  $\lambda$  allowing particles to interpenetrate each other. This functional form also “grows” particles gradually, reducing the fluctuations of the free energy estimator at small  $\lambda$ . Decomposing the

decoupling steps such that electrostatic interactions is turned off before Lennard-Jones has also been shown to improve convergence.

### C. *Binding Energy Distribution Analysis Method*

The binding energy distribution analysis method (BEDAM) (Gallicchio et al., 2010) is an absolute binding free energy alchemical method based on an implicit description of the solvent. It computes the binding free energy by means of Eq. (26) where the distribution of binding energies  $p_0(u)$  is computed numerically. The numerical difficulties in the application of Eq. (26) is illustrated in Fig. 2. Because low binding energies are very rarely sampled when the ligand is not guided by the interactions with the receptor, the accurate calculation of the important low-energy tail of  $p_0(u)$  cannot be accomplished by brute-force collection of binding energy values from a simulation of the complex in the uncoupled state. Instead, samples are collected from a series of biased MD simulations of the complex with biasing potential  $\lambda u$ . In going from  $\lambda=0$  to 1, the system progressively samples more and more favorable binding energies. The replicas collectively sample a wide range of unfavorable, intermediate, and favorable binding energies which are unbiased and combined together by means of the WHAM to yield the unbiased probability density  $p_0(u)$  (Gallicchio et al., 2005), which is then used in Eq. (26) to compute the binding free energy. The ladder of  $\lambda$  values is chosen so that uniform coverage of the range of binding energies important for binding is achieved. In particular, the low binding energy tail of  $p_0(u)$ , although small in magnitude, is reliably estimated because the relative precision of the binding energy distribution  $p_0(u)$  computed by WHAM depends mainly on the number of samples collected at binding energy  $u$ , rather than the value of  $p_0(u)$  itself.

Although, as discussed in Section II.D.1, the binding energy distribution formalism on which BEDAM is based is valid in general, in practice, it is only applicable with implicit solvation. This is because in BEDAM the effective binding energy is part of the potential energy of the system, requiring fast evaluation of  $u$  and its gradients for MD conformational sampling. With explicit solvation, however, each evaluation of the effective binding energy would entail a costly and impractical binding free energy calculation (see discussion near Eq. (24)).

In a recent study (Gallicchio et al., 2010) using the OPLS force field with the AGBNP2 (Gallicchio et al., 2009) solvation model, BEDAM was shown to accurately identify ligand binders from nonbinders in a challenging set of candidate ligands to T4 lysozyme receptors (Fig. 3) failed by docking programs. In addition, the standard binding free energies of the binders were found to be in good agreement with experimental measurements. In contrast, energy-only estimators, which do not include entropic and energy reorganization effects, did not correctly reproduce the experimental rankings. As with other full free energy models of binding, BEDAM implicitly incorporates entropic and reorganization effects. In this study, the reorganization free energies were evaluated using Eq. (46) and shown to be large and, in many cases, the discriminating factors between binders and nonbinders. Analysis of the binding energy distributions, as described in Section II.F.3, allowed the decomposition of the binding free energies into conformational contributions based on the orientation of the ligand within the binding pocket. It was found that in many cases, several binding modes contributed nearly equally to the total binding free energy.

There are clear parallelisms between BEDAM and conventional binding free energy methods such as double decoupling. They are both alchemical methods that utilize a hybrid potential of the form in Eq. (62) to build a thermodynamic path between the unbound and the bound states. The binding energies collected in BEDAM can yield directly the binding free energy by means of the  $f$  factors (Eq. (70)) returned by WHAM or MBAR. One advantage of BEDAM over double decoupling is that BEDAM estimates the binding free energy from a single perturbation leg rather than from the difference of two separate free energy calculations with double decoupling. This feature is potentially advantageous for more rapid convergence of the binding free energies of highly polar and charged ligands, which, in double-decoupling and end point approaches discussed below, are the result of a nearly complete cancellation between the large free energies of the unbound and the bound states (Deng and Roux, 2009).

The challenges in BEDAM calculations are similar to those discussed above in the context of double decoupling. In addition, BEDAM relies on the quality of the implicit solvent potential. To obtain accurate binding free energies, care should be taken to achieve the correct balance between direct interaction and hydration forces (Gallicchio et al., 2009). As discussed below to further enhance the conformational sampling of ligand–receptor conformations, BEDAM employs a  $\lambda$ -hopping replica



exchange (RE) algorithm. The problem of the convergence of free energy differences near  $\lambda=0$  is evidenced by the long tail of the  $p_0(u)$  distribution at large energies which is difficult to estimate accurately. Recent versions of BEDAM employ a soft-core hybrid potential of the form  $U(\lambda) = U_0 + \lambda f(u)$ , with  $f(u) = u_{\max} \tanh(u/u_{\max})$ , where  $u_{\max}$  is some maximum ceiling for the binding energy, which has been shown to improve convergence without appreciably affecting free energy estimates.

#### D. PMF Approach

The binding PMF approach described in Section II.C is an example of a nonalchemical transformation to the calculation of absolute binding free energies. Numerical applications of the PMF formula have a long history in the study of dimerization of simple solutes (Jorgensen, 1989; Payne et al., 1997), and few applications have been reported for protein–ligand binding free energy estimation (Woo and Roux, 2005; Lee and Olson, 2006; Deng and Roux, 2009). The main advantage of PMF calculations is that they can be conducted with explicit solvation, but, unlike double-decoupling methods, they do not suffer from the large cancellation between the solvation and binding components ( $\Delta G_1$  and  $\Delta G_2$  in Eqs. (12) and (10)). PMF calculations are therefore easier to converge for the binding between charged ligands and receptors whose solvation free energies can be of the order of  $\sim 100$  kcal/mol. The disadvantage of the PMF approach is that it relies on the presence of a physical unobstructed path for the ligand to reach the binding site from solution. This limitation basically prevents the application of the method to buried binding sites.

Computationally, it is impractical to obtain the PMF along all of the six external ligand coordinates. Typically, only one coordinate is used corresponding to a displacement distance  $d$  along an approach path from the bulk solution to the binding site. The other coordinates are either fixed (Woo and Roux, 2005) or averaged (Lee and Olson, 2006). In the former case, the work necessary to restrain the angular position and orientation of the ligand relative to the receptor is computed separately (Woo and Roux, 2005). The PMF is computed along the approach coordinated by biased sampling and reweighting, as discussed above. In the reported applications (Woo and Roux, 2005; Lee and Olson, 2006), harmonic biasing potentials were employed.

### E. Relative Binding Free Energies

Often in pharmaceutical applications (Reddy and Erion, 2001), we are interested in the difference of binding free energy between two related compounds to the same receptor. Computational methods designed to compute directly relative binding free energies, rather than the corresponding standard binding free energies, have been developed and resulted in some of the first applications of free energy methods to protein–ligand binding (Tembe and McCammon, 1984). Relative binding free energy calculations (commonly referred to as FEP calculations) constitute the majority of protein–ligand binding calculations conducted in academic and industrial settings, and a variety of techniques have been developed to improve their efficiency and accuracy. This body of work has been thoroughly reviewed (Oostenbrink and van Gunsteren, 2005; Chipot and Pohorille, 2007; Jorgensen and Thomas, 2008; Jorgensen, 2009; Knight and Brooks, 2009; Michel and Essex, 2010). In this section, we sketch out the foundations of the method based on the statistical mechanics theory presented above and point out connections between relative and absolute binding free energy calculations.

The difference of standard binding free energies,  $\Delta\Delta G_b^\circ = \Delta G_b^\circ(\text{B}) - \Delta G_b^\circ(\text{A})$ , between two ligands B and A is equivalently expressed as the ratio of the corresponding binding constants (Eq. (2)). Using Eq. (4), and assuming that both ligands bind to the same binding site of the receptor R, we arrive at the following expression

$$e^{-\beta\Delta\Delta G_b^\circ} = \frac{K_b(\text{B})}{K_b(\text{A})} = \frac{Z_{N,\text{RB}}}{Z_{N,\text{RA}}} \frac{Z_{N,\text{A}}}{Z_{N,\text{B}}} = e^{-\beta[\Delta\Delta G_{\text{R}}(\text{BA}) - \Delta\Delta G_{\text{solv}}(\text{BA})]}. \quad (74)$$

where  $\Delta\Delta G_{\text{R}}(\text{BA})$  is the difference in free energy of complexes RB and RA and  $\Delta\Delta G_{\text{solv}}(\text{BA})$  is the difference in solvation free energies between ligands B and A. We see that the relative free energy of binding is independent from the standard state concentration. Also, the ratios of partition functions in Eq. (74) can be expressed as averages, similar to those in Eqs. (10) and (12),<sup>4</sup> based on the difference in potential energy

<sup>4</sup>Note that these averages still contain the  $I(\zeta_{\text{L}})$  indicator functions (assumed to be the same for the two ligands). Like absolute binding free energies, therefore, relative binding free energies are dependent on the definition of the complexed state. This aspect is often overlooked in the literature.

between the ligands averaged over the ensembles of one of the ligands in the binding site and in solution, without resorting to intermediate gas phase state for the ligands. Given a suitable  $\lambda$ -dependent interpolation potential connecting the potential energies of the two ligands, these averages can be computed with the alchemical free energy estimators discussed in Section III.A. Two main mutation techniques, *single topology* and *dual topology* (Michel and Essex, 2010), exist to map the potential energy of one ligand to the other.

Relative binding free energy calculations are expected to be more efficient than computing the difference of the corresponding absolute binding free energies when the two ligands are similar to each other. Conversely, it is difficult to set up an interpolation potential and converge the relative binding free energy when the two ligands have very different chemical structures. However, ligand similarity alone is not a sufficient condition for obtaining reliable relative binding free energies. As in absolute binding free energy calculations, one of the main challenges is the extent of conformational sampling. It has been observed, for example (Boyce et al., 2009; Gallicchio et al., 2010), that even slight ligand modifications can cause large changes in the main ligand binding mode. In these cases, the sampling of both binding modes is required to yield reliable results, thereby reducing the computational advantage of relative binding free energy calculations over absolute ones. Relative binding free energy calculations are also considered less suitable than absolute ones to assess the reliability of algorithms and force fields against experimental data (Shirts et al., 2010; Chodera et al., 2011).

#### F. RE Conformational Sampling

Conformational equilibria relevant for the binding process occur on time scales which are unattainable with conventional MD even with the fastest supercomputers available. A commonly employed strategy to enhance sampling involves the application of biasing forces, and, as we discussed above, alchemical free energy methods employing hybrid potentials and PMF approaches employing umbrella potentials can be considered as belonging to this general class of methods. It has been shown in many contexts (Woods et al., 2003a,b; Murata et al., 2004; Liu et al., 2005, 2006; Bussi et al., 2006; Piana and Laio, 2007; Roitberg et al., 2007;

Hritz and Oostenbrink, 2008; Neale et al., 2008; Yeh et al., 2008; Jiang et al., 2009; Gallicchio et al., 2010; Jiang and Roux, 2010; Khavrutskii and Wallqvist, 2010; Meng and Roitberg, 2010; Mitsutake et al., 2010) that generalized ensemble conformational sampling methods based on parallel RE algorithms (Sugita and Okamoto, 1999) can speed up by orders of magnitude the convergence of biased simulations. The key aspect of parallel RE algorithms as applied to alchemical calculations is that simulations at different values of  $\lambda$ , which are executed in parallel, periodically exchange  $\lambda$  values, thereby allowing conformational transitions to occur at the value  $\lambda$  at which they are more likely to do so and, by so doing, to achieve more efficient exploration of conformational space. Some binding-induced conformational changes are more likely to occur at large  $\lambda$ s when the interaction between the ligand and the receptor is stronger, while others, such as reorientation of the ligand as a whole, are more likely to occur at small  $\lambda$ s when motion is less restricted. With RE, both kinds of conformational changes occur more easily in each individual replica causing a larger variety of conformations to appear at each  $\lambda$ , as opposed to, for example, conventional MD at fixed  $\lambda=1$  which is likely to explore only one or at most few conformations. Methods such as RETI (Woods et al., 2003a), FEP/REMD (Jiang et al., 2009), and BEDAM (Gallicchio et al., 2010) are examples of binding free energy methods that employ this  $\lambda$ -hopping strategy (Gallicchio and Levy, 2011).

### G. *Mining Minima*

Unlike the thermodynamic path methods discussed above, the mining minima (MM) binding free energy method (Chang and Gilson, 2004) is one of two examples of end point methods (the other being the MM/PBSA method below) that will be discussed in this review. The MM free energy estimator is unique in that it does not rely on MD/MC importance sampling of conformations. Instead, the method estimates configurational integrals by unweighted sampling of conformations around a set of selected low-energy states of the molecular system (Head et al., 1997). This feature constitutes both the main advantage and the main limitation of the method. On one hand, MM does not suffer from slow rates of conformational transitions typical of importance sampling algorithms. On the other hand, this advantage is counterbalanced by the challenge of performing a

sufficiently complete enumeration of the important stable minima of the system. Consequently, the method has been applied with implicit solvation and it has been most useful in the study of association equilibria, such as host guest systems (Chang and Gilson, 2004; Chang et al., 2007; Rekharsky et al., 2007; Moghaddam et al., 2009), with manageable number of degrees of freedom. Applications to protein–ligand binding equilibria have been also recently reported (Chen and Foloppe, 2010).

MM seeks to compute the binding free energy in the implicit solvent representation by explicitly computing each of the configurational integrals  $Z_{\text{RL}}$  and  $Z_{\text{R+L}}$  in Eq. (22) and expressing the standard binding free energy in terms of the end point of the equilibrium as the difference of the free energies of the binding partners:

$$\Delta G_{\text{I}} = G_{\text{RL}} - (G_{\text{R}} + G_{\text{L}}), \quad (75)$$

where  $G_{\text{RL}}$  is the free energy of the complex and the binding partners, where

$$G_{\text{RL}} = -kT \ln Z_{\text{RL}} \quad (76)$$

and similarly for  $G_{\text{R}}$  and  $G_{\text{L}}$ . Given a set of minima  $j$ , located by conformational sampling (Chang and Gilson, 2003), the configurational partition function,  $Z \simeq \sum_j z_j$ , of each state is approximated as the sum of local configurational partition functions  $z_j$  corresponding to each minimum defined schematically as

$$z_j = \int_j dx e^{-\beta[U(x)+W(x)]}, \quad (77)$$

where  $x$  represents the system coordinates and the integral is considered limited to the macrostate in the vicinity of the minimum. Local integrals are then computed by normal mode analysis assuming harmonic behavior augmented by numerical treatment of anharmonic deviations (Chang et al., 2003; Chang and Gilson, 2004). As mentioned above, the validity of the MM approach has been confirmed in several numerical applications (Chang and Gilson, 2004; Chang et al., 2007; Rekharsky et al., 2007; Moghaddam et al., 2009; Chen and Foloppe, 2010).

The MM method leads naturally to the study of the enthalpic and entropic components of the binding affinity (Chang et al., 2007; Zhou and Gilson, 2009). As described in Section II.F.1, the binding free energy in the implicit solvent representation is decomposable into the change of

average effective potential energy  $\Delta U_{\text{eff}}$  and the change in configurational entropy  $\Delta S_{\text{conf}}$  (Eq. (45)). These can be expressed in terms of the average energies and entropies of the end point states computed as sums over minima. For example,

$$\langle U_{\text{eff}} \rangle = \sum_j p_j \langle U_{\text{eff}} \rangle_j, \quad (78)$$

where  $p_j = z_j/Z$  is the population of the macrostate corresponding to minimum  $j$  and  $\langle U_{\text{eff}} \rangle_j$  is its average potential energy. Similarly, it can be shown from Eq. (44) that the configurational entropy can be expressed as (Zhou and Gilson, 2009)

$$S_{\text{conf}} = \sum_j p_j S_j - k \sum_j p_j \ln p_j, \quad (79)$$

where  $S_j$  is the configurational entropy of macrostate  $j$ , which can be estimated from the harmonic approximation discussed above. From Eq. (79), we see that contributions to the configurational entropy of binding come from both narrowing of energy well (changes in  $S_j$  upon binding) and redistribution of populations among the stable states (the second term in r.h.s. of Eq. (79)), with both being important, and, often, determinant factors in ligand binding (Chang et al., 2007; Gilson and Zhou, 2007).

#### H. MM/PBSA and MM/GBSA Approaches

MM/PBSA method (Kollman et al., 2000; Gouda et al., 2003; Chong et al., 2009) and its *generalized Born* variant (MM/GBSA) are, like the MM method above, an example of an end point approach to the calculation of binding free energies. Unlike the MM method, however, it is based on MD to sample conformational space. MD, like any other importance sampling-based method, is not suitable for computing directly configurational integrals, as in the MM method. Instead, MM-PBSA computes the binding free energy from using the enthalpy/entropy decomposition approach (Eq. (45)) with implicit solvation (the Poisson-Boltzmann (PB) model for MM/PBSA (Baker, 2005) and the generalized Born (GB) model for MM/GBSA (Bashford and Case, 2000; Chen et al., 2008)). In principle, a decomposition of this kind also applies to explicit representations of the

solvent (see e.g., Eqs. (33) and (36)); however, given the challenge of converging entropy and enthalpy changes with explicit solvation (Levy and Gallicchio, 1998), in practice, the method is limited to implicit solvent representations.

In MM/PBSA, the enthalpic term  $\Delta U_{\text{eff}}$  is computed as the difference between the average total potential energies in the bound and unbound states, collected from MD trajectories of the free ligand, free receptor, and their complex, which can be obtained from either explicit or implicit solvent MD simulations. The same approaches discussed above in the context of the MM method are applicable to the calculation of configurational binding entropies. So, while in principle, MM/PBSA is a rigorous formulation of the free energy of binding limited in principle only by the accuracy of the potential energy model, in practice, MM/PBSA applications have implemented the theory with varying degree of rigor.

Partly due to the limited extent of conformational sampling afforded by MD, the change in configurational entropy is often estimated from one of few conformational macrostates (Kollman et al., 2000; Foloppe and Hubbard, 2006) possibly neglecting contributions to the entropy change resulting from changes in populations of stable states (Eq. (79)). The quasiharmonic approximation (Levy et al., 1984) has also been employed to estimate the configurational entropy change; however, its accuracy for systems with multiple occupied energy wells has been questioned (Chang et al., 2005; Lee and Olson, 2006). In some MM/PBSA applications, the entropic terms have been neglected (Brown and Muchmore, 2007).

Difficulties in converging potential energy differences due to noise originating from the bulk of receptor–receptor interactions have led to single-trajectory approaches (Lee and Olson, 2006; Brown and Muchmore, 2007) in which the conformational ensembles for the free ligand and receptor are taken from the ensemble of the bound complex. This effectively replaces  $\Delta U_{\text{eff}}$  in Eq. (40) with the average binding energy  $\langle u \rangle_{\text{RL}}$  neglecting therefore reorganization energy contributions (Eq. (48)). When, in addition, entropic effects are neglected, the binding free energy is equated to the average binding energy (Brown and Muchmore, 2006). At this level of theory, all entropic and reorganization effects are neglected potentially leading to gross overestimation of binding affinities and lack of ability to discriminate binders from nonbinders (Gallicchio et al., 2010).

### *I. Studies of Ligand and Receptor Reorganization*

The binding free energy (Eq. (46)) is often the result of a large cancellation between the favorable work,  $\langle u \rangle_{\text{RL}}$ , of forming receptor–ligand interactions and the unfavorable work  $\Delta G_{\text{reorg}}$  to localize and reorganize the conformational ensembles of the ligand and receptor to their bound conformational states. While drug design is often concerned with strengthening receptor–ligand interactions, the reorganization component can play a fundamental role in regulating binding specificity in cases where variations of binding energies  $\langle u \rangle_{\text{RL}}$  are expected to be small. In such cases, optimization of binding affinity can proceed by strategies aimed at preorganizing the ligand for binding, that is by minimizing  $\Delta G_{\text{reorg}}$ .

For example, reorganization has been successfully used as the design principle for the optimization of the presentation of HIV epitopes for vaccine development (Lapelosa et al., 2010). This particular application was concerned with identifying modes of display of an HIV epitope on the surface of a rhinovirus vaccine vehicle in such a way that it would bind strongly to a known neutralizing antibody. Because the displayed epitope needs to necessarily reproduce the interaction of the antibody with HIV target, the binding interface between the epitope and the antibody is biologically restrained. In thermodynamic terms, the binding energy can be regarded as fixed and therefore preorganization of the epitope to the bound conformation is the only viable route for optimizing the binding affinity. Based on these reorganization concepts, molecular simulations were conducted which identified those presentation constructs with the highest fraction of epitope conformations compatible with antibody complexation (Lapelosa et al., 2009). Subsequent biochemical work confirmed the computational prediction and, remarkably, yielded some of the most antigenic vaccine constructs of this kind to date (Lapelosa et al., 2010).

In another recent example (DeLorbe et al., 2009), optimization of a class of inhibitors was achieved by chemical rigidification of the ligands into their bound conformations. In this case, structural analysis indicated that enhanced binding was indeed solely due to smaller reorganization penalties rather than stronger receptor–ligand interactions. Interestingly, in this work, it was regarded as paradoxical the fact that enhanced binding was not due to a reduced entropic penalty as expected, but rather to a more favorable enthalpic gain. However, this should not be regarded as surprising considering that (see Eq. (47)) reorganization has both entropic and enthalpic signatures.



Evidently, before rigidification, the ligands had to surmount an energetic penalty to form their bound conformations from their predominant solution conformations. The rigidified ligands instead did not suffer this penalty to the same extent, resulting in a more favorable binding enthalpy.

A number of recent studies have focused on ligand reorganization, which is simpler to model than receptor reorganization. Both [Yang et al. \(2009\)](#) and, on a more extensive set of systems, [Gao et al. \(2010\)](#) observed better correlation with experimental affinities when single-trajectory MM/GBSA scores were combined with ligand reorganization free energy estimates. As discussed above, the single-trajectory MM/GBSA model approximates the binding free energy with the ligand–receptor average binding energy,  $\langle u \rangle_{\text{RL}}$ , which, although easier to converge, omits sometimes critical reorganization free energy components (Eq. (46)). By introducing the ligand reorganization free energy, some of these effects are recaptured without substantially compromising the quality of the convergence, as most of the fluctuations in the MM/GBSA estimators come from the much more numerous degrees of freedom of the receptor. The ligand reorganization is defined as the sum of the ligand reorganization defined as (see Eq. (48))

$$\Delta U_{\text{reorg}}(\mathbf{L}) = \langle U_{\text{eff}}(x_{\text{L}}) \rangle_{\text{RL}} - \langle U_{\text{eff}}(x_{\text{L}}) \rangle_{\text{R+L}}, \quad (80)$$

and the change of ligand configuration entropy  $-T\Delta S_{\text{conf}}(\mathbf{L})$ . The latter is evaluated using the harmonic and quasiharmonic approaches discussed above. [Gao et al. \(2010\)](#) adopted a particularly rigorous entropic model incorporating both multiple minima (Eq. (79)) and anharmonic corrections ([Kolossvary, 1997](#); [Chang and Gilson, 2004](#)). It has been recently confirmed ([Okumura et al., 2010](#)) that MD sampling aided by temperature RE can also be used to accurately compute ligand reorganization free energies. Interestingly, it is observed ([Yang et al., 2009](#)) that the ligand configurational entropy does not always oppose binding. In a number of cases, there is a gain of entropy (positive  $\Delta S_{\text{conf}}(\mathbf{L})$ ) counterbalanced by an unfavorable reorganization energy. The same conclusion is suggested by the experimental work of [DeLorbe et al. \(2009\)](#) discussed above. This phenomenon might be quite general as it is known ([Perola and Charifson, 2004](#)) that ligands tend to form more extended, and possibly more flexible, conformations when bound to the receptor ([Perola and Charifson, 2004](#)) than in solution, where hydrophobicity causes them to adopt more compact conformations.

Binding modeling studies explicitly incorporating receptor reorganization effects are also beginning to appear. Major challenges exist due to the size of conformational space and the rarity of conformational transitions. Some recent studies have focused on the role of protein side-chain motion. [Mobley et al. \(2007a\)](#) have introduced a confine and release method to model the free energy associated with the conformation variability of a selected set of side chains in the binding site region. The technique consists of evaluating the binding free energy with the receptor side chains placed in various rotamer states. These are then combined, based on Eq. (51), with the free energy differences between rotamer states with and without the ligand present to yield the total binding free energy. In a number of cases, it was shown that including these terms improved the accuracy of binding affinity predictions ([Mobley et al., 2007a,b](#); [Boyce et al., 2009](#)) Similarly, a two-dimensional Hamiltonian RE FEP approach has been proposed to soften side-chain torsional barriers ([Jiang and Roux, 2010](#)).

#### IV. CONCLUSIONS

The accurate estimation of protein–ligand affinities remains one of the most difficult problems in computational biophysics. Atomistic free energy models of binding are progressively improving and will continue to represent important tools to further our understanding of molecular recognition phenomena and contribute to pharmaceutical research. Better potential models, more efficient computational algorithms, and faster computers are driving this progress forward. As this is happening, it is important that the relationships between theory and calculations remain clear and well understood. We have reviewed the statistical mechanics theory of binding, and we have shown how current computational methods and applications relate to the fundamental theory. These models have different features and limitations, and their ranges of applicability vary correspondingly. Yet their origins can all be traced back to a single fundamental theory. It is our hope that finding these commonalities will be useful to novices and experts alike to help them navigate the expanding universe of binding free energy methodologies and find novel ways to use them to study complex molecular recognition problems.

## ACKNOWLEDGMENT

This work has been supported in part by a research grant from the National Institute of Health (GM30580).

## REFERENCES

- Baker, N. A. (2005). Improving implicit solvent simulations: a poisson-centric view. *Curr. Opin. Struct. Biol.* **15**, 137–143.
- Barbieri, C. M., Kaul, M., Pilch, D. S. (2007). Use of 2-aminopurine as a fluorescent tool for characterizing antibiotic recognition of the bacterial rRNA A-site. *Tetrahedron* **63**(17), 3567–3574.
- Bashford, D., Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129–152.
- Beck, T. L., Paulaitis, M. E., Pratt, L. R. (2006). The Potential Distribution Theorem and Models of Molecular Solutions. Cambridge University Press, New York.
- Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22**(2), 245–268.
- Boresch, S., Tettinger, F., Leitgeb, M., Karplus, M. (2003). Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B* **107**(35), 9535–9551.
- Boyce, S. E., Mobley, D. L., Rocklin, G. J., Graves, A. P., Dill, K. A., Shoichet, B. K. (2009). Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.* **394**(4), 747–763.
- Brooijmans, N., Kuntz, I. D. (2003). Molecular recognition and docking algorithm. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 335–373.
- Brown, S. P., Muchmore, S. W. (2006). High-throughput calculation of protein-ligand binding affinities: modification and adaptation of the MM-PBSA protocol to enterprise grid computing. *J. Chem. Inf. Model.* **46**(3), 999–1005.
- Brown, S. P., Muchmore, S. W. (2007). Rapid estimation of relative protein-ligand binding affinities using a high-throughput version of MM-PBSA. *J. Chem. Inf. Model.* **47**(4), 1493–1503.
- Bussi, G., Gervasio, F. L., Laio, A., Parrinello, M. (2006). Free-energy landscape for  $\beta$  hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **128**(41), 13435–13441.
- Chang, C.-E., Gilson, M. K. (2003). Tork: conformational analysis method for molecules and complexes. *J. Comput. Chem.* **24**(16), 1987–1998.
- Chang, C.-E., Gilson, M. K. (2004). Free energy, entropy, and induced fit in host-guest recognition: calculations with the second-generation mining minima algorithm. *J. Am. Chem. Soc.* **126**(40), 13156–13164.
- Chang, C.-E., Potter, M. J., Gilson, M. K. (2003). Calculation of molecular configuration integrals. *J. Phys. Chem. B* **107**(4), 1048–1055.
- Chang, C.-E., Chen, W., Gilson, M. K. (2005). Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory Comput.* **1**(5), 1017–1028.

- Chang, C.-e.A., Chen, W., Gilson, M. K. (2007). Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. USA* **104**(5), 1534–1539.
- Chen, I.-J., Foloppe, N. (2010). Drug-like bioactive structures and conformational coverage with the LigPrep/ConfGen suite: comparison to programs MOE and catalyst. *J. Chem. Inf. Model.* **50**(5), 822–839.
- Chen, J., Brooks, C. L., III, Khandogin, J. (2008). Recent advances in implicit solvent based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **18**, 140–148.
- Chipot, C. and Pohorille, A. (Eds.) (2007). Free Energy Calculations. Theory and Applications in Chemistry and Biology. Springer Series in Chemical Physics. Springer, Berlin, Heidelberg.
- Chodera, J. D., Mobley, D. L., Shirts, M. R., Dixon, R. W., Branson, K., Pande, V. S. (2011). Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.* **21**(2), 150–160.
- Chong, L. T., Pitera, J. W., Swope, W. C., Pande, V. S. (2009). Comparison of computational approaches for predicting the effects of missense mutations on p53 function. *J. Mol. Graph. Model.* **27**(8), 978–982.
- Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**(19), 5179–5197.
- DeLorbe, J. E., Clements, J. H., Teresk, M. G., Benfield, A. P., Plake, H. R., Millsbaugh, L. E., et al. (2009). Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradox associated with ligand preorganization. *J. Am. Chem. Soc.* **131**(46), 16758–16770.
- Deng, Y., Roux, B. (2006). Calculation of standard binding free energies: aromatic molecules in the t4 lysozyme 199a mutant. *J. Chem. Theory Comput.* **2**(5), 1255–1273.
- Deng, Y., Roux, B. (2009). Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **113**(8), 2234–2246.
- Eriksson, A. E., Baase, W. A., Wozniak, J. A., Matthews, B. W. (1992). A cavity-containing mutant of t4 lysozyme is stabilized by buried benzene. *Nature* **355**(6358), 371–373.
- Ferrenberg, A. M., Swendsen, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63**, 1195–1198.
- Foloppe, N., Hubbard, R. (2006). Towards predictive ligand design with free-energy based computational methods? *Curr. Med. Chem.* **13**(29), 3583–3608.
- Fujitani, H., Tanida, Y., Ito, M., Jayachandran, G., Snow, C. D., Shirts, M. R., et al. (2005). Direct calculation of the binding free energies of FKBP ligands. *J. Chem. Phys.* **123**(8), 084108.
- Gallicchio, E., Andrec, M., Felts, A. K., Levy, R. M. (2005). Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B* **109**, 6722–6731.
- Gallicchio, E., Paris, K., Levy, R. M. (2009). The agbnp2 implicit solvation model. *J. Chem. Theory Comput.* **5**(9), 2544–2564.
- Gallicchio, E., Lapelosa, M., Levy, R. M. (2010). Binding energy distribution analysis method (BEDAM) for estimation of protein-ligand binding affinities. *J. Chem. Theory Comput.* **6**(9), 2961–2977.

- Galicchio, E., Levy, R. M. (2011). Advances in all atom sampling methods for modeling protein-ligand binding affinities. *Curr. Op. Struct. Biol.* **21**, 161–166.
- Gao, C., Park, M.-S., Stern, H. A. (2010). Accounting for ligand conformational restriction in calculations of protein-ligand binding affinities. *Biophys. J.* **98**(5), 901–910.
- Ge, X., Roux, B. (2010). Absolute binding free energy calculations of sparsomycin analogs to the bacterial ribosome. *J. Phys. Chem. B* **114**(29), 9525–9539.
- Gilson, M. K., Zhou, H.-X. (2007). Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42.
- Gilson, M. K., Given, J. A., Bush, B. L., McCammon, J. A. (1997). The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **72**, 1047–1069.
- Gouda, H., Kuntz, I. D., Case, D. A., Kollman, P. A. (2003). Free energy calculations for theophylline binding to an RNA aptamer: comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers* **68**(1), 16–34.
- Graves, A. P., Brenk, R., Shoichet, B. K. (2005). Decoys for docking. *J. Med. Chem.* **48**(11), 3714–3728.
- Groot, R. D. (1992). The association constant of a flexible molecule and a single atom: theory and simulation. *J. Chem. Phys.* **97**(5), 3537–3549.
- Guvench, O., Mackerell, A. D. (2009). Computational evaluation of protein-small molecule binding. *Curr. Opin. Struct. Biol.* **19**(1), 56–61.
- Head, M. S., Given, J. A., Gilson, M. K. (1997). Mining minima: direct computation of conformational free energy. *J. Phys. Chem. A* **101**(8), 1609–1618.
- Holt, D. A., Luengo, J. I., Yamashita, D. S., Oh, H. J., Konialian, A. L., Yen, H. K., et al. (1993). Design, synthesis, and kinetic evaluation of high-affinity FKBP ligands and the X-ray crystal structures of their complexes with FKBP 12. *J. Am. Chem. Soc.* **115**(22), 9925–9938.
- Hritz, J., Oostenbrink, C. (2008). Hamiltonian replica exchange molecular dynamics using soft-core interactions. *J. Chem. Phys.* **128**(14), 144121.
- Jayachandran, G., Shirts, M. R., Park, S., Pande, V. S. (2006). Parallelized-over-parts computation of absolute binding free energy with docking and molecular dynamics. *J. Chem. Phys.* **125**(8), 084901.
- Jiang, W., Roux, B. (2010). Free energy perturbation Hamiltonian replica-exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *J. Chem. Theory Comput.* **6**, 2559–2565.
- Jiang, W., Hodoscek, M., Roux, B. (2009). Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics. *J. Chem. Theory Comput.* **5**(10), 2583–2588.
- Jiao, D., Golubkov, P. A., Darden, T. A., Ren, P. (2008). Calculation of protein-ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci. USA* **105**(17), 6290–6295.
- Jiao, D., Zhang, J., Duke, R. E., Li, G., Schnieders, M. J., Ren, P. (2009). Trypsin-ligand binding free energies from explicit and implicit solvent simulations with polarizable potential. *J. Comput. Chem.* **30**(11), 1701–1711.
- Jorgensen, W. L. (1989). Interactions between amides in solution and the thermodynamics of weak binding. *J. Am. Chem. Soc.* **111**(10), 3770–3771.

- Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science* **303** (5665), 1813–1818.
- Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**(6), 724–733.
- Jorgensen, W. L., Thomas, L. L. (2008). Perspective on free-energy perturbation calculations for chemical equilibria. *J. Chem. Theory Comput.* **4**(6), 869–876.
- Jorgensen, W. L., Buckner, J. K., Boudon, S., Tirado-Rives, J. (1988). Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J. Chem. Phys.* **6**, 3742.
- Jorgensen, W. L., Maxwell, D. S., Tirado-Rives, J. (1996). Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225–11236.
- Khavrutskii, I. V., Wallqvist, A. (2010). Computing relative free energies of solvation using single reference thermodynamic integration augmented with Hamiltonian replica exchange. *J. Chem. Theory Comput.* **6**(11), 3427–3441.
- Knight, J. L., Brooks, C. L. (2009). Lambda-dynamics free energy simulation methods. *J. Comput. Chem.* **30**(11), 1692–1700.
- Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**(12), 889–897.
- Kolossvary, I. (1997). Evaluation of the molecular configuration integral in all degrees of freedom for the direct calculation of conformational free energies: prediction of the anomeric free energy of monosaccharides. *J. Phys. Chem. A* **101**(51), 9900–9905.
- Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A., Rosenberg, J. M. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011–1021.
- Lapelosa, M., Gallicchio, E., Arnold, G. F., Arnold, E., Levy, R. M. (2009). In silico vaccine design based on molecular simulations of rhinovirus chimeras presenting hiv-1 gp41 epitopes. *J. Mol. Biol.* **385**(2), 675–691.
- Lapelosa, M., Arnold, G. F., Gallicchio, E., Arnold, E., Levy, R. M. (2010). Antigenic characteristics of rhinovirus chimeras designed in silico for enhanced presentation of HIV-1 gp41 epitopes. *J. Mol. Biol.* **397**(3), 752–766.
- Lazaridis, T., Karplus, M. (1999). Effective energy function for protein in solution. *Proteins* **35**, 133–152.
- Lee, M. S., Olson, M. A. (2006). Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. *Biophys. J.* **90**(3), 864–877.
- Levy, R. M., Gallicchio, E. (1998). Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.* **49**, 531–567.
- Levy, R. M., Karplus, M., Kushick, J., Perahia, D. (1984). Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an  $\alpha$ -helix. *Macromolecules* **17**, 1370–1374.
- Liu, P., Kim, B., Friesner, R. A., Berne, B. J. (2005). Replica exchange with solute tempering: a method for sampling biological systems in explicit solvent. *Proc. Natl. Acad. Sci. USA* **102**, 13749–13754.

- Liu, P., Huang, X., Zhou, R., Berne, B. J. (2006). Hydrophobic aided replica exchange: an efficient algorithm for protein folding in explicit solvent. *J. Phys. Chem. B* **110** (38), 19018–19022.
- Lu, N., Kofke, D. A. (2001). Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling. *J. Chem. Phys.* **114**(17), 7303–7311.
- Lu, N., Singh, J. K., Kofke, D. A. (2003). Appropriate methods to combine forward and reverse free-energy perturbation averages. *J. Chem. Phys.* **118**(7), 2977–2984.
- Luo, H., Sharp, K. (2002). On the calculation of absolute macromolecular binding free energies. *Proc. Natl. Acad. Sci. USA* **99**(16), 10399–10404.
- MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**(18), 3586–3616.
- McInnes, C. (2007). Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.* **11**(5), 494–502.
- Meng, Y., Roitberg, A. E. (2010). Constant pH replica exchange molecular dynamics in biomolecules using a discrete protonation model. *J. Chem. Theory Comput.* **6**(4), 1401–1412.
- Michel, J., Essex, J. W. (2010). Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided Mol. Des.* **24**(8), 639–658.
- Mihailescu, M., Gilson, M. K. (2004). On the theory of noncovalent binding. *Biophys. J.* **87**(1), 23–36.
- Mitsutake, A., Mori, Y., Okamoto, Y. (2010). Multi-dimensional multicanonical algorithm, simulated tempering, replica-exchange method, and all that. *Phys. Procedia* **4**, 89–105.
- Miyamoto, S., Kollman, P. A. (1993). Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins* **16**(3), 226–245.
- Mobley, D. L., Dill, K. A. (2009). Binding of small-molecule ligands to proteins: “what you see” is not always “what you get”. *Structure* **17**(4), 489–498.
- Mobley, D. L., Chodera, J. D., Dill, K. A. (2006). On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *J. Chem. Phys.* **125** (8), 084902.
- Mobley, D. L., Chodera, J. D., Dill, K. A. (2007a). The confine-and-release method: obtaining correct binding free energies in the presence of protein conformational change. *J. Chem. Theory Comput.* **3**(4), 1231–1235.
- Mobley, D. L., Graves, A. P., Chodera, J. D., McReynolds, A. C., Shoichet, B. K., Dill, K. A. (2007b). Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.* **371**(4), 1118–1134.
- Moghaddam, S., Inoue, Y., Gilson, M. K. (2009). Host-guest complexes with protein-ligand-like affinities: computational analysis and design. *J. Am. Chem. Soc.* **131**(11), 4012–4021.
- Morton, A., Baase, W. A., Matthews, B. W. (1995). Energetic origins of specificity of ligand binding in an interior nonpolar cavity of t4 lysozyme. *Biochemistry* **34**(27), 8564–8575.

- Murata, K., Sugita, Y., Okamoto, Y. (2004). Free energy calculations for DNA base stacking by replica-exchange umbrella sampling. *Chem. Phys. Lett.* **385**(1–2), 1–7.
- Neale, C., Rödinger, T., Pomès, R. (2008). Equilibrium exchange enhances the convergence rate of umbrella sampling. *Chem. Phys. Lett.* **460**(1–3), 375–381.
- Okumura, H., Gallicchio, E., Levy, R. M. (2010). Conformational populations of ligand-sized molecules by replica exchange molecular dynamics and temperature reweighting. *J. Comput. Chem.* **31**, 1357–1367.
- Oostenbrink, C., van Gunsteren, W. F. (2005). Free energies of ligand binding for structurally diverse compounds. *Proc. Natl. Acad. Sci. USA* **102**(19), 6750–6754.
- Payne, V. A., Matubayasi, N., Reed Murphy, L., Levy, R. M. (1997). Monte Carlo study of the effect of pressure on hydrophobic association. *J. Phys. Chem. B* **101**, 2054–2060.
- Perola, E., Charifson, P. S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **47**(10), 2499–2510.
- Piana, S., Laio, A. (2007). A bias-exchange approach to protein folding. *J. Phys. Chem. B* **111**(17), 4553–4559.
- Pohorille, A., Pratt, L. R. (1990). Cavities in molecular liquids and the theory of hydrophobic solubilities. *J. Am. Chem. Soc.* **112**(13), 5066–5074.
- Pohorille, A., Jarzynski, C., Chipot, C. (2010). Good practices in free-energy calculations. *J. Phys. Chem. B* **114**(32), 10235–10253.
- Reddy, M. R. and Erion, M. D. (Eds.) (2001). *Free Energy Calculations in Rational Drug Design*. Springer-Verlag, New York.
- Rekharsky, M. V., Mori, T., Yang, C., Ko, Y. H., Selvapalam, N., Kim, H., et al. (2007). A synthetic host-guest system achieves avidin-biotin affinity by overcoming enthalpy-entropy compensation. *Proc. Natl. Acad. Sci. USA* **104**(52), 20737–20742.
- Roitberg, A. E., Okur, A., Simmerling, C. (2007). Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. *J. Phys. Chem. B* **111**(10), 2415–2418.
- Roux, B., Simonson, T. (1999). Implicit solvent models. *Biophys. Chem.* **78**, 1–20.
- Schuler, L. D., Daura, X., van Gunsteren, W. F. (2001). An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* **22**(11), 1205–1218.
- Serdyuk, I. N., Zaccai, N. R., Zaccai, G. (2007). *Methods in Molecular Biophysics: Structure, Dynamics, Function* Cambridge University Press, Cambridge, New York.
- Shirts, M. R., Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**(12), 124105.
- Shirts, M. R., Mobley, D. L., Chodera, J. D. (2007). Alchemical free energy calculations: ready for prime time? *Annu. Rep. Comput. Chem.* **3**, 41–59.
- Shirts, M. R., Mobley, D. L., Brown, S. P. (2010). *Drug Design—Structure- and Ligand-Based Approaches Chapter Free-Energy Calculations in Structure-Based Drug Design* Cambridge University Press. pp. 61–86, New York.
- Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature* **432**(7019), 862–865.



- Steinbrecher, T., Mobley, D. L., Case, D. A. (2007). Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* **127**(21), 214108.
- Sugita, Y., Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151.
- Swanson, J. M. J., Henchman, R. H., McCammon, J. A. (2004). Revisiting free energy calculations: a theoretical connection to mm/pbsa and direct calculation of the association free energy. *Biophys. J.* **86**(1 Pt. 1), 67–74.
- Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *J. Am. Stat. Assoc.* **99**(468), 1027–1036.
- Tembe, B. L., McCammon, J. A. (1984). Ligand-receptor interactions. *Comput. Chem.* **8**(4), 281.
- Wagoner, J., Baker, N. A. (2006). Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. USA* **103**, 8331–8336.
- Wang, J., Deng, Y., Roux, B. (2006). Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **91**(8), 2798–2814.
- Wei, B. Q., Baase, W. A., Weaver, L. H., Matthews, B. W., Shoichet, B. K. (2002). A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **322**(2), 339–355.
- Widom, B. (1963). Some topics in the theory of fluids. *J. Chem. Phys.* **39**(11), 2808–2812.
- Widom, B. (1982). Potential-distribution theory and the statistical mechanics of fluids. *J. Phys. Chem.* **86**(6), 869–872.
- Woo, H.-J., Roux, B. (2005). Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA* **102**(19), 6825–6830.
- Woods, C. J., Essex, J. W., King, M. A. (2003a). The development of replica-exchange-based free-energy methods. *J. Phys. Chem. B* **107**(49), 13703–13710.
- Woods, C. J., Essex, J. W., King, M. A. (2003b). Enhanced configurational sampling in binding free-energy calculations. *J. Phys. Chem. B* **107**(49), 13711–13718.
- Yang, C.-Y., Sun, H., Chen, J., Nikolovska-Coleska, Z., Wang, S. (2009). Importance of ligand reorganization free energy in protein-ligand binding-affinity prediction. *J. Am. Chem. Soc.* **131**(38), 13709–13721.
- Yeh, I.-C., Olson, M. A., Lee, M. S., Wallqvist, A. (2008). Free-energy profiles of membrane insertion of the m2 transmembrane peptide from influenza a virus. *Biophys. J.* **95**(11), 5021–5029.
- Zhou, H.-X., Gilson, M. K. (2009). Theory of free energy and entropy in non-covalent binding. *Chem. Rev.* **109**(9), 4092–4107.
- Zhou, Z., Felts, A. K., Friesner, R. A., Levy, R. M. (2007). Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model.* **47**(4), 1599–1608.
- Zwanzig, R. W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **22**(8), 1420–1426.

# HYBRID SCHEMES BASED ON QUANTUM MECHANICS/ MOLECULAR MECHANICS SIMULATIONS: GOALS TO SUCCESS, PROBLEMS, AND PERSPECTIVES

By SILVIA FERRER,<sup>\*</sup> JAVIER RUIZ-PERNÍA,<sup>\*</sup> SERGIO MARTÍ,<sup>\*</sup> VICENT MOLINER,<sup>\*</sup>  
IÑAKI TUÑÓN,<sup>†</sup> JUAN BERTRÁN,<sup>\*</sup> AND JUAN ANDRÉS<sup>\*</sup>

<sup>\*</sup>Departamento de Química Física y Analítica, Universitat Jaume I, Castellón, Spain

<sup>†</sup>Departamento de Química Física, Universidad de Valencia,

Dr. Moliner 50, Burjassot, Valencia, Spain

<sup>‡</sup>Departamento de Química, Universidad Autónoma de Barcelona, Bellaterra, Spain

I. Introduction: State of Art.....	83
A. Quantum Mechanics/Molecular Mechanics (QM/MM) Approaches.....	91
II. Potential of Mean Force/Free-Energy Calculations.....	97
III. Applications.....	98
A. Chorismate Mutase as a Working Example.....	98
B. Designing CAs: TSA.....	105
C. Toward Understanding of the Promiscuity in Enzyme Catalysis.....	113
IV. Conclusions and Outlook.....	120
References.....	129

## ABSTRACT

The development of characterization techniques, advanced synthesis methods, as well as molecular modeling has transformed the study of systems in a well-established research field. The current research challenges in biocatalysis and biotransformation evolve around enzyme discovery, design, and optimization. How can we find or create enzymes that catalyze important synthetic reactions, even reactions that may not exist in nature? What is the source of enzyme catalytic power? To answer these and other related questions, the standard strategies have evolved from trial-and-error methodologies based on chemical knowledge, accumulated experience, and common sense into a clearly multidisciplinary science that allows one to reach the molecular design of tailor-made enzyme catalysts. This is even more so when one refers to enzyme catalysts, for which the detailed structure and composition are known and can be manipulated to introduce well-defined residues which can be implicated in the chemical rearrangements taking place in the active site. The methods and techniques of theoretical and computational chemistry are becoming more and

more important in both understanding the fundamental biological roles of enzymes and facilitating their utilization in biotechnology. Improvement of the catalytic function of enzymes is important from scientific and industrial viewpoints, and to put this fact in the actual perspective as well as the potentialities, we recommend the very recent report of Sanderson [Sanderson, K. (2011). Chemistry: enzyme expertise. *Nature* **471**, 397.].

Great fundamental advances have been made toward the *ab initio* design of enzyme catalysts based on molecular modeling. This has been based on the molecular mechanistic knowledge of the reactions to be catalyzed, together with the development of advanced synthesis and characterization techniques. The corresponding molecular mechanism can be studied by means of powerful quantum chemical calculations. The catalytic active site can be optimized to improve the transition state analogues (TSA) and to enhance the catalytic activity, even improve the active site to favor a desired direction of some promiscuous enzymes. In this chapter, we give a brief introduction, the state of the art, and future prospects and implications of enzyme design. Current computational tools to assist experimentalists for the design and engineering of proteins with desired catalytic properties are described. The interplay between enzyme design, molecular simulations, and experiments will be presented to emphasize the interdisciplinary nature of this research field. This text highlights the recent advances and examples selected from our laboratory are shown, of how the applications of these tools are a first attempt to *de novo* design of protein active sites. Identification of neutral/advantageous/deleterious mutation platforms can be exploited to penetrate some of Nature's closely guarded secrets of chemical reactivity.

In this chapter, we give a brief introduction, the state of the art, and future prospects and implications of enzyme design. The first part describes briefly how the molecular modeling is carried out. Then, we discuss the requirements of hybrid quantum mechanical/molecular mechanics molecular dynamics (QM/MM MD) simulations, analyzing what are the basis of these theoretical methodologies, how we can use them with a view to its application in the study of enzyme catalysis, and what are the best methodologies for assessing its catalytic potential. In the second part, we focus on some selected examples, taking as a common guide the chorismate to prephenate rearrangement, studying the corresponding molecular mechanism *in vacuo*, in solution and in an enzyme environment. In addition, examples involving catalytic antibodies (CAs) and promiscuous enzymes will be presented. Finally, a special emphasis is made to

provide some hints about the logical evolution that can be anticipated in this research field. Moreover, it helps in understanding the open directions in this area of knowledge and highlights the importance of computational approaches in discovering specific drugs and the impact on the rational design of tailor-made enzymes.

## I. INTRODUCTION: STATE OF ART

Catalysis is at the heart of almost every chemical transformation process, and it remains at the core of chemical research, with its far-reaching impact on both applied and basic research, being detailed understanding of the active species and their related reaction mechanism of great interest. This insight helps to refine the fundamentals of catalysis and, in a very practical manner, can help researchers to optimize existing catalyst formulations or develop completely new ones. The field can be subdivided into three areas of homogeneous, heterogeneous, and enzyme catalysis. Homogeneous catalysts operate in the same phase as the reactants, while heterogeneous catalysts are present in a phase different from that of the reactants; usually, the catalyst is a solid surface. Enzyme catalysts are specialized proteins. For both homogeneous and enzyme catalysis, it has been possible to reach molecular-scale insight into the structure of the active site and the reaction mechanism for a multitude of catalysts and chemical reactions (Herrmann and Cornils, 2002; Ferreira et al., 2004; Garcia-Viloca et al., 2004; Siegbahn et al., 2007). This has been achieved by combining various structural characterization techniques, kinetic investigations, and computational studies for each system.

Enzymes are molecular machines which catalyze chemical reactions in living organisms. They can be considered as highly evolved machinery developed by nature, they catalyze reactions with formidable efficiency and specificity under mild conditions. Enzymes become a source of inspiration to chemists, demonstrating what could be achieved with a full understanding of the underlying principles of nature, and they have long provided a stimulus for researchers to make artificial equivalents. Our quest to understand the physical basis of this catalytic power, which is pivotal to our understanding of biological reactions and our exploitation of enzymes in chemical, biomedical, and biotechnological processes, is challenging and has involved sustained and intensive research efforts for

more than 100 years (for reviews see, e.g., Cannon and Benkovic, 1998; Cleland et al., 1998; Neet, 1998; Warshel, 1998; Herrmann and Cornils, 2002; Benkovic and Hammes-Schiffer, 2003; Ferreira et al., 2004; Garcia-Viloca et al., 2004; Siegbahn et al., 2007). Considerable efforts have been devoted for several decades on the development of enzyme-like catalysts with tailored properties by rationally manipulating natural and artificially synthesized biomolecules. One of the great challenges is to design artificial systems with catalytic efficiencies and specificities rivaling natural components. It is of great scientific interest and practical need to construct enzymes with new catalytic properties and enhanced stabilities.

The transition state (TS) is of strategic importance within the field of chemical reactivity, and it has been remarked recently by Williams (2010), “The key to understanding the fundamental processes of catalysis is the transition state (TS): indeed, catalysis is a transition-state molecular recognition event. Practical objectives, such as the design of TS analogues as potential drugs, or the design of synthetic catalysts (including catalytic antibodies), require prior knowledge of the TS structure to be mimicked.” Therefore, a challenging aspect of chemical reactions is the need to obtain an accurate description of the energy and the configuration of the TS.

Enzymes can achieve rate enhancements of up to 21 orders of magnitude relative to uncatalyzed reactions. To explain this enormous catalytic power, Pauling proposed the concept of transition-state stabilization, in which the role of the enzyme is to reduce the height of the potential-energy barrier that must be overcome for the reaction to occur (Pauling, 1948a,b) as it was quoted in his 1948 *New Scientist* article (Pauling, 1948a, b): “I believe that an enzyme has a structure closely similar to that found for antibodies, but with one important difference, namely that the surface configuration of the enzyme is not so closely contemporary to its specific substrate as is that of an antibody to its homologous antigen, but is instead complementary to a unstable molecule with only transient existence—namely the ‘activated complex’ for the reaction that is catalyzed by the enzyme. The mode of action of an enzyme would then be the following: the enzyme would show a small power of attraction for the substrate molecule or molecules, which would become attached to it in its active surface region. This substrate molecule, or these molecules, would then be strained by the forces of attraction to the enzyme, which would tend to deform it into the configuration of the activated complex, for which the power of attraction of the enzyme is the greatest.”

The concept of transition-state stabilization implies that enzymes have evolved active-site structures that are more complementary to the TS than the ground state of their cognate substrates. Now, it has been possible to test the details of active-site interaction directly through site-specific mutagenesis, which historically has been focused on active-site residues that function in acid–base catalysis, electrostatic stabilization, H-bonding, and so on; electrostatic stabilization of the TS is the predominate mechanism for tighter binding of the TS than the substrate to the enzyme, although other contributions such as hydrophobic interactions would facilitate proper orientation of the substrate/TS in the active-site pocket, which is preorganized in effect by the structure of the enzyme.

Improvement of the catalytic function of enzymes is essential from scientific and industrial viewpoints. Enzymes are key targets for drug discovery, and they are increasingly used in industrial processes such as bioenergy production. Thus, it is important to understand how they achieve their remarkable efficiency. Computational studies offer an alternative method toward obtaining key structural as well as chemical reactivity information (Brent and Bruck, 2006). Computational modeling can be used to resolve the reaction mechanisms and analyzing the causes of catalysis (Garcia-Viloca et al., 2004; Mulholland, 2005; Warshel et al., 2006; Alexandrova et al., 2008; Lonsdale et al., 2010). The methods and techniques that are the basis of these simulations allow calculating the energy profile of the chemical reaction to be catalyzed, analyze unstable species in the reaction path, study alternative mechanism when a specific mutation is carried out, and calculate specific contributions to catalysis.

The use of molecular modeling techniques is capable to provide with the relevant amount of atomic information for the design and optimization of both natural and synthetic enzymes. Computational techniques are nowadays broadly used in many areas of chemistry and biology to understand and predict the atomic behavior of molecules. As it was remarked by Carter and Rosky (2006), “The detailed description from theory of the complex chemical processes driving the sequence of events in the molecular machines of biology and the design of those targeted by modern nanoscience is a reasonable goal. The expectation that an in-depth understanding of such complex systems is on the horizon is supported by recent history. At the outset of the 21st century, TCC has arrived at a position of central importance not only for theorists but also in the laboratories of most experimentalists and in many disciplines. These disciplines include

not only chemistry but also biochemistry, chemical engineering, molecular biology, biomedical engineering, geophysics, and materials science. The prevalence of molecular calculations via quantum chemistry and the models of molecular mechanics as guidance and support for experimental research is a result of the maturation of concepts, methods, and algorithms developed over many decades within theoretical Chemistry. Theoretical chemists have adapted their tools for use in industry and by experimentalists. It is then interesting to ask what new tools and deeper insights one might expect to be routinely accessible to researchers in the not too distant future.”

The use of computational techniques for the design of *de novo* proteins (Bolon and Mayo, 2001; Garcia-Viloca et al., 2004; Mulholland, 2005; Warshel et al., 2006; Alexandrova et al., 2008; Lodola et al., 2008; Lonsdale et al., 2010; Nanda and Koder, 2010) or metal binding proteins (Summa et al., 2002; Maglio et al., 2007; Lu et al., 2009) has already showed interesting applications. Quantum chemistry is an enormous field of study that consists of three main approaches: *ab initio* (Hehre et al., 1986; Szabo and Ostlund, 1996), density functional theory (DFT) (Parr and Yang, 1989; Koch and Holthausen, 2001), and semiempirical quantum mechanical methods (Elstner et al., 1998; Clark, 2000; Thiel, 2007). The main strength of quantum chemistry lies in its ability to quantitatively describe chemical structures, energetics, and reactions. Although experimental techniques such as X-ray crystallography, NMR, and other spectroscopic methods are critical for studying enzyme structure, they are sometimes unable to answer questions concerning detailed catalytic mechanisms. Computational approaches allow the direct assessment and characterization of the enzyme–substrate and enzyme–product complexes, as well as metastable intermediates and transition structures. Simulation also enables enzyme reaction energetics to be dissected into individual contributions, and numerous ways to analyze the data. Thus, computer simulation can provide important information which is complementary to experiments.

Enzymes are highly versatile and proficient catalysts. Optimized by Darwinian evolution over millions of years, they can greatly accelerate chemical reactions while ensuring high substrate specificity, as well as exquisite enantioselectivity and stereoselectivity. However, there are often significant discrepancies between an enzyme’s function in nature and the specific requirements for *ex vivo* applications envisioned by

scientists and engineers. Computer-based methods are becoming increasingly important and complementary to wet laboratory experiments in studying the structure and function of biomolecules. Molecular docking is a frequently used tool in structure-based rational drug design. Although early efforts were hindered by limited possibilities in computational resources, due to recent advances in high performance computing, virtual screening methods became more and more efficient. These methods contributed to the development of several drugs and drug candidates that advanced to clinical trials. Examples include lead compounds to prevent myocardial infarction, to treat HIV infection, Alzheimer's disease, rheumatoid arthritis, and many other diseases (Thomas, 2007; Clark, 2008; Jorgensen, 2009). Docking programs simulate how a target macromolecule (receptor, enzyme, or nucleic acid) interacts with small molecule ligands, such as substrates, inhibitors, or other drug candidates. To model the binding between the ligand and the target molecule, their known three-dimensional structures are superimposed and the fit between the key sites of the target molecule and the ligand is then analyzed.

Although structure-based computational simulations are useful tools in drug discovery (Alonso et al., 2006), enzyme function is closely linked to enzyme dynamics, and several techniques have been developed to probe this relationship (Karplus et al., 2005; Boehr et al., 2009; Ma and Nussinov, 2010). NMR, X-ray crystallography, single-molecule experiments, and simulations clearly demonstrate that the free enzyme dynamics already encompass all the conformations necessary for substrate binding, preorganization, transition-state stabilization, and product release.

Fast and inexpensive docking protocols combined with accurate but more costly molecular dynamics (MD) techniques are a logical approach to predict reliable protein–ligand complexes (Karplus and McCammon, 2002). The advantage of this combination lies in their complementary strengths and weaknesses, where docking is used to find the correct conformation of a ligand neglecting receptor flexibility, and MD simulation is then applied to optimize complex structures by treating both ligand and receptor in a flexible way. To obtain reliable docking results, three complementary docking methods, Autodock, FlexX, and Genetic Optimization for Ligand Docking (GOLD; Bursulaya et al., 2003; Kontoyianni et al., 2004), can be used. Presently, structure-based sequence comparison provides valuable hints for the rational optimization of inhibitors, by identifying less conserved surrounding residues and structural differences.



Enzyme engineering by directed evolution has become the strategy of choice for tailoring the catalytic, biophysical, and molecular recognition properties of target proteins (Lutz and Bornscheuer, 2009). The methods of directed evolution, based on several rounds of mutagenesis in combination with efficient screening or selection, have been particularly successful in this effort owing to the high complexity of protein structures and our limited understanding of the protein structure–function relationships. Long-term efforts to assist directed evolution in focusing on the regions in protein structures relevant for the function as well as to design enzymes *de novo* led to development of a large variety of computational tools.

The ultimate goal of enzyme engineering is a true rational design, which aims at *de novo* engineering of enzymes (Kaplan and DeGrado, 2004). The scope of this approach is obvious; instead of using experimental approaches that are time, money, and intensive resource, enzyme engineering can be performed entirely *in silico* using fast computational algorithms (Zanghellini et al., 2006). Computational *de novo* design relies on the introduction of amino acid residues essential for catalysis into existing scaffolds (Gerlt and Babbitt, 2009). The underlining idea, based on the seminal hypothesis of Pauling (Pauling, 1946, 1948a,b), is that enzymes enhance chemical reactions by lowering an activation barrier due to stabilization of the TS by the residues of the active site. This principle implies that all proteins capable of binding to the TS could function as enzymes. Pauling's concept forms the basis of a computational approach that has recently yielded several *de novo* enzymes (Damborsky and Brezovsky, 2009). Initially, TS of the reaction and the idealized active-site geometry is modeled using quantum mechanics. Libraries of protein scaffolds are then searched to identify potential binding pockets that bind tightly to the TS and retain the desired geometry of the functional groups. Using geometry-based identification, the TS is matched with the binding site and the position of the TS and the catalytic side chains is optimized. Finally, the remaining residues for tight binding of the TS are designed and the designs are ranked on the basis of TS binding energy and catalytic geometry. The first step in this approach is the generation of an *in silico* model of TS. Next, individual amino acids are positioned around it to create an active site that stabilizes the TS in a computational process that uses quantum mechanical calculations.

Computer packages have been pioneering to the *de novo* design of enzymes, such as the programs DEZYMER (Hellings and Richards,

1991), ORBIT (Dahiyat and Mayo, 1996), and ROSETTA (Zanghellini et al., 2006). In particular, Siegel et al. (2010) have used Rosetta methodology to computational design of enzyme catalysts for a stereoselective bimolecular Diels–Alder reaction. The design methodology starts from three-dimensional atomic models of minimal active sites (theozymes) consisting of the reaction TS and protein functional groups involved in binding and catalysis. In fact, some Web sites using internet resources, such as Folding@home (<http://folding.stanford.edu>) and Rossetta@home (<http://boinc.bakerlab.org/rosetta>), can be cited in this context. Various protein scaffolds are evaluated for their ability to accommodate the *de novo* active site using MM modeling software such as RosettaMatch (Zanghellini et al., 2006; Das and Baker, 2008; Jiang et al., 2008; Rothlisberger et al., 2008; Murphy et al., 2009). These scaffolds are generated by taking a high-resolution structure of different natural proteins and virtually removing the amino acid side chains from the ligand binding pocket. In the final step, the remaining amino acid side chains in the pocket are computationally redesigned for high substrate specificity and tight TS binding. In another case of computational design, Faiella et al. computationally not only designed the active site but also calculated the scaffold to accommodate it from first principles (Faiella et al., 2009).

In addition to the sequence and structure-based design strategies, quantum mechanical (QM) and MD calculations, as well as machine-learning algorithms, have become invaluable tools to effectively explore the impact of amino acid substitutions on protein structure and stability. Together, these concepts offer promising predictors for altering protein features such as substrate specificity, stereoselectivity, and stability by enzyme redesign (but leaving the catalytic machinery of the native biocatalyst intact), as well as the creation of new function by *de novo* design. Molecular modeling techniques seem quite promising to provide with the relevant amount of atomic information for the design and optimization of natural or synthetic enzymes. Computational techniques are nowadays broadly used in many areas of chemistry and biology to understand and predict the atomic behavior of molecules. In fact, the use of computational techniques for the design of *de novo* proteins (Nanda and Koder, 2010) has already showed interesting applications. In addition, increasing insight in the mode of action of enzymes and the advances in computational modeling techniques have led to the *de novo* design of artificial enzymes with unnatural catalytic activities (Jiang et al., 2008). To date, several

technologies have been developed to achieve this goal: namely, computational design, catalytic antibodies (CAs), and mRNA display. These methods rely on different principles, trading off rational protein design against an entirely combinatorial approach of directed evolution of vast protein libraries, and very recently, [Golynskiy and Seelig \(2010\)](#) have reviewed and compared these methods and their potential for generating truly *de novo* biocatalysts. However, a note of caution is mandatory here: more stabilization of the (often putative) TS is not sufficient to create efficient catalysts, because in many cases the rate determining step of the catalytic activity is not associated to the chemical event (bond breaking/forming processes); then, the entire pathway including the diffusion of substrate or the release of the product needs to be included in the simulation.

An appropriate approach to study enzymatic reactions is to combine the highest accuracy QM level description of active center with an approximate description, for example, using classical molecular mechanics (MM), of its surrounding. These methods are known under the generic name of QM/MM techniques ([Warshel and Levitt, 1976](#)). In such simulations, the QM algorithm is typically called millions of times to generate the energy, forces, and charge distribution of the reaction center in the presence of electrostatic interactions with the MM environment ([Senn and Thiel, 2009](#)). The advantages of such an approach have long been recognized, and they are today standard methods for the treatment of reactivity in complex systems. Several reviews on the method are available ([Warshel, 2003](#); [Vreven and Morokuma, 2006](#); [Senn and Thiel, 2007a,b, 2009](#); [Hu and Yang, 2008](#); [Sousa and Ramos, 2008](#); [Truhlar, 2008](#); [Kamerlin et al., 2009](#); [Acevedo and Jorgensen, 2010](#); [Mata, 2010](#)). This QM/MM approach has been extensively applied to study enzymatic systems ([Warshel and Levitt, 1976](#); [Monard and Merz, 1999](#); [Martí et al., 2004a,b](#); [Riccardi et al., 2006](#); [Zhang, 2006](#); [Gao, 2007](#); [Hu and Yang, 2008](#); [Senn and Thiel, 2009](#)), indicating the reliability of the existing QM/MM methods for describing real-world chemical reactions ([Field et al., 1990](#); [Gao and Xia, 1992](#); [Thompson, 1996](#); [Gao and Freindorf, 1997](#); [Cui and Karplus, 2000a,b](#); [Martin et al., 2000](#)).

Computational designs can provide the close to atomic resolution predictions. The optimum way to understand the function of enzymes is to accomplish a perfect structural and time resolution of their catalytic chemical reactions to obtain results that explain experimental studies but also to provide complementary information not accessible in

experiment. Exactly, this is at least in principle possible by a quantum mechanical description-based QM/MM plus MD treatment. Reliability and predictability of calculations on such systems depend crucially on the accuracy of these necessarily approximate electronic structure methods. Unfortunately, to date, there still is not an universal and well-established computational protocol for the study of reaction mechanisms in enzymes.

Over the past years, a vast empirical knowledge of catalysis and catalysts for an enormous number of reactions has been accumulated. However, it is only recently that we are moving away to attempt a rational design of biological systems tailored to specific reactions. In this chapter, we address the problem of the first-principles design of catalytic functions. In general terms, we are applying MD and free-energy simulations with hybrid QM/MM potentials to study several enzyme-catalyzed reactions. The aim of this report is to give the interested reader a realistic overview on the current state of art in QM/MM MD simulations applied to the phenomena of enzyme catalysis, emphasizing the potential without disguising its present limitations, and to provide illustrative examples from our own work. Herein, we highlight the recent progress of this exciting field and mainly focus on the theoretical simulations and application explorations. Some perspectives are also given to illustrate the opportunities as well as challenges in the future.

#### A. *Quantum Mechanics/Molecular Mechanics (QM/MM) Approaches*

Quantum mechanics/molecular mechanics (QM/MM) approaches are today well-established methods for the study of chemical reactivity in complex systems, in which a wide range of environments can be simulated, from aqueous solutions, to enzymes or zeolites. Several techniques have been developed and applied for inorganic (Sushko et al., 2000; Sulimov et al., 2002; Sherwood et al., 2003; Bandura et al., 2004; Mysovsky et al., 2004; Sokol et al., 2004; Danyliv et al., 2007) and bio/organic systems (Colombo et al., 2002; Laio et al., 2002a,b; Laio and Parrinello, 2002; Piana et al., 2004; Sebastiani and Rothlisberger, 2004), metal catalytic centers in proteins (Sherwood, 1998; VandeVondele et al., 2002; Magistrato et al., 2004; Kastner et al., 2007), molecular crystals (Kimmel et al., 2008), bioinorganic systems (Deeth, 2004; Neese, 2006; Senn and Thiel, 2009; Siegbahn and Himø, 2009; Robles et al., 2011), and solutions

(Stefanovich and Truong, 1997; Rohrig et al., 2003; Sherwood et al., 2003; Dal Peraro et al., 2004; Dahlke and Truhlar, 2007a,b; Lin and Truhlar, 2007; Zhang et al., 2007).

Different reviews on this hybrid methodology have been published, showing their amazing development and application (Warshel, 2003; Vreven and Morokuma, 2006; Lin and Truhlar, 2007; Senn and Thiel, 2007a,b, 2009; Hu and Yang, 2008; Kamerlin et al., 2009; Acevedo and Jorgensen, 2010). The system is divided into regions and treated with accurate and computationally expensive methods only the part where necessary, whereas the reminder is treated at a lower level of theory and less demanding computational cost. The first will correspond to the active site and some selected amino acids, where the phenomena of chemical interest (forming/breaking bond processes) are taking place. This region can be treated with accurate and computationally QM expensive methods only the part where necessary, while the remainder is treated at a lower and less demanding level of theory as MM. The role of the latter is mainly to introduce environmental effects. This combined approach allows for the study of chemical reactivity in large complex systems which would otherwise be computationally prohibitive. In this QM/MM methods, the QM algorithm is typically called millions of times to generate the energy, forces, and charge distribution of the reactive region in the presence of electrostatic interactions with the MM environment (Senn and Thiel, 2009). Despite the use of QM methods only in the reactive region, the QM computations are often the bottleneck in such simulations. The computational savings of QM/MM stem from the local nature of chemistry, such that QM can be used on only a small locus of the system. The total Hamiltonian is given by

$$\hat{H} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \quad (1)$$

where the first term is the Hamiltonian for the QM system *in vacuo* and the second represents the MM energy for the remaining atoms. The interaction between the two regions is given by the last term. Several different QM/MM schemes have been proposed, basically differing on the form of this term. They can be grouped into three distinct formulations (Bakowicz and Thiel, 1996): mechanical, electrostatic, and polarization coupling. Very recently, Warshel et al. have been presented a sound discussion on the inclusion of electrostatic effects in the context of QM/MM calculations

(Kamerlin et al., 2009). In this way, a reasonable potential-energy surface (PES) can be obtained, which includes the more important effects on the energetics of the process under study, at a moderate computational cost.

However, knowledge of the PES is not enough to correctly describe chemical reaction in condensed media. When studying chemical reactivity in solution, solids, or enzymatic active sites, one is faced with a substantial difference with respect to gas phase reactivity. In this last case, the reactant and the TS usually correspond to single structures, and the thermodynamic properties can be obtained by applying statistical thermodynamics to the energy levels of these structures. In solution or enzymatic environments, there are a large number of conformations accessible to the environment and then one could find a myriad of stationary structures that could be assigned as reactants or TSs for a particular process. Thus, the statistical treatment must include the exploration of a significant ensemble of minima and transition structures appearing on the PES to properly define the reactant and TSs. This ensemble can be generated using different simulation techniques as Monte Carlo or MD. Information obtained from these simulations can be then used to derive thermodynamic information and in particular the activation free energy, which can be related to the rate of a chemical reaction through the use of TST. Several approximations can be used to solve the dilemma. Roughly speaking, one must renounce to include the quantum subsystem flexibility in the simulation, or alternatively one is then compelled to use a low-level electronic description, which usually means semiempirical Hamiltonians such as AM1, PM3, etc. (Dewar et al., 1985).

In order to overcome the quantitative limitations imposed by the use of semiempirical Hamiltonians in the description of the PES, several methodologies have been proposed. One obvious solution is to develop a new parametrization exclusively for the process under study (specific reaction parameters, SRP; Rivail et al., 1991; Tomasi and Persico, 1994). This has been adopted in a number of cases, but it is not always easy to improve simultaneously the reaction and activation energies. Moreover, continuity problems can arise if one is interested in multistep reactions, as far as different parameterizations may be required. Other possibility is to include correction terms to the PES based on valence-bond theory (Cramer and Truhlar, 1995). This has been successfully used provided the semiempirical Hamiltonian gives a good qualitative PES. Other strategies have been proposed to obtain *ab initio* QM/MM free-energy profiles using a simple reference potential (e.g., empirical valence bond; Kollman et al., 2001).

One powerful strategy to correct low-level energy functions for potential of mean force (PMF) calculations is the use of interpolated corrections at highest quantum description (Ruiz-Pernia et al., 2004). This proposal is an extension of the interpolated corrections methodology developed by Truhlar and coworkers (Nguyen et al., 1995; Corchado et al., 1998; Chuang et al., 1999) for gas-phase dynamical calculations. In this method, the energy difference between structures computed at the low level and a chosen high level is written as a function of the distinguished reaction coordinate used to follow the chemical transformation and conveniently interpolated through the use of cubic splines (Renka, 1993). While this procedure has been shown to introduce a systematic improvement in the results at a very low additional computational cost, it still relies on the assumption that the low-level energy surface is qualitatively reasonable. Effectively, one of the most important limitations for this method is the dependence on the AM1/MM minimum energy path, which can sometimes dramatically differ from the one obtained using Hamiltonians of higher level of theory. In other words, this simple one-dimensional correction scheme allows for displacements of the TS along the reaction coordinate but not in other directions. An approach to introduce corrections beyond this one-dimensional scheme, in the context of enzymatic reactions, was made by Field and coworkers (Proust-De Martín et al., 2000).

An extension of this methodology and following the interpolated correction scheme was proposed in our group (Ruiz-Pernia et al., 2006). In this strategy, the correction energy is expressed as a function of two geometrical coordinates relevant in the description of the chemical reaction. For example, if the process under study can be described using an antisymmetric combination of bond forming and bond breaking distances ( $d_{\text{BB}} - d_{\text{MB}}$ ), then the correction energy is mapped on a two-dimensional PES obtained as a function of these distances ( $d_{\text{MB}}; d_{\text{BB}}$ ). Figure 1 illustrates this situation, showing the minimum energy paths and transition structures corresponding to a QM/MM calculation using a low-level quantum treatment (LL/MM) or a high level (HL/MM). As it can be seen, in general, the transition structure described on the high-level surface can be found in an advanced or delayed position along the reaction coordinate ( $d_{\text{BB}} - d_{\text{MB}}$ ) or along an orthogonal direction. These displacements can be especially important when the high-level TS is considerably more associative or dissociative than the low-level one, that is, when the low-level energy surface has important qualitative drawbacks. In principle, this correction scheme is obviously more general than the previous procedure and it can

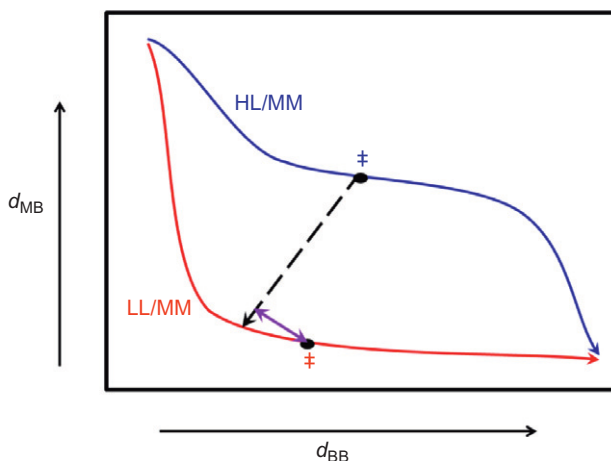


FIG. 1. Qualitative illustration of the minimum energy paths and transition structures on the potential energy surface obtained as a function of the breaking bond ( $d_{\text{BB}}$ ) and making bond ( $d_{\text{MB}}$ ) distances using a high- (HL, in blue) or a low-level (LL, in red) description of the QM subsystem. The HL transition structure (in black on the blue line) is displaced with respect to the low-level one (in black on the red line) essentially in a direction orthogonal to the antisymmetric combination  $d_{\text{BB}}-d_{\text{MB}}$ . A possible transition structure (in blue) corresponding to the use of one-dimensional corrections along the distinguished reaction coordinate is also shown. Adapted from Ruiz-Pernia et al., 2006.

be employed for PES-based applications (localization of stationary points and calculation of minimum energy paths, for example).

A most accurate but more computationally expensive methodology is the dual level strategy (Martí et al., 2005a,b). In this modified QM/MM approach, the “low-level” QM description of the quantum region is corrected during the optimization procedure by means of a “high-level” calculation *in vacuo*, keeping the QM–MM interaction contribution at a quantum “low-level.” This allows computation of energies, gradients, and Hessians including the polarization of the QM subsystem and its interaction with the MM environment, both terms are calculated using the low-level method at a reasonable computational cost. This methodology is based in the *Micro–Macro Iteration Optimization Algorithm* (Moliner et al., 1997; Turner et al., 1999; Monard et al., 2003; Prat-Resina et al., 2003,



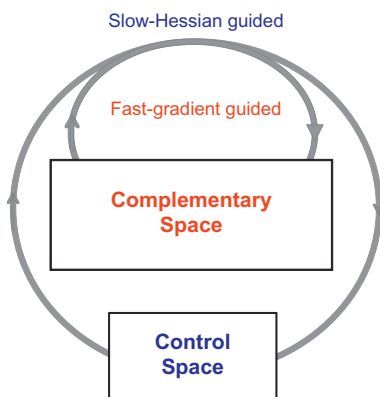


FIG. 2. Representation of the micro-macro iteration scheme for geometry optimizations in very large systems.

2004; Vreven et al., 2003). In this algorithm (see Fig. 2), a partition of the full space of coordinates of the system into a control space and complementary space subsets is done: those atoms or molecules directly involved in the reaction process (plus may be some rounding molecules or residues) are included in the control space, while the rest defines the complementary space. Then, optimization of structures can be efficiently carried out in coupled iterations over these two subspaces: at each step of the control space Hessian guided optimization, the rest of the system is fully relaxed merely using gradient vectors. This strategy leads to stationary structures with the adequate number of negative eigenvalues for a reduced Hessian matrix (the one defined for the control space). In a typical application of the micro/macro iteration approach, the control space usually contains the coordinates of up to  $\sim 100$  atoms, while the complementary space can include a number of atoms two orders of magnitude larger. When searching stationary structures, this is usually translated into about  $10^1$  Hessian guided optimization steps in the control space and up to  $10^2$ – $10^3$  gradient-based optimization steps in the complementary space at each control space movement. This last number of cycles is of course highly dependent on the size of the system and of the gradient of the considered structure. Anyway, a typical application in enzymes or condensed media can amount up to  $10^3$  or even more energy and gradient

evaluations to fully relax a stationary structure, which means that only low cost computational methods can be employed to describe the QM region (except in those cases where the QM subsystem contains very few atoms). Whereas upcoming computational power allows using higher electronic Hamiltonian approaches such as *ab initio* or based in DFT to describe the quantum region, the usually large amount of gradient vector evaluations needed during the macrosystem minimization procedure turns this way almost impracticable for hybrid QM/MM MD simulations, still making semiempirical methods the most suitable one. In this way, the HL term must be evaluated only during the optimization of the control space, while the LL terms are devoted to avoid the more time consuming complementary space optimization. This method has shown to provide considerably better descriptions than standard QM/MM calculations at the semiempirical level, while the computational cost is still reasonable but higher than the interpolation methodology one (Martí et al., 2005a,b). The dual level method is based on two different quantum treatments for the QM region, one LL method (usually a semiempirical Hamiltonian) and the other HL method (usually, DFT or *ab initio* correlated methods; Ruiz-Pernia et al., 2004, 2006; Martí et al., 2005a,b).

## II. POTENTIAL OF MEAN FORCE/FREE-ENERGY CALCULATIONS

Minimum (potential) energy pathways can in favorable cases provide a reasonably picture of the chemical processes in enzymatic reactions. However, it is desirable, and often necessary, to compute free-energy pathways to understand these complex processes in more realistic detail. Free-energy simulations based on statistical sampling can provide reasonable estimates of the free-energy changes associated with a chemical reaction. The free-energy change along a given reaction coordinate ( $\xi$ ) is called the PMF ( $W$ ) and can in principle be generated from MD simulations:

$$W(\xi) = C' - kT \ln \langle \rho(\xi) \rangle \quad (2)$$

where  $C'$  is an arbitrary constant and  $\rho(\xi)$  the probability distribution function of the selected coordinate. During a molecular simulation, the probability density can be evaluated measuring the number of times that the system has a value of the coordinate between  $\xi$  and  $\xi + \Delta\xi$ ,  $N(\xi)$ . If the total number of configurations collected in the simulation is  $M$ , the histogram of the coordinate can be constructed as

$$\langle \rho(\xi) \rangle \Delta \xi = \frac{\langle N(\xi) \rangle}{M} \quad (3)$$

When high-barrier processes are considered, direct simulation will result in getting stuck in the low energy regimes of the end states and it is unlikely that the transition of interest would ever be observed in the limited simulation time. In other words, the whole range of values of interest of the reaction coordinate cannot be covered in reasonable simulation times for regions exceeding few times  $kT$ . Several approaches have been developed in order to address the sampling problem, such as using biasing potentials or reducing the phase space to sample the relevant degrees of freedom (Amadei et al., 1993; Hamelberg et al., 2004). One of the most successful approaches used to improve the sampling of all the configurational space of interest is the umbrella sampling technique. In this method, the simulation is carried out in the presence of an additional biasing potential  $V_{\text{umb}}(\xi)$ , introduced to enhance the sampling in the neighborhood of a particular value of the coordinate  $\xi$  (Torrie and Valleau, 1974; Roux, 1995). A very common choice is a harmonic function of the form:

$$V_{\text{umb}} = \frac{1}{2} K_{\text{umb}} (\xi - \xi_{\text{ref}})^2 \quad (4)$$

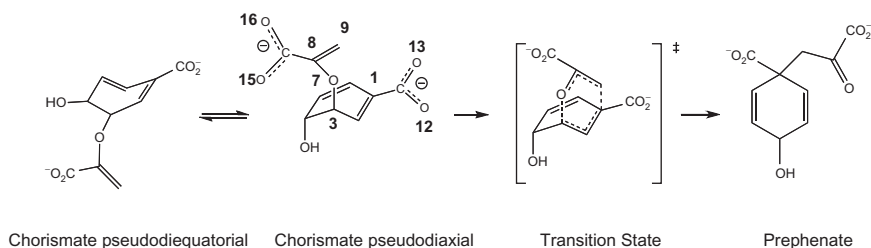
The question is now to recover the full distribution function of the unbiased system, from the distribution of the biased simulation windows. The weighted histogram analysis method (WHAM) provides an optimal way to combine the data collected in the simulations (Kumar et al., 1992).

Comparisons of various approaches for calculating QM/MM free energies can be found in different references (Kastner et al., 2006; Ytreberg et al., 2006; Senn and Thiel, 2007a,b, 2009; Hu and Yang, 2008; Acevedo and Jorgensen, 2010).

### III. APPLICATIONS

#### A. Chorismate Mutase as a Working Example

The conversion of (-)-chorismate to prephenate catalyzed by chorismate mutase (CM; Haslam, 1993) is shown in Scheme 1. This reaction is part of the shikimate pathway which produces aromatic amino acids in plants, fungi, and bacteria (Haslam, 1993), making CM a potential target



SCHEME 1. Pseudodiaxial-pseudodiequatorial conformational equilibrium of chorismate and Claisen rearrangement of chorismate to prephenate.

for herbicides, fungicides, and antibiotics. CM provides an important test of theories of enzyme catalysis, and of modeling methods, due to four main advantages: (i) the rearrangement of chorismate to prephenate catalyzed by the enzyme has its counterpart in solution, and experimental (Andrews et al., 1973; Copley and Knowles, 1987; Kast et al., 1996a,b; Gustin et al., 1999) and theoretical (Lyne et al., 1995; Carlson and Jorgensen, 1996; Hall et al., 2000; Martin et al., 2000; Guo et al., 2001; Kangas and Tidor, 2001; Bruice, 2002; Barbany et al., 2003; Hur and Bruice, 2003; Mandal and Hilvert, 2003; Martí et al., 2003a,b) studies have demonstrated that the reaction takes place following the same molecular mechanism. This is a very important feature as it allows to directly compare the results obtained in both media and to get insights into the role of the enzyme. We have to keep in mind that it is quite frequent that catalysts accelerate the chemical rate by changing the mechanism, and in such a case, the comparisons would not give information of the generic aspects of enzyme catalysis; (ii) there are data available in the literature (Andrews et al., 1973; Copley and Knowles, 1987) that offers the opportunity to compare theoretical and experimental results. Further, although some debate appeared in the literature as to what step was the rate limiting in CM, more recent studies based on kinetic isotope effects have demonstrated that the chemical reaction is preponderantly rate limiting in this enzyme (Gustin et al., 1999); (iii) no covalent bonds are formed between the substrate and the protein, avoiding technical problems of frontier treatments between QM and MM regions, such as the use of boundary atoms. This allows a simple division of both subsystems, that is, schematically depicted for the enzymatic reaction in Fig. 3; (iv) since the

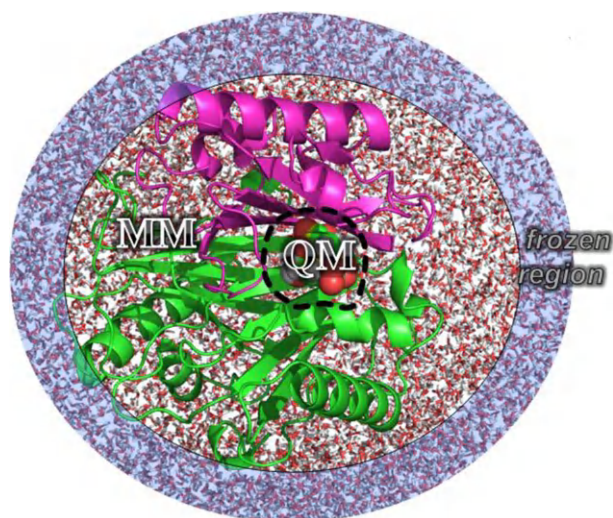


FIG. 3. The full system is divided into a QM region and an MM region. The blue shaded area shows the MM frozen atoms.

rearrangement of chorismate to prephenate is an unimolecular reaction, the first step of the energy profile depicted in [Scheme 1](#), the contribution of bringing two separated reactant species together to form the Michaelis complex (MC) in a bimolecular process, is simplified into a conformational problem: the work of changing a nonreactive chorismate conformer structure into a new one which is ready to proceed the rearrangement to prephenate (see [Scheme 1](#)). This enzyme is relatively simple, but nonetheless is at the center of controversies regarding the origin of its catalytic efficiency ([Martí et al., 2001](#)). In order to simulate the step of conversion from chorismate to prephenate, the PMF profiles in aqueous solution and in the enzyme were obtained in our laboratory using the antisymmetric combination of the forming and breaking bond distances as the distinguished reaction coordinate (see [Fig. 4](#)). From the analysis of this figure, it appeared that the difference between the solvent and enzyme theoretical free-energy barriers obtained at AM1/MM level is in very good agreement with the experimental data (8.7 vs. 9.1 kcal mol<sup>-1</sup>, respectively), although

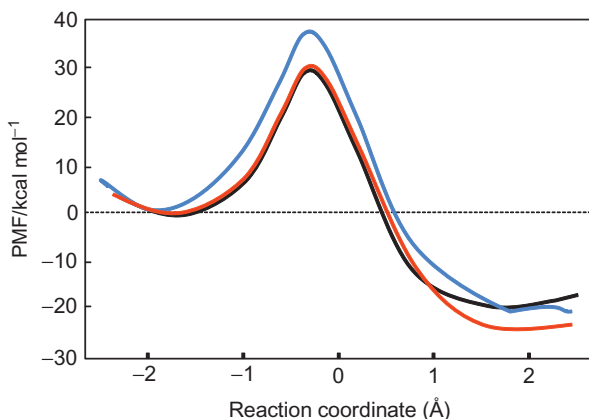


FIG. 4. Free-energy profiles, obtained in terms of QM/MM PMFs, for the chorismate to prephenate rearrangement obtained in solution (blue line), catalyzed by *B. subtilis* chorismate mutase, BsCM (black line) and catalyzed by *E. coli*, EcCM (red line).

the absolute barriers are overestimated due to the use of the AM1 Hamiltonian to describe the QM region.

As mentioned above, this enzymatic system presents, as one of the advantages, the fact that no covalent bonds exist between the QM and the MM regions. This feature allows one to carry out a very interesting decomposition analysis of the potential-energy barrier in aqueous solution and in the enzymatic environment. The total QM/MM activation energies obtained in aqueous solution or in the presence of the enzyme environment can be written as the following sum:

$$\Delta E^\ddagger = \Delta E_{\text{QM}}^\ddagger + \Delta E_{\text{int}}^\ddagger + \Delta E_{\text{MM}}^\ddagger \quad (5)$$

where  $\Delta E_{\text{QM}}$  is the *in vacuo* energy relative to R using the solute/substrate structures obtained in the QM/MM calculations,  $\Delta E_{\text{int}}$  is the solvent-solute or enzyme-substrate interaction energy relative to R, and  $\Delta E_{\text{MM}}$  is the MM energy relative to R structure. According to this decomposition, the energy barrier of the reaction is the sum of three contributions, which are given in Table I for the reaction in water and in BsCM.

TABLE I  
 Experimental and AM1/MM Averaged Values for the Free Energy and Potential-Energy Barrier and Its Components (see Eq. (5)) for the Chorismate Rearrangement in Water Solution and in BsCM (from Martí et al., 2001)

	$\Delta G_{\text{exp}}^{\ddagger}$	$\Delta G^{\ddagger}$	$\Delta E$	$\Delta E_{\text{QM}}$	$\Delta E_{\text{int}}$	$\Delta E_{\text{MM}}$
Aqueous solution	24.5 <sup>a</sup>	38.0	39.0	40.4	-2.7	1.4
BsCM	15.4 <sup>b</sup>	29.3	27.1	42.1	-16.2 <sup>b</sup>	1.2

All values are in kcal mol<sup>-1</sup>.

<sup>a</sup>From Andrews et al. (1973).

<sup>b</sup>From Kast et al. (1996a).

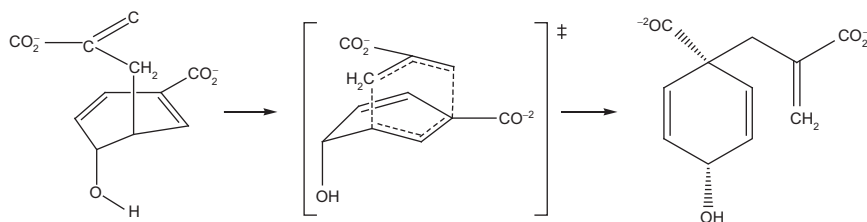
It must be kept in mind that the reported values in Table I come from the average of the structures appearing in QM/MM MD generated at the maximum (TS) and the minimum (reactant state, RS) of the free-energy profiles in Fig. 4. The MD simulations were calculated until convergence of the differences in energies. Nevertheless, as previously mentioned, these results have to be treated with caution and the conclusions have to be considered as qualitative because they can be affected by large statistical errors, especially the change in the MM environment energy. As can be seen in the table, the main contribution to the potential-energy barrier lowering does not come from the solute or the substrate energy, which in fact is in the opposite trend, but from the preferential interaction of the enzyme with the TS. It must be emphasized that the nature of this contribution is essentially electrostatic. It can be surprising that while the interaction contribution is much more important (in an absolute sense) in the enzyme than in water, the energetic change of the environment is very similar in both media. In aqueous solution, this last term is very close to one half the value of the interaction energy as predicted by linear response solvent models, while in the enzyme, this represents a much lower relative contribution (7.5%). Inside the enzyme, there is a large electrostatic effect associated with a very small reorganization. This analysis is in agreement with the preferential stabilization of the TS by electrostatic interactions in the enzyme than in water as previously pointed out by Warshel (1991, 1998) and Strajbl et al. (2003). Most recently, they have evaluated the binding free energy of the ground state and the TS in CM, demonstrating that the enzyme works by transition-state stabilization (TSS; Strajbl et al.,

2003). The evaluation of the different contributions to the reduction of the activation energy, using linear response methods, established that TSS resulted from electrostatic effects.

The geometrical and energetic analysis of the TS and the MC reveals the importance of the Glu78 residue; the Arg90 that activates the ether bond and stabilizes the TS; or the role of Arg7, Tyr108, and Arg115 that present direct ionic interactions to the substrate being catalytically significant, in addition to their obvious role in binding, as has been emphasized also by Lee et al. (2002). Lee et al. arrived to this conclusion from QM/MM PESs in BsCM along the reaction coordinate using a Hartree–Fock wavefunction not only to treat the substrate but also the side chains of Glu78 and Arg90 (Lee et al., 2002).

Hilvert et al. (Aemissegger et al., 2002) studied the Cope rearrangement of carbachorismate to carbaprephenate (see Scheme 2) and tested the catalytic activity of BsCM against this reaction. In the carba analogues, the ether oxygen is substituted by an apolar methylene group. No significant catalytic activity was found in this case. To elucidate the origin of the enzymatic ability to speed up the chorismate rearrangement to prephenate, we carried out a comparative analysis of the TS of the Claisen and Cope reactions in the active site of BsCM (Martí et al., 2004a,b).

The electrostatic interaction established between residues of the enzyme active site and the carboxylate groups of the substrate considerably reduces the calculated free-energy barriers for both reactions in the BsCM active site. Effectively, the two carboxylate groups must be approached during the reaction processes, and the electrostatic repulsion associated with the diminution of the distance between the negative charges is largely compensated by the interactions established with



SCHEME 2. Cope rearrangement of carbachorismate to carbaprephenate.



positively charged residues of the active site (Arg7, Arg63, Arg90, and Arg116; see Fig. 5). However, the ability of the enzyme to reduce the energy barrier is larger (by about  $3 \text{ kcal mol}^{-1}$ ) in the case of the Claisen reaction. This differential catalytic effect is mainly attributed to the enhanced electrostatic interaction established between the ether oxygen atom and Arg90 in the TS of the chorismate to prephenate rearrangement.

The electrostatic interactions established between the substrate and the enzyme play a primary effect reducing the free-energy barriers, but in addition, they also have consequences on the reactant structures. If the transition structures are going to be stabilized, then those reactant structures closer to these would be also relatively favored. In the active site, reactant structures have smaller distances between the carboxylate group than in solution, and thus a more pseudodiaxial character, becoming then more similar to the transition structures. The enzyme structure, whose active site would be exquisitely complementary to the TS thus stabilizing it more than the substrate and reducing the barrier (Strajbl et al., 2003), has a considerable effect on the reactants. In the global energy balance, the equilibrium among reactant substrate conformers is displaced toward those reactive conformations that are geometrically closer to the TS,

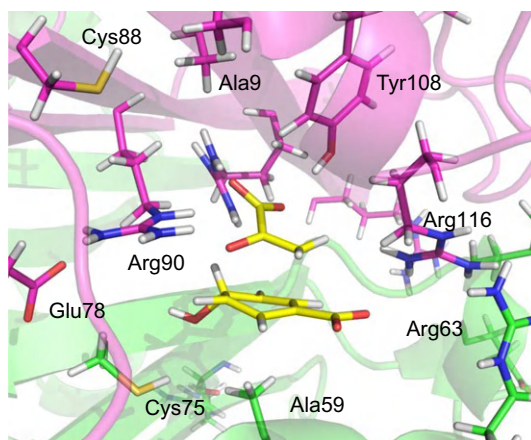


FIG. 5. Snapshot of the chorismate to prephenate TS in the active site of BsCM.

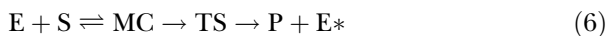
thus avoiding the energetic penalty associated with the deformation of the full enzyme–substrate system. Consequently, the final conclusion for this system is that the origin of the catalysis has to be searched for in the enzyme structure. This structure, designed to stabilize the TS relative to the in-solution process, obviously has consequences on the RS, favoring structures which are more similar to those appearing in the TS. Then, the two effects of the enzyme on the reaction process, stabilizing the TS and deforming the RS geometry, can be viewed as two faces of the same coin (Martí et al., 2004a,b). Another integrated vision was proposed by Warshel (1998) and Villà and Warshel (2001) According to these authors, the apparent NAC effect proposed by Bruice (2002) is not the reason for the catalytic effect but the result of the TS stabilization; the key catalytic effect is electrostatic in nature. However, since the charge distribution of the TS and the reactive reactants is similar, the stabilization of the TS leads to reduction in the distance between the reacting atoms in the RS.

### *B. Designing CAs: TSA*

Pauling’s seminal idea concerning enzymatic catalysis is that an enzyme lowers the energy of TS (Pauling, 1946, 1948a,b, 1960). Evidence in support of this proposal is the fact that stable compounds that resemble the TS, TSAs, are competitive inhibitors of enzymes (Wolfenden, 1972). One approach to TSA design is to establish the nature of the enzymatic TS and to synthesize chemically stable analogues with similar features (Schramm, 2003). Raso and Stollar pioneered the use of TSAs as happens in immunization processes to synthesize new catalysts, CAs (Raso and Stollar, 1975). The study of processes associated with the activity of CAs provides opportunity to examine and understand enzyme catalysis and vice versa. In-depth knowledge of enzyme activity can be used to improve the specificity, selectivity, and efficiency of these new catalysts (Schultz and Lerner, 1995). CAs are especially interesting as catalysts for those reactions for which no enzyme is known (Schultz and Lerner, 1993).

The simplest kinetic scheme used to understand enzymatic and CA processes is that proposed by Michaelis and Menten, which proceeds with the formation of a substrate–catalyst complex (the MC) before the product-forming step during which the catalyst is recovered (Mader and Bartlett, 1997).

In the simplest version, the reaction takes place through a single TS (Eq. (6)).



The activation free energy of the catalyzed reaction step can be related to that of a counterpart uncatalyzed process through the binding energies of the MC and the TS (Eq. (7)).

$$\Delta G_{\text{cat}}^{\ddagger} = \Delta G_{\text{uncat}}^{\ddagger} + \Delta G_{\text{bind}}^{\text{TS}} - \Delta G_{\text{bind}}^{\text{MC}} \quad (7)$$

Here,  $\Delta G_{\text{cat}}^{\ddagger}$  and  $\Delta G_{\text{uncat}}^{\ddagger}$  are the free energies of activation for the catalyzed and the uncatalyzed reactions, while  $\Delta G_{\text{bind}}^{\text{TS}}$  and  $\Delta G_{\text{bind}}^{\text{MC}}$  reflect the affinities of the protein for the TS and the MC, respectively. According to this scheme, the catalytic power of enzymes comes from the larger affinity of the enzyme for the TS than for the MC, since  $\Delta G_{\text{bind}}^{\text{TS}} - \Delta G_{\text{bind}}^{\text{MC}}$  is a negative quantity. Antibodies are synthesized on the basis of their affinity for a TSA ( $\Delta G_{\text{bind}}^{\text{TSA}}$ ), a quantity expected to be correlated with the binding energy of the true TS of the reaction to be catalyzed ( $\Delta G_{\text{bind}}^{\text{TS}}$ ). Thus, CAs are expected to provide a lower activation free energy.

However, initial expectations for CAs as catalysts have not been fully met. First of all, CAs are not as efficient as the enzymes. Second, not all antibodies that stabilize TSAs are catalysts of the reaction. Finally, there are cases where after a process of maturation (which results in an increased affinity of the CA for the TSA), a paradoxical decrease in the catalytic power is observed, relative to the initial or germline CA (Ulrich et al., 1997).

Different arguments have been proposed to explain these findings. First, the fact that CAs present modest rate enhancements relative to those of enzymes could be due to the low affinity between the CA and the TSA that can be developed by the immune system. This affinity is apparently not enough to result in the TS affinity required for a substantial increase of the reaction rate. During the process of maturation, the TSA-CA affinity (in terms of the dissociation constants calculated for the TSA) increases to ca.  $10^{-10}$ , while the TS-enzyme affinity increases to  $10^{-23}$  (Mader and Bartlett, 1997). Second, even when the CA-TSA affinity could be improved enough, to date, it has been difficult to elicit antibodies that are as effective at differentiating the ground state from the TS (Mader and Bartlett, 1997). In this regard, Schultz et al. have stressed the lack of flexibility of the CAs (Mundorff et al., 2000). The introduction of somatic mutations, which lead to an increase in binding affinities, can cause a

significant restriction in the relative orientation of the substrate, thereby decreasing the rate constant (Mundorff et al., 2000). The increased binding affinity of the affinity-matured antibody stabilizes the substrate in a catalytically unfavorable conformation. Thus, it would be possible to rationalize why some CAs do not act as catalysts. Finally, as also suggested by Mader and Bartlett (1997), it is not possible to devise very accurate TSAs; that is, the TSA cannot be similar enough to the true TS. As a consequence, an improvement in  $\Delta G_{\text{bind}}^{\text{TSA}}$  would not be directly translated into an improvement in  $\Delta G_{\text{bind}}^{\text{TS}}$ . All these problems arise because of an important issue of timescales. In its continuous processes of evolution, nature has brought forth countless mechanisms for complex biochemical reactions, but the biological selection process that produces the antibody differs from that governing enzyme evolution by an enormous factor: a timescale of weeks in the former in contrast to a process occurring over millions of years in the latter.

The methods and techniques of computational chemistry provide excellent tools for obtaining molecular details of catalytic processes (Marti et al., 2008). With the aim of increasing the stability of the TSs, scientists design CAs based on their affinity to TSAs. In this process, the effect of the CA on the RS is completely lost. According to our work, CAs usually present low catalytic efficiency (relative to that of enzymes) and even inverse correlations between maturation process and catalytic power not only because the TSA does not properly represent the TS but mainly because the MC is not considered in the improvement process. Enzymes and CAs have evolved with different purposes: the former to decrease the activation free energy of the reaction and the latter to increase the binding energy for the TSA (and for the TS). In the first case, the target is the difference between the binding energy of the TS and the MC ( $\Delta G_{\text{bind}}^{\text{TS}} - \Delta G_{\text{bind}}^{\text{MC}}$ ), while in the second case, attention is focused on  $\Delta G_{\text{bind}}^{\text{TSA}}$  (which is related to  $\Delta G_{\text{bind}}^{\text{TS}}$ ), and the MC is not considered at all.

QM/MM simulations can be used as a computer-aided rational-design protocol to overcome some of the limitations of standard rational-design techniques and that is tested for the chorismate to prephenate simple metabolic reaction (Villà and Warshel, 2001; Bruice, 2002; Benkovic and Hammes-Schiffer, 2003; Garcia-Viloca et al., 2004) (see Scheme 1). CMs from different organisms, such as *Bacillus subtilis* (BsCM; Chook et al., 1993) or *Escherichia coli* (EcCM; Lee et al., 1995), exhibit similar kinetic properties, although they may share little sequence similarity. Further,

a CA with modest CM activity was prepared against a TSA of CM (Hilvert et al., 1988; Hilvert and Nared, 1988), and its three-dimensional structure was determined at 3.0-Å resolution (Haynes et al., 1994). We studied the chorismate to prephenate rearrangement on this system and compared to the results obtained for natural enzymes such as *B. subtilis* (BsCM; Chook et al., 1993) or *E. coli* (EcCM; Lee et al., 1995).

The free-energy profiles obtained for the chorismate to prephenate rearrangement in aqueous solution, in the two CM enzymes (BsCM and EcCM), and in the 1F7 CA are depicted in Fig. 6, whereas our best estimation of the free-energy barriers is listed in Table II. The resulting profiles are in accordance with the expected results: the catalytic efficiency of the 1F7 appears between the catalytic power of both enzymes (which are very close to each other) and the reaction studied in solution.

The most remarkable result in Table II is that the computed catalytic power of the different tested proteins ( $\Delta\Delta G_{\text{theo}}^{\ddagger}$ ) is in very good agreement with that of experimental ( $\Delta\Delta G_{\text{exp}}^{\ddagger}$ ). This fact validates the employed methodology, giving encouragement for its use in obtaining a deeper

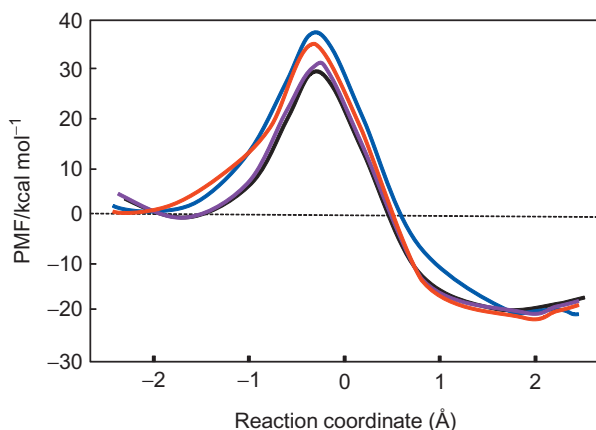


FIG. 6. Free-energy profiles (in terms of potential of mean force, PMF) for the chorismate to prephenate rearrangement obtained in the different environments: BsCM (black line), 1F7 (red line), 1F7 (N33S) mutant (purple line), and in aqueous solution (blue line). The reaction coordinate is the antisymmetric combination of the interatomic distances of the breaking and forming bonds, C3···O4 and C1···C6, respectively. Adapted from Martí et al. (2007).

TABLE II  
Theoretical Free-Energy Barriers (in kcal mol<sup>-1</sup>) for the Uncatalyzed Chorismate to Prephenate Rearrangement Compared with the Catalyzed Reaction by BsCM, EcCM, 1F7, and 1F7 (N33S) Mutant CA

	Water	BsCM	EcCM	1F7	1F7 (N33S)
$\Delta G_{\text{theo}}^{\ddagger}$	29.3 <sup>a</sup>	20.6 <sup>a</sup>	20.9	27.5	23.0
$\Delta\Delta G_{\text{theo}}^{\ddagger}$	0.0	-8.7	-8.4	-1.8	-6.3
$\Delta G_{\text{expt}}^{\ddagger}$	24.5 <sup>b</sup>	15.4 <sup>b</sup>	17.2 <sup>c</sup>	21.6 <sup>d</sup>	-
$\Delta\Delta G_{\text{exp}}^{\ddagger}$	0.0	-9.1	-7.3	-2.9	-

Data from [Hilvert et al. \(1988\)](#).

<sup>a</sup>Values are taken from [Martí et al. \(2007\)](#).

<sup>b</sup>Values are taken from [Martí et al. \(2001\)](#).

<sup>c</sup>Values are taken from [Chook et al. \(1993\)](#).

<sup>d</sup>Values are taken from [Lee et al. \(1995\)](#).

insight into the catalysis in the enzymes and 1F7. Further to free-energy barriers, [Fig. 6](#) can be used to identify the position of the TS and the MC along the reaction coordinate, defined as the antisymmetric combination of the breaking and forming bonds, C3—O4 and C1—C6, respectively. Considering that the values of the reaction coordinate of the different TSs are very close (see [Fig. 6](#)), if the difference of the reaction coordinates between the TS and the MC is small, the pre-equilibrium of the substrate (see [Scheme 1](#)) will be displaced toward the chair-like structure of the chorismate. In this regard, although the free-energy profile is rather flat in the minimum region, we could deduce a direct relationship between the difference in the reaction coordinate at the MC and the TS and the value of the free-energy barrier ([Martí et al., 2007](#)). The smallest differences are obtained in the enzymatic processes (1.4 and 1.5 Å for the BsCM and EcCM, respectively), whereas the largest difference is obtained in the solvent environment (2.3 Å). An intermediate value of 1.8 Å was determined for 1F7, thus fitting the order in the free-energy barriers. Roughly speaking, water molecules fit to the substrate structure in solution, while the protein induces conformational changes in the substrate.

The analysis of the averaged structures obtained in the different biological systems allows determination of which interactions favor the stabilization of the TS. [Fig. 7](#) presents the averaged interaction energy, electrostatic and van der Waals contributions, of individual residues with

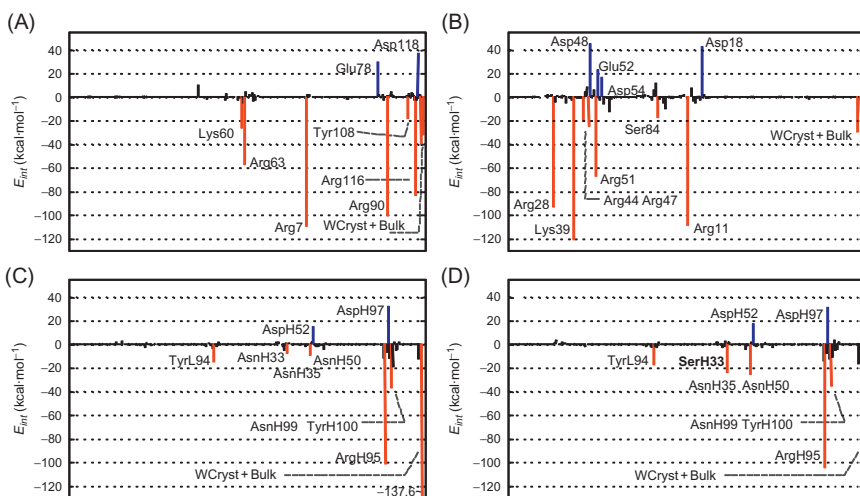


FIG. 7. Contributions of individual amino acid residues (ordered along the  $x$  axis) to the TS interaction (in  $\text{kcal mol}^{-1}$ ) of the (A) BsCM, (B) EcCM, (C) 1F7, and (D) 1F7 (N33S) mutant.  $E_{\text{int}}$  = substrate–protein interaction energy, WCryst + bulk = crystallization and bulk water molecules. Adapted from Kast et al. (1996a,b), Taylor et al. (2001), Lee et al. (2002), Kienhöfer et al. (2003), Martí et al. (2004), and Szczyk et al. (2004).

the substrate at the corresponding TSs. Several conclusions can be obtained from this kind of analysis: for both enzymes, BsCM and EcCM, the favorable interactions take place through the positively charged residues (mostly arginine residues) with the two negatively charged carboxylate groups of the substrate and the negative charge that develops on the ether oxygen (Khanjin et al., 1999). Concerning the 1F7, the magnitude of all the interactions is dramatically smaller than in the enzymes except for the interaction established with ArgH95, which presents similar values to the enzymatic ones. In fact, it seems that the 1F7 does not properly interact with the two carboxylate groups of the substrate, leaving them partially exposed to the solvent. Therefore, these results suggest that the substrate fits better in the enzyme active sites than in the CA pocket, which is in agreement with the previous observation concerning the reaction coordinate values in the MC: much closer to the TS in the CMs than in the 1F7 case.

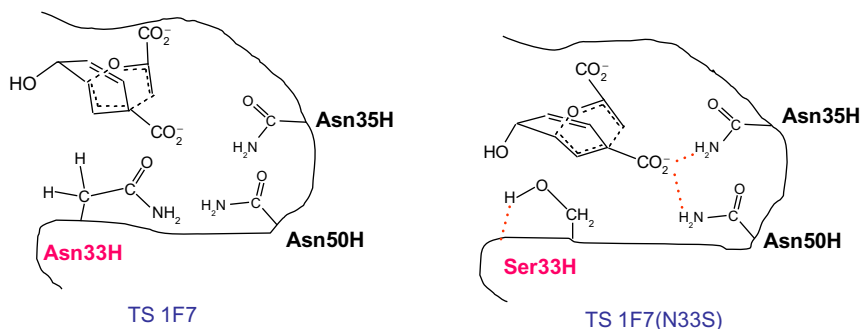
It is also important to point out that the pattern of interactions obtained in the TS of the 1F7 is not equal to the TSA–CA structure determined by the X-ray diffraction study. Thus, for instance, AsnH33 presents a noticeably different orientation in the TS–CA with respect to the TSA–CA complex: although the hydrogen atoms of the amino group interact with the hydroxy group of the inhibitor in the later, in the TS–CA complex, there are steric interactions between this residue and the aliphatic hydrogen atoms of the substrate that impede optimum positioning of the substrate into the cavity. As a result, strong interactions between the carboxylate group and amino acid residues of the inner part of the cavity are prevented, as is depicted in [Scheme 2](#). The low capability of the 1F7 to enhance the rate constant of the chorismate to prephenate rearrangement can be understood from this analysis. The strong stabilizing interactions observed in the enzyme between both carboxylate groups and the protein are not reproduced by the immune-system process when eliciting antibodies against a stable molecule that resembles, but is not equal to, the TS of the desired chemical transformation.

From the conclusions obtained by comparing BsCM and EcCM with 1F7 (see above), we proposed and checked *in silico* mutations that may improve the efficiency of the 1F7 CA. Thus, we changed the AsnH33 residue to a serine that would facilitate a better accommodation of the substrate in the cavity of the CA due to its smaller size, presumably enhancing the interactions of the substrate with the residues located in the inner part of the cavity. Once this mutation was carried out, the free-energy profile of the new 1F7-N33S, also presented in [Fig. 6](#), was obtained by using the same procedure as in the previous calculations. This mutation yields a noticeable decrease in the free-energy barrier, in comparison with the PMF obtained for 1F7. The corresponding activation free energy reported in [Table II](#) is  $4.5 \text{ kcal mol}^{-1}$  lower than the original 1F7 CA and only  $2.4 \text{ kcal mol}^{-1}$  above the most efficient BsCM enzyme. This diminution would imply an increase in the rate constant by a factor of  $10^3$  at room temperature, compared with 1F7 CA. Stabilization of the TS, as a consequence, preferentially selects and optimizes those reactant conformers that resemble the TS, thereby displacing the pre-equilibrium to the reactive reactant conformers. The reaction-coordinate difference between reactants and TS is now  $1.6 \text{ \AA}$ , a value closer to the enzymes than to that calculated for the 1F7. The analysis of the substrate–protein interactions in the TS, presented in [Fig. 7D](#), reveals that our predictions have been



confirmed, at least in the computed structures and energies; the new CA presents a more favorable pattern of interactions than the 1F7. The most important effect of the mutation is the extra space generated in the cavity, allowing the ring of the substrate to slightly rotate and its carboxylate groups to optimize the interactions with the available residues between the TS and the cavity (see [Scheme 3](#)). In particular, the interactions established between the carboxylate group and residues such as AsnH35 and AsnH50 are stronger in the mutated CA. Simultaneously, the interactions with the water molecules are reduced, which is similar to the situation observed in the CMs. The steric hindrance of the AsnH33 with the substrate in 1F7 prevents this movement, whereas in the mutated CA, a combination of the smaller size of the residue and the weaker interaction established with the substrate facilitates a more favorable relative orientation in the CA cavity, thus reducing the free-energy barrier of the chemical step.

It has been suggested that the limited structural diversity of the immune system imposes inherent limitations on catalytic efficiency ([Backes et al., 2003](#)). This work shows how our methodology, combined with other experimental strategies, may be used to determine whether the antibody scaffolds are evolutionary dead ends or can be further improved, as seems to be the case for 1F7. The study of TS–protein complexes, which have been demonstrated not to be equal to the TSA–protein structures obtained from experimental techniques, can be used to decide which residues should be changed in the active site of the CA to reduce the free-energy barrier of the catalyzed chemical transformation. Computer-



SCHEME 3. The substrate–protein interactions in the active site of the 1F7 and the 1F7 (N33S) CAs at their respective TSs. Adapted from [Martí et al. \(2007\)](#).

aided rational design might be used, not only as a first step for directed laboratory evolution experiments but also to shed some light on the divergent evolution of enzyme superfamilies.

### *C. Toward Understanding of the Promiscuity in Enzyme Catalysis*

Traditional views on enzymatic activity usually remark on their high efficiency and specificity. However, it has been recently suggested that this paradigm, which has dominated thinking in this field, could be too simplistic. Many enzymes have been found to present more than one catalytic activity, thus being capable of catalyzing secondary reactions at an active site that was, in principle, specialized to favor a primary reaction (Copley and Knowles, 1987; O'Brien and Herschlag, 1999; Aharoni et al., 2005; Khersonsky et al., 2006; Toscano et al., 2007). This phenomenon is known as catalytic promiscuity (Zanghellini et al., 2006). Nevertheless, Hult and Berglund (2007) recognized that, apart from this catalytic promiscuity (based on the ability of a single enzyme active site to catalyze several chemical transformations), there are two more major classes of promiscuity: condition promiscuity (enzymatic activity in various reaction conditions different from their natural ones), and substrate promiscuity (enzymes with a broad substrate specificity) and catalytic promiscuity (based on the ability of a single enzyme active site to catalyze several chemical transformations; Jensen, 1976; O'Brien and Herschlag, 1999; Glasner et al., 2006; Bershtein and Tawfik, 2008).

In most cases, catalytic proficiencies<sup>1</sup> for the promiscuous activities are much lower than that for the parent reaction, not exceeding  $(k_{\text{cat}}/K_m)/k_2$  values of  $10^{13}$ – $10^{15}$  and typically much less than that (O'Brien and Herschlag, 1999; Jonas and Hollfelder, 2008). Nevertheless, kinetic analyses of a significant amount of promiscuous enzymes reveal large rate accelerations for their secondary activities. Babbitt et al. (2010) suggest that these large values imply that binding and catalysis can be highly efficient for more than one reaction, challenging the notion that proficient catalysis requires specificity. Growing numbers of reported promiscuous activities indicate that catalytic versatility is an inherent property of many

<sup>1</sup>Catalytic proficiency,  $(k_{\text{cat}}/K_m)/k_2$ , is a ratio of the second-order rate constants for the enzyme-catalyzed reaction and the uncatalyzed reaction in solution; it is a measure of TS stabilization by an enzyme.

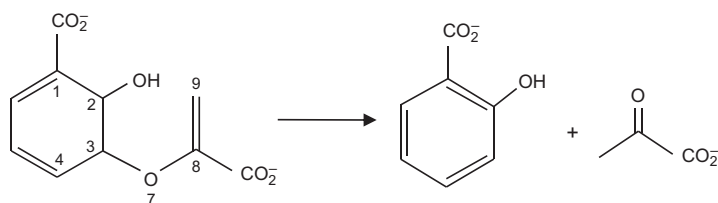
enzymes, showing that this is an important research field, as stressed by [Wu et al. \(2010\)](#) and [Busto et al. \(2011\)](#). Understanding protein promiscuity is becoming increasingly important, not only for the structural, mechanistic implications of this manifestation of infidelity of molecular recognition, as marked by [Tawfik and Khersonsky \(2010\)](#) but also for providing a raw starting point for the evolution of enzymes, as a new duplicated gene presenting low activity would provide a start for adaptative evolution ([O'Brien and Herschlag, 1999](#)). Thus, the introduction of new enzyme activities by protein engineering can be extremely useful in biotechnology ([Nobeli et al., 2009](#)). The use of new enzymes can be applied as an elegant and power synthetic methodology allowing the development of eco-friendly processes and the development of catalysts (proteins) that accelerate the rate of more than one chemical transformation in the same active site, ([Busto et al., 2011](#); [Wu et al., 2010](#)). Obtaining a protein scaffold, from a wild-type enzyme or a modified one by protein engineering, capable of catalyzing more than one step of a full synthetic process, even at ambient conditions of pressure and temperature, could reduce the costs associated with the productions of chemicals. Further, according to previous studies, promiscuous activities exhibit high plasticity as they can be readily increased by means of one or few mutations, allowing reaching the threshold to be improved under selective pressure ([Khersonsky et al., 2006](#)). Instead, primary activity presents a large robustness against mutations ([O'Brien and Herschlag, 1999](#)). In this context, enzyme promiscuity has been proposed to play a role in the divergent evolution of enzymes by providing a head start in activity and a possible selective advantage to a duplicated gene ([O'Brien and Herschlag, 1999](#); [Bornscheuer and Kazlauskas, 2004](#); [Khersonsky et al., 2006](#); [Hult and Berglund, 2007](#)). Then, the study of enzyme promiscuity can help to understand how new enzymatic functions may have emerged from a molecular scaffold during natural divergent evolution. In fact, new enzymatic functions can evolve in the period of years or even months, as happened recently with new synthetic chemicals or drugs ([O'Brien and Herschlag, 1999](#)).

There are at least two different strategies that can be used to obtain novel enzymes by means of mutations on existing enzymes. The first one is based on directed evolution, which consists of successive rounds of random mutations or recombinations followed by screening or selection ([Arnold, 1998](#)). This powerful tool does not require a deep knowledge of the details of the catalytic mechanism. A second strategy is the rational

design that implies directed mutation on particular residues of the active site (Morley and Kazlauskas, 2005). This strategy requires details about the process and the effect of the enzymatic environment on the reaction mechanism. The methods and techniques of computational chemistry have become a promising complementary tool to assist in the design of new enzymes. Thus, combinatorial optimization algorithms that integrate ligand docking and placement of amino side-chain rotamer libraries have been used to identify sequences that form complementary ligand-binding surfaces. Nevertheless, although impressive results have been obtained (O'Brien and Herschlag, 1999), some drawbacks are behind these methodologies. First, the structure of the backbone of the protein remains frozen during the functional design modeling, not introducing its inherent flexibility and lacking dynamic effects; second, the real TS of the catalyzed chemical reaction step, including the protein environment, has not been taken into account.

We will explore the advances evident in the catalytic promiscuity, chemical transformations that may differ in the functional group involved (type of bond formed or cleavage) and/or in the catalytic mechanism of bond making and breaking (Dwyer et al., 2004; Lassila et al., 2005). In this section, we present a computational approach to improve secondary catalytic activities of promiscuous enzymes. In particular, we take as test cases the PchB and the MbtI, two enzymes capable of catalyzing the chorismate to prephenate rearrangement. As previously demonstrated for the study of this reaction catalyzed by natural enzymes and CA, the method based on the use of MD simulations employing hybrid QM/MM methods (Warshel and Levitt, 1976) is a powerful *in silico* tool to get a detailed knowledge of the reaction mechanisms and, in particular, the TS and the free-energy profiles of the reaction, taking into account the effect of the protein environment (Martí et al., 2007). The results rendered by this kind of study provide clues to propose mutations on active side residues that reduce the free-energy barrier of the chemical reaction step, which hopefully should be reflected in an increase of the rate constant.

Isochorismate pyruvate lyase (IPL), from *Pseudomonas aeruginosa*, PchB, catalyzes isochorismate transformation into pyruvate and salicylate (Scheme 4), but it also presents secondary activity catalyzing the transformation of chorismate into prephenate (see Scheme 1; Gaille et al., 2002; DeClue et al., 2005; Künzler et al., 2005). In fact, this promiscuous CM



SCHEME 4. Isochorismate pyruvate lyase catalyzes isochorismate transformation into pyruvate and salicylate.

activity was used to ascribe a 1,5-sigmatropic reaction mechanism to its native or primary activity, because CMs are well known to catalyze pericyclic reactions (DeClue et al., 2005; Künzler et al., 2005). The recently obtained X-ray structure reveals that PchB is a structural homolog of some CMs, despite the low sequence identity (Siegel et al., 2010). Thus, this enzyme is an excellent candidate to improve its secondary activity by means of few mutations.

The second candidate, salicylate synthase (SS) from *Mycobacterium tuberculosis*, MbtI, initiates the biosynthesis of siderophores by converting chorismate to salicylate SS. Nevertheless, three more distinct activities have been described for wild-type MbtI *in vitro*: isochorismate synthase (IS), IPL, and CM (see Scheme 5). It has been observed that SS, IS, and IPL activities require the presence of  $Mg^{2+}$  in the active site, while the last one is observed for wild type in the absence of this cation in its active site (see Scheme 5). A note of caution has to be introduced at this point, as these results have been recently questioned by Ziebart and Toney, who have found that, after doubly purification, the CM activity was abolished or significantly reduced (Ziebart and Toney, 2010). Anyway, we can use this protein scaffold as a benchmark to study the CM activity and to check the viability of the conversion of chorismate to prephenate in its active site. Moreover, by comparison with the most efficient CM catalysts, EcCM and BsCM, we could try to suggest mutations that improve this activity, as previously done with the 1F7 CA.

Free-energy profiles, in terms of PMF, for the chorismate to prephenate reaction carried out in water and in the active site of BsCM, PchB, and MbtI are shown in Fig. 8. As expected, the catalytic efficiency of BsCM are much higher than that of PchB (Martí et al., 2008) and MbtI (Ferrer et al.,



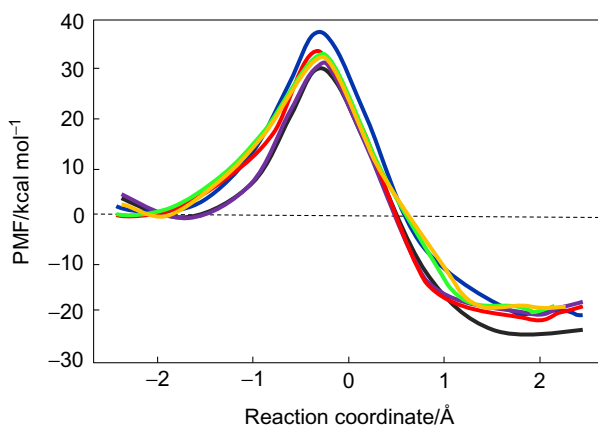


FIG. 8. Free-energy profiles (in terms of PMFs) for the chorismate to prephenate rearrangement obtained in different environments: BsCM (black line), PchB (red line), PchB-A38I (purple line), MbtI (green line), MbtI-I207F (yellow line), and in aqueous solution (blue line). The reaction coordinate is the antisymmetric combination of the interatomic distances of the breaking and forming bonds,  $C3 \cdots O7$  and  $C1 \cdots C9$ , respectively. EcCM renders a profile that almost superimpose to the one of BsCM, as shown in Fig. 6.

structure) active sites. Thus, a deeper insight into the substrate–protein pattern of interaction in EcCM reveals that there is a hydrophobic valine residue (Val35) that constrains the position of the ether bridge. The equivalent residue in PchB is a smaller alanine (Ala38) that cannot perform the same role. Thus, we decided to carry out an *in silico* mutation of this residue to a larger one (from Ala to Ile) and to repeat the PMF for the chorismate to prephenate reaction. The result of this mutation is that the mentioned dihedral angle evolution on the mutated PchB enzyme is closer to the CMs than to the native PchB or water. This geometrical behavior is reflected in the energetics, as shown in Fig. 8; the free-energy barrier in the mutated PchB is  $4.4 \text{ kcal mol}^{-1}$  lower than the native PchB, being only  $2.0 \text{ kcal mol}^{-1}$  above the primary reaction catalyzed by BsCM.

In fact, Mayo and coworkers, after performing 19 possible amino acid substitutions applied over six different positions of the engineered CM domain of the EcCM, obtained a Val35Ile mutation that renders an increase in the  $k_{\text{cat}}$  of about 1.5 times (Lassila et al., 2007). Moreover,

mutation of Val to Ala (the same residue present in PchB) reduces the  $k_{\text{cat}}$  by a factor of 2. Thus, the overall effect for a Ala35Ile mutation in EcCM is an increase of the  $k_{\text{cat}}$  by a factor of 3. Obviously, the predicted effect for the same mutation in the equivalent position in PchB (Ala38Ile) is much larger, which is consistent with the fact that this enzyme is not specialized in the catalysis of the chorismate to prephenate rearrangement. Our results are then both a prediction about the effect of a mutation on PchB and an interpretation about the success of Val35Ile mutation in EcCM. A mutation of Val by a similar but larger amino acid at position 35 would keep the enol pyruvyl moiety in a diaxial conformation, closer to the TS geometry, and reducing then the free-energy barrier. In this sense, QM/MM MD simulations of the Val35Ile variant of EcCM have been carried out, verifying that the proposed mutation increases the diaxial character of reactants.

In addition, we can check whether the primary activity of PchB (the IPL) is more robust than the secondary activity (CM) under selected mutations. In our simulations, the PchB-A38I mutant also displays an improved rate constant with respect to its primary activity, but the free-energy barrier diminution observed upon mutation is significantly lower than that obtained for its secondary activity (Martí et al., 2009).

In order to improve the CM activity of the highly promiscuous MbtI, and trying to increase the diaxial character of reactants in its active site, the 207 position initially occupied by an isoleucine has been exchanged by a phenylalanine, a much larger amino acid which, presumably, should be translated into a more constrained substrate, thus keeping the ether bridge into a more diaxial conformation and then favoring population of more reactive conformations. Our AM1/MM MD simulations show how this I207F mutation renders a diminution of the active site volume by about 13% (using a probe radius of 1.4 Å, we estimated that the averaged volumes of the active sites are 542 and 470 Å<sup>3</sup> for wild type and mutant, respectively). The resulting averaged structure after performing this mutation was superposed of reactant structures of wild type and mutant. The comparison shows how the two carbon atoms to be bounded, C1 and C9, appear at a shorter distance in the mutant than in the wild type (3.25 and 3.30 Å, respectively).

Another interesting effect of the I207F mutation is observed in the behavior of the O7–Arg405 interaction. Thus, while this distance is substantially increased from reactants to the TS in the wild-type enzyme (from



2.68 to 2.86 Å), this distance remains almost unchanged in the mutant (from 2.83 to 2.87 Å). As the interaction between an arginine residue (Arg90) and the ether oxygen stabilized the TS with respect to reactants in the BsCM (Martí et al., 2001, 2003a,b, 2004a,b), and considering that similar charge on O7 is obtained in reactants and TS in wild type and I207F, the smaller change of O7–Arg405 distance from reactants to the TS of the latter would contribute to reduce the free-energy barrier with respect to the wild-type. Obviously, the lengthening of the O7–Arg405 distance in the MC could also lead to a slight increase of  $K_M$ . The resulting free energy of activation render differences between the uncatalyzed reaction and the two catalyzed reactions, by BsCM and MbtI, of 8.7 and 3.8 kcal mol<sup>-1</sup>, respectively (Ferrer et al., 2011). These results predict that the catalytic rate constant for the CM activity of MbtI would be about 4000 times lower than that for BsCM, in very good agreement with the values deduced from the experimentally measured rate constants (Andrews et al., 1973; Kast et al., 1996a,b), from which activation free-energy differences of 9.1 and 4.7 kcal mol<sup>-1</sup> can be obtained applying transition state theory (TST) in its simplest version. The I207F mutant was 1.2 kcal mol<sup>-1</sup> lower than that for the wild-type MbtI, which would represent an enhancement of the rate constant by a factor of 7 at 310 K.

The agreement of our theoretical predictions with the experimental data obtained on the PchB (Lassila et al., 2005, 2007) and on the MbtI (Andrews et al., 1973; Kast et al., 1996a,b) allows being confident in our computational protocol. In this sense, we have been able to explain the origin of the effects observed in native enzymes after single mutations. As a consequence, improvement of the role of promiscuous enzymes can be guided by computational protein engineering. This method can be directly applied to the design of new enzymes, and the benchmark provides a powerful *in silico* test for guiding improvements in computational enzyme design.

#### IV. CONCLUSIONS AND OUTLOOK

In the past years, theoretical and technological advancements have produced an impressive improvement of computational facilities providing a wide range of methodologies, economically and conceptually accessible for a huge number of researchers in different fields of molecular sciences. Molecular modeling has established itself as an important

component of applied research in areas such as drug discovery, catalysis, and (bio)polymers. Improvements in the methods and techniques of theoretical and computational chemistry coupled with the increasing speed of computational hardware are making possible to perform predictive modeling on ever larger systems. Electronic structure calculations (Martin, 2004) represent nowadays one of the most commonly used approaches by the physical–chemical community, allowing highly accurate description of systems with a large number of atoms, that is, systems with an order of atoms of  $10^2$ – $10^3$  and more (Hung and Carter, 2009), and the results can have a significant impact on real-world problems. Simulation allows researchers to explore temporal and/or spatial domains that are not accessible by present experimental methods. For example, different chemical reaction pathways not directly accessible by experiment can be explored to learn why they are not favorable or to find missing steps in a complex multistep mechanism. Accurate simulations can actually replace experimental measurements that are too costly, too difficult, or too dangerous to perform. Computational chemistry has become an enabling tool for the design of processes for controlling and enabling chemical transformations, leading to higher selectivity and lower environmental impact and energy consumption. However, there is still a lot of work to do. As a matter of fact, modeling at the electronic level of systems with high configurational complexity is still challenging. In this sense, as Truhlar (2008) says, “Computations on complex systems are, in my opinion, the current frontier of theoretical chemistry.”

The main problem is both practical and conceptual as the different observables to be modeled depend on processes occurring at different length, energy, and time scales. Computational tools typically employed for systems of such dimensions are classical simulations which, however, produce reliable results as far as transitions in quantum degrees of freedom do not take place. However, when the observables of interest explicitly involve quantum degrees of freedom, for example, chemical reactions, their modeling should be derived from statistical averages of genuine quantum states interacting with fluctuating perturbing environments. In the past years, our group has been focusing its efforts in this direction, producing a theoretical–computational methodology, whose main feature is to describe at electronic level a portion of a large molecular system maintaining the complexity of the overall system.

After more than a half century of investigation into the origins of enzyme catalysis, it is gratifying to see the extensive new insights that can be gleaned. It is important to emphasize that the conceptual basis for enzyme catalysis has moved to a straightforward correlation among structure, dynamics, and function (Zalatan and Herschlag, 2009). Computer simulations can provide important insights into the energetic origins of substrate specificity and help to predict the effects of mutations quantitatively, as demonstrated from numerous previous investigations. Working in the area of computational and theoretical chemistry, we have, in the past decades, focused on the application of modern computational techniques, that is, quantum chemical calculations, to studies of structure, bonding, and chemical reactivity and have had opportunities to collaborate with a number of experimental chemists. We have applied MD simulations with hybrid QM/MM potentials to study several enzyme-catalyzed reactions earlier, related with CM.

This chapter reviews some of the modeling methods currently in use, providing illustrative examples of applications of the hybrid QM/MM MD simulations to challenges in CA and promiscuity of enzymes, and finally discusses prospects for future modeling approaches. This study has dealt with the development and application of QM/MM schemes into internal energy optimizations and MD simulations to study chemical rearrangement involved in the enzyme catalysis. The basis of the investigations presented in this chapter is the chorismate to prephenate rearrangement, and we have shown how to compute its electronic properties within complex environments. In this review, we have demonstrated how realistic and accurate simulations of reactions in solution and in enzymes become feasible by using QM/MM MD methods. The application of such methods will provide a comprehensive understanding of reactions in solution and in enzymes: enzyme mechanisms, the effect of mutations to understand the differential reactivity and selectivity, are becoming a key part of assessing the validity of mechanistic proposals. Our results are in good agreement with existing experimental findings. Moving forward, a number of extensions of this work should be explored. Most importantly, to make simulations of this type into a predictive tool, these ideas should be applied to a wider array of enzyme systems. The present juncture is an exciting one, providing intellectually provocative models of enzyme catalysis that suggest new avenues for experimental and computational investigations.

In the spirit of Richard Feynman's adage "what I cannot create, I do not understand" (Feynman, 1989), the ultimate challenge is to find a fundamental understanding of the intricacies of protein structure and function, that is, we need to answer the question: How does enzyme work? Specifically from a physical chemistry perspective, and of direct relevance to this review, we are starting to unravel some of the important aspects in enzyme catalysis; that is, how the characterization of reactants, products, possible intermediates, and TS encodes the chemical reactivity. Pursuing and understanding this is important because, with such relationships in hand, we can begin to build peptides and proteins to order. This is known as rational, or *de novo* protein design. Of course, this is the basic instinct of chemists, physicists, and engineers, but importantly and along with *ab initio* protein-structure prediction, it also provides the acid test of our understanding of the enzyme catalysis problem.

Zhang et al. point out that *ab initio* QM/MM MD with umbrella sampling can be considered as a state-of-the-art approach to simulate enzyme reactions (Zhou et al., 2010). Nevertheless, challenges remain, especially for further improvement of fast QM methods, more rapid mapping of free-energy surfaces and enhanced configurational sampling of biomolecules. The methodology presented above can be further extended to explore an even wider range of essential properties. Nonetheless, the present computational study provides a detailed characterization of the structure, dynamics, and reactivity of a series of important mutations. Additionally, rigorous experimental verification of designed models is essential in improving potential functions. The exclusion of conformational heterogeneity in proteins and conformational changes associated with catalysis in current enzyme designs could also explain the similar catalytic performance of *de novo* enzymes and their corresponding CAs. Focusing largely on the chemical step in the catalytic cycle, the reliance on TS models in the form of actual analogues or simulated theozymes was argued not to properly account for events such as substrate binding, product release, and conformational changes, hence capping the performance of present designs.

In summary, theoretical contributions to the design of catalysts from first principles come at many levels: (i) Modeling and understanding catalytic processes at the electronic/atomistic level, that is, proposing atomic structures, suggesting reaction pathways, computing reaction energetics, modeling reaction dynamics, characterization of stationary points

on PESs: minima (reactants, products, and possible intermediates) and TSs, and identifying key parameters controlling a catalytic process. (ii) Developing methods for bridging the large gaps in temporal and spatial scales that separate elementary molecular processes from the statistical behavior that governs chemical kinetics. (iii) Identifying general trends and unifying principles common or specific to various classes of catalytic phenomena: heterogeneous, homogeneous, and biological. At the level of electronic structure theory, there is no distinction between solid, molecular, and biological catalysts. (iv) Theory will therefore be an important component in the integration of the different subfields of catalysis. The development of a common language in heterogeneous, homogeneous, and enzyme catalysis. (v) Creating databases of both theoretical and experimental results and developing methodologies to perform data mining and optimization approaches to guide mutations as well as a design of new catalytic systems.

As noted above, a key goal for catalysis research is the integration of skills across a wide range of areas, including catalyst synthesis, catalyst characterization, determination of reaction pathways and the dynamics of elementary processes, and theoretical methods for predicting the structure of active centers and their catalytic properties. The direct coupling of theory and experiment is an extremely strong combination and is needed to advance catalytic science and our understanding of how to control chemical transformations. No single experiment reveals every detail and no calculation is perfect, but the combination provides the most profound and detailed insights into how chemical reactions proceed and how we can control their finest details. In general, a multidisciplinary approach has to be taken, comprising (i) first-principles description of the actual chemical reaction, not only of the TS but also of the complete reaction pathway including the physical processes (diffusion of substrate and/or product), (ii) more dynamic molecular modeling algorithms also accounting for protein flexibility. Future developments in this area, including new algorithms and simplified models, are expected to have a major impact on the rational design of tailor-made enzymes. These results will fuel the establishment of quantitative relationships between enzyme structure and its catalytic activity. In the long term, we expect that this will enable predictive enzyme engineering and truly *de novo* design of biocatalysts, the “holy grail” of biomolecules. Techniques like the *in silico* mutations shown above will be more common, which in turn is likely to lead to applications

in rational computational protein design. In this respect, the replay of [Lonsdale et al. \(2011\)](#) on a comment of [Canepa \(2010\)](#) emphasizing that energy barriers for enzyme-catalyzed reactions calculated with QM/MM methods can be in excellent agreement with activation energies derived from experiments, supporting the applicability of TST for enzymatic reactions can be remarked. [Roos et al. \(2009\)](#) have been published a key review giving a detailed description of the principles and concepts of conceptual DFT and highlighting its success to study enzymatic catalysis.

Combined experimental and theoretical model studies are successful in disentangling structure–reactivity relationships as demonstrated for enzyme-catalyzed reactions. Model systems may be characterized at the atomic level experimentally, which allows for direct comparison with theoretical modeling and allows useful correlations with systems of practical relevance. It is important to note that it is crucial to have appropriate experimental techniques at ones disposal to look at this. Although experimental methods usually tend to become more expensive with time, computational methods will become cheaper as computers become faster. In combination with new developments in the methods and techniques of theoretical and computational chemistry, this suggests that computational approaches for the discovery and development of catalysts hold great promise for the future. In this respect, the very recent work of Shaik and coworkers ([Cho et al., 2011](#)) on the reaction mechanism of allene oxide synthase (AOS) can be cited. [Difley et al. \(2010\)](#) have recently reviewed the basic concepts in first-principles modeling of the electronic properties of disordered organic semiconductors, by means of QM/MM techniques, allowing to incorporate the influence of the heterogeneous environment on the diabatic states. [Sushko et al. \(2010\)](#) have presented a QM/MM method for metal–organic interfaces, which incorporates contributions from long-range electron correlation, characteristic to metals and nonbonded interactions in organic systems. This method can be used to study structurally irregular systems. It is also important to cite the articles of Norskov et al. ([Christensen and Norskov, 2008](#); [Norskov et al., 2009](#)) on the molecular view of heterogeneous catalysis toward the computational design of solid catalysts by using a combination of theoretical methods, mainly DFT methods, detailed experiments on model systems, and synthesis and *in situ* characterization of catalysts, a complete atomic-scale insight into the structure and mechanism of surface-catalyzed reactions is provided. In addition, Norskov et al. used DFT to calculate the

formation energies of (2,2) nanorods, (3,3) nanotubes, and the (110) and (100) surfaces for the case of 15 different rutile and 8 different perovskite metal oxides (Mowbray et al., 2010). In this sense, the recent paper of Bligaard (2009) on the linear energy relations and the computational design of selective hydrogenation/dehydrogenation catalysts need to be cited, as well as the paper of Feibelman (2010) in which this author conclude that “applying DFT to decipher the meaning of well-characterized experimental data is apt to be more successful than to predict molecular level structure.”

Over the past century, we have accumulated a vast empirical knowledge of catalysis and catalysts for an enormous number of reactions. However, it is only recently that we are moving away from this empirical approach and we are attempting a rational design of materials tailored to specific reactions. Paul and Nardelli (2010) address the problem of the first-principles design of catalytic surfaces for the activation and reduction of carbon dioxide. Computational simulations have become a very important tool that complements experiments by bringing important information not easily obtained from experimental measures. We believe that this type of approach will expand in the coming years, enhancing the performance of conventional catalysis research.

At the heart of catalysis is the control of chemical transformations, and the ability to predict reaction rates is key to gaining a fundamental understanding of catalytic processes. Thus, computational studies need to put structure and dynamics on an equal footing. The ability to reliably predict reaction rates is lagging the ability to predict thermodynamics and kinetics. This is still the case for simple gas-phase processes and is certainly true for complex reactions in solution, in the large inorganic molecules relevant to homogeneous catalysis, in enzymes, and on surfaces for which even the dynamics of molecules moving on the surface are difficult to predict reliably. Characterization of the long-time and rare-event dynamics typical of catalyzed reactions is seriously limited for computational approaches and presents a challenge. In some cases (e.g., photocatalysis and charge transfer catalysis), quantum effects in dynamics are important, and methods are just now becoming available to treat such processes.

We stress that there are still several issues to be addressed by using QM/MM methods. One is the accurate calculation of excited electronic states using hybrid methods which are not present for the ground state and for conventional methods. Accurate modeling of excited states of large

molecular systems represents one of the greatest challenges to modern quantum chemistry. For example, in order to gain appropriate insight in catalytic phenomena involving transition metals, that is, metalloenzymes, where different electronic states can present similar energies, the nature of excited-state structures must be evaluated. In these cases, the ordering of the states may not be the same between different levels of theory, and one must ensure to combine corresponding states. Therefore, a careful investigation of the strengths and weaknesses of such methods is extremely important before their wide application in production calculations. In addition, a reaction could also be triggered by a transition of the enzymatic complex to a significantly different state from what is reached through thermal fluctuations such as electronic excited states in photoactivated enzymatic reactions. The basic argument for the importance of excited-state dynamics in enzyme processes is the lowering of the activation barrier. The ways of achieving the excited state, such as photoabsorption and ligand binding, have been discussed and exemplified in various cases of enzymes by [Petersen and Bohr \(2010\)](#). Very recently, [Deuss et al. \(2011\)](#) have been reviewed the progress in the design and application of ligand systems based on peptides and DNA and the development of artificial metalloenzymes.

Another research line to be developed is the investigation of the effect of pressure on the physical/chemical properties of biomolecules using QM/MM MD methods. This is essential for the knowledge of the intermolecular interactions and their effects on the molecular geometries, the electronic distributions, and ultimately, the chemical stability and reactivity. Indeed, the application of static pressure allows a fine-tuning of the intermolecular distances, without changing the temperature and composition of the system. This may have important applications in the study of biological systems at high pressure ([Meersman et al., 2006](#); [Winter et al., 2007](#); [Mishra and Winter, 2008](#)). Moreover, enzymes are not naked; they are dressed in a solvent. What is the explicit effect of the solvent on the dynamics of the reaction to be catalyzed? Are the various steps controlled by activation enthalpies or activation entropies? It is well recognized that QM/MM methods are best to describe short-ranged, explicit interactions, whereas continuum models are a better alternative for long-ranged, mediated interactions. The combination of the two methods by a three-layered QM/MM/PCM structure could therefore incorporate the strong points of both models by allowing an explicit description of solvation in



the vicinity of the solute and an implicit description through the continuum beyond a given solute–solvent distance. This goal has been achieved very recently by Kongsted and coworkers (Steindal et al., 2011). The novelty of this approach is in the development and implementation of a QM linear response formalism of a fully polarizable QM/MM/PCM interface, where both the MM and the PCM layers are self-consistently polarized.

Another open question is: How will the enzyme reaction change as a function of temperature and pressure? The application of pressure to molecular systems is known to produce both reversible and irreversible changes of the covalent bonds (Hemley, 2000; Schettino et al., 2005), when the intermolecular distances are sufficiently reduced. High-pressure chemistry may be very selective and may give unexpected products due to the constraints by which molecules are bound in the high-density phases: the high viscosity in liquids and the relative orientations and distances in crystals. However, in comparison with the temperature, the effect of pressure on both reaction mechanics and reaction rates in aqueous solvent and/or enzyme catalysis has not been explored thoroughly. Application of hydrostatic pressure allows tuning of both intermolecular and intramolecular interactions, and it can be used to understand the changes of chemical reactivity under the condition of external stimuli such as pressure. Therefore, accurate theoretical treatment and computational methods to study chemical reactivity of complex chemical systems, such as biomolecules, at different external pressure are necessary to support, complement, and also predict the outcome of experimental data (Heremans and Smeller, 1998; Ludwig, 1998; Kato and Hayashi, 1999; Silva et al., 2001).

Computers will not replace chemists, and data mining methods will not replace mechanistic studies. These methods will simply be part of the chemist's toolbox in the twenty-first century. As with this initial report, continued collaboration among experimentalists and theorists will be essential as we continue our research for understanding of enzyme catalysis phenomena. To close this section, we select the last paragraph of a very recent paper by Zaera (2010): "New catalytic materials should always be designed with particular applications in mind. Ideally, that design should be guided by the basic chemical principles extracted from mechanistic studies using surface science and theoretical tools, which should provide an indication of the type of active sites required to improve on the activity and, perhaps more importantly, the selectivity of the processes being

addressed. It is via this symbiotic relationship between fundamental mechanistic studies and new synthetic methodology that true advances can be expected in the quest to design highly selective catalysts in an effective and rational way from first principles.”

As theoretical and computational chemists continue to grapple with an ever-changing landscape of funding opportunities and justifications for our art, we do not hesitate to strongly advocate the case for exploiting enzyme for connecting chemistry as a discipline, to a broader swath of science. These research efforts, aided by design, will dictate how scientists and engineers design next-generation of biomolecules. Advances in merging catalysis are of course not possible without understanding catalytic systems on an atomic level. Molecular modeling is nowadays an indispensable research tool for catalyst tailoring (Catlow, 1996; Rothenberg, 2008), providing also a basis for the unification of all branches of catalysis. We have been additionally fortunate to have been able to interest and engage a wide range of collaborators in some of these projects, and their skills and expertise have made our journeys both fruitful and intellectually satisfying.

#### ACKNOWLEDGMENTS

This work is supported by Generalitat Valenciana (*Prometeo/2009/053* project) and by *Ministerio de Ciencia e Innovación* (project CTQ2009-14541-C02). The authors also acknowledge the Servei Informàtica of the Universitat Jaume I for generous allotment of computer time.

#### REFERENCES

- Acevedo, O., Jorgensen, W. L. (2010). Advances in quantum and molecular mechanical (QM/MM) simulations for organic and enzymatic reactions. *Acc. Chem. Res.* **43**(1), 142–151.
- Aemissegger, A., Jaun, B., et al. (2002). Investigation of the enzymatic and nonenzymatic cope rearrangement of carbaprephenate to carbachorismate. *J. Org. Chem.* **67** (19), 6725–6730.
- Aharoni, A., Gaidukov, L., et al. (2005). The ‘evolvability’ of promiscuous protein functions. *Nat. Genet.* **37**(1), 73–76.
- Alexandrova, A. N., Rothlisberger, D., et al. (2008). Catalytic mechanism and performance of computationally designed enzymes for Kemp elimination. *J. Am. Chem. Soc.* **130**(47), 15907–15915.
- Alonso, H., Bliznyuk, A. A., et al. (2006). Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **26**(5), 531–568.

- Amadei, A., Linssen, A. B. M., et al. (1993). Essential dynamics of proteins. *Proteins Struct. Funct. Bioinform.* **17**(4), 412–425.
- Andrews, P. R., Smith, G. D., et al. (1973). Transition-state stabilization and enzymic catalysis. Kinetic and molecular orbital studies of the rearrangement of chorismate to prephenate. *Biochemistry* **12**(18), 3492–3498.
- Arnold, F. H. (1998). Design by directed evolution. *Acc. Chem. Res.* **31**(3), 125–131.
- Babtie, A., Tokuriki, N., et al. (2010). What makes an enzyme promiscuous? *Curr. Opin. Chem. Biol.* **14**(2), 200–207.
- Backes, A. C., Hotta, K., et al. (2003). Promiscuity in antibody catalysis: esterolytic activity of the decarboxylase 21D8. *Helv. Chim. Acta* **86**(4), 1167–1174.
- Bakowies, D., Thiel, W. (1996). Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.* **100**(25), 10580–10594.
- Bandura, A. V., Sykes, D. G., et al. (2004). Adsorption of water on the TiO<sub>2</sub> (rutile) (110) surface: a comparison of periodic and embedded cluster calculations. *J. Phys. Chem. B* **108**(23), 7844–7853.
- Barbany, M., Gutiérrez-de-Terán, H., et al. (2003). On the generation of catalytic antibodies by transition state analogues. *Chembiochem* **4**(4), 277–285.
- Benkovic, S. J., Hammes-Schiffer, S. (2003). A perspective on enzyme catalysis. *Science* **301**(5637), 1196–1202.
- Bershtein, S., Tawfik, D. S. (2008). Ohno's Model revisited: measuring the frequency of potentially adaptive mutations under various mutational drifts. *Mol. Biol. Evol.* **25**(11), 2311–2318.
- Bligaard, T. (2009). Linear energy relations and the computational design of selective hydrogenation/dehydrogenation catalysts. *Angew. Chem. Int. Ed.* **48**(52), 9782–9784.
- Boehr, D. D., Nussinov, R., et al. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**(11), 789–796.
- Bolon, D. N., Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. USA* **98**, 14274–14279.
- Bornscheuer, U. T., Kazlauskas, R. J. (2004). Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways. *Angew. Chem. Int. Ed. Engl.* **43**(45), 6032–6040.
- Brent, R., Bruck, J. (2006). 2020 computing: can computers help to explain biology? *Nature* **440**(7083), 416–417.
- Bruice, T. C. (2002). A view at the millennium: the efficiency of enzymatic catalysis. *Acc. Chem. Res.* **35**(3), 139–148.
- Bursulaya, B. D., Totrov, M., et al. (2003). Comparative study of several algorithms for flexible ligand docking. *J. Comput. Aided Mol. Des.* **17**(11), 755–763.
- Busto, E., Gotor-Fernandez, V., et al. (2011). ChemInform abstract: hydrolases: catalytically promiscuous enzymes for non-conventional reactions in organic synthesis. *ChemInform* **42**(8), 4504–4523.
- Canepa, C. (2010). A stationary-wave model of enzyme catalysis. *J. Comput. Chem.* **31**(2), 343–350.
- Cannon, W. R., Benkovic, S. J. (1998). Solvation, reorganization energy, and biological catalysis. *J. Biol. Chem.* **273**(41), 26257–26260.

- Carlson, H. A., Jorgensen, W. L. (1996). Monte Carlo investigations of solvent effects on the chorismate to prephenate rearrangement. *J. Am. Chem. Soc.* **118**(35), 8475–8484.
- Carter, E. A., Rossky, P. (2006). Computational and theoretical chemistry. *J. Acc. Chem. Res.* **39**, 71–72.
- Catlow, R. (1996). Modelling of catalysts and catalysis. *J. Comput. Aided Mater. Des.* **3**(1), 56–60.
- Cho, K.-B., Lai, W., et al. (2011). The reaction mechanism of allene oxide synthase: interplay of theoretical QM/MM calculations and experimental investigations. *Arch. Biochem. Biophys.* **507**(1), 14–25.
- Chook, Y. M., Ke, H., et al. (1993). Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog. *Proc. Natl. Acad. Sci. USA* **90**, 8600.
- Christensen, C. H., Norskov, J. K. (2008). A molecular view of heterogeneous catalysis. *J. Chem. Phys.* **128**(18), 182503.
- Chuang, Y., Corchado, J. C., et al. (1999). Mapped interpolation scheme for single-point energy corrections in reaction rate calculations and a critical evaluation of dual-level reaction path dynamics methods. *J. Phys. Chem. A* **103**, 1140.
- Clark, T. (2000). Quo Vadis semiempirical MO-theory? *J. Mol. Struct.: THEOCHEM* **530** (1–2), 1–10.
- Clark, D. E. (2008). What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discov.* **3**(8), 841–851.
- Cleland, W. W., Frey, P. A., et al. (1998). The low barrier hydrogen bond in enzymatic catalysis. *J. Biol. Chem.* **273**(40), 25529–25532.
- Colombo, M. C., Guidoni, L., et al. (2002). Hybrid QM/MM Car-Parrinello simulations of catalytic and enzymatic reactions. *Chimia* **56**(1–2), 13–19.
- Copley, S. D., Knowles, J. R. (1987). The conformational equilibrium of chorismate in solution: implications for the mechanism of the non-enzymic and the enzyme-catalyzed rearrangement of chorismate to prephenate. *J. Am. Chem. Soc.* **109**, 5008.
- Corchado, J. C., Coitiño, E. L., et al. (1998). Interpolated variational transition-state theory by mapping. *J. Phys. Chem. A* **102**, 2424.
- Cramer, C. J., Truhlar, D. G. (1995). Continuum solvation models: classical and quantum mechanical implementations. In: *Reviews in Computational Chemistry*, Lipkowitz, K. B. and Boyd, D. B. (Eds.), 6, pp. 1–72. VCH Publishers, New York.
- Cui, Q., Karplus, M. (2000a). Molecular properties from combined QM/MM methods. 2. Chemical shifts in large molecules. *J. Phys. Chem. B* **104**(15), 3721–3743.
- Cui, Q., Karplus, M. (2000b). Molecular properties from combined QM/MM methods. I. Analytical second derivative and vibrational calculations. *J. Chem. Phys.* **112**(3), 1133–1149.
- Dahiyat, B. I., Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**(5), 895–903.
- Dahlke, E. E., Truhlar, D. G. (2007a). Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller-Plesset perturbation theory energies for large water clusters. *J. Chem. Theory Comput.* **3**(4), 1342–1348.

- Dahlke, E. E., Truhlar, D. G. (2007b). Electrostatically embedded many-body expansion for large systems, with applications to water clusters. *J. Chem. Theory Comput.* **3**(1), 46–53.
- Dal Peraro, M., Llarrull, L. I., et al. (2004). Water-assisted reaction mechanism of monozinc beta-lactamases. *J. Am. Chem. Soc.* **126**(39), 12661–12668.
- Damborsky, J., Brezovsky, J. (2009). Computational tools for designing and engineering biocatalysts. *Curr. Opin. Chem. Biol.* **13**(1), 26–34.
- Danyliv, O., Kantorovich, L., et al. (2007). Treating periodic systems using embedding: Adams-Gilbert approach. *Phys. Rev. B* **76**(4), 045107.
- Das, R., Baker, D. (2008). Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77**, 363–382.
- DeClue, M. S., Baldrige, K. K., et al. (2005). Isochorismate pyruvate lyase: a pericyclic reaction mechanism? *J. Am. Chem. Soc.* **127**(43), 15002–15003.
- Deeth, R. J. (2004). Computational bioinorganic chemistry. Principles and Applications of Density in Inorganic Chemistry II **113**, Springer, Berlinpp. 37–69.
- Deuss, P. J., den Heeten, R., et al. (2011). Bioinspired catalyst design and artificial metalloenzymes. *Chem. Eur. J* **17**, 4680–4698.
- Dewar, M. J. S., Zoebisch, E. G., et al. (1985). Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107**, 3902.
- Difley, S., Wang, L.-P., et al. (2010). Electronic properties of disordered organic semiconductors via QM/MM simulations. *Acc. Chem. Res.* **43**(7), 995–1004.
- Dwyer, M. A., Looger, L. L., et al. (2004). Computational design of a biologically active enzyme. *Science* **304**(5679), 1967–1971.
- Elstner, M., Porezag, D., et al. (1998). Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **58**(11), 7260.
- Faiella, M., Andreozzi, C., et al. (2009). An artificial di-iron oxo-protein with phenol oxidase activity. *Nat. Chem. Biol.* **5**(12), 882–884.
- Feibelman, P. (2010). DFT versus the “real world” (or, waiting for Godft). *Top. Catal.* **53**(5), 417–422.
- Ferreira, K. N., Iverson, T. M., et al. (2004). Architecture of the photosynthetic oxygen-evolving center. *Science* **303**(5665), 1831–1838.
- Ferrer, S., Martí, S., et al. (2011). Molecular mechanism of chorismate mutase activity of promiscuous MbtI. *Theor. Chem. Acc. Theory Comput. Model. (Theor. Chim. Acta)* **128**(4), 601–607.
- Feynman, R. (1989). Feynman’s office; the last blackboards. *Phys. Today* **42**(2), 88.
- Field, M. J., Bash, P. A., et al. (1990). A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **11**, 700–733.
- Gaille, C., Kast, P., et al. (2002). Salicylate biosynthesis in *Pseudomonas aeruginosa*. *J. Biol. Chem.* **277**(24), 21768–21775.
- Gao, J. (2007). Methods and Applications of Combined Quantum Mechanical and Molecular Mechanical Potentials. John Wiley & Sons, Inc., Hoboken, NJ.

- Gao, J., Freindorf, M. (1997). Hybrid ab initio QM/MM simulation of N-Methylacetamide in aqueous solution. *J. Phys. Chem. A* **101**(17), 3182–3188.
- Gao, J., Xia, X. (1992). A priori evaluation of aqueous polarization effects through Monte Carlo QM-MM simulations. *Science* **258**(5082), 631–635.
- Garcia-Viloca, M., Gao, J., et al. (2004). How enzymes work: analysis by modern rate theory and computer simulations. *Science* **303**(5655), 186–195.
- Gerlt, J. A., Babbitt, P. C. (2009). Enzyme (re)design: lessons from natural evolution and computation. *Curr. Opin. Chem. Biol.* **13**(1), 10–18.
- Glasner, M. E., Gerlt, J. A., et al. (2006). Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* **10**(5), 492–497.
- Golynskiy, M. V., Seelig, B. (2010). De novo enzymes: from computational design to mRNA display. *Trends Biotechnol.* **28**, 340–345.
- Guo, H., Cui, Q., et al. (2001). Substrate conformational transitions in the active site of chorismate mutase: their role in the catalytic mechanism. *Proc. Natl. Acad. Sci. USA* **98**(16), 9032–9037.
- Gustin, D. J., Mattei, P., et al. (1999). Heavy atom isotope effects reveal a highly polarized transition state for chorismate mutase. *J. Am. Chem. Soc.* **121**, 1756.
- Hall, R. J., Hindle, S. A., et al. (2000). Aspects of hybrid QM/MM calculations: the treatment of the QM/MM interface region and geometry optimization with an application to chorismate mutase. *J. Comput. Chem.* **21**(16), 1433–1441.
- Hamelberg, D., Mongan, J., et al. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**(24), 11919–11929.
- Haslam, E. (1993). Shikimic Acid: Metabolism and Metabolites. Wiley, New York.
- Haynes, M., Stura, E., et al. (1994). Routes to catalysis: structure of a catalytic antibody and comparison with its natural counterpart. *Science* **263**(5147), 646–652.
- Hehre, W. J., Radom, L., et al. (1986). Ab Initio Molecular Orbital Theory. Wiley, USA.
- Hellinga, H. W., Richards, F. M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. *J. Mol. Biol.* **222**(3), 763–785.
- Hemley, R. J. (2000). Effects of high pressure on molecules. *Annu. Rev. Phys. Chem.* **51**(1), 763–800.
- Heremans, K., Smeller, L. (1998). Protein structure and dynamics at high pressure. *Biochim. Biophys. Acta* **1386**(2), 353–370.
- Herrmann, W. A., Cornils, B. (2002). Applied Homogeneous Catalysis with Organometallic Compounds. Wiley, New York.
- Hilvert, D., Carpenter, S. H., et al. (1988). Catalysis of concerted reactions by antibodies: the Claisen rearrangement. *Proc. Natl. Acad. Sci. USA* **85**(14), 4953–4955.
- Hilvert, D., Nared, K. D. (1988). Stereospecific Claisen rearrangement catalyzed by an antibody. *J. Am. Chem. Soc.* **110**(16), 5593–5594.
- Hu, H., Yang, W. (2008). Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods. *Annu. Rev. Phys. Chem.* **59**, 573–601.

- Hu, H., Yang, W. (2009). Development and application of ab initio QM/MM methods for mechanistic simulation of reactions in solution and in enzymes. *J. Mol. Struct.: THEOCHEM* **898**(1–3), 17–30.
- Hult, K., Berglund, P. (2007). Enzyme promiscuity: mechanism and applications. *Trends Biotechnol.* **25**(5), 231–238.
- Hung, L., Carter, E. A. (2009). Accurate simulations of metals at the mesoscale: explicit treatment of 1 million atoms with quantum mechanics. *Chem. Phys. Lett.* **475**(4–6), 163–170.
- Hur, S., Bruice, T. C. (2003). Enzymes do what is expected (chalcone isomerase versus chorismate mutase). *J. Am. Chem. Soc.* **125**, 1472.
- Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**(1), 409–425.
- Jiang, L., Althoff, E. A., et al. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**(5868), 1387–1391.
- Jonas, S., Hollfelder, F. (2008). Mechanism and Catalytic Promiscuity: Emerging Mechanistic Principles for Identification and Manipulation of Catalytically Promiscuous Enzymes. In: *The Protein Engineering Handbook*, Bornscheuer, U. and Lutz, S.(Eds.). Vol. 1. Wiley, Chichester.
- Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**(6), 724–733.
- Kamerlin, S. C., Haranczyk, M., et al. (2009). Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of pKa, redox reactions and solvation free energies. *J. Phys. Chem. B* **113**(5), 1253–1272.
- Kangas, E., Tidor, B. (2001). Electrostatic complementarity at ligand binding sites: application to chorismate mutase. *J. Phys. Chem. B* **105**(4), 880–888.
- Kaplan, J., DeGrado, W. F. (2004). De novo design of catalytic proteins. *Proc. Natl. Acad. Sci. USA* **101**, 11566–11570.
- Karplus, M., Gao, Y. Q., et al. (2005). Protein structural transitions and their functional role. *Philos. Transact. A Math. Phys. Eng. Sci.* **363**(1827), 331–355 discussion 355–336.
- Karplus, M., McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**(9), 646–652.
- Kast, P., Asif-Ullah, M., et al. (1996a). Is chorismate mutase a prototypic entropy trap? - Activation parameters for the *Bacillus subtilis* enzyme. *Tetrahedron Lett.* **37**, 2691.
- Kast, P., Asif-Ullah, M., et al. (1996b). Exploring the active site of chorismate mutase by combinatorial mutagenesis and selection: the importance of electrostatic catalysis. *Proc. Natl. Acad. Sci. USA* **93**(10), 5043–5048.
- Kastner, J., Senn, H. M., et al. (2006). QM/MM free-energy perturbation compared to thermodynamic integration and umbrella sampling: application to an enzymatic reaction. *J. Chem. Theory Comput.* **2**(2), 452–461.
- Kastner, J., Thiel, S., et al. (2007). Exploiting QM/MM capabilities in geometry optimization: a microiterative approach using electrostatic embedding. *J. Chem. Theory Comput.* **3**(3), 1064–1072.

- Kato, M., Hayashi, R. (1999). Effects of high pressure on lipids and biomembranes for understanding high-pressure-induced biological phenomena. *Biosci. Biotechnol. Biochem.* **63**(8), 1321–1328.
- Khanjin, N. A., Snyder, J. P., et al. (1999). Mechanism of chorismate mutase: contribution of conformational restriction to catalysis in the claisen rearrangement. *J. Am. Chem. Soc.* **121**, 11831.
- Khersonsky, O., Roodveldt, C., et al. (2006). Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10**(5), 498–508.
- Kienhöfer, A., Kast, P., et al. (2003). Selective stabilization of the chorismate mutase transition state by a positively charged hydrogen bond donor. *J. Am. Chem. Soc.* **125**(11), 3206–3207.
- Kimmel, A. V., Sushko, P. V., et al. (2008). Effect of molecular and lattice structure on hydrogen transfer in molecular crystals of diamino-dinitroethylene and triamino-trinitrobenzene. *J. Phys. Chem. A* **112**(19), 4496–4500.
- Koch, W., Holthausen, M. C. (2001). *A Chemist's Guide to Density Functional Theory*. Wiley, Weinheim, Germany.
- Kollman, P. A., Kuhn, B., et al. (2001). Elucidating the nature of enzyme catalysis utilizing a new twist on an old methodology: quantum mechanical–free energy calculations on chemical reactions in enzymes and in aqueous solution. *Acc. Chem. Res.* **34**, 72.
- Kontoyianni, M., McClellan, L. M., et al. (2004). Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **47**(3), 558–565.
- Kumar, S., Rosenberg, J. M., et al. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**(8), 1011–1021.
- Künzler, D. E., Sasso, S., et al. (2005). Mechanistic insights into the isochorismate pyruvate lyase activity of the catalytically promiscuous PchB from combinatorial mutagenesis and selection. *J. Biol. Chem.* **280**(38), 32827–32834.
- Laio, A., Parrinello, M. (2002). Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**(20), 12562–12566.
- Laio, A., VandeVondele, J., et al. (2002a). D-RESP: dynamically generated electrostatic potential derived charges from quantum mechanics/molecular mechanics simulations. *J. Phys. Chem. B* **106**(29), 7300–7307.
- Laio, A., VandeVondele, J., et al. (2002b). A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. *J. Chem. Phys.* **116**(16), 6941–6947.
- Lassila, J. K., Keeffe, J. R., et al. (2005). Computationally designed variants of *Escherichia coli* chorismate mutase show altered catalytic activity. *Protein Eng. Des. Sel.* **18**(4), 161–163.
- Lassila, J. K., Keeffe, J. R., et al. (2007). Exhaustive Mutagenesis of six secondary active-site residues in *Escherichia coli* chorismate mutase shows the importance of hydrophobic side chains and a helix N-Capping position for stability and catalysis. *Biochemistry* **46**(23), 6883–6891.
- Lee, A. Y., Karplus, P. A., et al. (1995). Atomic structure of the buried catalytic pocket of *Escherichia coli* chorismate mutase. *J. Am. Chem. Soc.* **117**(12), 3627–3628.



- Lee, Y. S., Worthington, S. E., et al. (2002). Reaction mechanism of chorismate mutase studied by the combined potentials of quantum mechanics and molecular mechanics. *J. Phys. Chem. B* **106**(46), 12059–12065.
- Lin, H., Truhlar, D. G. (2007). QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **117**(2), 185–199.
- Lodola, A., Mor, M., et al. (2008). Identification of productive inhibitor binding orientation in fatty acid amide hydrolase (FAAH) by QM/MM mechanistic modeling. *Chem. Commun. (Camb.)* (2), 214–216.
- Lonsdale, R., Harvey, J. N., et al. (2011). Comment on "A Stationary-Wave Model of Enzyme Catalysis" By Carlo Canepa. *J. Comput. Chem.* **32**(2), 368–369.
- Lonsdale, R., Ranaghan, K. E., et al. (2010). Computational enzymology. *Chem. Commun. (Camb.)* **46**(14), 2354–2372.
- Lu, Y., Yeung, N., et al. (2009). Design of functional metalloproteins. *Nature* **460**(7257), 855–862.
- Ludwig, H. (1998). *Advances in High Pressure Bioscience and Biotechnology*. Springer-Verlag, Berlin.
- Lutz, S., Bornscheuer, U. (2009). *Protein Engineering Handbook*. Wiley-VCH, Weinheim.
- Lyne, P. D., Mulholland, A. J., et al. (1995). Insights into chorismate mutase catalysis from a combined QM/MM simulation of the enzyme reaction. *J. Am. Chem. Soc.* **117**(45), 11345–11350.
- Ma, B., Nussinov, R. (2010). Enzyme dynamics point to stepwise conformational selection in catalysis. *Curr. Opin. Chem. Biol.* **14**(5), 652–659.
- Mader, M. M., Bartlett, P. A. (1997). Binding energy and catalysis: the implications for transition-state analogs and catalytic antibodies. *Chem. Rev.* **97**(5), 1281–1302.
- Magistrato, A., Woo, T. K., et al. (2004). Enantioselective palladium-catalyzed hydro-silylation of styrene: detailed reaction mechanism from first-principles and hybrid QM/MM molecular dynamics simulations. *Organometallics* **23**(13), 3218–3227.
- Maglió, O., Nistri, F., et al. (2007). Diiron-containing metalloproteins: developing functional models. *C. R. Chim.* **10**(8), 703–720.
- Mandal, A., Hilvert, D. (2003). Charge Optimization Increases the Potency and Selectivity of a Chorismate Mutase Inhibitor. *J. Am. Chem. Soc.* **125**, 5598.
- Martí, S., Andrés, J., et al. (2008). Computational design of biological catalyts. *Chem. Soc. Rev.* **37**(12), 2634–2643.
- Martí, S., Andrés, J., et al. (2001). A hybrid potential reaction path and free energy study of the chorismate mutase reaction. *J. Am. Chem. Soc.* **123**(8), 1709–1712.
- Martí, S., Andrés, J., et al. (2003). Preorganization and reorganization as related factors in enzyme catalysis: the chorismate mutase case. *Chem. Eur. J.* **9**(4), 984–991.
- Martí, S., Andrés, J., et al. (2004). A comparative study of claisen and cope Rearrangements catalyzed by chorismate mutase. An insight into enzymatic efficiency: transition state stabilization or substrate preorganization? *J. Am. Chem. Soc.* **126**(1), 311–319.
- Martí, S., Andrés, J., et al. (2007). Computer-aided rational design of catalytic antibodies: the 1F7 case. *Angew. Chem. Int. Ed.* **46**(1–2), 286–290.

- Martí, S., Andrés, J., et al. (2008). Predicting an improvement of secondary catalytic activity of promiscuous isochorismate pyruvate lyase by computational design. *J. Am. Chem. Soc.* **130**(10), 2894–2895.
- Martí, S., Andrés, J., et al. (2009). Mechanism and plasticity of isochorismate pyruvate lyase by computational study. *J. Am. Chem. Soc.* **131**(44), 16156–16161.
- Martí, S., Moliner, V., et al. (2003). QM/MM calculations on kinetic isotope effects in the chorismate mutase active site. *Org. Biomol. Chem.* **1**, 483.
- Martí, S., Moliner, V., et al. (2005a). Improving the QM/MM description of chemical processes: a dual level strategy to explore the potential energy surface in very large systems. *J. Chem. Theory Comp.* **1**(5), 1008–1016.
- Martí, S., Moliner, V., et al. (2005b). Computing kinetic isotope effects for chorismate mutase with high accuracy. A new DFT/MM Strategy. *J. Phys. Chem. B* **109**, 3707.
- Martí, S., Roca, M., et al. (2004). Theoretical insights in enzyme catalysis. *Chem. Soc. Rev.* **33**(2), 98–107.
- Martin, R. M. (2004). *Electronic Structure. Basic Theory and Practical Methods*. Cambridge University Press, Cambridge, UK.
- Martin, M. E., Sanchez, M. L., et al. (2000). A multiconfiguration self-consistent field/molecular dynamics study of the ( $n \rightarrow \pi^*$ ) transition of carbonyl compounds in liquid water. *J. Chem. Phys.* **113**(15), 6308–6315.
- Mata, R. A. (2010). Application of high level wavefunction methods in quantum mechanics/molecular mechanics hybrid schemes. *Phys. Chem. Chem. Phys.* **12**(19), 5041–5052.
- Meersman, F., Dobson, C. M., et al. (2006). Protein unfolding, amyloid fibril formation and configurational energy landscapes under high pressure conditions. *Chem. Soc. Rev.* **35**(10), 908–917.
- Mishra, R., Winter, R. (2008). Cold- and pressure-induced dissociation of protein aggregates and amyloid fibrils. *Angew. Chem. Int. Ed.* **47**(35), 6518–6521.
- Moliner, V., Turner, A. J., et al. (1997). Transition-state structural refinement with GRACE and CHARMM: realistic modelling of lactate dehydrogenase using a combined quantum/classical method. *J. Chem. Soc. Chem. Commun.* 1271.
- Monard, G., Merz, K. M. (1999). Combined quantum mechanical/molecular mechanical methodologies applied to biomolecular systems. *Acc. Chem. Res.* **32**(10), 904–911.
- Monard, G., Prat-Resina, X., et al. (2003). Determination of enzymatic reaction pathways using qm/mm methods. *Int. J. Quantum. Chem.* **93**, 229–244.
- Morley, K. L., Kazlauskas, R. J. (2005). Improving enzyme properties: when are closer mutations better? *Trends Biotechnol.* **23**(5), 231–237.
- Mowbray, D. J., Martínez, J. I., et al. (2010). Trends in metal oxide stability for nanorods, nanotubes, and surfaces. *J. Phys. Chem. C* **115**(5), 2244–2252.
- Mulholland, A. J. (2005). Modelling enzyme reaction mechanisms, specificity and catalysis. *Drug Discov. Today* **10**(20), 1393–1402.
- Mundorff, E. C., Hanson, M. A., et al. (2000). Conformational effects in biological catalysis: an antibody-catalyzed oxy-cope rearrangement. *Biochemistry* **39**(4), 627–632.

- Murphy, P. M., Bolduc, J. M., et al. (2009). Alteration of enzyme specificity by computational loop remodeling and design. *Proc. Natl. Acad. Sci. USA* **106**(23), 9215–9220.
- Mysovsky, A. S., Sushko, P. V., et al. (2004). Calibration of embedded-cluster method for defect studies in amorphous silica. *Phys. Rev. B* **69**(8), 085202.
- Nanda, V., Koder, R. L. (2010). Designing artificial enzymes by intuition and computation. *Nat. Chem.* **2**(1), 15–24.
- Neese, F. (2006). A critical evaluation of DFT, including time-dependent DFT, applied to bioinorganic chemistry. *J. Biol. Inorg. Chem.* **11**(6), 702–711.
- Neet, K. E. (1998). Enzyme catalytic power minireview series. *J. Biol. Chem.* **273**(40), 25527–25528.
- Nguyen, K. A., Rossi, I., et al. (1995). A dual-level Shepard interpolation method for generating potential energy surfaces for dynamics calculations. *J. Chem. Phys.* **103**, 5522–5531.
- Nobeli, I., Favia, A. D., et al. (2009). Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol.* **27**(2), 157–167.
- Norskov, J. K., Bligaard, T., et al. (2009). Towards the computational design of solid catalysts. *Nat. Chem.* **1**(1), 37–46.
- O'Brien, P. J., Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**(4), R91–R105.
- Parr, R. G., Yang, W. (1989). Density Functional Theory of Atoms and Molecules. Oxford University Press, New York.
- Paul, S., Nardelli, M. B. (2010). Rational computational design of optimal catalytic surfaces. *Appl. Phys. Lett.* **97**(23), 233108.
- Pauling, L. (1946). Molecular architecture and biological reactions. *Chem. Eng. News* **24**, 1375–1377.
- Pauling, L. (1948a). Molecular architecture and the processes of life. *Am. Sci.* **36**, 51.
- Pauling, L. (1948b). Nature of forces between large molecules of biological interest. *Nature* **161**, 707–709.
- Pauling, L. (1960). The Nature of Chemical Bond. Cornell University Press, Ithaca, NY.
- Petersen, F., Bohr, H. (2010). The mechanisms of excited states in enzymes. *Theor. Chem. Acc. Theory Comput. Model. (Theor. Chim. Acta)* **125**(3), 345–352.
- Piana, S., Bucher, D., et al. (2004). Reaction mechanism of HIV-1 protease by hybrid carparinello/classical MD simulations. *J. Phys. Chem. B* **108**(30), 11139–11149.
- Prat-Resina, X., Bofill, J. M., et al. (2004). Geometry optimization and transition state search in enzymes: different options in the micro-iterative method. *Int. J. Quantum. Chem.* **98**, 367–377.
- Prat-Resina, X., González-Lafont, A., et al. (2003). How important is the refinement of transition state structures in enzymatic reactions? *J. Mol. Struct. Theochem* **632**, 297.
- Proust-De Martín, F., Dumas, R., et al. (2000). A hybrid-potential free-energy study of the isomerization step of the acetoxyhydroxy acid isomeroreductase reaction. *J. Am. Chem. Soc.* **122**, 7688.
- Raso, V., Stollar, B. D. (1975). Antibody-enzyme analogy. Characterization of antibodies to phosphopyridoxyltyrosine derivatives. *Biochemistry* **14**(3), 584–591.
- Renka, R. (1993). ALGORITHM 716. TSPACK: tension spline curve fitting package. *J. ACM Trans. Math. Software* **19**, 81–94.

- Riccardi, D., Schaefer, P., et al. (2006). Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. *J. Phys. Chem. B* **110**(13), 6458–6469.
- Rivail, J. L., Rinaldi, D., et al. (1991). Theoretical and Computational Models for Organic Chemistry. Kluwer, Dordrecht.
- Robles, V. M., Ortega-Carrasco, E., et al. (2011). What can molecular modelling bring to the design of artificial inorganic cofactors? *Faraday Discuss.* **148**, 137–159.
- Rohrig, U. F., Frank, I., et al. (2003). QM/MM Car-Parrinello molecular dynamics study of the solvent effects on the ground state and on the first excited singlet state of acetone in water. *Chemphyschem* **4**(11), 1177–1182.
- Roos, G., Geerlings, P., et al. (2009). Enzymatic catalysis: the emerging role of conceptual density functional theory. *J. Phys. Chem. B* **113**(41), 13465–13475.
- Rothenberg, G. (2008). Catalysis: Concepts and Green Applications. Wiley-VCH, Weinheim.
- Rothlisberger, D., Khersonsky, O., et al. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**(7192), 190–195.
- Roux, B. (1995). The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **91**, 275.
- Ruiz-Pernia, J. J., Silla, E., et al. (2004). Hybrid QM/MM potentials of mean force with interpolated corrections. *J. Phys. Chem. B* **108**(24), 8427–8433.
- Ruiz-Pernia, J. J., Silla, E., et al. (2006). Hybrid quantum mechanics/molecular mechanics simulations with two-dimensional interpolated corrections: application to enzymatic processes. *J. Phys. Chem. B* **110**(35), 17663–17670.
- Sanderson, K. (2011). Chemistry: enzyme expertise. *Nature* **471**, 397.
- Schettino, V., Bini, R., et al. (2005). Chemical Reactions at Very High Pressure. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Schramm, V. L. (2003). Enzymatic transition state poise and transition state analogues. *Acc. Chem. Res.* **36**(8), 588–596.
- Schultz, P. G., Lerner, R. A. (1993). Antibody catalysis of difficult chemical transformations. *Acc. Chem. Res.* **26**(8), 391–395.
- Schultz, P., Lerner, R. (1995). From molecular diversity to catalysis: lessons from the immune system. *Science* **269**(5232), 1835–1842.
- Sebastiani, D., Rothlisberger, U. (2004). Nuclear magnetic resonance chemical shifts from hybrid DFT QM/MM calculations. *J. Phys. Chem. B* **108**(9), 2807–2815.
- Senn, H., Thiel, W. (2007a). QM/MM methods for biological systems. In: Atomistic Approaches in Modern Biology, Reiher, M. (Ed.), 268, pp. 173–290. Springer, Berlin.
- Senn, H. M., Thiel, W. (2007b). QM/MM studies of enzymes. *Curr. Opin. Chem. Biol.* **11**(2), 182–187.
- Senn, H. M., Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed Engl.* **48**(7), 1198–1229.
- Sherwood, P. (1998). QM/MM approaches for metal oxide, zeolite, and enzyme systems. *J. Mol. Graph. Model.* **16**(4–6), 275–284.

- Sherwood, P., de Vries, A. H., et al. (2003). QUASI: a general purpose implementation of the QM/MM approach and its application to problems in catalysis. *J. Mol. Struct. Theochem* **632**, 1–28.
- Siegbahn, P. E., Himo, F. (2009). Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *J. Biol. Inorg. Chem.* **14**(5), 643–651.
- Siegbahn, P. E. M., Tye, J. W., et al. (2007). Computational studies of [NiFe] and [FeFe] hydrogenases. *Chem. Rev.* **107**(10), 4414–4435.
- Siegel, J. B., Zanghellini, A., et al. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329**(5989), 309–313.
- Silva, J. L., Foguel, D., et al. (2001). Pressure provides new insights into protein folding, dynamics and structure. *Trends Biochem. Sci.* **26**(10), 612–618.
- Sokol, A. A., Bromley, S. T., et al. (2004). Hybrid QM/MM embedding approach for the treatment of localized surface states in ionic materials. *Int. J. Quantum Chem.* **99**(5), 695–712.
- Sousa, S. F., Ramos, M. J. (2008). In: Computational Proteomics, Ramos, M. J. (Ed.), p. 101. Transworld Research Network, India, Kerala.
- Stefanovich, E. V., Truong, T. N. (1997). Theoretical approach for modeling reactivity at solid-liquid interfaces. *J. Chem. Phys.* **106**(18), 7700–7705.
- Steindal, A. H., Ruud, K., et al. (2011). Excitation energies in solution: the fully polarizable QM/MM/PCM method. *J. Phys. Chem. B* **115**(12), 3027–3037.
- Strajbl, M., Shurki, A., et al. (2003). Apparent NAC effect in chorismate mutase reflects electrostatic transition state stabilization. *J. Am. Chem. Soc.* **125**(34), 10228–10237.
- Sulimov, V. B., Sushko, P. V., et al. (2002). Asymmetry and long-range character of lattice deformation by neutral oxygen vacancy in alpha-quartz. *Phys. Rev. B* **66**(2), 024108.
- Summa, C. M., Rosenblatt, M. M., et al. (2002). Computational de novo design, and characterization of an A(2)B(2) diiron protein. *J. Mol. Biol.* **321**(5), 923–938.
- Sushko, P. V., Shluger, A. L., et al. (2000). Relative energies of surface and defect states: ab initio calculations for the MgO(001) surface. *Surf. Sci.* **450**(3), 153–170.
- Sushko, M. L., Sushko, P. V., et al. (2010). QM/MM method for metal-organic interfaces. *J. Comput. Chem.* **31**(16), 2955–2966.
- Szabo, A., Ostlund, N. S. (1996). Modern Quantum Chemistry. Dover publications, Inc., Mineola, New York.
- Szefczyk, B., Mulholland, A. J., et al. (2004). Differential transition-state stabilization in enzyme catalysis: quantum chemical analysis of interactions in the chorismate mutase reaction and prediction of the optimal catalytic field. *J. Am. Chem. Soc.* **126**(49), 16148–16159.
- Tawfik, D. S., Khersonsky, O. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**(1), 471–505.
- Taylor, S. V., Kast, P., et al. (2001). Investigating and engineering enzymes by genetic selection. *Angew. Chem. Int. Ed.* **40**(18), 3310–3335.
- Thiel, W. (2007). Perspectives on Semiempirical Molecular Orbital Theory. John Wiley & Sons, Inc., Hoboken, NJ.
- Thomas, G. (2007). Medicinal Chemistry. Wiley, Chichesterpp. 127–129.

- Thompson, M. A. (1996). QM/MMpol: a consistent model for solute/solvent polarization. Application to the aqueous solvation and spectroscopy of formaldehyde, acetaldehyde, and acetone. *J. Phys. Chem.* **100**(34), 14492–14507.
- Tomasi, J., Persico, M. (1994). Molecular interactions in solution: an overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **94**, 2027.
- Torrie, G. M., Valleau, J. P. (1974). Monte Carlo free energy estimates using non-Boltzmann sampling: application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.* **28**(4), 578–581.
- Toscano, M. D., Woycechowsky, K. J., et al. (2007). Minimalist active-site redesign: teaching old enzymes new tricks. *Angew. Chem. Int. Ed.* **46**(18), 3212–3236.
- Truhlar, D. G. (2008). Molecular modeling of complex chemical systems. *J. Am. Chem. Soc.* **130**(50), 16824–16827.
- Turner, A. J., Moliner, V., et al. (1999). Transition-state structural refinement with GRACE and CHARMM: flexible QM/MM modelling for lactate dehydrogenase. *J. Phys. Chem. Chem. Phys.* **1**, 1323.
- Ulrich, H. D., Mundorff, E., et al. (1997). The interplay between binding energy and catalysis in the evolution of a catalytic antibody. *Nature* **389**(6648), 271–275.
- VandeVondele, J., Colombo, M. C., et al. (2002). QM/MM study of the copper binding site of prion protein. *Biophys. J.* **82**(1), 2377.
- Villà, J., Warshel, A. (2001). Energetics and dynamics of enzymatic reactions. *J. Phys. Chem. B* **105**, 7887.
- Vreven, T., Morokuma, K. (2006). Chapter 3 hybrid methods: ONIOM(QM:MM) and QM/MM. In: Annual Reports in Computational Chemistry, David, C. S. (Ed.), Vol. 2, pp. 35–51. Elsevier, USA.
- Vreven, T., Morokuma, K., et al. (2003). Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints. *J. Comput. Chem.* **24**, 760.
- Warshel, A. (1991). Computer Modeling of Chemical Reactions in Enzymes and Solutions. Wiley, New York.
- Warshel, A. (1998). Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **273**(42), 27035–27038.
- Warshel, A. (2003). Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 425–443.
- Warshel, A., Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**(2), 227–249.
- Warshel, A., Sharma, P. K., et al. (2006). Electrostatic basis for enzyme catalysis. *Chem. Rev.* **106**(8), 3210–3235.
- Williams, I. H. (2010). Catalysis: transition-state molecular recognition? *Beilstein J. Org. Chem.* **6**, 1026–1034.
- Winter, R., Lopes, D., et al. (2007). Towards an understanding of the temperature/pressure configurational and free-energy landscape of biomolecules. *J. Non-Equilib. Thermodyn.* **32**(1), 41–97.
- Wolfenden, R. (1972). Analog approaches to the structure of the transition state in enzyme reactions. *Acc. Chem. Res.* **5**(1), 10–18.

- Wu, Q., Liu, B. K., et al. (2010). Enzymatic promiscuity for organic synthesis and cascade process. *Curr. Org. Chem.* **14**, 1966–1988.
- Ytreberg, F. M., Swendsen, R. H., et al. (2006). Comparison of free energy methods for molecular systems. *J. Chem. Phys.* **125**(18), 184114.
- Zaera, F. (2010). The new materials science of catalysis: toward controlling selectivity by designing the structure of the active site. *J. Phys. Chem. Lett.* **1**(3), 621–627.
- Zalatan, J. G., Herschlag, D. (2009). The far reaches of enzymology. *Nat. Chem. Biol.* **5**(8), 516–520.
- Zanghellini, A., Jiang, L., et al. (2006). New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**(12), 2785–2794.
- Zhang, Y. (2006). Pseudobond ab initio QM/MM approach and its applications to enzyme reactions. *Theor. Chem. Acc. Theory Comput. Model. (Theor. Chim. Acta)* **116**(1), 43–50.
- Zhang, Y., Lin, H., et al. (2007). Self-consistent polarization of the boundary in the redistributed charge and dipole scheme for combined quantum-mechanical and molecular-mechanical calculations. *J. Chem. Theory Comput.* **3**(4), 1378–1398.
- Zhou, Y., Wang, S., et al. (2010). Catalytic reaction mechanism of acetylcholinesterase determined by Born–Oppenheimer ab initio QM/MM molecular dynamics simulations. *J. Phys. Chem. B* **114**(26), 8817–8825.
- Ziebart, K. T., Toney, M. D. (2010). Nucleophile specificity in anthranilate synthase, aminodeoxychorismate synthase, isochorismate synthase, and salicylate synthase. *Biochemistry* **49**(13), 2851–2859.

# EXPLORING MEMBRANE AND PROTEIN DYNAMICS WITH DISSIPATIVE PARTICLE DYNAMICS

By GERNOT GUIGAS,<sup>\*</sup> DIANA MOROZOVA,<sup>†</sup> AND MATTHIAS WEISS<sup>\*,†</sup>

<sup>\*</sup>Experimental Physics I, University of Bayreuth, Bayreuth, Germany

<sup>†</sup>Cellular Biophysics Group, German Cancer Research Center, Heidelberg, Germany

I.	Introduction .....	144
II.	Setting Up DPD Simulations .....	145
	A. Why Use DPD for Simulating Biomembranes? .....	145
	B. Basic Ideas of DPD .....	147
	C. Forces in DPD .....	148
	D. Connecting Particles to Larger Structures .....	149
	E. Integrating the Equations of Motion .....	153
	F. Barostat .....	155
	G. Initial and Boundary Conditions .....	156
	H. Choosing Parameters .....	157
	I. Testing and Calibrating the Simulation .....	157
III.	Investigating Structure and Dynamics of Membranes with DPD .....	162
	A. Physical Properties of Model Lipid Bilayers .....	163
	B. Multicomponent Membranes .....	165
	C. Structure of Lipid Aggregates .....	166
	D. Phase Diagrams of Lipid Bilayers .....	167
	E. Membrane Fission and Fusion .....	169
	F. Dynamics of Membrane Proteins .....	172
	G. Altering Biomembrane Properties by Exogenous Factors .....	178
	H. Conclusion .....	179
	References .....	179

## ABSTRACT

In this chapter, we review recent approaches and results when studying membrane and protein dynamics by means of dissipative particle dynamics (DPD). First, we introduce and discuss DPD as a method, for example, the choice of the thermostat, which is of interest when constructing a DPD code. Then, we review important results on pure membranes and lipid-water systems that have been obtained with DPD. Finally, we focus on simulations of membranes with associated or embedded model proteins that may trigger future research on the fundamental interactions of lipids and proteins in the context of living cells.



## I. INTRODUCTION

Biological systems, for example, cells and tissues, are highly complex in structure and function. Investigating their dynamic properties therefore is a major challenge for biologists, chemists, and physicists. Modern experimental methods allow one to observe complex cellular structures and processes, sometimes even with single-molecule precision. Moreover, chemical reactions and interactions as well as mechanical and thermodynamic quantities can be quantified thoroughly even *in vivo*. We can nowadays access quantities and properties from the visible macroworld down to the nanolevel of individual molecules and atoms. However, every experimental technique is naturally limited in its applicability and its capability to report on information of the system of interest. Light microscopy methods, for instance, are very powerful in picturing structural and dynamic features of biological matter on length scales above the optical resolution limit. However, they can typically not discriminate objects beyond the diffraction limit, with the notable exception of specialized single-molecule techniques that trade in a higher spatial resolution for a loss in temporal resolution. Another example is methods of structural biology, for example, NMR, X-ray crystallography, or cryo-electron microscopy. All of these allow one to determine atomic structures and features of complex molecules but require a large ensemble of identical particles and a considerable acquisition time for revealing the desired information.

When experiments fail to yield the desired information about a system, computer simulations lend themselves as a powerful alternative. Simulations allow one to access parts of a system in detail, and the emerging quantitative results yield a link to the data from experimental approaches. Simulation models for science and technology have been developed already in the early days of the computer and became—with the increasing performance of computers—a standard tool in physics, chemistry, and applied sciences and engineering. Research on biological matter and biological processes can rely today on a wide choice of well-established simulation methods, from detailed all-atom descriptions to simplified procedures that model events on supramolecular length scales. With the latter, it is possible to access a variety of different problems that cannot be addressed with all-atom approaches, for example, fluid dynamics, complex chemical reactions, or tissue structure. Depending on whether one is concerned with the structure and internal motion of single molecules, or the dynamics and interactions of

a larger number of molecules, or the behavior of larger volumina where single-molecule effects can be neglected, one can choose between micro-scale models (covering nanometers and nanoseconds), coarse-grained meso-scale models (micrometers and microseconds), or continuum models that are based on (partial) differential equations.

A general bottleneck for all simulations is the limited processor speed of computers. Bound by current technology, only a fairly limited number of calculations can be performed within a reasonable period. Therefore, the number of individual particles and their interactions is restricted. To be more specific, simulations in atomic detail are typically bound to some nanometers and some 10 ns. Aiming at structures and processes on larger length and time scales, coarse-grained or mesoscopic simulations have to be employed in which a certain number of atoms is combined to single beads that interact via effective forces. The loss of structural details is hence compensated by the gain in system size and simulation time.

In this chapter, we review a powerful mesoscopic simulation tool, dissipative particle dynamics (DPD), as well as the results obtained with this approach. Indeed, DPD had originally been designed to model simple and complex fluids and was later extended to include also the description of polymers and lipid membranes. Going beyond technical details of the simulations, we will focus, in particular, on results obtained for (bio)membranes. We will first give a comprehensive and detailed description of DPD as a method. Subsequently, we will explain how fluids or soft matter systems are modeled by DPD and how physical theory and experiment can be related. We will put some emphasis on how DPD can be implemented into a program code, that is, which integration scheme for the equations of motion is best, and how the model can be calibrated. Then, we will report on the structural and dynamic features of lipid membranes that have been investigated by DPD so far. In this context, our focus is on showing the wide range of questions that can be treated with DPD and how one practically proceeds when employing DPD for a specific problem.

## II. SETTING UP DPD SIMULATIONS

### A. *Why Use DPD for Simulating Biomembranes?*

When studying (bio)membranes by simulation approaches, collective phenomena like membrane-mediated protein–protein interactions or protein-induced shape changes are the most interesting subjects, as they are

hard to tackle experimentally. Time scales involved in these processes are given by the requirement that lipids and proteins can freely diffuse within microseconds. Additionally, we have to consider that also the solvent is responsible for some characteristic properties of our system, for example, the fluctuation dynamics of the membrane (Lin and Brown, 2006).

To gain a microscopic understanding of such processes, the straightforward way would be to include all atoms and calculate the interactions between them. Such an approach is realized in all-atom molecular dynamics (MD) simulations. This widely used simulation technique is a powerful tool, for example, when studying conformational changes of proteins in the context of drug design. Usually, MD takes into account a variety of geometrical (bonds, angles) and physical (electrostatics, Van der Waals) interactions. All atoms are modeled as rigid bodies with a finite size. The use of a hardcore potential for excluded-volume interactions nonetheless demands short integration steps to prevent particle overlaps. Also, the long-range interactions are computationally very expensive given the smallest scales (about 1 Å) and the number of particles in the system. As a result, MD simulations typically are limited to scales of tens of nanometers and tens of nanoseconds. Still, full atomistic simulations of membrane systems or membrane proteins have been used to study, for example, water passage through a membrane pore (de Groot and Grubmuller, 2001) or the formation of a small vesicle (de Vries et al., 2004).

Much larger length and time scales can be addressed by continuum methods with the membrane being represented by an elastic sheet. Here, the physics of the membrane is modeled by sets of partial differential equations that are fueled by hydrodynamic equations and the famous Helfrich Hamiltonian (Helfrich, 1973) for elastic deformations of membranes. However, for continuum methods, it is difficult to handle individual molecules or complex spatial structures. Also, topological changes (membrane rupture and fission) are demanding topics.

The systems on which we focus in the remainder are best characterized on scales between the molecular dimension of lipids and the micron size of cells. They are hence best suited for a mesoscopic simulation technique. Here, the number of degrees of freedom of the system is reduced with respect to MD by coarse graining, that is, grouping several atoms into effective beads. Only those properties that are expected to influence the collective phenomenon of interest, for example, the amphiphilic nature of

lipids, are taken into account. An example of such a method is coarse-grained MD (see, e.g., [Ollila et al., 2009](#)) where several molecules are treated as a single particle with a given hydrophobicity so that a lipid molecule consisting of more than 100 atoms can be represented with only 10 particles of two types: hydrophilic heads and hydrophobic tails. Still, the particles interact via a hardcore potential in this approach which requires small integration steps for the equations of motion. This constraint is softened by DPD.

### B. Basic Ideas of DPD

DPD is a well-established explicit solvent mesoscale simulation technique that is related to coarse-grained MD methods ([Shillcock, 2008](#)). In DPD, several atoms are combined to larger beads that represent a small bulk of material as illustrated in [Fig. 1](#). To account for the friction due to internal (hidden) degrees of freedom, a dissipative force is introduced. The positions of single atoms in a bead are smeared out, and thus, a softcore potential allowing for an overlap of beads can be used here. In this way, one loses atomic details, yet one can achieve a substantially larger temporal range as compared to the above-mentioned methods.

DPD was introduced by [Hoogerbrugge and Koelman, 1992](#) for simulations of hydrodynamic phenomena. The method was further developed to the currently used formalism by [Español and Warren \(1995\)](#) who introduced the fluctuation-dissipation theorem into the original algorithm to couple frictional and stochastic forces. In this way, the statistical mechanics of the beads is

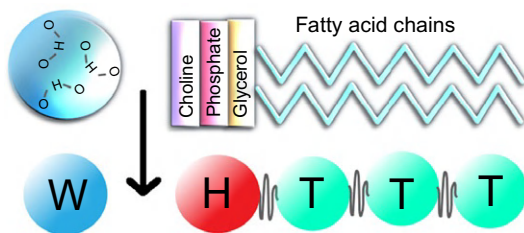


FIG. 1. Scheme of DPD coarse graining. One water bead (W) represents about three water molecules, DPD lipids are formed from a hydrophilic head bead (H), and a chain of hydrophobic tails (T) connected by Hookean springs.

consistent with the Gibbs canonical ensemble, and it yields the correct thermodynamics at sufficiently long time and length scales.

The motion of the  $i^{\text{th}}$  DPD bead is governed by Newton's equations of motion

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i, \quad (1)$$

$$m \frac{d\mathbf{v}_i}{dt} = \mathbf{F}_i^{\text{T}}, \quad (2)$$

where  $\mathbf{r}_i$  is the position of the bead's center of mass,  $\mathbf{v}_i$  is the bead velocity, and  $\mathbf{F}_i^{\text{T}}$  is the total force acting on the bead. The total force  $\mathbf{F}_i^{\text{T}}$  exerted on a free bead is given by three contributions: the dissipative force  $\mathbf{F}_i^{\text{D}}$ , the random force  $\mathbf{F}_i^{\text{R}}$ , and a conservative linear repulsive force  $\mathbf{F}_i^{\text{C}}$  that mimics excluded-volume interactions,

$$\mathbf{F}_i^{\text{T}} = \sum_{j \neq i}^N \left( \mathbf{F}_{ij}^{\text{C}} + \mathbf{F}_{ij}^{\text{D}} + \mathbf{F}_{ij}^{\text{R}} \right). \quad (3)$$

Here,  $N$  denotes the number of all beads in the system. The potentials in DPD are assumed to be short ranged, that is, all forces are nonzero only if the distance between two beads  $i, j$  is  $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j| < r_0$ . The cutoff radius  $r_0$  therefore defines the effective bead size and sets the smallest internal length scale. Typically,  $r_0 = 1$  is used for convenience; translation to SI units is possible (see below).

### C. Forces in DPD

#### 1. Character of Beads

The repulsive conservative force that mimics excluded volume interactions (Fig. 2) is usually given by

$$\mathbf{F}_{ij}^{\text{C}} = a_{ij} \left( 1 - \frac{r_{ij}}{r_0} \right) \mathbf{e}_{ij}. \quad (4)$$

Here  $\mathbf{e}_{ij} = \mathbf{r}_{ij}/r_{ij}$  is the unit vector pointing from particle  $j$  to  $i$ . The repulsion parameter  $a_{ij}$  depends on the combination of the two interacting particles and hence defines the interaction of different bead types. In the

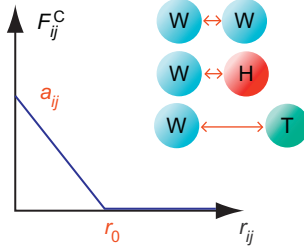


FIG. 2. Two beads  $i, j$  repel each other by a conservative force  $F_C$  that is linear and vanishes at the cutoff distance  $r_0$ . The strength of the repulsion is determined by the two bead types via the repulsion parameter  $a_{ij}$ .

most basic setup, the system consists of three DPD bead types only: hydrophobic tail beads (T), hydrophilic head beads (H), and water beads (W). The values of the repulsion constants depend on the system but frequently are assumed to obey  $a_{WW} = a_{HH} = a_{TT} \approx a_{HW}$  and  $a_{WT} = a_{HT} \approx 2 \dots 5 a_{WW}$  to implement amphiphatic lipids in a water environment. A concrete example of the interaction matrix used by [Shillcock and Lipowsky \(2002a\)](#) is:

$$a_{ij} = \frac{k_B T}{r_0} \begin{pmatrix} & \text{W} & \text{H} & \text{T} \\ \text{W} & 25 & 35 & 75 \\ \text{H} & 35 & 25 & 50 \\ \text{T} & 75 & 50 & 25 \end{pmatrix}. \quad (5)$$

#### D. Connecting Particles to Larger Structures

Larger entities like lipids and proteins are usually constructed by connecting individual beads  $i, j$  via a harmonic potential with equilibrium distance  $l_0$ :

$$U_{\text{harm}}(r_{ij,i+1}) = \frac{k_{\text{harm}}}{2} (r_{ij,i+1} - l_0)^2. \quad (6)$$

In bead-spring polymer models, also finite extensible nonlinear elastic (FENE) bonds with a potential  $U_{\text{FENE}} \sim r_0^2 \ln[1 - (r/r_0)^2]$  are used. The force derived from the FENE potential is approximately linear at small and intermediate distances, that is, it is equivalent to Eq. (6) for short

distances. It grows, however, dramatically when the interparticle distance gets large, hence preventing a too large separation. In many simulations, no drag force is applied to lipids and proteins and the connecting springs oscillate around an equilibrium length only due to thermal motion. The harmonic potential is here sufficient to maintain the integrity of the structures.

The rigidity of a hydrocarbon chain model is considered by a three-point bending potential assigned to three consecutive beads

$$U_{\text{bend}}(r_{i,i+1}, r_{i+1,i+2}) = k_{\text{bend}}(1 - \cos(\theta - \theta_0)), \quad (7)$$

where  $\cos(\theta) = \mathbf{e}_{i,i+1} \cdot \mathbf{e}_{i+1,i+2}$ . For a straight linear arrangement,  $\theta_0 = 0$ . In some simulations,  $\theta_0 \neq 0$  may be required, and here the force acting on bead  $i$  has the form

$$\mathbf{F} = \frac{k_{\text{bend}} \sin(\theta - \theta_0) [\mathbf{e}_{jk} - \mathbf{e}_{ij}(\mathbf{e}_{ij} \cdot \mathbf{e}_{jk})]}{r_{ij} |\sin \theta|}. \quad (8)$$

### 1. DPD Thermostat

The dissipative force  $\mathbf{F}^{\text{D}}$  represents the viscous drag on a bead due to the atomistics friction with neighboring molecules. The random force  $\mathbf{F}^{\text{R}}$ , however, encodes thermal kicks that a bead receives from its neighbors. These two forces commonly assume a form:

$$\mathbf{F}_{ij}^{\text{D}} = -\gamma \omega^{\text{D}}(\mathbf{r}_{ij}) (\mathbf{e}_{ij} \cdot \mathbf{v}_{ij}) \mathbf{e}_{ij}, \quad (9)$$

$$\mathbf{F}_{ij}^{\text{R}} = -\sigma \omega^{\text{R}}(\mathbf{r}_{ij}) \mathbf{e}_{ij} \xi_{ij}. \quad (10)$$

The weight function  $\omega^{\text{D}}(r) = (1 - r/r_0)^2$  is nonzero only for  $r \in [0, r_0]$ , and the Gaussian white-noise term  $\xi_{ij}$  is a random variable with zero mean and unit variance, uncorrelated for different pairs of particles at different times.

The random and dissipative forces act as a heat source and sink, respectively, and hence they are often referred to as the DPD thermostat. These two forces are further coupled via the fluctuation–dissipation theorem (Español and Warren, 1995), that is, the parameters  $\sigma$  and  $\gamma$  (which represent the friction coefficient and the noise amplitude) satisfy the relation

$$\sigma = \sqrt{2k_{\text{B}}T\gamma}, \quad (11)$$

which provides a correct NVT ensemble (fixed particle number  $N$ , volume  $V$ , and temperature  $T$ ). In addition, the weight functions are related as  $\omega^{\text{D}}(r) = [\omega^{\text{R}}(r)]^2$ . Commonly chosen parameters are  $\sigma=3$ , giving  $\gamma=9/2$  for imposing a thermal energy of  $k_{\text{B}}T=1$  (Nikunen et al., 2003).

As mentioned already above, different bead types are distinguished only via the conservative force  $\mathbf{F}_{\text{C}}$ , that is, via the repulsion parameter  $a_{ij}$ . Owing to Newton's third law ( $\mathbf{F}_{ij} = -\mathbf{F}_{ji}$ ), the sum of all forces in the system (including the thermostat forces) vanishes and (angular) momentum is conserved (Groot, 2004a). Moreover, the total force between all particles in a subset of the system vanishes, too. The total acceleration of any such volume of liquid is then given only by the sum of forces that cross its boundary, which is the starting condition for the derivation of the Navier–Stokes equation. This intrinsic implementation of hydrodynamics is a major advantage of DPD as compared to other simulation techniques. In Brownian motion approaches, for example, the random force is not pairwise, but it is related to a fixed heat bath so that momentum is no longer conserved.

## 2. Alternative Methods of Temperature Control

Besides the DPD thermostat, several other methods can be used to control the system's temperature. The Nosé–Hoover thermostat (Nosé, 1984; Hoover, 1985) is widely used in MD. Here, a heat bath is introduced as an integral part of the Hamiltonian by adding an extra term—an artificial variable with an artificial mass. The disadvantage of the Nosé–Hoover thermostat is that it does not satisfy Galilean invariance, that is, a preferential inertial reference frame is singled out by the implementation of the thermostat. Hence, the motion of the center of mass of the system has to be explicitly corrected for, otherwise temperature increases. This is, in particular, problematic for nonequilibrium simulations. Further, the Nosé–Hoover thermostat is a global algorithm with a uniform, unrealistic dissipation of energy in the system. Hence, it does not allow for a local temperature control. Another approach often used in MD simulations is the Berendsen thermostat (Berendsen et al., 1984) that efficiently imposes a desired temperature. However, it suppresses fluctuations of the kinetic energy and hence does not reproduce the canonical ensemble, especially



for small systems. Usually, it is used in combination with the Nosé–Hoover thermostat for an initial equilibration.

The DPD thermostat introduced above was worked out as a modification of the Langevin thermostat (Grest and Kremer, 1986). In both methods, a random force and a constant friction are applied to all particles. The two forces are related via a fluctuation–dissipation theorem. Similar to the DPD thermostat, the Langevin thermostat is a stochastic, local method where the energy dissipation in the system is spatially localized. In contrast to the DPD implementation, each particle has its own heat bath that is independent of all other particles. Another stochastic realization is Andersen’s scheme (Andersen, 1980) which consists in a velocity rescaling. The velocity of a particle is periodically exchanged with that of a bath particle. This procedure mimics collisions with bath particles at a specified temperature  $T$ . The strength of the coupling to the heat bath is given by a collision frequency  $\Gamma$ . The drawback of the Andersen and Langevin thermostats is the absence of momentum conservation and thus a lack of hydrodynamics. In both stochastic thermostats, propagation of momentum is disturbed due to uncorrelated random forces, that is, a reliable reproduction of viscosity is problematic. In case that transport properties play a role in the system of interest, the DPD thermostat is hence the method of choice.

A variation of the DPD technique based on the Andersen thermostat was introduced by Lowe (Lowe, 1999; Koopman and Lowe, 2006). Similar to Andersen’s thermostat, an exchange frequency  $\Gamma$  is used to assign new particle velocities from a Maxwell distribution. This stochastic velocity is, nevertheless, imposed on pairs of neighboring particles in such a way that the overall momentum is conserved. This method shares many above mentioned advantages of the DPD thermostat (locality, Galilean invariance, conservation of momenta) and satisfies the detailed balance condition<sup>1</sup> as in Andersen’s method. The viscosity of the fluid in this method is proportional to the exchange frequency  $\Gamma$  which, in turn, determines the thermostat efficiency. Therefore, when aiming at a good thermostat sampling, a low viscosity regime cannot be accessed.

<sup>1</sup>The principle of detailed balance describes the relation of transition probabilities  $P$  between two states  $A$  and  $B$  for a system in equilibrium:  $N_A P(A \rightarrow B) = N_B P(B \rightarrow A)$ , where  $N_{A,B}$  denotes the number of particles in each state. The transition processes must be reversible, which in the DPD thermostat is violated by the dissipative force.

In this context, an important characteristics of the system is the dimensionless Schmidt number,  $Sc$ . It is defined as the ratio of kinematic viscosity  $\mathcal{V}$  and the diffusion coefficient  $D$ . In a fluid, momentum transport is rapid via interparticle forces while mass transport happens on a different time scale. In other words, the displacement of particles is slow in comparison to momentum transport. As a result, a rather high Schmidt number is obtained for fluids,  $Sc \sim 10^3$ . In DPD, the soft-core potential does not allow for such an efficient momentum transport. In a DPD fluid, the intrinsic viscosity is of the same order as the diffusion coefficient and the Schmidt number hence has a low, gas-like value,  $Sc \sim 1$ . Thus, when the viscous time scale is matched to experimental data, the diffusion in the DPD fluid is overestimated. Using the Lowe–Andersen method, one can achieve a much higher viscosity. This can be crucial when the correct representation of viscous flow is essential. However, in some cases, a simulation technique with fast diffusion is useful. For molecular processes that are diffusion controlled, the DPD thermostat allows to observe the phenomena of interest within a shorter simulation time (Groot, 2004b).

### *E. Integrating the Equations of Motion*

Knowing the initial conditions and the forces acting on each bead, we are interested in how the system will evolve in time. For this purpose, various numerical methods for integrating the equations of motion exist. The most intuitive one, the Euler method, is based on an approximation of first derivatives. While it is a straightforward way to calculate the trajectories, it is inaccurate and numerically unstable especially for larger time-steps  $\Delta t$ . In contrast, Velocity-Verlet (VV) approaches fall into the class of second order methods of numerical integration (similar to the basic Verlet or the Leapfrog algorithms). Even more precise higher order algorithms, for example, Runge–Kutta, can be used, but their accuracy is counterbalanced by the increased computational demand.

The standard algorithm for numerically integrating the equations of motion for DPD is a modified Velocity-Verlet algorithm (DPD-VV) (Nikunen et al., 2003). The VV algorithm assumes the acceleration of a particle to depend on its positions only and not on velocity. This is not fulfilled for nonconservative systems. In DPD, the dissipative force is velocity dependent and the velocities, in turn, are governed by the

dissipative forces. In DPD-VV, it is accounted for that in an approximate manner by updating the dissipative forces additionally at the end of every iteration step.

In case of the NVT ensemble, where a fixed simulation box is used, the integration can be performed according to [Nikunen et al. \(2003\)](#). The calculation of the dynamics of the system is performed for  $N_{\text{steps}}$  iterations where the positions and velocities of all particles are updated. In every step, a new velocity is calculated for each bead according to the total force acting on it. Subsequently, the new positions are calculated from the velocities, and the forces are reevaluated in accordance with the new conformation. At the end, the velocities are updated and the dissipative force is changed accordingly. The DPD-VV integration scheme hence reads:

1. Calculate velocities  $v_i \leftarrow v_i + \frac{1}{2m} (\mathbf{F}_i^C \Delta t + \mathbf{F}_i^D \Delta t + \mathbf{F}_i^R \sqrt{\Delta t})$
2. Update positions  $x_i \leftarrow x_i + v_i \Delta t$
3. Calculate all forces  $\mathbf{F}_i^T = \sum_{i \neq j}^N (\mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R)$
4. Calculate velocities
  - (a)  $v_i^0 \leftarrow v_i + \frac{1}{2m} (\mathbf{F}_i^C \Delta t + \mathbf{F}_i^R \sqrt{\Delta t})$
  - (b)  $v_i \leftarrow v_i^0 + \frac{1}{2m} (\mathbf{F}_i^D \Delta t)$
5. update the dissipative force  $\mathbf{F}_i^D$
6. calculate physical quantities of interest.

Please note here that the contribution of the random force scales as  $\sqrt{\Delta t}$  due to the imposed Wiener process ([Español and Warren, 1995](#); [Groot, 2004a](#)). In the self-consistent version of the integrator, the loop over steps (4b) and (5) is repeated until the instantaneous temperature has reached its limiting value. A more efficient scheme can be used that only includes these steps once, as it was shown to give a sufficiently good performance ([Nikunen et al., 2003](#)). The random variable  $\xi_{ij}$  is supposed to exhibit a Gaussian distribution, the production of which is computationally fairly expensive. Using the central limit theorem, a more efficient way is possible via the use of appropriately normalized random numbers drawn from a uniform (box) distribution. No statistical difference was found between simulations using the two types of random variables ([Español and Warren, 1995](#)).

Including a barostat into the equations of motion (cf. below) increases the computational costs significantly. Therefore, a barostat is typically used only in the initial equilibration phase of the simulation. After the system has achieved equilibrium, the basic DPD algorithm with fixed box dimensions is used instead.

### F. Barostat

Natural (bio)membranes are considered to have zero surface tension, and there are several strategies to achieve this in simulations. A commonly used technique is combining DPD with a Monte Carlo algorithm to update the box size at random time points (de Meyer et al., 2008a). Another method is to find the lipid density for a tensionless membrane within a box of fixed dimensions by trial and error, for example, by successively adding more lipids into the system.

Here, we would like to present a real-time relaxation method—the barostat algorithm introduced by Jakobsen (2005). This method is an analogy of the Langevin piston barostat used in MD simulations (Feller et al., 1995). Here, the simulation box is allowed to shrink and expand due to a virtual piston, so that in every step, all particle positions are rescaled according to the “breathing” of the simulation box. The size of the simulation box changes upon the action of a piston force  $F_\beta$  described by the Langevin equation. The force  $F_\beta$  acts along the edges of the simulation box, and its size depends on several contributions: the difference between the actual and the target pressure ( $P - P_0$ ) in the corresponding direction, DPD bead momenta  $p_i$ , a dissipative force proportional to the piston force  $v_\beta$ , and a random force dependent on a random variable  $\xi_\beta$

$$F_\beta = \Delta V(P - P_0) + \frac{d}{N_f} \sum_i \frac{p_i^2}{m} - \frac{\gamma_\beta v_\beta}{M_\beta} + \frac{\sigma_\beta \xi_\beta}{2M_\beta \sqrt{\Delta t}}. \quad (12)$$

Except for the degrees of freedom  $N_f$  of  $N$  beads in  $d=3$  dimensions ( $N_f = dN - d$ ), there are three additional degrees of freedom that represent the three edge lengths of an orthorhombic box. The mass of the piston  $M_\beta$  is given by  $M_\beta = (N_f + d) k_B T \tau^2$ , where  $\tau$  is the characteristic barostat time which is set to  $\tau = 2$  (cf. discussion in Jakobsen, 2005). As there are in general different fluid phases in the system, the pressure  $P$  is a tensor

$$P^{xy} = \frac{1}{V} \left( \sum_i \frac{u_i^x u_i^y}{m} + \sum_i F_i^{Cx} r_i^y \right), \quad (13)$$

where the indices  $x, y$  denote any pair of the three space coordinates. The target pressure  $P_0$  has the value of  $P_0 = 23.649 k_B T / r_0^3$  (Jakobsen, 2005). The dissipation of the piston,  $\sigma_\beta$ , and the coefficient for the random force,  $\gamma_\beta$ , are again related via the dissipation–fluctuation theorem

$$\sigma_\beta^2 = 2\gamma_\beta M_\beta k_B T \quad (14)$$

Here,  $\gamma_\beta = 10/\tau = 5$ .

When using the barostat, the volume  $V$  of the simulation box is slightly fluctuating around the initial value. The Langevin framework, however, does not lead to unphysical oscillations of the simulation box as observed in case of some other barostat implementations. The barostat algorithm introduced in Jakobsen (2005) also requires a shorter equilibration time and is characterized by shorter correlation times of various system parameters as compared to other methods. For practical use, it is important that the coupling of the pressure to the system does not enforce the use of smaller time-steps  $\Delta t$ . For details on how to include the barostat into the integration of the equations of motion, we refer the reader to Jakobsen (2005).

### G. Initial and Boundary Conditions

DPD simulations are typically performed in an orthorhombic box with periodic boundaries, that is, a particle that escapes the box at one boundary reappears via the opposing boundary of the box with the same velocity. The bilayer is spread in a horizontal plane parallel to the base of the box, and it virtually continues in all the images of the simulation box forming an infinite yet periodic membrane (cf. the analogy to crystals in solid state physics). The box size has to be chosen sufficiently large in order to suppress finite size effects, especially interaction of the bilayer with its periodic images in vertical direction. For some applications, one also has to consider the influence of the box on bilayer fluctuations, as periodic boundaries impose a planar noncurved arrangement on the membrane.

If a proper concentration of lipids and water beads is randomly distributed in the box, a membrane will spontaneously self-assemble after a sufficient time. For speeding up the equilibration part of the simulation,

we usually used a predefined membrane setting where concrete starting positions in the midplane of the box are assigned to lipid beads. In particular, lipids are arranged initially as a regular crystal which prevents flipping of lipids from one leaflet to the other within the first simulation steps. The values for all initial velocities are taken from a Maxwell–Boltzmann distribution with the desired temperature as an input parameter.

### H. Choosing Parameters

The above described potentials and forces are sufficient for setting up a model membrane in an explicit water solvent. Due to the amphiphilic nature of the lipids, a membrane forms spontaneously (Venturoli and Smit, 1999). Membrane properties in the DPD model are comparable to experimental results, for example, concerning the lateral stress profile, area compression modulus, or the bending rigidity (cf. below).

A typical set of parameters used in simulations is spring stiffness  $k_{\text{harm}} = 128k_{\text{B}}T/r_0$ , equilibrium bond length  $l_0 = 0.5r_0$ , and bending rigidity  $k_{\text{bend}} = 20k_{\text{B}}T/r_0$  in conjunction with the interaction matrix Eq. (5) (Shillcock and Lipowsky, 2002a). Lipids are typically modeled either as a single chain consisting of one hydrophilic head bead and several hydrophobic tail beads ( $\text{HT}_n$ ) or as a double-chain ( $\text{H}_m(\text{T}_n)_2$ ) (cf. also Fig. 7). The mass of all beads as well as the temperature is usually set to unity, while the bead density of the whole system is  $\rho = 3/r_0^3$ . The initial lipid area density in the membrane is  $\rho \approx 2.8/r_0^2$ , and the distribution of particle velocities is well captured by a Maxwell–Boltzmann distribution. For integrating the equations of motion, it is safe to use time-steps  $< 0.05$  in order to avoid deviations from the imposed temperature beyond 2% (Groot and Warren, 1997). When dealing with membranes, this estimate has to be even more careful (Jakobsen et al., 2005a), a reasonable choice is  $\Delta t = 0.01$ .

### I. Testing and Calibrating the Simulation

To confirm the functionality of the program code and to calibrate the simulation, physical quantities can be measured and compared to data in the literature. Here, we discuss how to concretely determine the relevant observables and which values indicate a properly operating simulation code.

### 1. Temperature

During the whole simulation, the temperature has to be constant on average with fluctuations around the imposed temperature. The temperature of a system with  $N$  free particles of mass  $m$  is given via the kinetic energy and the equipartition theorem, that is,

$$\langle k_B T \rangle = \frac{m}{3N} \sum v_i^2. \quad (15)$$

Due to the DPD thermostat, the temperature fluctuates only slightly around the imposed temperature (defined here via  $k_B T = 1$ ; see Fig. 3). After a very short period of equilibration, the temperature becomes constant with small fluctuations. The peak in the first steps of the simulation is caused by the random initial positioning of water beads, which can occasionally overlap and thus experience an extremely strong repulsion. This effect relaxes quickly, and owing to the choice of a sufficiently small time-step, such an overlap does not occur any more after the initial equilibration.

### 2. Bead Velocity

The average velocity of all beads, that is, the velocity of the system's center of mass,

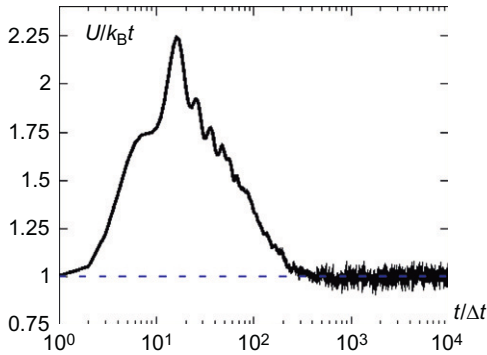


FIG. 3. Temperature fluctuates around unity as requested by the parameter settings. Note the logarithmic  $x$ -axis that highlights the initial relaxation in more detail.

$$\langle \mathbf{v} \rangle = \frac{1}{N} \sum_{i=1}^N \mathbf{v}_i \quad (16)$$

is zero throughout the simulations. The distribution of the absolute velocities  $v = \sqrt{v_x^2 + v_y^2 + v_z^2}$  assumes the form of a Maxwell–Boltzmann distribution (Fig. 4)

$$p(v) = \sqrt{\frac{2}{\pi}} \left( \frac{m}{k_B T} \right)^{3/2} v^2 \exp \left( \frac{-mv^2}{2k_B T} \right). \quad (17)$$

### 3. Density Profiles

In a system of water beads and lipids (which form a membrane), the density of different bead types is not uniform. One characteristic of the model is obtained by measuring the average density of beads in  $0.25r_0$  thick slices of the simulation box parallel to the membrane plane. As can be seen in Fig. 5, below and above the membrane, there is a water layer of a homogeneous density  $\rho = 3n_0^3$ , whereas the water density is zero in the hydrophobic core of the membrane due to strong hydrophobic interactions between water and hydrophobic tail beads. Lipid heads show up as peaks at the solvent–membrane interface and lipid tail beads fill the space between the two lipid head peaks.

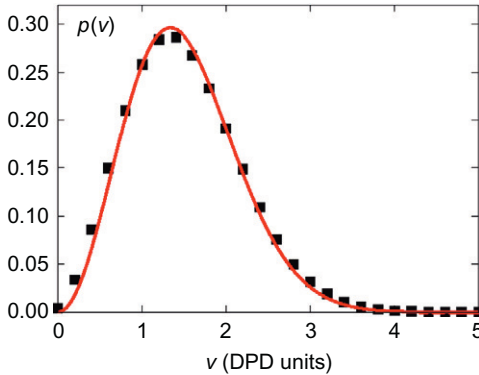


FIG. 4. The Maxwell–Boltzmann distribution (red line) characterizes the distribution of absolute velocities of beads (filled squares).



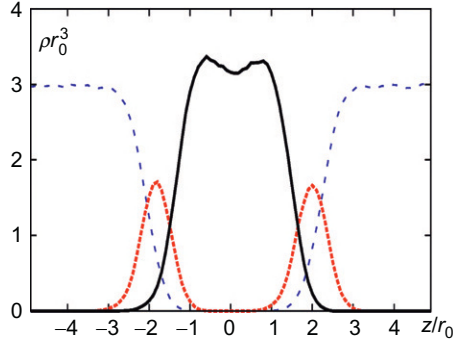


FIG. 5. Density profile of beads in the simulation box along the bilayer normal ( $z$ -axis). Hydrophobic tail beads are depicted with a full black line, hydrophobic heads with a red dotted line, and water with a blue dashed line.

Lipid tails are slightly compressed, and the membrane core is filled up with an almost homogeneous density. Thus, the lipid chains are disordered in a way that is characteristic for the fluid phase of a lipid bilayer. There is a visible dip in the bead density in the midplane of the bilayer indicating that all lipid tails terminate near to the bilayer midplane. Yet, this dip is not very deep as the lipids of opposite layers are slightly interdigitated. From Fig. 5, we can also infer the membrane thickness  $h$  by measuring the distance between the centers of mass of the head beads in opposing leaflets. The value  $h \sim 3.8r_0$  yields a means to link the intrinsic length scale  $r_0$  to the membrane thickness observed in experiments ( $h = 30\text{--}40 \text{ \AA}$ ).

#### 4. Barostat

The action of the barostat can be probed by calculating the dimensionless compressibility of water

$$\kappa^{-1} = \frac{V}{\langle dV^2 \rangle \rho}. \quad (18)$$

To this end, a DPD system consisting of only free water beads (repulsion parameter  $a_{ij} = 25k_B T$  and density  $\rho = 3/r_0^3$ ) is simulated and the fluctuations of the box volume  $dV$  are recorded. The parameter choice should

reproduce correctly the compressibility of water (Groot and Warren, 1997), and indeed, using a simulation box of  $(15r_0)^3$  for  $10^6$  time-steps, the measurement yielded  $\kappa^{-1} = 15.95$ , in a good agreement with Groot and Warren (1997) where  $\kappa^{-1} = 15.98$  was found.

Fluctuations of the box edges caused by the barostat piston can be seen in Fig. 6. Here, the barostat was used for the whole simulation ( $10^6$ ). The membrane is placed in the  $xy$  plane. The initial planar density of lipids was too low and so the box  $xy$ -dimension had to decrease. This is compensated by an expansion in  $z$  direction so that the volume and bead density are conserved. We can see that approximately after  $2 \times 10^5$  time-steps, the system fluctuates only slightly around the steady-state value indicating that this is a sufficient period for equilibration.

The effect of the piston on membrane surface tension is also shown in Fig. 6. Surface tension can be computed from the pressure tensor as (Marrink and Mark, 2001)

$$\sigma = L_z(P_{\text{norm}} - \langle P_{\text{lat}} \rangle), \quad (19)$$

where  $P_{\text{norm}}$  is the component of pressure in the direction normal to the bilayer plane (here:  $z$  direction) and  $\langle P_{\text{lat}} \rangle$  is the average of tangential components,  $P_{xx}$  and  $P_{yy}$ . After the equilibration, the mean surface tension is zero.

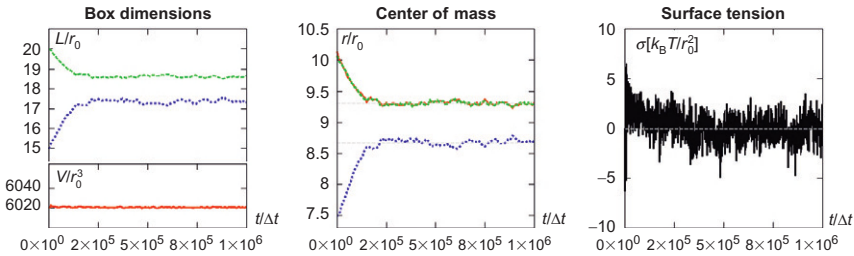


FIG. 6. Fluctuations of the box due to the barostat action. (A) While the coupled  $x$ - and  $y$ -edges (green) are shrinking, the  $z$ -edge (blue) is expanding to keep the volume  $V$  constant. (B) The center of mass stays in the middle of the simulation box ( $x$ ,  $y$ ,  $z$ —red, green, blue). (C) The surface tension of the membrane fluctuates around zero. Data for all graphs come from the same simulation run.

### 5. Conversion to SI Units

Simulation units are defined by the simulation time  $\Delta t=0.01$  and the cutoff radius for interaction between two beads,  $r_0=1$ . For a conversion to SI units, we will take into account that one water bead represents  $N_w=3$  water molecules (Groot and Rabone, 2001) and the volume occupied by a single water molecule is  $V_w=30 \text{ \AA}^3$  (Lu et al., 1993). The overall density of beads set up in typical simulations is  $\rho=3r_0^{-3}$ . In a simulation box containing only water, a unit volume  $r_0^3$  contains three water beads, that is, nine water molecules of  $30 \text{ \AA}^3$  each. In this way, we find

$$r_0 = \sqrt[3]{\rho N_w V_w} = 6.43 \text{ \AA}. \quad (20)$$

In case of a heterogeneous composition of the system, where beads of different types form structures, the density varies throughout the simulation box. The effective volume (how much space can be occupied by one bead) depends on the local density of the given bead type, and it is influenced by the interactions with the surrounding. Knowing the number of lipids in the bilayer of a given equilibrated area, we obtain a surface area of  $\sim 65 \text{ \AA}^2$  per lipid which corresponds to the value for lipids in a biological membrane (Lantzsch et al., 1994). One hydrophobic tail bead then corresponds to roughly 3.8 hydrocarbon groups.

To match the time units, we can similarly compare the diffusion coefficient of a single lipid with experimental values. Doing so, the system yields roughly  $\sim 90$  ps for a single time-step  $\Delta t=0.01$  (Schmidt et al., 2008).

## III. INVESTIGATING STRUCTURE AND DYNAMICS OF MEMBRANES WITH DPD

Using the outlined DPD method, a variety of biologically and biophysically important problems can be and have been addressed. In this section, we will first review some results on the physical chemistry of lipid bilayers. This will include the interdigitation of lipids, that is, the coupling of membrane leaflets, the phase behavior of bilayers with multiple lipid species, and topological changes due to fusion and fission events. The second part of this section is devoted to the interaction of proteins and membranes. In particular, we will discuss membrane-mediated interactions between proteins that are likely to play a pivotal role in protein sorting in eukaryotes.

### A. *Physical Properties of Model Lipid Bilayers*

The earliest simulation of a surfactant bilayer with DPD was performed in 1999 by [Venturoli and Smit \(1999\)](#). Here and in later papers by [Groot and Rabone \(2001\)](#) and [Shillcock and Lipowsky \(2002b\)](#), it was shown that within the DPD formalism, the self-assembly of amphiphilic surfactant molecules like lipids to a bilayer can be modeled. Shillcock showed that both single-chain  $HT_n$  and double-chain  $H_m(T_n)_2$  lipid models can be used without compromising the gross results ([Shillcock and Lipowsky, 2002b](#)). Both models yield a bilayer self-assembly with similar density profiles for water, lipid head, and tail beads. As expected, water is excluded from the bilayer core for both settings due to the strong repulsion of water beads and lipid tails. [Figure 7](#) displays sketches of the two basic lipid models and a simulation snapshot of a membrane after completed self-assembly.

For single-chain lipids, the bending stiffness of the chain turned out to be a crucial parameter for membrane stability while this was much less an issue for double-chain lipid models. Imposing a chain stiffness helped to orient single-chain lipids along the bilayer normal, reduced the event of an upside-down lipid, and also prevented an interdigitation of the leaflets. Further, it was found that when using double-chain lipids,  $m \geq 3$  was necessary to protect the tail from water and yield stable and well-ordered bilayers.

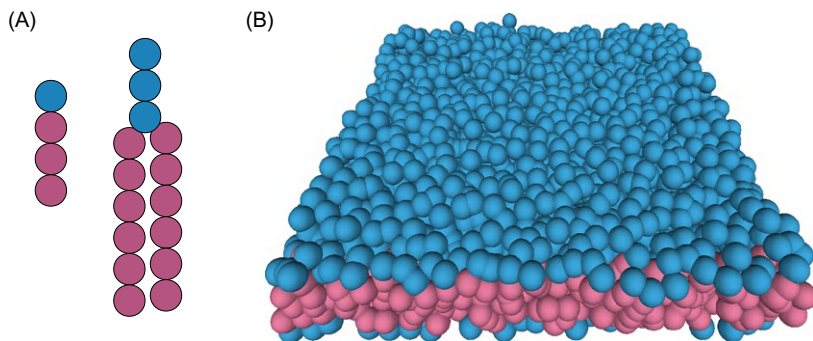


FIG. 7. (A) Single chain  $HT_n$  and double-chain  $H_m(T_n)_2$  lipid models have been considered in DPD. Shown examples are  $HT_3$  and  $H_3(T_6)_2$  with hydrophilic heads (H) in light blue and hydrophobic tails (T) in light red. (B) Self-assembled membrane consisting of  $HT_3$  lipids (water not shown for better visibility).

An important parameter characterizing a lipid bilayer is its lateral pressure profile,  $p(z)$ . This quantity has been suggested to influence structure and function of membrane proteins (de Kruiff, 1997) and modulate the action of anesthesia (Cantor, 1997). However, there is no experimental technique to quantify the lateral pressure profile while it can be calculated by mean field approaches or via membrane simulations (Goetz and Lipowsky, 1998; Venturoli and Smit, 1999; Shillcock and Lipowsky, 2002b). The pressure profile  $p(z)$  is defined as the difference of the normal and the lateral components of the pressure tensor summed over all potentials. When determining  $p(z)$  in simulations, one averages the contribution from all bead–bead interactions (i.e., repulsion, bond potential, and chain stiffness) over thin slices along the membrane normal. The general structure of the lateral pressure profile is characterized by maxima at the monolayer–water interfaces, which appear due to the repulsion of hydrophilic beads (water and lipid head) with the hydrophobic lipid tail beads. The adjacent minima are due to the Hookean bond potential of the lipids, while the inner maxima near the monolayer mid-plane are caused by the chain stiffness potential. The general features of the lateral pressure profile of a DPD membrane hence agree well with the results of more detailed MD simulations (Goetz and Lipowsky, 1998). In fact, the use of hardcore Lennard–Jones potentials in MD (in contrast to soft-core potentials in DPD) neither causes a major difference in  $p(z)$  nor changes the pressure profile for single and double-chain lipid models in a gross way (Shillcock and Lipowsky, 2002b).

By integrating the pressure profile across the bilayer, one can calculate the surface tension,  $\sigma$ . The surface tension increases with the projected area per lipid  $A$ , as this affects the relative contributions of the repulsion, bond potential, and chain stiffness in  $p(z)$ . For the preferred area per lipid,  $A_0$ , the surface tension vanishes. Enlarging the head group or reducing the lipid tail length shifts the zero surface tension to larger values of  $A$  for both single- and double-chain lipids. The dependence on head group size is, however, less pronounced for the double-chain model (Shillcock and Lipowsky, 2002b).

Surface tension, area per lipid, and the lipid length are related to the area stretch modulus  $K$  via  $\sigma = K(A - A_0)/A_0$ . Choosing the right DPD parameters, experimentally reported values can be obtained for  $K$  (Shillcock and Lipowsky, 2002b). Moreover, the bilayer bending rigidity  $\kappa$  can be estimated via  $K$ , and the bilayer thickness  $d$  via  $\kappa = Kd^2/48$  or by

monitoring the undulations of the membrane. Reasonable values of a few  $10k_{\text{B}}T$  can be found for DPD simulations, with the precise value being mainly influenced by the lipid chain stiffness. For increasing lipid length, also  $\kappa$  increases. It is worth noting, however, that changing DPD parameters will affect not only a single observable, for example,  $\kappa$ , but it will also affect other quantities. Determining parameter sets that help to mimic an experimental data set therefore can become quite tedious. In addition, lipid chain length and chain asymmetry were found to influence the physical structure of a lipid bilayer (Illya et al., 2005).

### B. Multicomponent Membranes

DPD offers the possibility to study membranes composed of more than one lipid species. A lipid species carrying long saturated chains together with another lipid species carrying one or more double bonds was considered, for example, in Illya et al. (2006). Both lipid types had a structure  $\text{H}_3(\text{T}_6)_2$ , yet with a weaker repulsion between the tail beads of the second lipid type in order to mimic the tighter packing of unsaturated fatty acid chains in the bilayer core. These two lipid species were seen to separate in the simulations, that is, they formed domains. A single-component bilayer made of the unsaturated lipids had a bending rigidity almost twice as large as a bilayer composed solely of the saturated species, due to the closer packing. In membranes consisting of both species, the bending rigidity monotonically increased with an increasing concentration of unsaturated lipids.

When combining two lipid species of similar preferred packing area but different lengths ( $\text{H}_3(\text{T}_6)_2$  and  $\text{H}_3(\text{T}_8)_2$ ), no domain formation was observed. The area stretching modulus and the bending rigidity changed nonmonotonically (Illya et al., 2006).

Aiming at a phase separation and the formation of lipid domains, also other models were proposed (Yamamoto and Hyodo, 2003; Laradji and Sunil Kumar, 2004, 2005, 2006). Here, two lipid species formed domains due to an increase in the repulsive force between the beads of the different species. This approach is a simple and effective way to induce domain formation and to set a line tension between the two emerging phases. Starting with an initial random configuration in a vesicle, the dynamics and time course of domain formation and coalescence were studied (Laradji and Sunil Kumar, 2004, 2005, 2006) (see also below for details).

### C. Structure of Lipid Aggregates

Lipids immersed in water do form different types of aggregates, depending on the lipid type. Micelles, vesicles with a bilayer membrane, or inverted structures of cubic or hexagonal phases can be obtained. In DPD simulations, it is also possible to induce the formation of different lipid aggregate structures. Due to the finite size of the simulation box, it depends on the number fraction of lipids and the preferred area of the used lipid species which structure may form. Typically, a lipid number fraction in the range 3–6% results in a flat bilayer (Shillcock and Lipowsky, 2002b). With lower numbers, the formation of micelles is observed, whereas a higher lipid number induces the formation of complex three-dimensional structures.

It was demonstrated that DPD simulations can also model the self-assembly of lipids to vesicles (Yamamoto et al., 2002). The lipid concentrations necessary for this to happen were 5–10 vol%. Vesicle formation was observed with different initial configurations, that is, with lipids being randomly dispersed in the box and with lipids being already preassembled to a bilayer. In both cases, the pathway of vesicle formation included an intermediate step in which lipids formed an oblate bilayer, a kind of stretched bilayer-like micelle. This object subsequently changed to a vesicle by collapsing, enclosing water, and sealing itself to a vesicle. Single chain lipids with a short tail had a higher potential for forming vesicles compared to lipids with a longer chain. The aggregation time of double-chain lipid was seen to be faster than that of single-chain lipids, due to the larger radius of gyration of the former and the related larger probability to meet another lipid.

Inverted hexagonal and cubic phase also have been studied with DPD (Da-Wei et al., 2004). Here, a bond angle potential with a strength of  $5k_{\text{B}}T$  instead of  $20k_{\text{B}}T$  (Shillcock and Lipowsky, 2002b) was used, bearing in mind that the dependence of the area stretch modulus  $K$  with decreasing area per molecule was then more consistent with experimental data. Notably, the lateral pressure profile did not change its gross features with this modification. Using a single-chain lipid  $\text{HT}_n$  with  $n=1, \dots, 12$  and various lipid head sizes ( $1 - 1.2r_0$ ) at various lipid concentrations and temperatures, different normal micellar phases (spherical, rodlike, and disklike) were found at low lipid concentrations while stacks of lipid bilayers and inverted hexagonal phases were observed (depending on  $n$ )

for higher concentrations. These results show that DPD is fully capable of reproducing the rich lipid phase behavior known from theory and experiments despite strong approximations underlying the model.

#### D. Phase Diagrams of Lipid Bilayers

Lipid bilayers were experimentally found to adopt different phases depending on the lipid composition and temperature. A typical example, phosphatidylcholine (PC) membranes are in the gel state at low temperatures and in a fluid state at higher temperatures. The difference between the phases is the ordering of the lipids' fatty acid chains. At low temperatures, most lipid bilayers are in the subgel phase  $L_c$  where the hydrocarbon tails exhibit a high order. Here, lipids are tilted with respect to the bilayer. At higher temperatures, membranes adopt a lamellar gel phase, depending on the structure of the lipids head group. This can be either the  $L_\beta$  phase in which lipids are oriented parallel to the membrane normal (e.g., phosphatidylethanolamine, PE) or the  $L_\beta'$  phase in which lipids are tilted with respect to the membrane normal (e.g., PC). In both phases, the order of the lipid tails is still high albeit lower than in the  $L_c$  phase. At higher temperatures, the liquid crystalline phase  $L_\alpha$  or other fluid phases are found in which lipids are disordered. Also, phase coexistence is possible, and its consequence for cells is currently an active area of research.

A series of DPD studies by Smit and coworkers is devoted to the phase behavior of membranes (Kranenburg et al., 2003a,b, 2004a,b; Kranenburg and Smit, 2004, 2005; de Meyer and Smit, 2009). In this context, the considered observables are the area per lipid ( $A$ ), the bilayer thickness ( $h$ ), and the chain overlap and interdigitation characterized by  $D_{\text{overlap}} = (2L_z - D_c)/L_z$  with  $D_c$  being the thickness of the hydrophobic core and  $L_z$  being the distance between the first and the terminal chain bead projected onto the bilayer normal. Moreover, the angular order parameter of lipids can be considered, that is,  $S = \langle 3 \cos^2 \theta - 1 \rangle / 2$  with  $\theta$  being the angle between the lipid and the bilayer normal. Also, the in-plane radial distribution function of lipid head beads may be examined as an observable.

In Kranenburg et al. (2003a,b), the phase behavior of membranes of single-chain lipids HT<sub>5</sub> and double-chain lipids H<sub>3</sub>(T<sub>4</sub>)<sub>2</sub> was studied. Membranes consisting of single-chain lipids were found to be in the  $L_\beta$  phase at low temperature ( $k_B T = 0.8, 0.9$ ), that is, a high order parameter



of the lipid chains ( $S > 0.8$ ) was observed. This indicates that the chains aligned along the bilayer normal. The radial distribution function also showed pronounced peaks as compared to higher temperatures, indicating a more structured organization of lipid head groups in the bilayer plane. At higher temperature ( $k_B T \geq 1$ ), lipids were in the liquid crystalline phase  $L_\alpha$  ( $S \leq 0.5$ ) and the membrane thickness decreased. While for single-chain lipids there was no tilt of lipids observed in the gel phase ( $L_\beta$ ), double-chain lipids were tilted ( $L_{\beta'}$ ).

For some lipid species with large head sizes or strong head-head repulsion parameters, membranes were found to be in the interdigitated gel phase  $L_{\beta 1}$  at low temperatures, where the termini of the lipids penetrated the opposing monolayer (visible in bead density plots). The interdigitation was also reflected in a reduced thickness of the lipid bilayer and in an increased area per lipid.

In addition to the above, the existence of a rippled phase  $P_{\beta'}$  at the transition between  $L_{\beta'}$  and  $L_\alpha$  was observed for strong head-head repulsion (Kranenburg et al., 2004a; Kranenburg and Smit, 2005). In the rippled phase, the membrane is segmented in thick and thin regions in an alternating fashion. While in thick regions lipids of the opposing monolayers are strictly separated, they overlap in the thinner regions. Moreover, lipids are tilted in thick zones but not in thin zones. The stability of this particular membrane organization grew with increasing lipid tail length. From their observations, the authors interpreted the rippled phase as a 50–50% coexistence of the  $L_{\beta'}$  (or the  $L_c$ ) and the  $L_\alpha$  phase.

Alcohol-induced interdigitation was also examined by DPD (Kranenburg and Smit, 2004; Kranenburg et al., 2004b). The addition of alcohol to membrane systems is known to influence the transition temperature in dependence on the alcohol molar fraction and temperature. In the simulations, a lipid model  $H_3(T_7)_2$  was chosen that can be mapped onto DSPC; alcohol was modeled by short single-chains  $HT_n$  with  $n = 1, 2, 3$ . At low alcohol concentrations, a noninterdigitated phase formed, whereas at high concentrations, a fully interdigitated phase was stabilized. Between these two extremes, a coexistence region was observed. The interdigitation was explained by the formation of voids in the hydrophobic core of the bilayer upon alcohol insertion.

Another study of membrane phase behavior with DPD inspected the action of cholesterol (de Meyer and Smit, 2009). Here, double-chain lipids  $H_3(T_4)_2$  (corresponding to DMPC) were combined with a model of

cholesterol that consisted of a single head group followed by a ring structure representing the sterol part of the molecule. The DMPC–cholesterol phase diagram had a rich structure with various phases. Further, the so-called condensation effect was observed in these simulations in quantitative agreement with experiments. This effect reflects that lipid and cholesterol do not mix ideally. Instead, the area per molecule is much lower as compared to an ideal mixing. The main phase transition temperature was found to increase upon adding cholesterol to the membrane due to the conical-shaped cholesterol being incapable of protecting the hydrophobic core of the membrane as efficiently as normal lipids. At high temperature, this could be compensated by fluctuations of the lipids, yet at lower temperatures, it forced the membrane to adopt the more ordered  $L_o$  state. Indeed, the condensation effect was seen to vanish when changing the cholesterol model's shape toward a more cylindrical form.

In summary, DPD also provides an excellent tool to faithfully explore the rich phase behavior of membranes.

### *E. Membrane Fission and Fusion*

#### *1. Membrane Budding and Fission*

A large variety of biological phenomena like endo- and exocytosis or intracellular protein trafficking rely on budding and fission events of lipid membranes. The length scales relevant for these events are below 100 nm, and DPD simulations hence are an adequate tool to investigate these dynamic topological membrane deformations.

The release of small vesicles from a large paternal vesicle consisting of two different lipid species with the same geometry (single-chains  $HT_3$ ) was studied quite extensively (Yamamoto and Hyodo, 2003; Laradji and Sunil Kumar, 2004, 2005, 2006). The repulsion between the two lipid types was set to be stronger than the repulsion between two lipids of the same species, that is, a segregation of lipids was enforced. The initial setting of the simulations was taken as a preformed vesicle of heterogeneous composition with a small, self-healing hole to prevent an initial osmotic pressure (Laradji and Sunil Kumar, 2004, 2005, 2006), or alternatively as a self-forming vesicle with homogenous composition derived from an initially flat bilayer (Yamamoto and Hyodo, 2003). In the latter approach, about 30% of the lipids were changed to the second lipid species after the

vesicle had equilibrated. In both approaches, a demixing of the two lipid species was observed and the emerging domains coalesced with time. Under appropriate conditions, the domains formed buds and pinched off from the paternal vesicle as microvesicles.

The coalescence dynamics was characterized via the net interface length  $L$  between the two phases and the domain number  $N_C$  (Laradji and Sunil Kumar, 2004, 2005, 2006). Both parameters showed a power-law decrease in time according to  $L \sim t^{-0.3}$  and  $N_C \sim t^{-2/3}$ . Moreover, the dynamics of domain coalescence was seen to depend on the state of the vesiculation process. Relevant parameters for vesicle formation were the lateral tension of the paternal membrane, its bending modulus, the line tension between the domains (set by the repulsion strength between the two lipid species), and the spontaneous curvature of the buds (Laradji and Sunil Kumar, 2005). A low line tension was seen to inhibit the pinch-off of even large buds. Successful vesiculation events happened within very short times (500–1000 DPD time units) (Laradji and Sunil Kumar, 2005).

Details of the fission event have been investigated in Yamamoto and Hyodo (2003). The authors observed two possible pathways, depending on the repulsion between the two lipid species and on the membrane bending rigidity. In the first fission pathway, a domain formed a bud with a neck at the domain edge. The neck then tightened until the microvesicle was released from the paternal membrane. This process of fission is in agreement with predictions from continuum theories like the bending elastic model for the case of a small bending rigidity or a large interfacial energy. It also fits to experimental observations. The second fission scenario was observed for strong segregation of the two lipid species (induced by an increased repulsion), or when the bending rigidity was reduced via tuning the water–lipid repulsion for one lipid species. In this fission process, a cut along the bud boundary induced the fission, that is, the paternal vesicle was left with a hole that slowly closed again. This fission pathway was also observed when the thermal undulations of the membrane were increased.

Further, vesiculation from a planar membrane was studied (Hong et al., 2007). While standard DPD simulations fix the total number  $N$  of beads in the system and employ periodic boundary conditions, the authors chose here a variable number of beads with fixed boundary conditions. Adding/removing lipids to/from the membrane according to a density criterion (add/remove when density is low/high) allowed for the observation of vesiculation even in fairly small systems. Periodic boundary conditions with

a fixed number of lipids, in general, suppress large deformations (budding or fission events) as they try to maintain a flat membrane. Only if the size of the bud is considerably smaller than the total size of the membrane, the perturbation becomes small enough to allow for budding and/or pinch-off events. Using their nonstandard approach, [Hong et al. \(2007\)](#) observed budding and fission of a preformed circular domain of one lipid species from a planar membrane of a second lipid species. Again, the interplay of line tension, membrane bending modulus, and surface tension was found to influence the emergence and duration of the budding event. A higher line tension did speed up the budding process considerably. Further, the time of bud formation increased with the size of the domain.

## 2. *Membrane Fusion*

The inverse process of vesiculation, that is, fusion events, has been investigated in a series of papers by Shillcock, Lipowsky, and coworkers. The fusion of a single vesicle with a planar membrane ([Shillcock and Lipowsky, 2005](#); [Grafmüller et al., 2007, 2009](#)) was examined as well as the fusion of two vesicles with each other ([Gao et al., 2008](#)). The lipid model in these studies was  $H_3(T_4)_2$  which can be mapped onto DMPC.

As a result, the authors found that fusion of a vesicle with a planar membrane is possible only if both membranes are under tension, that is, both had to have an increased area per molecule. No fusion was observed when vesicle and planar membrane were initially relaxed ([Shillcock and Lipowsky, 2005](#)). Instead, the vesicle adhered to and spread onto the planar membrane. With a growing membrane tension (i.e., molecular area  $A$ ), the fusion probability increased linearly to a maximal value of about 90% and decreased thereafter ([Grafmüller et al., 2009](#)). Fusion was observed in only 50–70% of all simulations, rendering the event unreliable and stochastic. Membranes that did not undergo fusion relaxed their tension by other pathways, for example, via rupture of the vesicle or the planar membrane (at high tensions), or via a hemifusion of the two membranes bilayers at the contact region (at low tensions). Making fusion a more reliable event thus requires additional forces, for example, via specialized proteins like SNAREs.

The observed fusion pathway was as follows (Grafmüller et al., 2007, 2009): as an initial event, the vesicle adhered to the planar membrane. Then lipids started to perform an interbilayer flip-flop from the vesicle to the membrane. Lipid tails moving through the head groups connected the hydrophobic cores of the adherent two bilayers. A partially fused, strongly disordered membrane patch established within the contact region. Next, lipids started to reorder and to form a small domain of a single hemifused bilayer that expanded and ruptured to form a full fusion pore. The fusion time depended on the tension. For high tension, the fusion time  $t_f$  (measured from first contact of the two membranes to the opening of the fusion pore) was well below 1  $\mu$ s, while smaller tensions yielded  $t_f \leq 12 \mu$ s (Grafmüller et al., 2007). Long fusion times corresponded to an intermediate adherent state of the vesicle. Comparing two different vesicle sizes (15 and 30 nm), fusion was faster and happened more likely for smaller (15 nm) vesicles. However, it seems that rather the ratio of vesicle radius to the planar bilayer area difference than the actual vesicle size primarily influences the fusion time (Grafmüller et al., 2007).

During the fusion process, energetic barriers had to be overcome between the vesicle adhesion to the membrane and the first lipid flip, and between the first flip and the formation of the hemifused domain (Grafmüller et al., 2007, 2009). The corresponding times were seen to decrease exponentially with increasing membrane tension, suggesting that the energy barriers should depend linearly on tension.

### *F. Dynamics of Membrane Proteins*

Biological membranes are not pure lipid bilayers but rather are crowded with membrane proteins that associate with the bilayer. Membrane proteins contribute roughly 30% of the net weight of a typical biomembrane, and about one-third of a cell's proteome is membrane proteins. Membrane proteins are involved in a multitude of tasks like cell-cell communication, signal transfer, enzymatic reactions, or connection of the membrane to the cell's cytoskeleton.

DPD was used in a number of studies to examine the physical properties and dynamics of membrane proteins and their interactions with the embedding lipid bilayer. In particular, the problem of hydrophobic matching of transmembrane proteins and its consequences were addressed.

The simplest transmembrane proteins consist of two hydrophilic moieties at either end of an extended hydrophobic domain (typically an  $\alpha$ -helix). When the hydrophobic part of the protein does not match the thickness of the hydrophobic bilayer core, one speaks of a “hydrophobic mismatch”. Hydrophobic mismatch can be positive or negative, depending on whether the protein’s hydrophobic domain is longer or shorter, respectively, than the membrane’s hydrophobic core. In both cases, deformations of the local lipid bilayer structure have been predicted which compensate for the exposure of hydrophobic regions to the solvent. Effects like a changed mobility of proteins or nonspecific membrane-mediated forces between proteins can be expected.

In DPD, a transmembrane protein can be modeled as a cylinder consisting of hydrophobic beads with one or more layers of hydrophilic beads at the bases of the cylinder (see Fig. 8A). The cylindrical structure can be achieved by a hexagonal arrangement of bead chains, and the number of “shells” around the central chain determines the in-plane protein radius. Adjacent beads in the cylinder are interconnected by spring potentials, which make the cylinders fairly rigid. In such a model, both the protein radius and length are variable.

### *1. Perturbations of the Lipid Bilayer due to Membrane Proteins*

Several DPD studies (Venturoli et al., 2005; Guigas and Weiss, 2008; Schmidt et al., 2008) examined the local bilayer deformation due to an embedded transmembrane protein. In order to explore the effects of a positive or negative hydrophobic mismatch on the state of the lipid bilayer, the hydrophobic length of the protein was systematically varied. Venturoli et al. (2005) used in their study a two-chain lipid model ( $H_3(T_5)_2$ , similar to DMPC), while we have used a reduced lipid model, HT3 (Guigas and Weiss, 2008; Schmidt et al., 2008).

These studies found that transmembrane inclusions of different lengths differ in the average tilt angle of the protein with respect to the bilayer normal. A rather weak tilting was observed for proteins with a negative or a vanishing hydrophobic mismatch, whereas a monotonic increase of the tilting with an increasing positive mismatch was seen (Venturoli et al., 2005; Schmidt et al., 2008). The increase in tilting was stronger for proteins with small radii as compared to those with larger radii (Venturoli et al., 2005). The presence of a mismatched transmembrane

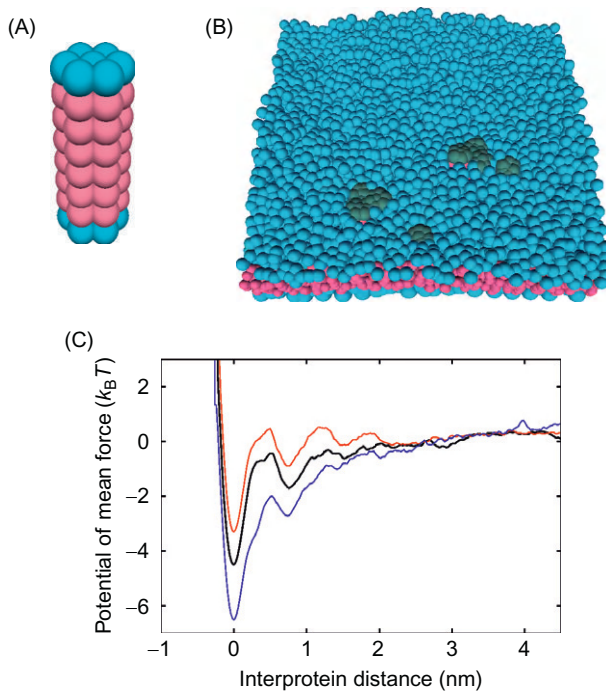


FIG. 8. (A) A model transmembrane protein consisting of a hexagonal cylinder of beads (Guigas and Weiss, 2006, 2008; Schmidt et al., 2008). Head groups are depicted in blue and tail groups in red. (B) Transmembrane proteins form clusters due to a hydrophobic mismatch with the embedding membrane (Schmidt et al., 2008). Two trimers can be seen as well as two monomers. For clarity, proteins are depicted in dark gray. (C) The interaction of two proteins can be characterized by the potential of mean force (PMF). The depth of the potential well indicates a bound state at short interprotein distances that increases with positive (blue curve) and negative (red curve) hydrophobic mismatch (data from Schmidt et al., 2008).

protein changed the local membrane thickness, that is, the bilayer thickness locally adapted to the protein length (Venturoli et al., 2005; Guigas and Weiss, 2008; Schmidt et al., 2008) as  $d = d_0 + d_1 \exp(-x/\xi)$  with  $d_0$  denoting the unperturbed bilayer thickness, and  $d_1$ ,  $\xi$  being the thickness increment and the typical relaxation distance.

The thickness change  $d_1$  grew monotonically with increasing mismatch, until it reached a constant value for large positive mismatches beyond

which the protein only increased its tilting angle. The decay length  $\xi$  of the thickness deformation was found to have typical values of about 1 nm.

Inspecting the order parameter  $S$  of the lipids in the vicinity of the protein, lipids showed a higher ordering in an annulus of size  $\xi$  around the protein (Schmidt et al., 2008). In particular, lipids were found to be less tilted close to a protein with a positive hydrophobic mismatch and more strongly tilted close to a protein with a negative mismatch.

## 2. *Mobility of Membrane Inclusions*

DPD also allows one to determine stationary transport coefficients, for example, diffusion constants. In Guigas and Weiss (2006), the dependence of the lateral and the rotational diffusion coefficient on the radius of transmembrane proteins with no mismatch was examined. The authors tested proteins with radii up to 40 nm. Objects with a diameter larger than 10 nm would rather represent large membrane inclusions, for example, lipid microdomains or protein clusters. To be able to achieve system sizes that are large enough for these inclusions, the authors used for some of their simulations an implicit solvent variant of DPD. In contrast to standard DPD, the solvent-mediated attraction of lipids was replaced here by an attractive force between hydrophobic beads of neighboring lipids.

In all these simulations, rotational and translational diffusion coefficients were determined for different protein radii  $R$  from the protein's (angular) mean-square displacement. Both the lateral and the rotational diffusion coefficients were found to vary systematically with protein radius, showing a logarithmic decrease with  $R$  for lateral diffusion and a power-law decrease  $\sim 1/R^2$  for rotational diffusion. These results are in agreement with theoretical predictions derived in a mean-field perturbative approach (Saffmann and Delbruck, 1975). For large protein radii, where the perturbative prediction fails (Hughes et al., 1981), a crossover from the logarithmic behavior toward  $D \propto 1/R^2$  was found for the translational diffusion coefficient. This result may be explained by drawing the analogy of the protein's diffusion to the edgewise motion of a thin disk in which internal fluctuations are excited due to the surrounding lipids' impact.

It was shown later that the gross result does not change when the protein exhibits a hydrophobic mismatch (Guigas and Weiss, 2008). Only 20–30% lower values of the translational diffusion coefficient were observed while the radius dependence still followed the theoretical



prediction (Hughes et al., 1981). Indeed, the different protein lengths were reflected in varying effective membrane viscosities.

### 3. Membrane-Mediated Protein–Protein Interactions

Many DPD studies (Schmidt et al., 2008; de Meyer et al., 2008a,b; Morozova and Weiss, 2010; Schmidt and Weiss, 2010) dealt with nonspecific, lipid-mediated interactions of transmembrane proteins with varying degrees of hydrophobic mismatching. Such interactions have been predicted, for example, from elastic continuum theories and capillary forces.

In accordance with these theoretical predictions, Schmidt et al. (Schmidt et al., 2008) observed that membrane proteins with the same hydrophobic mismatch spontaneously formed stable clusters (see Fig. 8B). The strength of this interaction was determined from the potential of mean force (PMF) of two equally shaped transmembrane proteins embedded in a lipid bilayer. The PMF is defined as the negative logarithm of the distance distribution (i.e., the pair correlation function) of the proteins. A uniform sampling of all protein distances within the simulation box can be achieved by an umbrella sampling (de Meyer et al., 2008a). An alternative approach to obtain the same information is to monitor the distance and net forces acting on the two proteins (Schmidt et al., 2008). However, in cases of a strong protein interaction, this method may suffer from a lack of statistics for large protein distances.

The PMFs obtained for mismatched proteins had a deep minimum at short interprotein distance, indicating a bound state of the two proteins (see Fig. 8C) (Schmidt et al., 2008). Depth and width of the potential well increased for positive and negative hydrophobic mismatches, reaching binding energies of up to  $12k_bT$  for large mismatches. At interprotein distances corresponding to the thickness of one and two lipids, additional side minima were observed that mean field theories did not predict. Most likely, these minima reflect the discrete distances of ordered lipid “shells” around the protein. For larger distances, the potentials essentially became flat. From the PMF, one can determine the mean first passage time from the bound to the free state. These dimer lifetimes increased exponentially with increasing hydrophobic mismatching. Simulations of large ensembles of proteins further revealed a cluster formation that correlated with the strength of the hydrophobic mismatch. While for vanishing mismatch at best transient dimers and trimers were

observed, long-lived assemblies like octamers and nonamers appeared for large positive or negative mismatches.

In a follow-up study, it was shown that proteins with different hydrophobic mismatch can segregate into homo-oligomers (Schmidt and Weiss, 2010) if the difference in mismatch is sufficiently large. If this condition is not met, hetero-oligomers with a dipole-like arrangement formed, that is, proteins with the same mismatch aimed at being next neighbors and tried to reduce their contact to the second protein species. Indeed, both effects may be crucial for understanding protein trafficking in the early secretory pathway of eukaryotes (Schmidt and Weiss, 2010).

The PMF of membrane-mediated protein–protein interactions was also studied by De Meyer (de Meyer et al., 2008a). Here, comparable results were found for proteins with small radii and for large proteins with negative mismatch. As a slight difference, a weak repulsive barrier between proteins at intermediate protein distance was detected here. Proteins with large radii and a vanishing or positive mismatch gave potentials of a completely different shape without any pronounced minima but with weak repulsive and attractive elements at short and intermediate protein distances.

At the heart of the reported clustering of membrane proteins lies the system's wish to reduce its total energy. Indeed, in order to protect hydrophobic groups from water, lipids experience severe geometrical constraints near to a protein. This is ultimately reflected in the lipids' order parameter and the bilayer thickness (cf. above). To gain entropy for the whole system, a reduction of the lipid–protein interface is favorable and hence proteins start to cluster. This phenomenon is similar to the spontaneous formation of micelles when lipids are immersed in water.

A measure for the quality of the hydrophobic shielding was suggested in de Meyer et al. (2008a) via the ratio of lipid head fraction to lipid tail fraction at each point in the membrane plane. For values larger than unity, this observable indicates a high density of lipid heads and thus a good shielding. The authors show that minima and maxima (i.e., regions of mutual protein attraction and repulsion) found in the PMF are related to good and bad shielding of proteins separated by the respective distances.

Further, it was determined that the interaction between a single protein and a cluster or between two clusters differed from that between two single proteins (de Meyer et al., 2008b).

In addition to the membrane-mediated interaction of transmembrane proteins, also the partitioning behavior on heterogeneous membranes for various degrees of hydrophobic mismatching was studied (Schmidt and Weiss, 2010). Consistent with earlier results, proteins partitioned to the domains that matched best their hydrophobic length. Again, this partitioning may be key for understanding cellular protein sorting. Related to protein sorting, also the influence of acylation of transmembrane proteins was explored (Morozova and Weiss, 2010). Acylation was found to significantly enhance the tilting of transmembrane proteins and, as a consequence, to alter hydrophobic mismatch-induced clustering and the partitioning on phase-separated bilayers (Morozova and Weiss, 2010). Thus, acylation can be used as a posttranslational regulator of transmembrane length-induced sorting. These results support recent experimental findings that indicated a role of acylation in regulating the trafficking behavior of vital transmembrane proteins.

Finally, also the influence of cholesterol on protein aggregation was studied via DPD (de Meyer et al., 2008b). The authors found larger clusters of transmembrane proteins with positive mismatch in the presence of cholesterol as compared to cholesterol-free membranes. This effect mainly arose due to the formation of cholesterol-enriched and cholesterol-depleted zones surrounding proteins, which improved the hydrophobic shielding.

### *G. Altering Biomembrane Properties by Exogenous Factors*

One of the first applications of DPD to biomembranes highlighted the effect of nonionic surfactants, in particular, alcohol ethoxylates, on membrane morphology (Groot and Rabone, 2001). Such surfactants can be found in detergents and were, for example, shown to inhibit bacterial growth. By measuring the diffusion of water across the membrane, Groot and Rabone showed that small transient holes appear in the membrane at 40% mole fraction of  $C_6E_8$ . At 60% and 70% holes remained permanently. With the surfactant,  $C_{12}E_6$  permanent holes occurred only at 90% mole fraction. Thus, the size of surfactant head group determines the extent of membrane damage. In the same study also the rupture properties of bilayers with different surfactant fractions were investigated. The inclusion of surfactants considerably reduced the stress resistance of the bilayer, that is, rupturing the membrane was easier.

DPD is especially useful for studies of large membranes where the behavior on the level of single lipid molecules is of interest. In [Jakobsen et al. \(2005b\)](#), the influence of the action of the enzyme phospholipase PLA<sub>2</sub> on membrane structure was investigated. PLA<sub>2</sub> catalyzes the hydrolysis of phospholipids, producing a lysolipid and a fatty acid, that is, the enzyme works as scissors that cut a lipid with two hydrophobic tails into two unequal parts. Due to high demands on computational power, the dynamics of the enzyme activity and the concrete process of cleavage were not taken into account, but rather the mechanical properties after the cleavage were monitored. Varying the fraction of hydrolysis products in one of the membrane leaflets, these simulations showed that the bilayer integrity was not compromised, yet the fluctuations increased strongly due to a reduced bending stiffness. Concomitantly, the lateral diffusion of lipids was enhanced, especially in the leaflet where the enzyme activity had taken place. Cleavage products also showed an enhanced probability for flipping to the opposite leaflet (flip-flop).

#### H. Conclusion

The above examples highlight that DPD is a powerful and yet fairly easy method to study biomembrane properties with and without proteins on a mesoscale. Neglecting details on the subnanometer level, DPD allows for studying longer length and time scales than MD approaches. The mesoscale at which DPD can be efficiently used helps to bridge the gap between experimental and theoretical results and hence provides an excellent tool in a multiscale approach to biomembranes.

#### REFERENCES

- Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72**, 2384.
- Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A., Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684.
- Cantor, R. S. (1997). The lateral pressure profile in membranes: a physical mechanism of general anesthesia. *Biochemistry* **36**, 2339.
- Da-Wei, Li, Liu, X. Y., Feng, Y. P. (2004). Bond-angle-potential-dependent dissipative particle dynamics simulation and lipid inverted phase. *J. Phys. Chem. B* **108**, 11206.

- de Groot, B. L., Grubmüller, H. (2001). Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. *Science* **294**, 5550.
- de Kruijf, B. (1997). Biomembranes. Lipids beyond the bilayer. *Nature* **386**, 129.
- de Meyer, F., Smit, B. (2009). Effect of cholesterol on the structure of a phospholipid bilayer. *Proc. Natl. Acad. Sci. USA* **106**, 6654.
- de Meyer, F., Venturoli, M., Smit, B. (2008a). Molecular simulations of lipid-mediated protein-protein interactions. *Biophys. J.* **95**, 1851–1865.
- de Meyer, F., Rodgers, J. M., Willems, T. F., Smit, B. (2008b). Molecular simulation of the effect of cholesterol on lipid-mediated protein-protein interactions. *Biophys. J.* **99**, 3629.
- de Vries, A. H., Mark, A. E., Marrink, S. J. (2004). Molecular dynamics simulation of the spontaneous formation of a small DPPC vesicle in water in atomistic detail. *J. Am. Chem. Soc.* **126**, 4488.
- Español, P., Warren, P. (1995). Statistical mechanics of dissipative particle dynamics. *Europhys. Lett.* **30**, 191.
- Feller, S. E., Zhang, Y., Pastor, R. W., Brooks, B. R. (1995). Constant pressure in molecular dynamics: the Langevin piston method. *J. Chem. Phys.* **103**, 4613.
- Gao, L., Lipowsky, R., Shillcock, J. C. (2008). Tension-induced vesicle fusion: pathway and pore dynamics. *Soft Matter* **4**, 1208.
- Goetz, R., Lipowsky, R. (1998). Computer simulations of bilayer membranes: self-assembly and interfacial tension. *J. Chem. Phys.* **108**, 7397.
- Grafmüller, A., Shillcock, J., Lipowsky, R. (2007). Pathway of membrane fusion with two tension-dependent energy barriers. *Phys. Rev. Lett.* **98**, 218101.
- Grafmüller, A., Shillcock, J., Lipowsky, R. (2009). The fusion of membranes and vesicles: pathway and energy barriers from dissipative particle dynamics. *Biophys. J.* **96**, 2658.
- Grest, G. S., Kremer, K. (1986). Molecular-dynamics simulation for polymers in the presence of a heat bath. *Phys. Rev. A* **33**, 3628.
- Groot, R. D. (2004a). Applications of Dissipative Particle Dynamics, Volume 640, Springer, Berlin/Heidelberg.
- Groot, R. D. (2004b). Applications of dissipative particle dynamics. *Lect. Notes Phys.* **640**, 5.
- Groot, R. D., Rabone, K. L. (2001). Mesoscopic simulation of cell membrane damage, morphology change and rupture by nonionic surfactants. *Biophys. J.* **81**, 725.
- Groot, R. D., Warren, P. (1997). Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulations. *J. Chem. Phys.* **107**, 4423.
- Guigas, G., Weiss, M. (2006). Size-dependent diffusion of membrane inclusions. *Biophys. J.* **91**, 2393.
- Guigas, G., Weiss, M. (2008). Influence of hydrophobic mismatching on membrane protein diffusion. *Biophys. J.* **95**, L25–L27.
- Helfrich, W. (1973). Elastic properties of lipid bilayers—theory and possible experiments. *Z. Naturforsch.* **28**, 693–703.
- Hong, B., Qiu, F., Zhang, H., Yang, Y. (2007). Budding dynamics of individual domains in multicomponent membranes simulated by n-varied dissipative particle dynamics. *J. Phys. Chem. B* **111**, 5837.

- Hoogerbrugge, P., Koelman, J. (1992). Simulating microscopic hydrophobic phenomena with dissipative particle dynamics. *Europhys. Lett.* **19**, 155.
- Hoover, W. G. (1985). Canonical dynamics—equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695.
- Hughes, B. D., Pailthorpe, B. A., White, L. R. (1981). The translational and rotational drag on a cylinder moving in a membrane. *J. Fluid Mech.* **110**, 349.
- Illya, G., Lipowsky, R., Shillcock, J. C. (2005). Effect of chain length and asymmetry on material properties of bilayer membranes. *J. Chem. Phys.* **122**, 244901.
- Illya, G., Lipowsky, R., Shillcock, J. C. (2006). Two-component membrane material properties and domain formation from dissipative particle dynamics. *J. Chem. Phys.* **125**, 114710.
- Jakobsen, A. F. (2005). Constant-pressure and constant-surface tension simulations in dissipative particle dynamics. *J. Chem. Phys.* **122**, 124901.
- Jakobsen, A. F., Mouritsen, O. G., Besold, G. (2005a). Artifacts in dynamical simulations of coarse-grained model lipid bilayers. *J. Chem. Phys.* **122**, 204901.
- Jakobsen, A. F., Mouritsen, O. G., Weiss, M. (2005b). Dissipative particle dynamics is especially useful for studies of large membrane systems where the behavior of single lipid molecules is of interest. Close-up view of the modifications of fluid membranes due to phospholipase A(2). *J. Phys. Condens. Matt.* **17**, S4015–S4024.
- Koopman, E. A., Lowe, C. P. (2006). Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations. *J. Chem. Phys.* **124**, 204103.
- Kranenburg, M., Smit, B. (2004). Simulating the effect of alcohol on the structure of a membrane. *FEBS Lett.* **568**, 15.
- Kranenburg, M., Smit, B. (2005). Phase behavior of model lipid bilayers. *J. Phys. Chem. B* **109**, 6553.
- Kranenburg, M., Venturoli, M., Smit, B. (2003a). Phase behavior and induced interdigitation in bilayers studied with dissipative particle dynamics. *J. Phys. Chem. B* **104**, 11497.
- Kranenburg, M., Venturoli, M., Smit, B. (2003b). Molecular simulations of mesoscopic bilayer phases. *Phys. Rev. E* **67**, 060901(R).
- Kranenburg, M., Venturoli, M., Smit, B. (2004a). Mesoscopic simulations of phase transitions in lipid bilayers. *Phys. Chem. Chem. Phys.* **6**, 4531.
- Kranenburg, M., Vlaar, M., Smit, B. (2004b). Simulating induced interdigitation in membranes. *Biophys. J.* **87**, 1596.
- Lantzsch, G., Binder, H., Heerklotz, H. (1994). Surface area per molecule in lipid/C12EO<sub>n</sub> membranes as seen by fluorescence resonance energy transfer. *J. Fluoresc.* **4**(4), 339–343.
- Laradji, M., Sunil Kumar, P. B. (2004). Dynamics of domain growth in self-assembled fluid vesicles. *Phys. Rev. Lett.* **93**, 198105.
- Laradji, M., Sunil Kumar, P. B. (2005). Domain growth, budding, and fission in phase-separating self-assembled fluid bilayers. *J. Chem. Phys.* **123**, 224902.
- Laradji, M., Sunil Kumar, P. B. (2006). Anomalously slow domain growth in fluid membranes with asymmetric transbilayer lipid distribution. *Phys. Rev. E* **73**, 040901(R).

- Lin, L. C. L., Brown, F. L. H. (2006). Simulating membrane dynamics in nonhomogeneous hydrodynamic environments. *J. Chem. Theory Comput.* **2**, 472.
- Lowe, C. P. (1999). An alternative approach to dissipative particle dynamics. *Europhys. Lett.* **47**, 145.
- Lu, J. R., Li, Z. X., Thomas, R. K., Staples, E. J., Tucker, I., Penfold, J. (1993). Neutron reflection from a layer of monododecyl hexaethylene glycol adsorbed at the air-liquid interface—the configuration of the ethylene-glycol chain. *J. Chem. Phys.* **97**, 8012.
- Marrink, S. J., Mark, A. E. (2001). Effects of undulations on surface tension in simulated bilayers. *J. Phys. Chem. B* **105**, 6122.
- Morozova, D., Weiss, M. (2010). On the role of acylation of transmembrane proteins. *Biophys. J.* **98**, 800.
- Nikunen, P., Karttunen, M., Vattulainen, I. (2003). How would you integrate the equations of motion in dissipative particle dynamics simulations? *Comput. Phys. Commun.* **153**, 407–423.
- Nosé, S. (1984). A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52**, 255.
- Ollila, O. H. S., Risselada, H. J., Louhivouri, M., Lindahl, E., Vattulainen, I., Marrink, S. J. (2009). 3D pressure distribution in lipid membranes and membrane-protein complexes. *Phys. Rev. Lett.* **102**, 078101.
- Saffmann, P. G., Delbruck, M. (1975). Brownian motion in biological membranes. *Proc. Natl. Acad. Sci. USA* **72**, 1250.
- Schmidt, U., Weiss, M. (2010). Hydrophobic-mismatch induced clustering as a primer for protein sorting in the secretory pathway. *Biophys. Chem.* **151**, 34–38.
- Schmidt, U., Guigas, G., Weiss, M. (2008). Cluster formation of transmembrane proteins due to hydrophobic mismatching. *Phys. Rev. Lett.* **101**, 128104.
- Shillcock, J. C. (2008). Insight or illusion? Seeing inside the cell with mesoscopic simulations. *HFSP J.* **2**(1),1–6.
- Shillcock, J. C., Lipowsky, R. (2002). Equilibrium structure and lateral stress distribution of amphiphilic bilayers from dissipative particle dynamics simulations. *J. Chem. Phys.* **117**, 5048.
- Shillcock, J. C., Lipowsky, R. (2005). Tension-induced fusion of bilayer membranes and vesicles. *Nat. Mater.* **4**, 225–228.
- Venturoli, M., Smit, B. (1999). Simulating the self-assembly of model membranes. *Phys. Chem. Commun.* **10**, 45–49.
- Venturoli, M., Smit, B., Sperotto, M. M. (2005). Simulation studies of protein-induced bilayer deformations, and lipid-induced protein tilting, on a mesoscopic model for lipid bilayers with embedded proteins. *Biophys. J.* **88**, 1778.
- Yamamoto, S., Hyodo, S. (2003). Budding and fission dynamics of two-component vesicles. *J. Chem. Phys.* **118**, 7937.
- Yamamoto, S., Maruyama, Y., Hyodo, S. (2002). Dissipative particle dynamics study of spontaneous vesicle formation of amphiphilic molecules. *J. Chem. Phys.* **116**, 5842.

# COARSE-GRAINED REPRESENTATION OF PROTEIN FLEXIBILITY. FOUNDATIONS, SUCCESSES, AND SHORTCOMINGS

By MODESTO OROZCO,<sup>\*,†,\*</sup> LAURA ORELLANA,<sup>\*</sup> ADAM HOSPITAL,<sup>\*,†</sup>  
ATHI N. NAGANATHAN,<sup>\*</sup> AGUSTÍ EMPERADOR,<sup>\*</sup> OLIVER CARRILLO,<sup>\*</sup> AND  
J. L. GELPI<sup>\*,†,\*</sup>

<sup>\*</sup>Joint IRB-BSC Program on Computational Biology, Barcelona Supercomputing Center and  
Institute of Research in Biomedicine, Parc Científic de Barcelona, Barcelona, Spain

<sup>†</sup>National Institute of Bioinformatics, Structural Bioinformatics Node,  
Baldiri Reixac 10-12, Barcelona, Spain

<sup>\*</sup>Departament de Bioquímica i Biologia Molecular, Facultat de Biologia, Barcelona, Spain

I. Introduction .....	184
II. Coarse-Grained Potentials .....	188
A. Gō-Like Potentials .....	190
B. Harmonic Potentials .....	191
C. Flat Potentials .....	194
D. Physical and Pseudo-physical Potentials .....	197
III. Sampling Techniques .....	200
A. Normal Mode Analysis .....	200
B. Monte Carlo .....	203
C. Langevin Dynamics .....	205
D. Discrete Molecular Dynamics .....	208
IV. Conclusions .....	210
References .....	212

## ABSTRACT

Flexibility is the key magnitude to understand the variety of functions of proteins. Unfortunately, its experimental study is quite difficult, and in fact, most experimental procedures are designed to reduce flexibility and allow a better definition of the structure. Theoretical approaches have become then the alternative but face serious timescale problems, since many biologically relevant deformation movements happen in a timescale that is far beyond the possibility of current atomistic models. In this complex scenario, coarse-grained simulation methods have emerged as a powerful and inexpensive alternative. Along this chapter, we will review these coarse-grained methods, and explain their physical foundations and their range of applicability.



## I. INTRODUCTION

Biological macromolecules, and in particular proteins, are large and flexible entities, which perform their biological action when embedded in solvent, either water or the membrane phospholipids. Analysis of current version of the Protein Data Bank (PDB; Berman et al., 2000; <http://www.pdb.org>) illustrates that proteins of known experimental structure range typically between 500 and 7000 atoms, but in some cases, protein systems reach more than 16,000 atoms (see Fig. 1). As experimental resolution techniques advance, the size histogram in Fig. 1 is expected to displace to the right side due to the incorporation of large protein assemblies to the database. However, the real problem in protein simulation originates from the need to introduce solvent in the calculation, which dramatically increases the number of particles in the system. For example, in our MoDEL (Molecular Dynamics Extended Library) database, that contains atomistic molecular dynamics (MD) simulations of representative PDB proteins in water (largely enriched in domain-sized proteins), typical

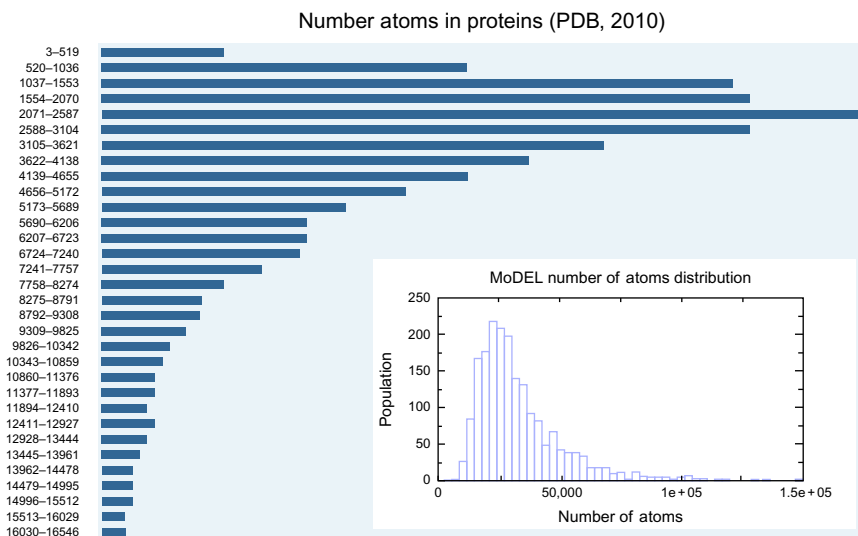


FIG. 1. Distribution of protein atoms in 2010 version of the Protein Data Bank (PDB). Inset corresponds to the distribution of atoms in solvated protein systems in our MoDEL database (<http://mmb.pcb.ub.es/MoDEL>).

simulation systems range from 10,000 to 50,000 atoms, but some systems have more than 150,000 atoms (see Fig. 1), that is, we are dealing with systems with up to half a million degrees of freedom. If we are interested in studying protein interactions, diffusion, or aggregation processes, simulated systems can easily reach many millions degrees of freedom, making atomistic simulation very complex.

As noted in the previous paragraph, size is a major limitation for the atomistic simulation of proteins, but often even more dramatic than the size problem is the time problem. Proteins are flexible, they move continuously, and therefore biological function cannot be understood without considering protein dynamics. Unfortunately, proteins move as a result of atomic vibrations happening in the nanosecond timescale, while most biologically relevant protein motions happen in the millisecond to second range. Thus, in order to follow, with atomistic detail, a biologically relevant protein motion, its energy (and associated forces) should be computed at least  $10^{12}$  times. For a typical system of 50,000, the calculation of just interatomic distances would require of the order of  $10^{21}$  floating point operations, not far from the Avogadro number.

Protein dynamics can be studied by different techniques, the most rigorous one being atomistic MD. In this approach, all atoms of the system are included at the same level of detail and their trajectories are determined by simple integration of Newton's (or closely related) equations of motion:

$$m_i \vec{a}_i = - \frac{dE}{d\vec{r}_i} \quad (1)$$

$$\vec{v}_i(t) = \vec{v}_i(t=0) + \int_{t=0}^{t=dt} \vec{a}_i(t) dt \quad (2)$$

$$\vec{r}_i(t) = \vec{r}_i(t=0) + \int_{t=0}^{t=dt} \vec{v}_i(t) dt \quad (3)$$

where  $t$  stands for time,  $r$  for the position,  $v$  for the velocity and  $a$  for the acceleration of atom  $i$ . The potential energy  $E$  is computed by using potential functions containing both bonded (stretching, bending, and torsion) and nonbonded interactions (van der Waals and electrostatic).

Stretching and bending are represented by harmonic expressions, torsions by Fourier series, electrostatics by Coulombic  $r^{-1}$  term and van der Waals by a Lennard–Jones  $r^{-12}$ ,  $r^{-6}$  term. These functional terms have been carefully parametrized, using experimental data and high-level quantum mechanical calculations as reference. It is not our purpose to comment these methods here and we just address the reader to suitable reviews of both atomistic MD and atomistic force fields (McCammon et al., 1977; Brooks et al., 1988; Karplus and McCammon, 2002).

Since its development in the seventies, MD has increased its popularity and it is now a technique used routinely by a large number of laboratories around the world. Several programs running highly optimized codes are available, often with free or almost free license scheme for (at least) academic groups. The improvement of MD codes and the development of more efficient and powerful computers have made MD simulations possible in the microsecond timescale for small proteins, while for the larger systems (more than million atom systems have been considered), “state-of-the art” simulations are at least one order of magnitude shorter. Databases such as MoDEL (Rueda et al., 2007a; Meyer et al., 2010) or Dymeomics (Van der Kamp et al., 2010) compile and make available to the community the near-equilibrium (10–100 ns range) dynamics of proteins in water for nearly 2000 representative proteins (see Fig. 2), covering a good percentage of unique-proteins PDB space (see Fig. 3).

In summary, MD is now a mature and widely used technique which provides results of high quality. Unfortunately, despite its successes we cannot ignore the existence of four fundamental problems that handicap its practical applicability: (i) the use of MD requires access to large computer resources and a notable degree of expertise in the setup of simulations; (ii) MD simulations are very costly and even with the best computer resources, human-collection time can extend into the months (or even years) timescale; (iii) timescale accessible to MD simulations is still far from that required to properly represent many biologically relevant transitions; and finally, (iv) data mining of hundreds of gigabytes of trajectories is complex, with a small signal/noise ratio and again requires significant experience and special computer equipment.

Development of tools for the automatization of the setup of MD simulations, which take care of completing missing atoms in crystal structure, of relaxing bad contacts, choosing suitable ionization states for titrable group of proteins, detecting the placement of structural waters and ions,

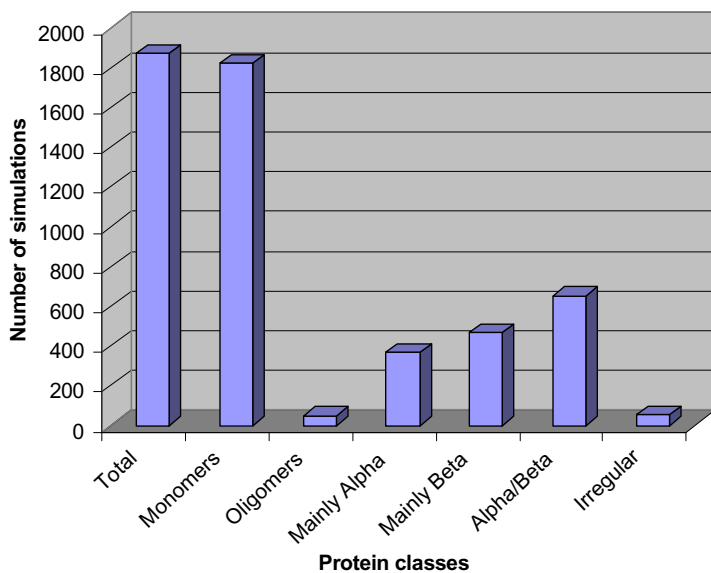


FIG. 2. Total number of protein simulations in 2010 version of MoDEL and distribution across structural classes.

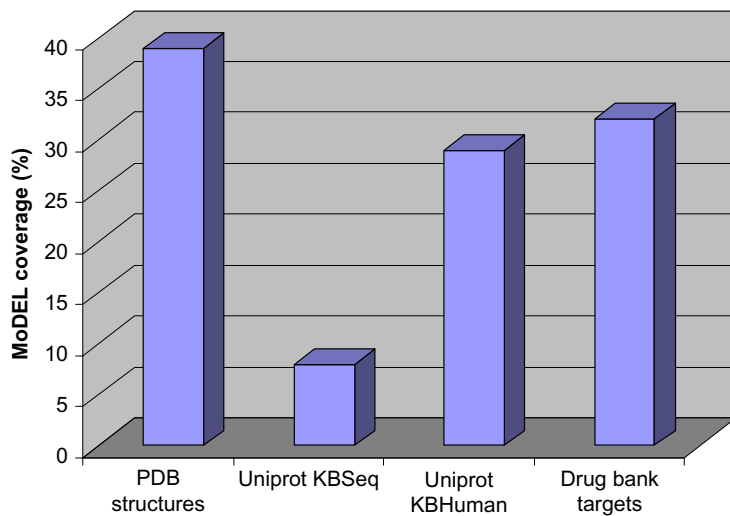


FIG. 3. Coverage (measured from BLAST sequences comparison using a limit  $e$ -value of  $10^{-5}$ ) of 2010 version of MoDEL on different structure and sequence databases.

defining topologies, creating the solvent environment, and performing the thermalization and equilibration, will surely open-up MD to a broader community. Initiatives such as MDWeb go in this direction (<http://mmb.pcb.ub.es/MDWeb>; see Fig. 4). Similar initiatives, but centered in the data mining of trajectories (Camps et al., 2009; see <http://mmb.pcb.ub.es/Model> and <http://mmb.pcb.ub.es/FlexServ>) will be of great help for facilitating trajectory analysis to nonexperts. Finally, many other initiatives, such as the Distributed European Infrastructure for Supercomputing Applications (DEISA; <http://www.deisa.eu>) or Scalalife (<http://www.scalalife.eu>), are now being developed to facilitate the access of MD users to high-performance computers. In parallel, software developers are making a tremendous effort to develop programs able to use parallel architectures (Phillips et al., 2005; Hess et al., 2008) and porting of all these codes to GPU architectures is an ongoing process (Harvey et al., 2009; Voelz et al., 2010). Major advance in the field will come from the use of MD-specific computers (<http://www.deshaw.com/>), which can increase by two to three orders of magnitude the size of the system or the length of the collected trajectory. However, even with all these spectacular technical improvements, MD will remain a technique far too slow and complicated to provide the interactivity that experimental biologists often require. This is the main motivation for the development of approximate coarse-grained models, where, in order to increase computer efficiency, we accept a certain loss of accuracy with a significant reduction in structural resolution. Such a loss of resolution might in fact be beneficial for deriving more intuitive description of many biologically processes (such as large conformational transitions or protein aggregation) occurring in the mesoscopic scale.

## II. COARSE-GRAINED POTENTIALS

The coarse graining of a protein implies always the compression of a series of atoms in a pseudo-particle and a simplification in the representation of the solvent that is: (i) neglected, (ii) simulated as a continuum, or (iii) represented by pseudo-particles which account for clusters of solvent molecules. In all cases, the simplification implies the need to recalibrate the potential function (the force field) or use information-based potentials to describe intraprotein interactions. The most common level of coarse graining for proteins involves the representation of every residue by a single particle located at the  $C_\alpha$ . Refinements of the model that have

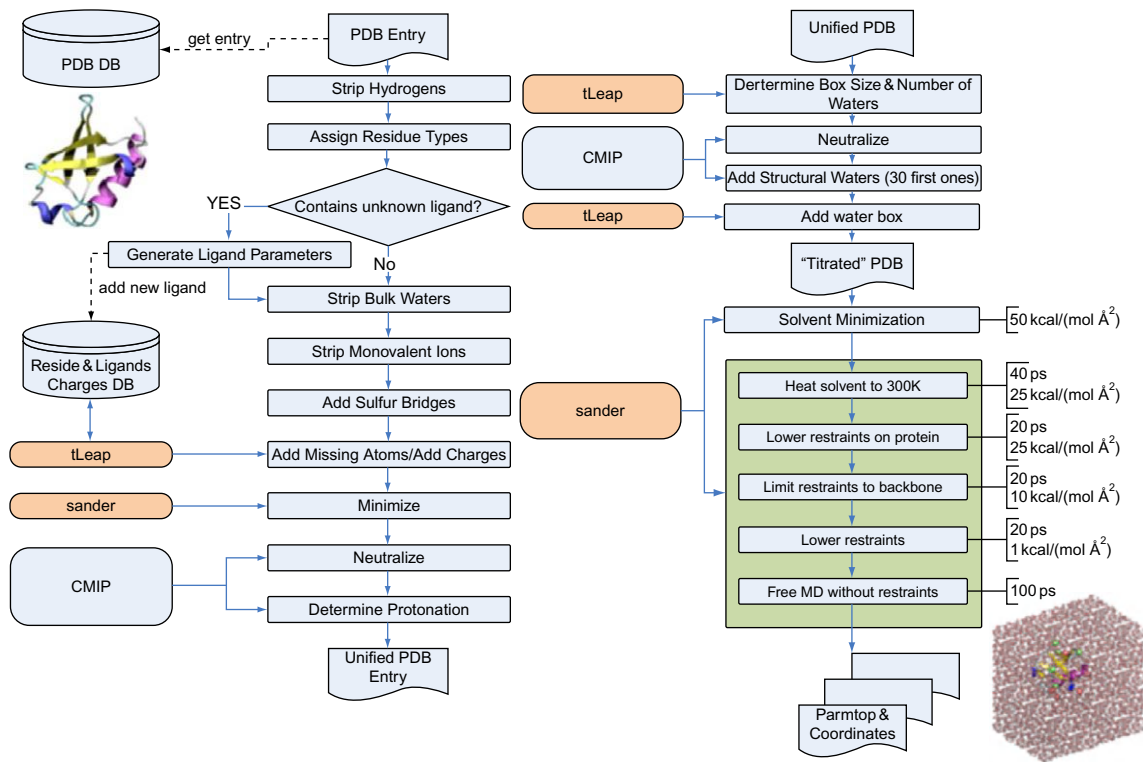


FIG. 4. General workflow of the MDWeb Server (<http://mmb.pcb.ub.es/MDWeb>) for automatic generation of MD trajectories. In the example shown, the workflow will submit an AMBER MD simulation using a PDB entry as input.

been explored by some authors consist of using additional particles to mimic the side chains or some backbone atoms.

### A. $G\bar{o}$ -Like Potentials

Originating from the early works of  $G\bar{o}$  and coworkers (Taketomi et al., 1975), these potentials are the basis of many of the currently used information-based potentials.  $G\bar{o}$  potentials are typically used in conjunction with a  $C_\alpha$  coarse-graining of the protein and consider that any two residues that are in contact in the three-dimensional structure of the protein have a favorable interaction, while if they are not in contact such interaction is none or unfavorable:

$$E = \sum_{i,j} \Delta_{i,j} \varepsilon_{ij} \quad (4)$$

where  $i$  and  $j$  stand for protein particles (typically residue  $C_\alpha$ ),  $\delta_{i,j}$  is a Dirac function which takes value  $-1$  if the two residues are in contact and  $0$  or  $+1$  otherwise, and  $\varepsilon_{ij}$  is a energy constant equal for all pairs (uncolored  $G\bar{o}$  potential;  $\varepsilon_{ij} = \varepsilon$ ) or different (colored  $G\bar{o}$  potentials). Despite its extreme simplicity  $G\bar{o}$  potentials have been quite successfully used to study protein folding and have been crucial in the development of some of today's most accepted theories of folding (see review in Go, 1983). Very recently, these potentials have increased in complexity adopting a formalism, which resembles that of atomistic physical models, like that in Onuchic's functional (Clementi et al., 2000).

$$V = V_{\text{bonded}} + V_{\text{angle}} + V_{\text{dihedral}} + V_{\text{nonbonded}}^{\text{native}} + V_{\text{nonbonded}}^{\text{nonnative}} \quad (5)$$

$$\begin{aligned} V = & \sum_{\text{bonds}} K_r (r - r_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 \\ & + \sum_{\text{dihedrals}} \left\{ K_\phi^{(1)} [1 - \cos(\phi - \phi_0)] + K_\phi^{(3)} [1 - \cos 3(\phi - \phi_0)] \right\} \\ & + \sum_{\text{native}} \varepsilon \left[ 5 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] + \sum_{\text{nonnative}} \varepsilon \left( \frac{\sigma_0}{r_{ij}} \right)^{12} \end{aligned} \quad (6)$$

where in the three first terms,  $r$ ,  $\theta$ , and  $\phi$  are the bond length, angle, and dihedral angle, respectively. The corresponding subscripts "0" stand for values in the experimental structure. The fourth term corresponds to the Lennard-Jones-like (LJ) stabilization energy represented as a 12-10

function that acts on only those contacts present in the native state. Here,  $r_{ij}$  and  $\sigma_{ij}$  identify the distance between atoms  $i$  and  $j$  in one snapshot and in the native state, respectively ( $r_{ij} = |\vec{r}_{ij}|$  and  $\sigma_{ij} = |\vec{r}_{ij}^0|$ ). The fifth term is an excluded volume function that energetically disfavors any close nonnative contact ( $\sigma_0 = 4 \text{ \AA}$ ). In typical implementations of this model, native contacts are identified with a  $5\text{-\AA}$  heavy-atom cutoff excluding up to  $i-i+3$  sequential neighbors. Such a contact calculation preserves the number of atomic contacts per residue that depends on the size of the amino acid. Nearest-neighbor energy terms are usually defined by:  $K_r = 100\varepsilon$ ,  $K_\theta = 20\varepsilon$ ,  $K_\phi^{(1)} = \varepsilon$ , and  $K_\phi^{(3)} = 0.5\varepsilon$ , where  $\varepsilon$  sets the energy scale.

Onuchic’s Gō-like potentials coupled, for example, to Langevin dynamics sampling algorithms (see below) are being extensively used to analyze experimental biophysical measures on protein folding and unfolding (Clementi et al., 2000, 2003).

### B. Harmonic Potentials

They can be understood as an evolution of Gō-like potentials, as they penalize the deviation on native inter-residue distances. These potentials have become very popular for the study of the “near-equilibrium” dynamics of proteins when implemented in sampling techniques derived from normal mode analysis (NMA). The basic assumption when using these potentials is that a protein behaves as an elastic network model (ENM; Tirion, 1996; Atilgan et al., 2001), where usually  $C_\alpha$ s act as network nodes which are connected by harmonic springs. Note that the number of springs runs with the number of residues in the protein ( $N$ ), as  $(N-1)!$ , and accordingly, direct application of ENM will result in an artifactual over-rigidification of the protein as the protein size is increased. This problem can be corrected by using, for example, a distance-dependent cutoff that annihilates the interactions between remote residues, leading to an energy functional as that developed in Eqs. (7)–(10), where the energy ( $E$ ) to distort a protein from its equilibrium conformation ( $r_{ij}^0$ ), considered a energy minimum, is given by the pairwise Hookean potential (Tirion, 1996):

$$E = \sum_{i \neq j} K_{ij} (r_{ij} - r_{ij}^0)^2 \quad (7)$$

where  $r_{ij}$  stands now for the distance between residues  $i$  and  $j$  (represented by the corresponding  $C_\alpha$  in the protein configuration), and  $K_{ij}$  stands for



the spring constant. The force of the spring restricting the motion of the  $ij$  residue pair is computed as:

$$K_{ij} = \frac{1}{2} \kappa \Gamma_{ij} \quad (8)$$

$\kappa$  being a phenomenological constant (in energy/distance<sup>2</sup> units) and  $\Gamma$  being a Kirchhoff topology matrix of inter-residue contacts, where  $ij$ th element for  $i \neq j = 1, \dots, N$ , is equal to 1 if residues  $i$  and  $j$  are within the cutoff distance  $r_c$ , or zero otherwise:

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } r_{ij} \leq r_c \\ 0 & \text{if } r_{ij} > r_c \end{cases} \quad (9)$$

The diagonal elements ( $ii$ th) are equal to the coordination number or residue connectivity taken as:

$$\Gamma_{ii} = - \sum_{j|j \neq i}^M \Gamma_{ij} \quad (10)$$

Despite their simplicity, functionals as those shown in Eqs. (7)–(10) are able to provide quite accurate representation of the near-equilibrium dynamics of many proteins but are extremely dependent on the selected cutoff for remote interactions, which can have different optimal values for each protein (often difficult to predict *a priori* (Sen and Jernigan, 2006)). This led to the derivation of new methods where the discrete Hamiltonian is replaced by continuous functions, typically dependent on the inverse exponential of the inter-residue distance ( $r_{ij} = |\vec{r}_{ij}|$ ). Thus, Hinsen et al. (2000) derived a function for the spring strength by fitting to a local minimum from a single MD simulation. This procedure leads to a force constant definition with stronger couplings for neighbors along the backbone, and a sixth power of distance for the rest of the interactions. The distinction of short- and long-range terms was dependent on a short cutoff, and the formulation also included a protein-fitted scaling factor for the global energetics, which limits its general applicability. Kovacs et al. (2004) proposed a simpler sixth-power exponential, which does not require any cutoff and has become very popular in current elastic network implementations:

$$K_{ij} = C \left( \frac{r^0}{r_{ij}} \right)^6 \quad (11)$$

where the proportionality constant  $C$  (usually taken as  $40 \text{ kcal}/(\text{mol } \text{Å}^2)$ ) controls the global rigidity of protein contacts, and  $r^0$  is normally taken as  $3.8 \text{ Å}$ , which is approximately the mean  $C_\alpha$ - $C_\alpha$  distance between any pair of consecutive residues.

Different authors have tried to improve harmonic potentials by, for example, defining rigid blocks (Tama et al., 2000), by scaling differently covalent and non-covalent contacts (Kondrashov et al., 2006), by adding short-range terms (Moritsugu and Smith, 2007), or by defining distinguished chain interactions by a bond-cutoff (Jeong et al., 2006). Following these directions, we have recently (Orellana et al., 2010) developed a hybrid approach calibrated using a large database of atomistic MD trajectories of representative proteins (Rueda et al., 2007a; Meyer et al., 2010). The analysis of MD simulations showed that the topology of nearest-neighbor interactions, the basis of the secondary structure, is the main component in the large motions traced by ENM (Rueda et al., 2007a; Meyer et al., 2010). Accordingly, the method (named essential-dynamics elastic network model, ed-ENM; Orellana et al., 2010) treats differentially the sequential and nonsequential (“Cartesian”) contacts. For the first  $M$  sequential contacts, a fully connected matrix is used, while Cartesian contacts are treated using a continuum distance-dependent function with a calibrated size-dependent cutoff, which helps to remove artifactual long-range interactions. Therefore, the elements of the topology matrix are defined as:

$$\Gamma_{ij} \begin{cases} \text{if } S_{ij} \leq M, = -1 \\ \text{otherwise} \begin{cases} = -1 & \text{if } r_{ij} \leq r_c \\ = 0 & \text{otherwise} \end{cases} \end{cases} \quad (12)$$

and the matrix  $\Gamma$  has always  $2M+1$  nonzero-diagonal entries defining neighbor chained contacts. Accordingly, the force constants  $K_{ij}$  are dependent, not only on the Cartesian but also on the sequential distance:

$$K_{ij} \begin{cases} = C^{\text{seq}} / S_{ij}^{n^s}, \text{ if } S_{ij} \leq M \\ \text{otherwise} \begin{cases} = \left( C^{\text{cart}} / r_{ij} \right)^{n^c} \\ = 0 & \text{otherwise} \end{cases} \end{cases} \quad (13)$$

where values for all terms ( $n^s=2$  and  $C^{\text{seq}}=60 \text{ kcal}/(\text{mol } \text{Å}^2)$ ;  $n^c=6$  and  $C^{\text{cart}}=6 \text{ kcal}/(\text{mol } \text{Å}^2)$ , in energy units) were obtained by fitting to apparent force constants and structural variance profiles obtained in a large

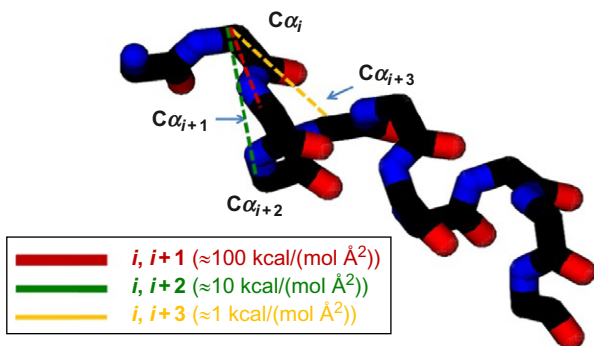


FIG. 5. Formulation of the ed-ENM model. The ed-ENM is a nearest-neighbor-based model, maintaining the secondary structure stereochemistry, where the three first-order constants acquire values close to a 100:10:1 ratio.

number of atomistic MD simulations. A value of  $M=3$  was used for sequential interactions based on MD simulations, which were also instrumental to define the cutoff radii ( $r_c$ ), which is computed using an empirical logarithmic relationship with the size of the protein. This formalism guarantees sequential contacts which decay quickly with the number of connecting bonds (see Fig. 5) and a continuum decay of the strength of Cartesian contacts up to a cutoff. Attempts to improve the formalism by adding “color” to the topological relations, that is, different spring constants for different physical interactions, or by adding differential weights to different secondary elements, did not yield to clear improvements in the results (Orellana et al., 2010).

The hybrid ENM outlined above can work coupled to any sampling technique (see below) and provides quite accurate representations of the near-equilibrium dynamics properties of proteins at both the global (essential dynamics, global variance) and local levels ( $B$ -factor distribution), representing a significant improvement with respect to simpler schemes (see Fig. 6).

### C. Flat Potentials

In recent years, the use of discontinuous flat potentials (also named stepwise potentials) has gained popularity due to its use in discrete MD sampling algorithms (Zhou and Karplus, 1999; Ding et al., 2005;

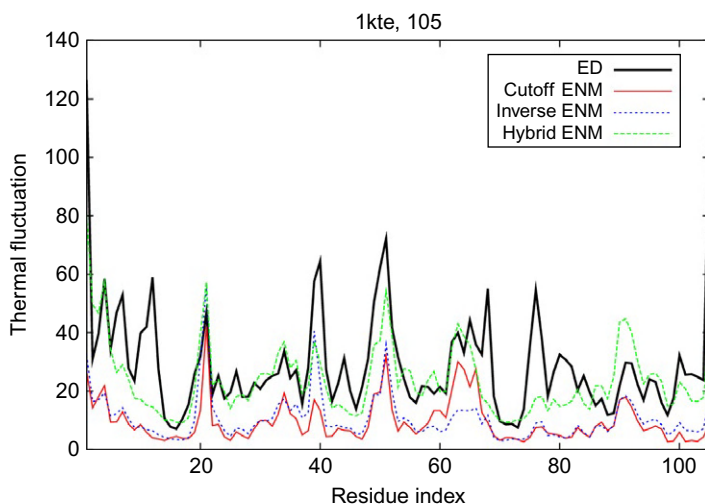


FIG. 6. Thermal fluctuations ( $B$ -factors) for protein 1 kte obtained from essential dynamics treatment of atomistic molecular dynamics simulations (ED) and from ENM models incorporating a simple cutoff scheme, the inverse exponential decay function proposed by Kovacs and the hybrid method developed in our group. The good qualitative value of all ENM calculations is evident, as is clear the better performance of our hybrid approach.

Emperador et al., 2008a). These potentials are based on the idea that for coarse-grained calculations, the continuum physical potentials can be approximated as a series of discontinuous potentials defined by square wells. The simplest flat square potential describes atoms as hard spheres undergoing hardcore collisions, which are then defined by an interaction potential with an infinite step at the distance corresponding to the sum of the hard sphere radii of two particles (Fig. 7A). Stepwise potentials can also be used to represent strong interactions between particles, which are described as square-well potentials with infinite walls and well amplitude taken from a typical particle–particle distance vibration at room temperature (Fig. 7B). This type of potentials can be easily implemented in Gō-like strategies using then the experimental inter-residue distance to define the center of the well (Emperador et al., 2008a).

Flat potentials can be also adapted to work within the physical or pseudo-physical strategies (see Section II.D and also Emperador et al., 2008b, 2010). In this case, nonbonded terms like Coulomb and Van der

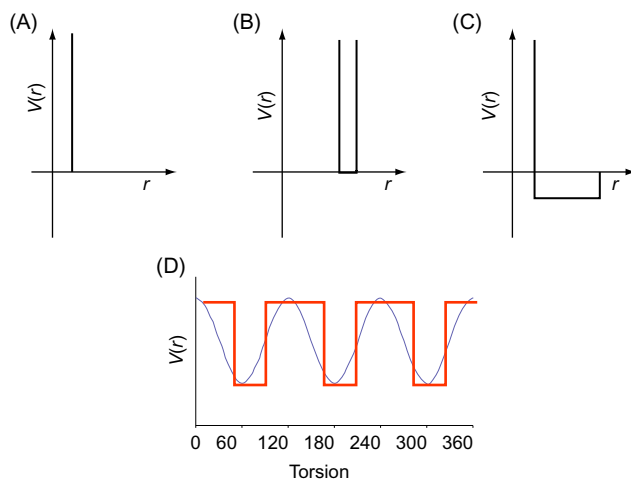


FIG. 7. Examples of flat potentials that can be used to approach physical potentials (see text for discussion).

Waals interactions can be also represented with square-well potentials with one or several steps depending on the accuracy required. For example, potential function shown in Fig. 7C can be used to represent a Van der Waals or an electrostatic interaction between two particles with opposite charge neglecting long-range effects. Introduction of additional wells can help in the representation of long-range attractive effects. Stretching and bending terms are easily approximated by a single square well centered at equilibrium values (Fig. 7C). Torsional terms needed to reproduce one to four particle interactions can be introduced by simply fitting Fourier expansions to discrete square-well potentials, using one to four distances as rotation variable and adjusting well barriers to potential torsional terms (see Fig. 7D).

Flat potentials have the advantage of allowing the treatment of trajectories within the ballistic regime which guarantees a good computational efficiency (see below). Besides, they can be used to reproduce not only near-equilibrium dynamics of proteins (as ENM-NMA) but also local rearrangements like those happening during protein-protein interactions, very large transitions, or even protein folding events (Ding and Dokholyan, 2008). Additionally, the functional allows a very simple implementation in mixed calculations with different degree of granularity in the representation of the protein.

### D. Physical and Pseudo-physical Potentials

There are a large variety of coarse-grained potentials that try to maintain a physical foundation while reducing the degrees of freedom of the system and accordingly increasing computational efficiency. One of the most popular ones is the MARTINI force field developed by Marrink and coworkers (Marrink et al., 2007, 2009). The force field, very popular in membrane simulation, has also an elegant protein implementation, which follows a four-to-one mapping, that is four heavy atoms are represented by a single bead, which can be annotated in four types: polar, nonpolar, apolar, and charged. Within each type, there are subclassifications based on the hydrogen bond donor/acceptor capabilities, or by the polarity. An interesting aspect of this force field is that water is explicitly included using also the four-to-one mapping strategy (i.e., one bead represents four water molecules). The interactions between beads are represented as the addition of “bonded” and “nonbonded” terms:

$$V = V_{\text{bonded}} + V_{\text{nonbonded}} \quad (14)$$

Bonded terms ( $V_{\text{bonded}}$ ) take care of keeping the covalent structure of the protein, maintaining the chirality of the protein and the secondary structure (note that MARTINI is really a pseudo-physical force field since require previous structural knowledge on the target protein).

$$V_{\text{bonded}} = \sum_{\text{bonds}} K_b (r_{ij} - r_{ij}^0)^2 + \sum_{\text{angles}} K_a (\cos\varphi_{ijk} - \cos\varphi_{ijk}^0)^2 + \sum_{\text{dihedral}} K_d \left[ 1 + \cos(\theta_{ijkl} - \theta_{ijkl}^0) \right] + \sum_{\text{imp dihedral}} K_{\text{id}} (\theta_{ijkl} - \theta_{ijkl}^0)^2 \quad (15)$$

where  $i, j, k$ , and  $l$  are “beads”;  $r_{ij}$  stands for the bond distance between atoms  $i$  and  $j$ ;  $\varphi_{ijk}$  stands for the bond angle ( $i-j-k$ ); and  $\theta_{ijkl}$  represents the dihedral ( $i-j-k-l$ ). The equilibrium values are always denoted by superindex “0.”

Nonbonded terms account for interactions between non-neighboring beads, which are modeled by means of a Lennard–Jones-like potential and a screened Coulombic term applied only to interactions between charged particles.

$$V_{\text{nonbonded}} = 4 \sum_{i,j} \varepsilon_{ij}^* \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \right] + \sum_{i,j}^{\text{charged}} \frac{q_i q_j}{\varepsilon r_{ij}} \quad (16)$$

where  $i$  and  $j$  stand here for nonbonded beads separated by a distance  $r_{ij}$ , and  $\varepsilon_{ij}^*$  and  $\sigma_{ij}$  are the standard Lennard–Jones parameters for

noninteracting particles. The Coulombic term (where  $q$  stands for bead charges) extends only for charged pairs and is screened by a relative dielectric constant ( $\epsilon=15$ ) and by a shifting procedure which reduces slowly interactions to zero for distant interactions. All the bonded and nonbonded terms appearing in Eqs. (15) and (16) have been carefully parametrized to reproduce experimental data.

The MARTINI force field was created to be implemented easily in the same MD simulation algorithms that are used for atomistic simulations, particularly GROMACS simulation package (Hess et al., 2008). Since beads weigh typically from 48 to 72 amu, intra-bead vibrations are slow. This allows the use of large integration steps (up to 40 fs), which combined with the drastic reduction in the degrees of freedom in the system leads to a dramatic increase in the performance of the MD simulation. This explains the popularity of MARTINI force field, especially in the study of membranes (Marrink et al., 2009).

Related force fields, which try to maintain potential functions similar to the “all-atoms” ones, have been developed by different groups. For example, Schulten and coworkers (Shih et al., 2006) have developed a force field where each residue is represented by two beads, one representing backbone and the other (different for each residue) the side chain. The potential energy function is represented by a CHARMM-like (McKerell et al., 1995) potential (similar to that in Eqs. (15) and (16)), where the different terms are calibrated to reproduce atomistic MD results on the same system (Shih et al., 2006). The same group has developed much more aggressive coarse-graining approaches, where beads (representing in some cases hundreds of atoms) are not centered in real residues but are spread around the protein to get an as accurate as possible reproduction of the protein shape (Arkhipov et al., 2006a). Again, effective potentials are fitted to reproduce the behavior found in atomistic MD simulations with the same system (Arkhipov et al., 2006a,b). These force fields have provided very interesting results in the study of very large viruses or in the analysis of cooperative effects of multiple protein aggregates (Arkhipov et al., 2006a,b; Shih et al., 2006). As with MARTINI, Schulten potentials are created to facilitate implementation in atomistic MD simulation codes, particularly NAMD (Phillips et al., 2005), using Langevin’s dynamics and continuum representation of solvent.

Scheraga and coworkers developed another coarse-grained model based on pseudo-physical potentials (Liwo et al., 2005). The method, which has

been implemented into a variety of sampling techniques, defines proteins as a combination of two types of beads, one to represent the center of the peptide bonds and the other, of different sizes, to represent the side chains. The corresponding force field (named UNRES) contains up to 10 terms accounting for different types of backbone and side-chain interactions, including coupling terms which are commonly ignored in other force fields. The different terms are parametrized using statistical data on folded structures from PDB and quantum mechanical and atomistic dynamics simulations of protein fragments (Liwo et al., 2005, 2007). The UNRES force field has provided quite encouraging results in the prediction of protein structure and in the representation of protein folding (Liwo et al., 2005, 2007; Khalili et al., 2006).

Sorensen and Head-Gordon (2002) developed another one-bead model, similar in structure to the MARTINI one, where each residue is represented by a  $C_\alpha$  centered bead and potential energy is determined as:

$$\begin{aligned}
 V_{\text{bonded}} = & \sum_{\text{bonds}} K_b \left( r_{ij} - r_{ij}^0 \right)^2 + \sum_{\text{angles}} K_a \left( -\varphi_{ijk} - \varphi_{ijk}^0 \right)^2 + \\
 & \sum_{\text{dihedral}} K_{d1} [1 + \cos\theta] + K_{d2} [1 - \cos\theta] + K_{d3} [1 + \cos 3\theta] + \\
 & K_{d4} [1 + \cos(\theta + \pi/4)] + \sum_{i,j} 4\epsilon_{ij}^* \left[ S_{ij}^1 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - S_{ij}^2 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]
 \end{aligned} \tag{17}$$

where the meaning and the purpose of the different terms are similar to that in MARTINI force field (see above) with the exception that the two scaling coefficients  $S^1$  and  $S^2$  increase the flexibility of the pseudo Lennard-Jones term to account for different type of nonbonded residue-residue interactions. As in many other models, solvent is not treated explicitly, and the parametrization of the constants in Eq. (17) comes from the statistical analysis of known proteins and/or from atomistic simulations. These potentials are typically run in conjunction with Langevin MD simulation protocols and a continuum representation of solvent (Sorensen and Head-Gordon, 2002).

Coarse-grained potentials outlined here are just a small fraction of the universe of continuum potentials which try to maintain a physical foundation but in reality introduce knowledge-based terms (e.g., in the torsional term, or in the parametrization of the nonbonded terms). As noted, discontinuous flat potentials designed to reproduce physical potentials have also



been used in the context of discrete MD simulations to study different aspects of protein structure and interactions (see previous paragraph).

### III. SAMPLING TECHNIQUES

Irrespective of the nature of the Hamiltonian used to reproduce the dependence of the energy on the protein conformation, the study of flexibility requires the use of sampling techniques, which evaluate the degree of structural variation predicted for the different protein particles at working temperature.

#### A. Normal Mode Analysis

It is possible to determine analytically the expected sampling of a single particle moving in one dimension, subject to very simple potential functions like the Hookean spring, since Newton's equation reads as:

$$M \frac{\partial^2 \vec{x}(t)}{\partial t^2} = -K \vec{x}(t) \quad (18)$$

where  $M$  is the mass of the particle,  $x$  stands for the displacement of the particle from equilibrium spring length, and  $K$  being the spring constant. General solution for this type of equations has the exponential form:  $x(t) = Ae^{i\omega t}$  ( $A$  being the constant and  $\omega$  the frequency of the oscillator), which means that after some algebra Newton's equation can be rewritten as:

$$\omega^2 MA = AK \quad (19)$$

and the sampling distribution along time becomes defined simply by:

$$\vec{x}(t) = A \cos \left( t \sqrt{\frac{K}{M}} \right) \quad (20)$$

For a system of  $N$  particles in Cartesian space subject to a series of connected springs, Eqs. (18)–(20) take matrix form, for example, Eq. (18) reads as:

$$M \ddot{\vec{R}}(t) + H \vec{R}(t) = 0 \quad (21)$$

where  $H$  is the Hessian matrix, defined as the matrix of second derivatives of the energy with respect to the coordinates, and the vector  $R(t)$  stands

for the particle coordinates at time  $t$ ; note that for compactness we have moved here to the compact notation for time derivative. This equation is solved by diagonalization of the mass-weighted Hessian matrix which yields a series of eigenvectors ( $v$ ) and a series of eigenvalues (the frequencies,  $\omega$ ), which define the normal modes, that is the movements expected for the different particles as:  $\vec{R}(t) = A \vec{v} \cos(\omega t)$ .

In summary, starting from a harmonic potential, NMA yields a series of eigenvectors and eigenvalues obtained by diagonalization of the Hessian matrix. The eigenvectors are a lineal combination of atomic movements, which indicate global movement of the proteins (the essential deformation modes), while the associated eigenvalues indicate the expected displacement along each eigenvector in frequencies (or distance units if the Hessian is not mass-weighted), that is, the impact of each deformation movement in the total protein motion. Thus, protein dynamics is represented as a simple combination of vibrations along the set of eigenvectors. The lowest frequency eigenvectors tend to represent more collective, largest amplitude motions, and can trace functional rearrangements and transitions (Sorensen and Head-Gordon, 2002; Moritsugu and Smith, 2007).

NMA can be applied in conjunction with any continuum and differentiable potential energy function. This is done by assuming that protein conformation is at one stationary state  $R$  and that deformations are going to be small and Gaussian, making it possible to expand any potential function as a Taylor series (with  $i$  and  $j$  being particles):

$$V(R) = V(R^0) + \sum_i \left( \frac{\partial V}{\partial r_i} \right)_0 (r_i - r_i^0) + \frac{1}{2} \sum_{i,j} \left( \frac{\partial^2 V}{\partial r_i \partial r_j} \right)_0 (r_i - r_i^0) (r_j - r_j^0) + \dots \quad (22)$$

where for the sake of simplicity we have skipped here the vector notation. Note that for a stationary point, assuming the reference value  $V(R^0)$  as zero and neglecting higher-order terms, Eq. (22) leads to a harmonic expression, which enters nicely into the NMA framework:

$$V(R) = \frac{1}{2} \sum_{i,j} k_{ij} (r_i - r_i^0) (r_j - r_j^0) \quad (23)$$

where  $k_{ij}$  stands for the elements of the second-derivative matrix (the Hessian matrix).

The practical use of NMA with all atom nonharmonic potentials presents three major problems: (i) the complex behavior of solvent, which cannot fit into the model implicitly assumed in Eq. (22), (ii) the need to be in an energy minimum (otherwise some frequencies will be imaginary), and (iii) the cost of evaluating the Hessian for complex potential functions. All these problems have hampered the use of NMA coupled to atomistic or even coarse-grained physical potentials, favoring its application in conjunction with ENMs (see above), where reference structure is by definition a minimum, water is ignored and force field is by definition fully harmonic.

ENM–NMA methods have been used quite extensively for the description of large deformation movements in proteins (Tirion, 1996; Hinsen et al., 2000; Tama et al., 2000; Atilgan et al., 2001; Krebs et al., 2002; Kovacs et al., 2004; Jeong et al., 2006; Kondrashov et al., 2006; Sen and Jernigan, 2006; Moritsugu and Smith, 2007; Orellana et al., 2010). If accurate ENM potentials are used, the method is able to reproduce with reasonable accuracy experimental *B*-factor distributions, and the pattern of flexibility detected in NMR experiments (Abseher et al., 1999; Yang et al., 2007) and also in atomistic MD simulations (Rueda et al., 2007b; Orellana et al., 2010). Using these techniques, different authors have proven that biologically relevant movements (i.e., those essential for protein function) are very often correlated with the most relevant deformation modes (i.e., those leading to large collective movements in the proteins). Methods to trace transitions between conformational states of the proteins based on displacements along the essential deformation modes have been developed and have been largely used to obtain first mechanical models of large conformational transitions (Krebs et al., 2002; Kong et al., 2006; Moritsugu and Smith, 2007; Dobbins et al., 2008; Rueda et al., 2009). Approaches based on ENM–NMA methods have been developed to improve fitting of protein structures to electron-density maps, especially those derived from electron microscopy (Phillips, 2006). Related methods have been derived to introduce flexibility in protein docking (Rueda et al., 2009). Recent efforts in the area are focussing on improving the formalism of ENM, in relaxing the definition of the reference structure and in performing the NMA using internal rather than Cartesian coordinates (Ma, 2009; Mendez and Bastolla, 2010), which reduces the cost of the calculation and guarantees that structures generated by considering activation of a reduced number of deformation modes maintain chemical sense.

### B. Monte Carlo

In the ergodic regime, protein flexibility can be simulated as a Markov chain of movements. According to the popular Metropolis algorithm the sampling is obtained by perturbing randomly an initial configuration of the protein ( $\vec{R}_0$ ) to generate a trial configuration ( $\vec{R}'_0$ ). This new configuration will be accepted, and accordingly considering part of the sampling (i.e.,  $\vec{R}_1 = \vec{R}'_0$ ) if its energy (as determined by force field) is smaller than that of the initial configuration (i.e.,  $E(\vec{R}'_0) \leq E(\vec{R}_0)$ ); otherwise, the probability of acceptance depends of a Boltzmann probability function which for a given temperature makes more likely the acceptance of ( $\vec{R}'_0$  as  $E(\vec{R}'_0)$  approaches to  $E(\vec{R}_0)$ ). When the trial movement is not accepted, the new configuration is considered equal to the previous one (i.e.,  $\vec{R}_1 = \vec{R}_0$ ). The process is then repeated millions of times to guarantee proper sampling of all the degrees of freedom. A flowchart of metropolis Monte Carlo (mMC) simulation in chemical systems can be then summarized as shown in Fig. 8, where we present the algorithm for a general configurational variable  $X$ , in our case, where we limit to consider only conformational movement ( $R=X$ ).

In order to improve computational efficiency, the practical use of mMC implies the generation of the trial configuration created by perturbing a previously accepted conformation. The magnitude of the perturbation is adjusted in trial calculations in such a way that the average acceptance rate will be around 40–50% (average acceptance rates out of this range lead to lower sampling efficiency). Monte Carlo methods are typically used in conjunction with representation of the systems by means of internal coordinates, focusing sampling in the torsional degrees of freedom. This strategy is however complex for its application to proteins and in general to any polymer, since small changes in a backbone torsion can introduce dramatic changes in the coordinates of distant residues and accordingly large steric clashes which would lead to a very low acceptance rate, and, accordingly, to a very poor sampling efficiency. This problem has been solved by introducing sampling strategies which couple several torsions to avoid large Cartesian movements in terminal residues (Umschneider and Jorgensen, 2003). Current computer programs for mMC simulations in proteins, like MCPRO (Jorgensen and Tirado-Rives, 2005), work with atomistic physical potentials but can easily be adapted to work with a variety of coarse-grained pseudo-physical potentials.

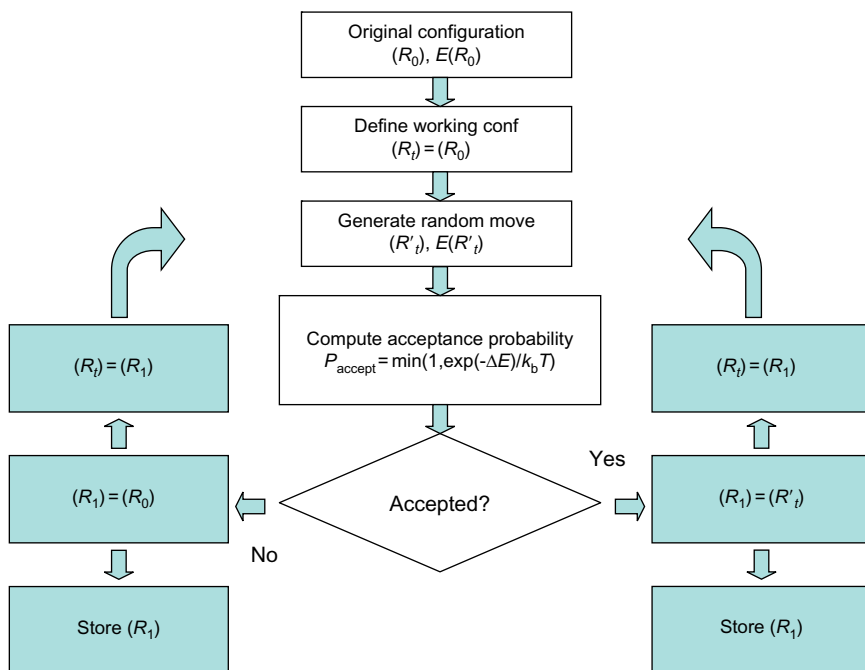


FIG. 8. Flowchart of the metropolis Monte Carlo sampling algorithm.

Monte Carlo simulations have the advantage with respect to NMA that no harmonicity is assumed for the derivation of sampling, allowing then the detection of large anharmonic conformational transitions. Unfortunately, this generality is gained at the expense of a much larger computational effort. Compared with sampling techniques in Cartesian space, such as molecular or Langevin dynamics, mMC has the advantage that the user can select the important degrees of freedom, focusing then the sampling effort in relevant coordinates. Unfortunately, the later advantage can become a problem in cases where selecting *a priori* the important degrees of freedom is not obvious. An additional shortcoming of MC that cannot be ignored is the loss of the time coordinate in the simulation, something which represents a major problem in nonequilibrium simulations.

Recently, strategies to couple Monte Carlo with ENM-NMA have been suggested (Rueda et al., 2007b), the idea is not to sample directly the residue movements on the ENM potential, but to activate the movements

of the proteins as displacements along the essential deformation modes. Trajectories of the particles can then be obtained using a back projection to Cartesian space. According to this procedure, a set of modes are selected and movements along them are randomly made and its associated energy is computed using Einstein's equation for harmonic oscillator:

$$V(\Delta \vec{R}) = \frac{1}{2} \sum_{i=1}^M \frac{k_B T}{\lambda_i} (\Delta \vec{R}_i)^2 \quad (24)$$

where the  $\Delta \vec{R}_i$ , multidimensional vector, represents the displacement along principal deformation mode  $i$ ,  $\lambda$  stands for eigenvectors (in distance<sup>2</sup> units) obtained by diagonalization of the nonmass-weighted Hessian matrix,  $k_B$  is Boltzmann constant, and  $T$  is the absolute temperature. The sum is typically extended to a limited set of essential deformation modes.

Equation (24) can be generalized by adding perturbational terms in any coordinate space, allowing then to escape from pure harmonic representation, or can be modified as to give higher weight to essential deformation modes that, for example, overlap better to a given transition vector or that correlate with a higher-order motion in a given region of the protein. The result is that we can bias the sampling to guarantee better representation of biologically relevant degrees of freedom.

### C. Langevin Dynamics

Just 3 years after the publication of Einstein's description of Brownian motion, Paul Langevin modeled the continual movement of particles suspended in a fluid with Newton's second law. He considered that the Brownian motion of a particle (of mass  $m$ ) in a fluid is due to the molecular-thermal agitation of the surrounding solvent (which lead to random collisions on the particle,  $\vec{\xi}$ ) and a dispersive force accounting for the viscous resistance the particle feels on going through the fluid ( $-\gamma \vec{v}$ ) at velocity  $\vec{v}$ . Mathematically, it can be set as:

$$m_i \frac{d\vec{v}_i}{dt} = m_i \dot{\vec{v}}_i = -\gamma \vec{v}_i + \vec{\xi}_i(t) \quad (25)$$

where for the sake of simplicity, the random force is supposed to satisfy two conditions (Kubo, 1959, 1965; Lemons, 1977): (i) the stochastic

process  $\vec{\xi}(t)$  is Gaussian with zero mean, and (ii) its autocorrelation function has the form

$$\left\langle \vec{\xi}_i(t) \vec{\xi}_j(t') \right\rangle = \sigma^2 \Delta_{ij} \Delta(t - t') \quad (26)$$

where  $\sigma^2$  is the standard deviation (also named noise intensity) associated with the Gaussian process  $\xi(t)$  whose expression is defined below (Eq. (29)),  $\delta_{ij}$  is the Kronecker's delta and  $\delta(t - t')$  is the Dirac's delta.

The considerations made above about the forces acting on the Brownian particle inspired physicists and biologists to apply them in the resolution of the equation of motion of proteins in the case that an explicit environment is substituted by a continuum media, which behaves in a stochastic way. Within the Langevin dynamics approach, the equations of motion of a protein become defined as:

$$m \dot{\vec{v}}_i = \vec{F}_i - \gamma \vec{v}_i + \vec{\xi}_i(t) \quad (27)$$

where for sake of simplicity we have used compact notation for time derivatives. The force  $\vec{F}_i$  acting on the different protein particles comes typically from a potential energy  $V(\vec{r})$ ,

$$\vec{F}_i = - \frac{\partial V(\vec{r})}{\partial \vec{r}_i} \quad (28)$$

Note that at the limit of high friction and low forces, Eq. (27) converges to Brownian's equations of motion, while in the absence of collisions with the continuum solvent, it converges to Newton's equations of motion. Note also that besides of representing the solvent, friction and noise terms play together to create a natural thermostat for the system. Thus, the random energy shots given by the noise term are balanced by the dissipative force given by friction, keeping then constant the temperature. Note that friction and stochastic collision dissipation terms are related by the so-called fluctuation–dissipation relation:

$$\sigma^2 = 2m_i k_B T \gamma \quad (29)$$

which means that the standard deviation of noise can be expressed as a function of the mass of the particle, the temperature of the thermal bath, and the factor of the dissipation force. If more particles have to be considered, one has to use a different noise term with such a different standard deviation with the corresponding particle mass.

Formulation of Eq. (27) in integral form largely facilitates the posterior numerical analysis. For this purpose, we should first define a characteristic time  $\tau = m\gamma^{-1}$  accounting for the loss of energy due to the dissipation term. Dividing all terms in Eq. (27) by  $\gamma$  and using the identity:

$$\tau \vec{v}_i + \vec{v}_i = \tau e^{-t/\tau} \frac{d}{dt} \left( e^{t/\tau} \vec{v}_i \right) \quad (30)$$

we obtain:

$$\frac{d}{dt} \left( e^{t/\tau} \vec{v}_i \right) = m_i^{-1} e^{t/\tau} \left( \vec{F}_i + \vec{\zeta}_i \right) \quad (31)$$

whose integration leads to the following integral expression for the velocity at time  $\Delta t$ :

$$\vec{v}_i = e^{\Delta t/\tau} \vec{v}_i^0 + m^{-1} e^{-\Delta t/\tau} \left[ \int_0^{\Delta t} \vec{F}_i e^{t'/\tau} dt' + \int_0^{\Delta t} \vec{\zeta}_i(t) e^{t'/\tau} dt \right] \quad (32)$$

Considering an integration time step  $\Delta t$  small enough as to assume that the force has a constant value  $F_i^0$  during the integration, we can rewrite Eq. (32) as:

$$\vec{v}_i = e^{-\Delta t/\tau} \vec{v}_i^0 + \gamma^{-1} \left( 1 - e^{-\Delta t/\tau} \right) \vec{F}_i^0 + m^{-1} \int_0^{\Delta t} \vec{\zeta}_i(t) e^{(t'-\Delta t)/\tau} dt' \quad (33)$$

Note that we are forced to maintain the explicit integral for the noise function (last term in Eq. (33)), since unlike the force, it is not a smooth and deterministic function, and, accordingly, we cannot assume that it takes a constant value at integration step. The updated position can be obtained by integrating both sides with respect to time to provide the new position:

$$\begin{aligned} \vec{r}_i &= \vec{r}_i^0 + \tau \left( 1 - e^{-\Delta t/\tau} \right) \vec{v}_i^0 + \frac{\Delta t}{\lambda} \left( 1 - \frac{\tau}{\Delta t} \left( 1 - e^{-\Delta t/\tau} \right) \right) \vec{F}_i^0 \\ &\quad + \gamma^{-1} m^{-1} \int_0^{\Delta t} \left( 1 - e^{(t'-\Delta t)/\tau} \right) \vec{\zeta}_i dt' \end{aligned} \quad (34)$$

where again we cannot skip the explicit integral formalism for the noise contribution to the new position. Fortunately, these stochastic contributions to the updating of positions and velocities have the shape of the so-called stochastic integral (Van Kampen, 1981; Gardiner, 1989):

$$\int_0^{\Delta t} G(t) dW(t) \approx \sigma G(0) \vec{u}(0) \Delta t^{1/2} + O(\Delta t^{3/2}) \quad (35)$$



where  $G(t)$  is an arbitrary analytical function and  $dW(t)$  corresponds to a differential Wiener process ( $dW(t) = \vec{\xi}_i dt$ ). Assuming a small time step, we can approximate the integral up to leading order in  $dW(t)$ , where  $u(0)$  is a Wiener process of variance equal to 1.

Implementation of Langevin equations with different types of pseudo-physical coarse graining is straightforward using in most cases standard highly optimized atomistic MD codes, or other more specifically developed codes in the case of simpler harmonic potentials (Emperador et al., 2008a; Camps et al., 2009). In these cases, Langevin dynamics offers the advantage with respect to NMA that the introduction of perturbation terms (not harmonic in nature) is straightforward and that updating of the reference coordinates for the different harmonic terms along time, or any sampling variable is also very simple.

#### D. Discrete Molecular Dynamics

This technique is related to standard MD but avoids the integration of Newton's equations of motion by assuming a ballistic regime, that is, by considering that the particles move at constant velocity in flat wells (see below). Under these conditions, the position of a particle after some period of time (the minimum collision time) is given by:

$$\vec{r}_i(t + t_c) = \vec{r}_i(t) + \vec{v}_i(t)t_c, \quad (36)$$

where  $\vec{r}_i$  and  $\vec{v}_i$  stand for positions and velocities, and  $t_c$  is the minimum among the collision times  $t_{ij}$  between each pair of particles  $i$  and  $j$ :

$$t_{ij} = \frac{-b_{ij} \pm \sqrt{b_{ij}^2 - v_{ij}^2(r_{ij}^2 - d^2)}}{v_{ij}^2}, \quad (37)$$

where interparticle distance  $r_{ij}$  is the modulus of  $\vec{r}_{ij} = \vec{r}_j - \vec{r}_i$ , the relative velocity  $v_{ij}$  is the modulus of  $\vec{v}_{ij} = \vec{v}_j - \vec{v}_i$ ,  $b_{ij} = \vec{r}_{ij} \cdot \vec{v}_{ij}$ , and  $d$  is the distance corresponding to the wall of the square well. Note that if the inner term of the square root is negative (i.e., negative collision times), the two particles will not collide.

When two particles collide (assuming elastic collision regime), there is a transfer of linear momentum in the direction of the vector  $\vec{r}_{ij}$  and accordingly:

$$m_i \vec{v}_i = m_i \vec{v}'_i + \Delta \vec{p} \quad (38)$$

$$m_j \vec{v}_j + \Delta \vec{p} = m_j \vec{v}'_j$$

where the prime denotes the variables after the collision.

In order to calculate the change in velocities upon collision, the velocity of each particle is projected in the direction of the vector  $\vec{r}_{ij}$  and conservation rules are applied:

$$m_i u_i + m_j u_j = m_i u'_i + m_j u'_j \quad (39)$$

$$\frac{1}{2} m_i u_i^2 + \frac{1}{2} m_j u_j^2 = \frac{1}{2} m_i u_i'^2 + \frac{1}{2} m_j u_j'^2 + \Delta V, \quad (40)$$

where  $u_i$ ,  $u_j$  are the projections of the velocities  $v_i$ ,  $v_j$  along the direction  $\vec{r}_{ij}$  and  $\Delta V$  stands for the height of the step in the interatomic potential.

The transferred momentum can be easily determined from:

$$\Delta p = \frac{m_i m_j}{m_i + m_j} \left\{ \sqrt{(u_j - u_i)^2 - 2 \left( \frac{m_i + m_j}{m_i m_j} \right) \Delta V} - (u_j - u_i) \right\}, \quad (41)$$

Note that the two particles can overcome the potential step as long as:

$$\Delta V < \frac{m_1 m_2}{2(m_1 + m_2)} (u_j - u_i)^2 \quad (42)$$

Otherwise, the particles rebound and Eq. (41) reduces to:

$$\Delta p = \frac{m_i m_j}{m_i + m_j} \left\{ \sqrt{(u_j - u_i)^2} - (u_j - u_i) \right\} \quad (43)$$

which taking the negative solution of the root leads to:

$$\Delta p = \frac{2 m_i m_j}{m_i + m_j} (u_i - u_j) \quad (44)$$

This corresponds to the transfer of linear momentum in the case of an infinite wall, like those used to prevent steric clashes or the infinitely deep square wells used to define covalent bonds.

In practice, a DMD calculation is an extremely simple process: (i) create a list of events (“collisions”), (ii) compute the changes in velocities after first collision, (iii) update collision list and repeat the process. Since no integration is made, there is no need to recompute energies, forces, positions and velocities every few femtoseconds, and trajectories progress from collision to collision, irrespective of the collision time. The technique is then ideal to represent systems with very slow dynamics (for example diffusion processes), where traditional MD (or related techniques) will be rather inefficient. As noted above, DMD requires the use of flat well potentials, which implies some intrinsic simplification, but using a reasonable combination of flat well potentials quite reasonable approximation to physical potentials can be achieved (Zhou and Karplus, 1999; Ding et al., 2005; Ding and Dokholyan, 2008; Emperador et al., 2008a,b, 2010). Despite its simplicity, the technique seems able to provide reasonable approximation to the real dynamics of proteins, with a very small computational cost. Practical applications of DMD in a variety of cases, from equilibrium dynamics of proteins to folding, docking or protein aggregation have been published (Zhou and Karplus, 1999; Ding et al., 2005; Ding and Dokholyan, 2008; Emperador et al., 2008a,b, 2010). In order to allow a more general use, the technique has been recently implemented in our webservice-based tool FlexServ (<http://mmb.pcb.ub.es/FlexServ>; see Fig. 9).

#### IV. CONCLUSIONS

The improvement in computer codes and the development of new hardware are dramatically increasing the range of applicability of atomistic models coupled to MD simulation protocols. However, there are and there will be many cases, in the near future, where atomistic MD simulation is not recommended for several reasons: (i) the timescale of the dynamic process is too large, (ii) the system is too large, or (iii) there is a strong requirement for a fast response. In these cases, the latest generation of coarse-grained modeling techniques provides us with powerful tools to gain qualitative insight into the protein dynamics.

Recent software developments are approaching coarse-grained modeling to general users, by allowing access to web pages and web applications which can be used remotely by nonexpert users (Fig. 9). We are then

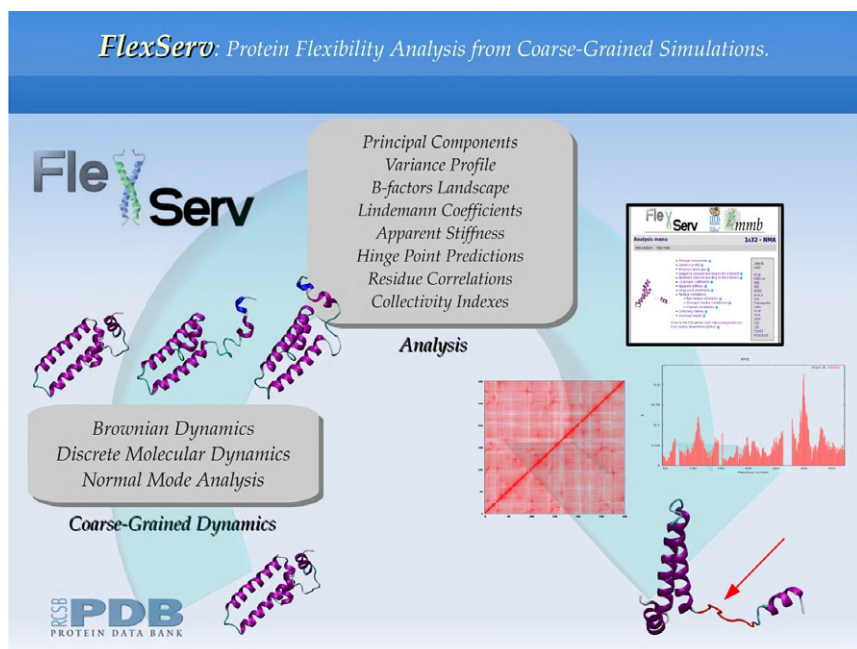


FIG. 9. General Scheme of our FlexServ webpage for accessing to a battery of coarse-grained methods using as input PDB entries.

facing a scenario, where coarse-grained analysis of protein flexibility will be done as a routine task in biology laboratories, even by nonexperts with little or no knowledge on the physical foundations on the analysis that is performed. We will also witness the systematic use of coarse-grained technique as a *prior* step to much more demanding atomistic MD simulations. Clear examples in the later area are going to be frequent in the representation of complex conformational transitions by means of state-of-the-art techniques such as steered MD, umbrella sampling, or meta-dynamics, which require to be effective some previous knowledge of the preferred transition pathway. Finally, even not discussed in this chapter, we can expect an explosion of hybrid methods combining low and high resolution of proteins and of methods based on the “open boundary” paradigm, where atomistic and coarse grain representation of parts of the protein will interchange as trajectory evolves.

## REFERENCES

- Abseher, R., Horstink, L., Hilbers, C. W., Nilges, M. (1999). Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins Struct. Funct. Gen.* **31**, 370–382.
- Arkhipov, A., Freddolino, P. L., Imada, K., Namba, K., Schulten, K. (2006a). Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum. *Biophys. J.* **91**, 4589–4597.
- Arkhipov, A., Freddolino, P. L., Schulten, K. (2006b). Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure* **14**, 1767–1777.
- Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
- Brooks, C. L., III, Karplus, M., Pettitt, B. M. (1988). Dynamical simulation methods. In *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics; Advances in Chemical Physics*, vol. LXXI, pp. 33–58. John Wiley & Sons, New York.
- Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., et al. (2009). FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* **25**(13), 1709–1710.
- Clementi, C., Nymeyer, H., Onuchic, J. N. (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.
- Clementi, C., Garcia, A. E., Onuchic, J. N. (2003). Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism all-atom representation study of protein. *J. Mol. Biol.* **326**, 933–954.
- Ding, F., Dokholyan, N. V. (2008). Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation. *Proc. Natl. Acad. Sci. USA* **105**, 19696–19701.
- Ding, F., Buldyrev, S. V., Dokholyan, N. V. (2005). Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.* **2005**(88), 147–155.
- Dobbins, S. E., Lesk, V. I., Sternberg, M. J. E. (2008). Expand+ Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–protein docking. *Proc. Natl. Acad. Sci. USA* **105**, 10390–10395.
- Emperador, A., Carrillo, O., Rueda, M., Orozco, M. (2008a). Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.* **95**, 2127–2138.
- Emperador, A., Meyer, T., Orozco, M. (2008b). United-atom discrete molecular dynamics of proteins using physics-based potentials. *J. Chem. Theory Comput.* **4**, 2001–2010.
- Emperador, A., Meyer, T., Orozco, M. (2010). Protein flexibility from discrete molecular dynamics simulations using quasi-physical potentials. *Proteins* **78**, 83–94.

- Gardiner, C. W. (1989). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Science*. second ed. Springer, Berlin.
- Go, N. (1983). Protein folding as a stochastic process. *J. Stat. Phys.* **30**, 413–423.
- Harvey, M. J., Giupponi, G., De Fabritiis, G. (2009). ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**(6), 1632–1639.
- Hess, B., van der Spoel, D., Lindahl, E. (2008). GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**(3), 435–447.
- Hinsen, K., Petrescu, A., Dellerue, S., Bellissent-Funel, M., Kneller, G. (2000). Harmonicity in slow protein dynamics. *Chem. Phys.* **261**, 25–37.
- Jeong, J. I., Jang, Y., Kim, M. K. (2006). A connection rule for  $\alpha$ -carbon coarse-grained elastic network models using chemical bond information. *J. Mol. Graph. Model.* **24**, 296–306.
- Jorgensen, W. L., Tirado-Rives, J. (2005). Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **26**, 1689–1700.
- Karplus, M., McCammon, J. A. (2002). Molecular dynamics simulations of proteins. *Nat. Struct. Biol.* **9**, 646–652.
- Khalili, M., Liwo, A., Scheraga, H. A. (2006). Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and UNRES model of polypeptide chains. *J. Mol. Biol.* **355**, 536–547.
- Kondrashov, D. A., Cui, Q., Philips, G. N. (2006). Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophys. J.* **91**, 2760–2767.
- Kong, Y., Ma, J., Karplus, M., Lipscomb, W. N. (2006). The allosteric mechanism of yeast chorismate mutase. A dynamics analysis. *J. Mol. Biol.* **356**, 237–247.
- Kovacs, J. A., Chacon, P., Abagyan, R. (2004). Predictions of protein flexibility, first-order measurements. *Proteins* **56**, 661–668.
- Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, L., Yu, H., Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework; developing mode concentration as a useful classifying statistic. *Proteins* **48**, 682–695.
- Kubo, R. (1959). Some Aspects of the Statistical Mechanical Theory of irreversible Processes. In: *Lectures in Theoretical Physics*, vol. 1, British, W. (Ed.), pp. 120–203. Interscience, New York.
- Kubo, R. (1965). Linear Response Theory of Irreversible Processes. In: *Statistical Mechanics of Equilibrium and Non-Equilibrium*, Proc. Int. Symposium, Aachen, Meixner, J. (Ed.), pp. 81–99. North-Holland, Amsterdam.
- Lemons, D. S. (1977). Paul Langevin's 1908 paper "On the Theory of Brownian Motion". *Am. J. Phys.* **65**(11), 1079–1081. (Sur la théorie du mouvement brownien, *C. R. Acad. Sci. (Paris)* 146, 530–533 (1908)).
- Liwo, A., Khalili, M., Scheraga, H. A. (2005). Molecular dynamics with the united-residue (UNRES) model of polypeptide chains: test of the approach on model. *Proteins* **102**, 2362–2367.
- Liwo, A., Khalili, M., Czaplewski, C., Kalinowski, S., Oldziej, S., Wachucik, K., et al. (2007). Modification and optimization of the UNRES potential energy function for canonical simulations. *J. Phys. Chem. B* **111**, 260–285.

- Ma, J. (2009). Coarse-grained elastic normal mode analysis and its applications in X-ray crystallographic refinement at moderate resolutions. In: *Coarse-Graining of Condensed Phase and Biomolecular Systems*, Voth, G. A. (Ed.), pp. 255–266. CRC Press, Boca Raton, FL.
- Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., de Vries, A. H. (2007). The MARTINI force-field: coarse Grained model for biomolecular simulations. *J. Phys. Chem. B* **111**, 7812–7824.
- Marrink, S. J., Fuhrmans, M., Risselada, H. J., Periolo, X. (2009). The MARTINI force-field. In: *Coarse Grained of Condensed Phase and Biomolecular Systems*, Voth, G. A. (Ed.), pp. 5–20. CRC Press, Boca Raton, FL.
- McCammom, J. A., Gelin, B. R., Karplus, M. (1977). Dynamics of folded proteins. *Nature* **267**, 585–590.
- McKerell, A., Jr., Wiorkiewicz-Kuczera, J., Karplus, M. (1995). An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.* **117**, 11946–11975.
- Mendez, R., Bastolla, U. (2010). Torsional Network Model: normal modes in torsion angle space better correlate with conformational changes in Proteins. *Phys. Rev. Lett.* **104**, 228103–228107.
- Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., Carrillo, O., Camps, J., Fenollosa, C., Repchevsky, D., Gelpí, J.L., Orozco, M. (2010). MoDEL (molecular dynamics extended library): a database of atomistic molecular dynamics trajectories. *Structure*. (In Press).
- Moritsugu, K., Smith, J. C. (2007). Coarse-grained biomolecular simulation with REACH, Realistic extension algorithm via covariance hessian. *Biophys. J.* **93**, 3460–3469.
- Orellana, L., Rueda, M., Ferrer-Costa, C., López-Blanco, J. R., Chacón, P., Orozco, M. (2010). Approaching elastic network models to molecular dynamics flexibility. *J. Chem. Theory Comput.* **6**, 2910–2923.
- Phillips, G. N., Jr. (2006). Normal mode analysis in studying protein motions with X-ray crystallography. In: *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, Cui, Q. and Bahar, I. (Eds.), pp. 155–170. Chapman & Hall/CRC, Boca Raton, FL.
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802.
- Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., et al. (2007a). A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA* **104**, 796–801.
- Rueda, M., Chacón, P., Orozco, M. (2007b). Thorough validation of protein normal mode analysis; a comparative study with essential dynamics. *Structure* **15**, 565–575.
- Rueda, M., Bottegoni, G., Abagyan, R. (2009). Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J. Chem. Inf. Model.* **49**(3), 716–725.
- Sen, T. Z., Jernigan, R. L. (2006). Optimizing the parameters of the Gaussian network model for ATP-binding proteins. In: *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, Cui, Q. and Bahar, I. (Eds.), pp. 171–186. CRC Press, Boca Raton, CA.

- Shih, A. Y., Arkhipov, A., Freddolino, P. L., Schulten, K. (2006). Coarse grained protein-lipid model with application to protein particles. *J. Phys. Chem. B* **110**, 3674–3684.
- Sorensen, M., Head-Gordon, T. (2002). Toward minimalist models of larger proteins: a ubiquitin-like protein. *Proteins* **46**, 368–379.
- Taketomi, H., Ueda, Y., Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulations. I. Effect of specific amino-acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Prot. Res.* **7**, 445–459.
- Tama, F., Gadea, F. X., Marques, O., Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* **41**, 1–7.
- Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908.
- Ulmschneider, J. P., Jorgensen, W. L. (2003). Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. *J. Chem. Phys.* **118**, 4261.
- Van der Kamp, M. W., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., et al. (2010). Dynameomics: a comprehensive database of protein dynamics. *Structure* **18**, 423–435.
- Van Kampen, N. G. (1981). *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam.
- Voelz, Vincent A., Bowman, Gregory R., Beauchamp, Kyle, Pande, Vijay S. (2010). Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9 (1–39). *J. Am. Chem. Soc.* **132**(5), 1526–1528.
- Yang, L., Eyal, E., Chennubhotla, C., Jee, J. G., Gronenborn, A., Bahar, I. (2007). Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure* **15**, 741–749.
- Zhou, Y. Q., Karplus, M. (1999). Interpreting the folding kinetics of helical proteins. *Nature* **401**, 400–403.



# RECENT ADVANCES IN THE MOLECULAR MODELING OF ESTROGEN RECEPTOR-MEDIATED TOXICITY

By IVANKA TSAKOVSKA,\* ILZA PAJEVA,\* PETKO ALOV,\* AND ANDREW WORTH†

\*Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, Sofia, Bulgaria

†Institute for Health and Consumer Protection, European Commission—Joint Research Centre, Ispra, Italy

I.	Introduction .....	218
II.	Structural Studies of the Estrogen Receptor and Its Ligands .....	220
	A. Structural Characterization of Estrogen Receptor Subtypes .....	220
	B. Estrogenic Endocrine Active Substances .....	224
III.	Molecular Modeling Approaches to Investigate Estrogen Receptor-Mediated Toxicological Effects .....	228
	A. Three-Dimensional Quantitative Structure–Activity Relationships .....	229
	B. Multidimensional Quantitative Structure–Activity Relationships .....	231
	C. Receptor-Based Approaches .....	233
IV.	Molecular Modeling of Estrogen Receptor-Mediated Toxicological Effects: Case Studies .....	234
	A. Ligand-Based and Combined Molecular Modeling Studies .....	234
	B. Docking and Virtual Screening Studies .....	242
V.	Conclusions .....	245
	References .....	246

## ABSTRACT

In the past two decades, there has been increasing concern about the potentially adverse effects of exogenous endocrine active substances (EAS) that alter the function of the endocrine system by interfering with hormone regulation. The mechanistic pathways by which EAS may elicit adverse effects, such as developmental and reproductive toxicity, often involve direct binding to nuclear hormone receptors. Certainly, the best studied nuclear receptor is the estrogen receptor (ER). Large-scale *in vitro* and *in vivo* methods have been developed to assess the estrogenic toxicity of chemicals. However, there are financial and animal welfare concerns related to their application. Quantitative structure–activity relationship (QSAR) approaches have proven their utility as a priority setting tool in the risk assessment of EAS. In addition, the models help to clarify the

binding mode of the interacting substances. As estrogen-mediated effects are usually related to ligand–receptor interactions, and as there have been comprehensive structural studies on the ER, molecular modeling together with other *in silico* approaches provide a suitable means of studying these estrogenic effects. This chapter provides an overview of the molecular modeling approaches applied to ligand–ER interactions. The progress in the field is outlined, and some critical issues are analyzed based on recently published models where these approaches are applied.

## I. INTRODUCTION

In the 1996 European Workshop on the Impact of Endocrine Disruptors on Human Health and Wildlife (EC, 1996), an endocrine-disrupting chemical (EDC) was defined as an exogenous substance or mixture that alters the function of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny. It is important to distinguish between endocrine active substances (EAS), that is, chemicals that interact with and affect the functioning of the endocrine system, but which do not necessarily trigger or contribute to the development of adverse effects in humans or wildlife, and endocrine disruptors (EDs), which according to the Weybridge definition, are EAS that are additionally associated with evidence of adverse effects at the *in vivo* level. Currently, there are no internationally harmonized criteria for determining the ED status of a chemical, which means that the label of ED is not consistently applied.

Over the past decade, there has been a focused international effort to identify the possible adverse effects of the EAS on humans and wildlife. Chemicals capable of acting as EAS include pesticides, pharmaceuticals, natural foodstuffs, and industrial chemicals. Ecological exposures to such substances are primarily from industrial and wastewater treatment effluents, whereas human exposures are mainly through the food chain (Kavlock et al., 2008). Regarding the possible human health effects of EAS, concerns have been related to effects on male reproductive health, hormone-dependent cancers, and effects on the immune system. There have even been links to rises in obesity and diabetes. In relation to the environment, concerns have been related to effects on development, growth, and reproduction of the living organisms.

In many cases, EAS act by direct binding to a nuclear hormone receptor (NR). This interaction, which triggers a cascade of molecular events

resulting in effects at the *in vivo* and population levels, is referred to as a molecular initiating event. NRs are ligand-inducible transcription factors involved in the regulation of specific target genes, and they are of critical importance for cellular processes such as cell growth, differentiation, and metabolic processes (McKenna and O'Malley, 2002). Members of the NR superfamily include receptors for various steroid hormones as estrogen, androgen, progesterone, several corticosteroids, retinoic acid, thyroid hormones, vitamin D, and dietary lipids (the peroxisome proliferator-activated receptor, PPAR).

The manner in which ligand binding regulates the activity of NRs is through a distinct ligand-inducible receptor conformation that triggers a number of downstream events resulting in the up- or downregulation of gene expression. Binding of EAS to the receptor may mimic the biological effect of a hormone, thus initiating the cell's normal response to the naturally occurring hormone at the wrong time or to an excessive extent (agonistic effect). Alternatively, EAS may bind to the receptor but not activate it. Instead, the presence of the chemical in the receptor prevents binding of the natural hormone (antagonistic effect). A third category of ligands, termed selective estrogen receptor modulators (SERMs), have the ability to act as both agonists and antagonists, depending on the cellular and promoter context.

The largest and best studied group of NRs is the estrogen receptor (ER) family. It mediates the effects of the steroid hormone estradiol (E2) in males and females. It is needed for the development and maintenance of reproductive tissues but is also present in a number of nonreproductive tissues, such as bone, liver, brain, and the cardiovascular system (Katzenellenbogen, 1996; Katzenellenbogen et al., 1997).

Since the discovery in the 1930s that compounds which were structurally unrelated to E2 could mimic its effects, scientists have investigated the estrogenic activity of a vast range of both naturally occurring and man-made chemicals. Such studies have revealed numerous structural motifs that are able to bind to ER and exert either estrogenic or antiestrogenic activities. Whereas most possess a phenol group—a prerequisite for high-affinity binding—these compounds exhibit, in addition, a huge variety of molecular scaffolds with diverse connectivities (Fang et al., 2001; Pike, 2006).

While traditional *in vitro* assays for detecting potential ED binding may be a suitable choice for prioritizing chemicals for additional *in vivo* animal studies, there are limitations related to the time and cost associated with

screening thousands. In contrast, the development and application of *in silico* models could be a more viable approach when setting priorities for further experimental evaluation (Tong et al., 2003). Beyond prediction, these models offer numerous additional benefits as (i) elucidation of existing structure–activity relationships, (ii) providing insights into mechanisms of action (e.g., agonist vs. antagonist), (iii) identifying key structural features associated with high/low activity, (iv) suggesting new design strategies, and (v) narrowing the dose range for planned assay (Fang et al., 2003).

Among the numerous *in silico* approaches that have been applied to the modeling of estrogenic effects, those that make use of the three-dimensional (3D) structures of both the receptor and the ligand, referred to here as 3D molecular modeling approaches, play an important role. The first reason is that the molecular modeling is directly applicable to receptor-mediated effects, as discussed above. Second, high-resolution crystallographic structures are available of the ER–ligand-binding domain (LBD) bound to a range of ligands (Fig. 1). This provides further opportunities to rationalize receptor–ligand structure–activity relationships using molecular modeling approaches.

In this study, we provide an overview of the recent advances in the modeling of the estrogen-related, receptor-mediated toxicity through application of molecular modeling tools, which are based on 3D structures of the ligands and the receptor. The three main sections describe the structural and functional characterization of the ER and its ligands, the available modeling methodology including 3D QSAR techniques, pharmacophore identification, docking, virtual screening (VS), as well as case studies on the modeling of ligand–ER interactions.

## II. STRUCTURAL STUDIES OF THE ESTROGEN RECEPTOR AND ITS LIGANDS

### A. Structural Characterization of Estrogen Receptor Subtypes

Two ERs, ER $\alpha$  and ER $\beta$ , regulate gene expression in response to estrogen exposure. They are encoded by distinct genes or separate chromosomes and display unique and overlapping physiological roles that are highly dependent on the tissue and cell type (Shanle and Xu, 2011).

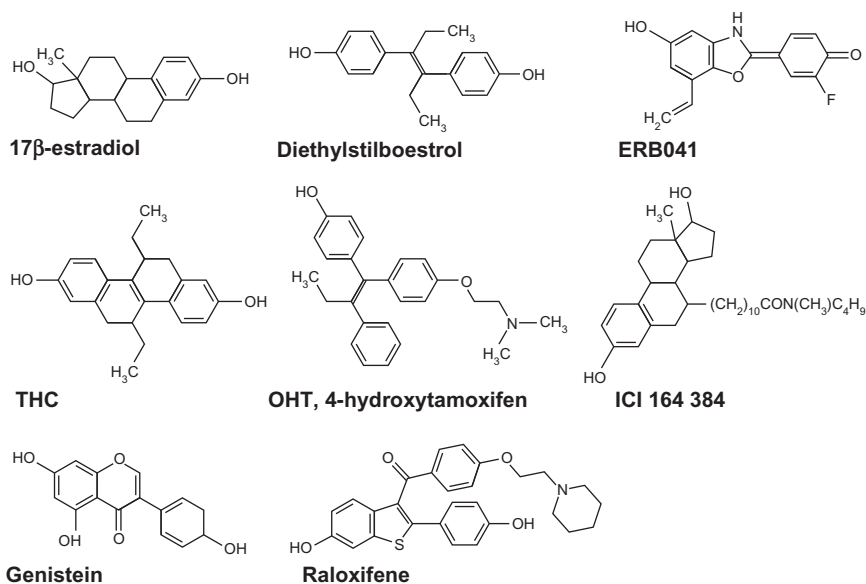


FIG. 1. Ligands commonly used in ER structural studies.

Both ER subtypes possess a modular organization that is characteristic of the NRs—five functional domains from the N- to C-termini, designated A/B, C (DNA-binding domain, DBD), D, E (LBD), and F (Fig. 2) (Evans, 1988).

There are two parts associated with distinct activation functions (AF1 and AF2) that facilitate transcriptional activation of target gene expression by promoting interactions with coregulator proteins. The first one is a part of A–B regulatory domain, and its action is independent of the presence of ligand. The transcriptional activation of AF1 is normally very weak, but it does synergize with AF2 in the LBD to produce a more robust upregulation of gene expression. The A–B domain is highly variable in sequence between both subtypes (~20%). C domain containing DBD is highly conserved (>90% sequence identity between both subtypes). It is responsible for binding to DNA at estrogen response elements. The site for hormone recognition is located in the carboxy-terminal LBD. It contains AF2 whose action is dependent on the presence of a bound ligand. AF2 in the LBD is localized in a conformationally flexible region and is involved in recruitment of coactivators and corepressors (Pike, 2006). ER $\alpha$  and ER $\beta$  have rather

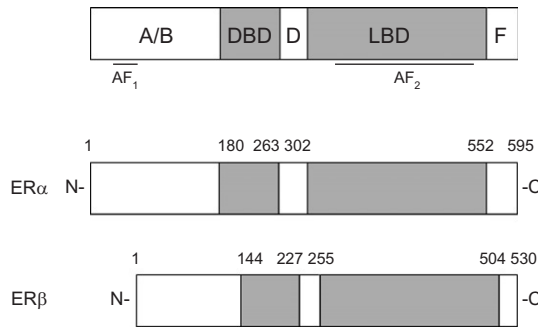


FIG. 2. Schematic presentation of ER $\alpha$  and ER $\beta$  with the functional domains (A/B, DBD, D, LBD) and activation functions (AF<sub>1</sub> and AF<sub>2</sub>) indicated.

different AF<sub>1</sub> domains: they share only 18% similarity, and the AF<sub>1</sub> domain of ER $\alpha$  enhances estrogen-induced expression of reporter genes to a greater extent than that of ER $\beta$  (Shanle and Xu, 2011). The AF<sub>2</sub> consists of  $\alpha$  helices that form a hydrophobic groove to which cofactors can bind.

Numerous crystal structures have been determined for the LBDs of both subtypes, and these have given a detailed insight into the structure and alterations during the ligand binding. The LBDs of the both subtypes ER share about 60% sequence identity. However, the residues composing the E2 binding sites in ER $\alpha$  and ER $\beta$  are identical except two pairs (see below).

Because of the significance of the receptor-mediated effects, the LBDs are the focus of scientific attention. The overall structure of ER LBD adopts the classical helical sandwich fold (Fig. 3A). Twelve  $\alpha$ -helices are arranged into three layers (Wurtz et al., 1996). The two parallel outer layers sandwich a central, orthogonal one that occupies the top half of the LBD fold. Thus, a cavity is formed in the lower part of the domain where the ligand binds. The bound ligand induces conformational changes allowing dimerization and DNA binding. The resultant LBD conformation depends on the steric parameters of the ligands and determines what type of coregulator is recruited: the binding of agonists favors coactivator recruitment, leading to transcriptional activation, whereas binding of antagonists favors corepressor interaction, leading to inhibition of transcriptional activation.

Many EAS compete with the endogenous estrogen at the ER-binding site, directly influencing the ER signaling. Therefore, the receptor-mediated mechanism of EAS action is probably the best studied disruption of the ER

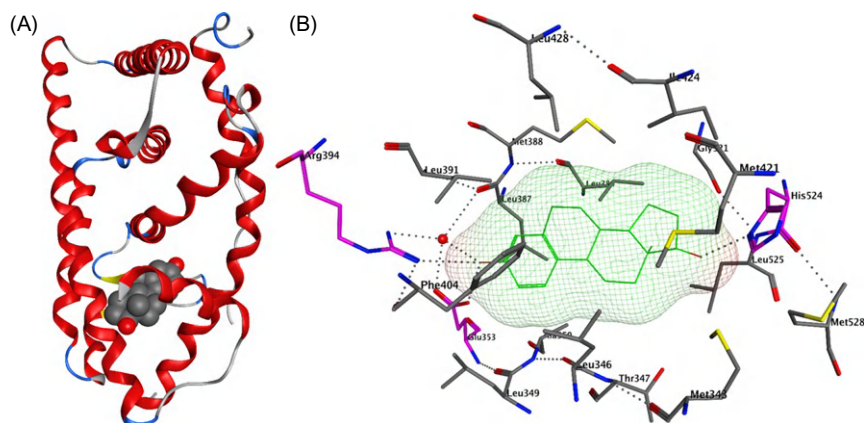


FIG. 3. (A) Structure of the ER $\alpha$  ligand-binding domain with bound estradiol (PDB ID 1ERE, Brzozowski et al., 1997). The upper part of the domain appears to be a stable, rather rigid structure, and the lower part, in which the ligand is accommodated, appears to be more flexible and dynamic (Katzenellenbogen and Muthyala, 2003). The protein 3D structure is drawn as a ribbon and colored according to the secondary structure (red—helices, yellow—strands, blue—turns, light gray—loops); E2 is rendered in a space-filling form with the C-atoms colored in dark gray and the O-atoms in red. (B) A closer view of the E2 binding pocket with the residues involved: Met343, Leu346, Thr347, Leu349, Ala350, Glu353, Leu384, Leu387, Met388, Leu391, Arg394, Phe404, Met421, Ile424, Leu428, Gly521, His524, Leu525, Met528. The amino acids are shown in stick rendering and colored by atom types: C, dark gray; O, red; N, blue; S, yellow; the C-atoms in the residues that make HBs with E2 (Arg394, Glu353, and His524) are colored in magenta. The water molecule is presented as a red ball. The ligand surface is colored by atom types and the structure is shown in green lines. The figures were generated by MOE (MOE 2010.10).

signaling. Crystal structures of the LBDs of the two ER subtypes reveal features of the binding pockets that are important to understand which compounds may display estrogenic activity through direct interaction with the receptor. However, even in one structural series of ligands, large changes in affinity often result from minor stereochemical changes. These unique ligand-binding properties are due to the size and flexibility of the ligand-binding pocket. The ER is unusual among the steroid receptors in that the size of the binding pocket is considerably larger than the ligand (450 and 390 Å<sup>3</sup> for ER $\alpha$  and ER $\beta$  pockets, respectively, and 245 Å<sup>3</sup> for E2 molecule) (Brzozowski et al., 1997; Pike et al., 1999), thus allowing a diverse set of small molecules to access the

pocket. [Figure 3A](#) illustrates the structure of the whole LBD of ER $\alpha$  and [Fig. 3B](#) the E2 binding site with the residues involved in interactions with the ligand (in a different orientation compared to [Fig. 3A](#) for a better view). The pocket is composed of about 20 amino acids and one water molecule. The residues interact with each other and with water through hydrogen bonds (HBs). The high affinity to E2 is a result of hydrophobic interactions and HBs between the hydroxyl group of E2, a water molecule in the cavity and amino acid residues in the ligand-binding pocket.

In particular, the network between the phenolic hydroxyl in A-ring of E2, water molecules, and amino acids Glu353 and Arg394 of ER $\alpha$  and Glu305 and Arg346 of ER $\beta$  is of critical importance for estrogenic activity ([Fig. 4](#)). On the other side of the E2 molecule, the hydroxy group of the D-ring shares a HB with His524 in ER $\alpha$  and His475 in ER $\beta$ . EAS binding directly to ER share structural similarities with E2: they have a phenolic group that mimics the steroidal A-ring of E2 and a rigid scaffold of high hydrophobicity. The second hydroxy group if present is recognized by a single histidine (His524 or His475). The orientation of these histidines is flexible and can accommodate different ligand-binding modes. Thus, the basic estrogen pharmacophore is well established and consists of two appropriately spaced hydroxy groups at either end of a near-planar hydrophobic scaffold ([Anstead et al., 1997](#)). Both receptors differ by two pairs of residues: Met421 corresponds to Ile373 and Leu384 to Met336 in ER $\alpha$  and ER $\beta$ , respectively. Even slight, such differences can contribute to subtype selectivity to some chemicals, although both receptors bind E2 with similar affinities.

LBD flexibility is another feature determining the ability of the ER to bind diverse chemicals. Rigid regions recognize the estrogenic features such as the phenolic ring and hydroxyl–hydroxyl separation. The flexible ones allow accommodation of bulky substituents like the long side chains of SERMs. Flexibility also allows the ER to shrink the volume of the cavity to better accommodate smaller ligands ([Pike, 2006](#)).

### *B. Estrogenic Endocrine Active Substances*

Estradiol, estrone, and estriol are the main endogenous mammalian estrogens. The exogenous compounds that interfere with the normal ER signaling include pharmaceuticals, industrial chemicals, pesticides, and phytoestrogens ([Fig. 5](#)).



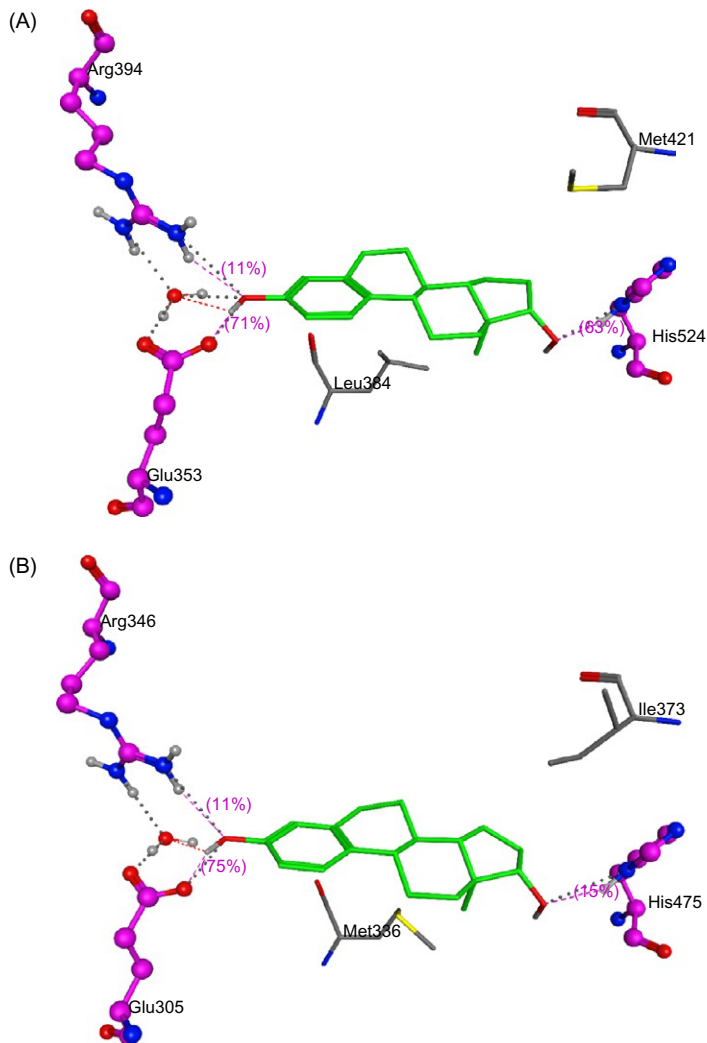


FIG. 4. Binding of the ligand E2 (green) in: (A) ER $\alpha$  (PDB ID 1ERE); (B) ER $\beta$  (PDB ID 3OLS, [Mocklinghoff et al., 2010](#)). The key residues involved in HB interactions (magenta dot lines) are rendered in balls and sticks with C-atoms colored in magenta; the differing residues Met421 and Ile373 in ER $\alpha$ ; and Leu384 and Met336 in ER $\beta$  are rendered in sticks and colored according to the atom types (C, dark gray; O, red; N, blue; S, yellow; polar H, light gray). HBs are scored in percents according to the distance and orientation of the polar atoms: the higher the score, the closer they are to the optimal values. This figure was generated by MOE.

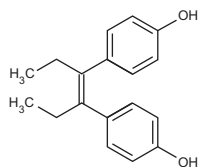
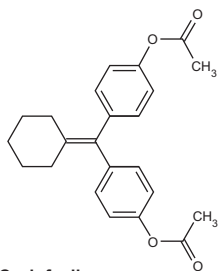
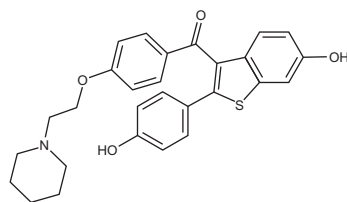
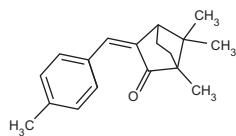
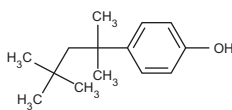
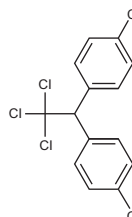
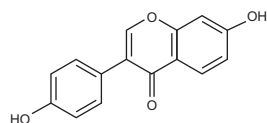
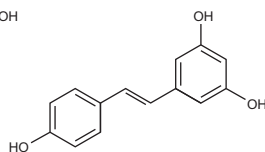
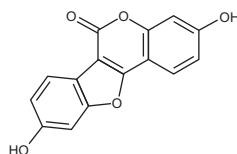
**Pharmaceuticals****Diethylstilbestrol****Cyclofenil****Raloxifene****Industrial chemicals and pesticides****4-Methylbenzylidene  
camphor****Tert-octyl phenol****Dichlorodiphenyltrichloroethane  
(DDT)****Phytoestrogens****Daidzein****Resveratrol****Coumestrol**

FIG. 5. Exogenous EAS: pharmaceuticals, industrial chemicals, pesticides, phytoestrogens.

### 1. *Pharmaceuticals*

Many synthetic estrogens have been designed for pharmacological purposes. They are constructed by certain substitutions on the steroidal estrogen skeleton that increases estrogenic potency by enhancing binding to the ER. However, substitution with large moieties at given positions may lead to antiestrogenic effects ([Katzenellenbogen and Muthyala, 2003](#)).

There are also many classes of synthetic, nonsteroidal estrogens. A number of nonsteroidal estrogens have been developed as tissue-selective estrogens—SERMs (Grese and Dodge, 1998). These agents are used for breast cancer prevention and treatment and for menopausal hormone replacement.

Because estrogen pharmaceuticals are rationally constructed for their hormonal effects in the body, these compounds are not typically thought as potential EDCs in humans. Nevertheless, some of these compounds have been found in rivers and streams where they can affect aquatic wildlife (Sumpter, 1998). The use of certain estrogen pharmaceuticals such as anabolic agents in livestock can result in residual levels in meat or animal by-products through which other animals and humans can be exposed.

## 2. *Industrial Chemicals and Pesticides*

Quite a number of man-made compounds, or their metabolites, have been found to have estrogenic activity. Although these are typically of low potency, some are very lipophilic and potentially environmentally persistent and bioaccumulative compounds. For example, certain components of plastics, dialkyl phthalates used as plasticizers, and bisphenol A, a component of thermostable polycarbonate polymers, have been reported to be weak estrogens (McLachlan, 2001).

A number of highly chlorinated aromatic substances have been found to have estrogenic activity. These include polychlorinated biphenyls (PCBs), which are typically mixtures of many isomers with varying degrees of chlorine substitution (Layton et al., 2002). The highest potency is found with PCB metabolites that have a parahydroxy group (Connor et al., 1997), but the binding affinities are still relatively low. Methoxychlor and DDT are organochlorine pesticides that can exhibit estrogenic activity through interaction with both subtypes of ER. In addition, a variety of nonaromatic chlorinated pesticides have been reported as estrogens. For instance, endosulfan and dieldrin are polycyclic and heavily chlorinated (Soto et al., 1994), and though they are not phenols, each of them has a polar moiety that might fulfill this function.

## 3. *Phytoestrogens*

These are naturally occurring compounds in plants that mimic steroidal estrogens. They have polyphenolic structures and include flavonoids, lignans, coumestans, and stilbenes. These natural estrogens are low-affinity,

low-potency ligands for the ER. Exposure occurs mainly through dietary intake of food, including fruits, herbs, vegetables, and especially soy which contains high levels of these agents. Among the different classes of phytoestrogens, flavonoids are one of the most widespread in the food.

Many of the phytoestrogens that are natural components in food are generally considered to confer a health benefit (Katzenellenbogen and Muthyala, 2003). Nevertheless, concerns have been raised about the possible health consequences of exposure to phytoestrogens in certain situations. For instance, many animal studies indicate that phytoestrogens can compete for ER binding and modulate its normal function; in transgenic estrogen reporter mice, genistein inhibited the estrogenic response of E2 in the liver (Shanle and Xu, 2011).

### III. MOLECULAR MODELING APPROACHES TO INVESTIGATE ESTROGEN RECEPTOR-MEDIATED TOXICOLOGICAL EFFECTS

QSARs correlate the change in the binding affinity within a series of ligands with the same mechanism of action and comparable manner of binding to the changes in their structures (Kubinyi, 1995). While their primary function in the drug design is lead discovery and optimization, in the field of toxicology, they have played an important role as a priority setting tool for risk assessment. In this context, it is important to note that the observation, or prediction, of receptor binding does not constitute proof of endocrine-disrupting effects. It is simply one molecular event among a whole series of events (mode-of-action pathway) which under certain circumstances may lead to endocrine-related effects.

QSARs based on three-dimensional models (3D QSARs) are of particular interest when discussing ER-mediated effects. These models rely on the 3D structures of the ligands and binding sites in order to quantify the ligand–receptor interactions and to help in clarifying the molecular mechanisms of interactions.

The ER is one of the main targets investigated using QSAR and molecular modeling approaches in the identification of potential EDCs. Usually, the models are based on *in vitro* data. Various tests have been developed to estimate the potential estrogenic effect of chemicals *in vitro*. They can be divided roughly into three groups: (i) ER competitive binding assays estimating the binding affinity of a ligand to ER; (ii) reporter-gene assays,

encoding not only ligand–ER binding but also transcriptional and translational effects; and (iii) cell proliferation assays (Devillers et al., 2006).

Taking into account the complexity of the ED phenomenon, it should be noted that while QSAR approaches provide a useful means of simulating specific key events, they cannot on their own address the entire phenomenon. To capture the entire sequence of events comprising the mode of action would require the development of a battery of *in silico* tools integrated with *in vitro* methods (Roncaglioni and Benfenati, 2008).

#### A. Three-Dimensional Quantitative Structure–Activity Relationships

As the binding of EAS to the ER receptor depends on the 3D structures of both ligand and receptor, different 3D QSAR models have been derived for predicting ER-binding affinity. Compared to classical QSAR approaches, they allow for better understanding and estimation of the ligand–receptor interaction. However, one should take into account that these methods are based on the assumption of the same binding mode of ligands. Thus their application to large databases with structurally heterogeneous compounds could be problematic or at least would require precise selection of subsets and development of a number of models for the different training sets.

These methods are based on the concept of molecular interaction fields (MIFs). The main idea is placing the structure of interest that is preliminarily optimized within a lattice that simulates the receptor environment and calculating interaction energies of the molecule with a probe (e.g., carbon atom) in each intercept of the grid.

The GRID method was the first MIF to calculate interaction energies between a target molecule and monoatomic probes such as a carbonyl oxygen atom or a negative carboxyl oxygen atom, or polyatomic probes such as water, an amino group, or a methyl group. Although fully empirical, it allows a precise evaluation of the total interaction energy as the sum of three components, namely (1) van der Waals interactions; (2) electrostatic interactions; and (3) hydrogen bonding (Goodford, 1985; Cruciani, 2006).

Among the MIF 3D QSAR approaches, comparative molecular field analysis (CoMFA) has been widely used to correlate the differences within 3D MIFs with the binding affinity to ER (Kubinyi, 1993). The more recent 3D QSAR approach, comparative molecular similarity indices analysis (CoMSIA), has also been widely applied in the field of ER modeling

(Klebe, 1998). Both approaches require the studied structures to be appropriately aligned and “placed” in a cubic grid that simulates the receptor space. CoMFA evaluates the interaction energy between a probe atom and each atom in the aligned molecules in regularly spaced grid points. In the standard implementation of the method, only Lennard-Jones and Coulomb potentials for calculation of the steric and electrostatic energies, respectively, are used to describe the enthalpic contribution to the free energy of binding. Because of the singularity of these potentials at the atomic positions, their applicability is restricted to regions outside the molecules. In addition to the standard CoMFA method, an empirical hydrophobic field-like 3D function has been calculated with the program HINT (hydrophobic interactions) and imported into the SYBYL implementation of CoMFA. The addition of hydrophobicity describes the entropic contribution to the ligand–receptor interactions and appears to offer increased interpretability of the CoMFA models (Kellogg et al., 1991). The CoMSIA method calculates steric, electrostatic, and HB donor and acceptor similarity indices of aligned molecules at regularly spaced grid points occupied by a common probe atom. The most important contributions responsible for binding affinity are considered to be covered by these properties. Gaussian-type functions with no singularities are used, so in contrast to CoMFA, no arbitrary definition of cutoff values is required. Both CoMFA and CoMSIA use the partial least squares analysis to correlate the structural parameters calculated in the 3D QSAR analysis with the biological activity of interest (Wold et al., 2001). The model built can be used to predict the activity of new structures. Based on the model, so-called contour maps can be built. They identify the structural regions corresponding to differences in the fields which contribute most (about 80% of the signal) to the differences in investigated activity.

The 3D QSAR analysis requires preliminary geometry optimization and energy minimization of the molecular structures as the bioactive conformation—the one that interacts with the receptor—is generally assumed to be a low-energy conformation. For this purpose, molecular mechanics, semiempirical, or *ab initio* quantum chemical methods are used.

As 3D QSAR is based on the assumption of the same mechanism of action and the same binding mode of the molecules, the alignment of the structures is a critical step in the modeling process. It is based on the potential pharmacophore points or the common structural skeleton in the absence of the pharmacophore hypothesis.

The pharmacophore modeling is an important step in the 3D QSAR analysis. In fact, the pharmacophore is a 3D representation of the arrangements of key structural features that are responsible for binding of the ligands to the receptor. These may include HB donor and acceptor groups, hydrophobic regions, ionizable groups, etc. A pharmacophore search is rather extensive procedure that requires energy minimization, conformational analysis of flexible ligands as well as alignment of potentially important structural features for various conformers using least squares fitting (Madden and Cronin, 2010). The selection of the molecules to be superimposed is very important in order to obtain significant results.

Despite the fact that there are many crystallographic studies of ER and that crystallographic data are considered, in general, as the most reliable structural information to be used for molecular design, pharmacophore modeling and other ligand-based molecular modeling techniques still have an important place in the field of ER modeling. This is because the docking studies could meet some general problems related to crystallization process: (i) possible inadequate resolution of crystallographic structures; (ii) possible structural distortions of the ligand–protein complex during crystallization; (iii) lack of information on hydrogen atoms in the crystallographic structures; (iv) crystallographic structures generally ignore structural heterogeneity related to protein anisotropic motion and discrete conformational substates (Taha et al., 2010). In addition, 3D QSAR models could be very informative for predictive purposes.

Another 3D approach, named COREPA COmmon REactivity PAttern, was developed by Mekenyan and Serafimova (2009). The COREPA approach is a probabilistic classification method which assesses the impact of molecular flexibility on stereo electronic properties of chemicals. Similarity between chemicals is analyzed by comparing their conformational distributions, and the system automatically identifies the parameter that best discriminate chemicals in groups. A Bayesian decision tree is then developed for classifying untested chemicals.

### *B. Multidimensional Quantitative Structure–Activity Relationships*

One of the critical points in CoMFA analysis is its limitation to a single conformation of each ligand. It is appropriately addressed in 4D QSAR (Hopfinger et al., 1997; Ekins et al., 1999; Vedani et al., 2000) in which

each ligand molecule is presented by an ensemble of conformations, orientations, and protonation states, respectively.

The basic concept of 5D QSAR reflects the situation that accommodation of ligands into the binding site is often facilitated by the adaptation of the protein to the ligand topology (induced fit)—a mechanism triggered by the interaction between the small molecule and the protein. The induced fit may be local including rearrangement of few side chains or global by including main chain alterations (Vedani et al., 2006). In addition to alteration in the topology of the binding pocket, the induced fit may change the hydrophobicity of subpockets or the solvent accessibility of the binding site. Based on experimental data, the impact of the induced fit on ligand binding has been investigated and consequently appropriate algorithms have been developed and implemented in the Quasar software by Vedani et al. (1998). Quasar generates a family of quasi-atomistic receptor surrogates that are optimized using a genetic algorithm. Thus it addresses the situation when the structure of the receptor is unknown. A hypothetical receptor is developed by means of 3D surface which surrounds the ligand structures at the van der Waals distance. Its topology reflects the shape of the binding site. The surface is populated with properties mapped onto it, representing important information such as hydrophobicity, electrostatic potential, and hydrogen-bonding ability (Lill et al., 2005).

Further, a 6D QSAR approach has been developed by the same group that allows for consideration of different solvation models, and it has been implemented in the Quasar software. Besides classical 3D structural presentations, the approach considers also the possibility for more than one bound conformations, the induced fit ligand–receptor interaction, and the solvation effect. The solvation terms (ligand desolvation and solvent stripping) are independently scaled for each different model within a surrogate family of receptors, reflecting varying solvent accessibility of the binding pocket.

Based on these multidimensional QSAR approaches, VirtualToxLab has been developed. This is a commercial tool for predicting endocrine-disrupting potential by simulating and quantifying the interactions with aryl hydrocarbon, estrogen alpha/beta, androgen, thyroid alpha/beta, glucocorticoid, liver X, mineralocorticoid, and PPAR gamma (Vedani and Smiesko, 2009; Vedani et al., 2009). It also includes metabolic considerations by simulating interactions with the enzymes CYP450 3A4 and 2A13. The tool is based on the combined use of automated flexible docking with multidimensional QSAR. Flexible docking is based on Yeti software (Vedani et al., 2005).



It aims at identifying all potential orientations and conformations of a small molecule within the binding pocket. The docking protocol includes two steps: (i) the simulation of induced fit, allowing the protein to adapt its shape to the different orientations and conformations of the small molecule during the search procedure and (ii) the consideration of solvent effects (typically water). Multidimensional QSAR is based on the Quasar software.

### C. Receptor-Based Approaches

These techniques are used when the structure of the receptor is known. They yield important information concerning the spatial orientation of the ligands in the binding site and predict the binding free energy thus evaluating the strength of the ligand–receptor interactions (Höltje et al., 2008). They are often used as a complementary tool to improve the quality of the developed 3D QSAR models. In particular, docking can be applied to generate potential bioactive conformations for further QSAR analysis.

As a source of the 3D receptors' structures often the Protein Data Bank (<http://www.pdb.org/pdb>) is used. It is a free access repository of experimentally determined protein structures which is continuously enriched and updated. Once the structures of the ligands and receptor are available, different docking algorithms can be applied to “place” the ligand structures into the receptor-binding site. Two main problems exist here: (i) which is the ligand conformation that interacts with the receptor and (ii) how to calculate the free energy of binding (scoring problem). Different algorithms use different scoring functions to quantify protein–ligand interactions. Most approaches rely on molecular mechanics force fields. Others use available experimental data to obtain parameters for some relatively simple functions that allow quick estimation of the binding energy (Höltje et al., 2008). The estimated values allow discrimination between active and inactive molecules. However, the developed scoring functions are still far from a precise description of the highly complex terms that need to be taken into account when the free energy is quantified.

Two main strategies exist for the placement of the molecules into the receptor pocket (Schneider and Baringhaus, 2008): either the whole molecule is docked (like DOCK [Kuntz et al., 1982], GLIDE [Friesner et al., 2004], GOLD [Jones et al., 1997] programs), or it is virtually dissected into structural fragments that are reconstructed in the binding site (e.g. FlexX program [Rarey et al., 1996]).

The protein structure should be treated as flexible in ligand docking. However, for a long time, this has been computationally impossible. The development of docking algorithms taking protein flexibility into account has started recently, aided by the development of hardware technologies. Thus, based on the consideration of flexibility, the docking procedures can be classified into three categories: (i) rigid body docking—both protein and ligand are treated as rigid; (ii) semiflexible docking—only the ligand is considered flexible; and (iii) fully flexible docking—both ligand and protein are treated as flexible.

Some programs take into account the flexibility of the receptor during docking, for instance the Surflex-Dock method (Jain, 2009); others perform postdocking optimization of the protein–ligand complexes considering different levels of protein atom flexibility, for instance AMMOS (Pencheva et al., 2008).

VS is designed for searching in large electronic databases of chemical structures by using computational analysis. It aims at selecting a limited number of candidate molecules that are likely to be active against a particular biological receptor. VS could be considered as a logical extension of 3D pharmacophore-based database searching or molecular docking. Thus, VS approaches can be classified into two categories: pharmacophore-based virtual screening (PBVS) and docking-based virtual screening (DBVS).

#### IV. MOLECULAR MODELING OF ESTROGEN RECEPTOR-MEDIATED TOXICOLOGICAL EFFECTS: CASE STUDIES

3D QSAR and receptor-based modeling are of particular interest when ER-mediated toxicological effects are investigated. There are several studies that report molecular modeling results of ligands–ER interactions. An extensive (but not exhaustive) list of recently developed models as published in the literature is provided in Table I, and the models are discussed in more details in the sections below.

##### A. *Ligand-Based and Combined Molecular Modeling Studies*

Various CoMFA and CoMSIA studies were performed with different datasets over the years, as described in a number of reviews (Devillers et al., 2006; Lill and Vedani, 2007; Roncaglioni and Benfenati, 2008;

TABLE I  
Examples of Ligand- and Receptor-Based Modeling Studies for Prediction of ER Binding (Ordered According to the Appearance in the Sections)

Bioeffect	Computational approach	Dataset: chemical class/size	Reference
Estrogenic potency to the hER $\alpha$	CoMSIA; docking	Polybrominated diphenyl ethers, para-hydroxylated polybrominated diphenyl ethers, and brominated bisphenol A compounds: 26 substances	Yang et al. (2010)
ER $\alpha$ and ER $\beta$ binding	CoMFA; docking	3-Arylquinazolinethione derivatives: 45 substances	Xiao et al. (2008)
ER $\alpha$ and ER $\beta$ binding	CoMFA; GRID	6-Phenylnaphthalenes and 2-phenylquinolines: 81 substances	Salum et al. (2008)
ER $\alpha$ binding	CoMFA; 2D; Hologram QSAR	Flavanoids, dihydrobenzoxathiins, and dihydrobenzodithiins: 127 substances	Salum et al. (2007)
MCF-7 inhibition	CoMFA; CoMSIA; GRIND QSAR	Raloxifene analogues: 15 substances	Menezes et al. (2006)
ER $\alpha$ binding	CoMSIA; docking	Rhenium complexes and organic ligands: 29 substances	Wolohan and Reichert, (2007)
ER $\alpha$ and ER $\beta$ binding	CoMFA	Diphenolic azoles: 104 substances	Demyttenaere-Kovacheva et al. (2005)
ER $\beta$ binding	Pharmacophore modeling	Diverse set of ER $\beta$ ligands: 119 substances	Taha et al. (2010)
MCF-7 inhibition	Pharmacophore modeling	Structurally different SERMs: 53 substances	Brogi et al. (2009)
ER binding	Pharmacophore modeling	Flavones, coumestans, isoflavones, triphenylethylenes, steroids, etc.: 137 substances	Islam et al. (2008)
hER binding	COREPA	Structurally diverse dataset: 645 substances	Serafimova et al. (2007)
hER $\alpha$ binding	Multidimensional QSAR; docking	Structurally diverse substances: 106 substances	Vedani et al. (2005)

*(Continued)*

TABLE I (Continued)

Bioeffect	Computational approach	Dataset: chemical class/size	Reference
hER $\alpha$ binding	Molecular dynamics; docking-based virtual screening	Structurally diverse dataset: 3500 substances	<a href="#">Sivanesan et al. (2005)</a>
rER binding	Docking-based virtual screening	Structurally diverse dataset: 281 substances	<a href="#">Rabinowitz et al. (2009)</a>
ER $\alpha$ and ER $\beta$ ER $\alpha$ binding	Docking Pharmacophore virtual screening; docking-based virtual screening	Structurally diverse dataset: 12 substances Database of 32 actives and 1990 nonactives	<a href="#">Kiss and Allen (2007)</a> <a href="#">Chen et al. (2009)</a>
hER $\alpha$ binding	Molecular dynamics; docking	PCBs, plasticizers, and pesticides: 43 substances	<a href="#">Celik et al. (2008)</a>
hER $\alpha$ binding ER $\alpha$ and ER $\beta$ binding	docking Shape signatures tool; docking	Phenol-related derivatives: 14 substances Commercially available organic chemicals: 200,000 substances	<a href="#">Nose et al. (2009)</a> <a href="#">Wang et al. (2006)</a>
hER $\alpha$ binding	Docking	Polychlorinated compounds: 7 substances	<a href="#">D'Ursi et al. (2005)</a>

Lo Piparo and Worth, 2010). They give useful information that is often combined with the X-ray crystallographic data and the results from other QSAR approaches, thus outlining a consistent picture of the important structural features of different structural classes to bind to the ER pocket. The results are used as 3D search queries to mine 3D libraries for new ER ligands, to predict the biological activities of the new ligands, to help in better understanding the binding mode of the ligands, and to outline the structural requirements for ligands selectivity to the ER subtypes.

Among the more recent studies is the 3D QSAR to predict estrogenicity of polybrominated diphenyl ethers, *para*-hydroxylated polybrominated diphenyl ethers, and brominated bisphenol A compounds to the human ER $\alpha$  (hER $\alpha$ ), using a training set of 26 compounds (Yang et al., 2010). Based on the molecular conformations developed from molecular docking, predictive CoMSIA models were developed with the following statistical parameters: correlation coefficient  $r^2=0.949$ ; standard error of estimate SEE=0.24; cross-validated correlation coefficient  $q^2=0.72$ , and predictive correlation coefficient  $r_{pr}^2=0.68$  (six compounds in the test set). Because of the limited size of the group studied, the particular benefit of the models needs to be further demonstrated.

The study of Xiao et al. (2008) explored the selectivity requirements of 3-arylquinazolinethione derivatives for binding with ER $\beta$  versus ER $\alpha$  using CoMFA and docking. Docking results indicated that the 3-arylquinazolinethione derivatives are of an ideal length for forming tight HBs between the 4'-hydroxyl and Glu305 and Arg346 at one end of the ER $\beta$  pocket and between the 7-hydroxyl and His475 at the other end. CoMFA models successfully predicted the inhibitory activity against ER $\beta$  ( $r^2=0.97$ ; SEE=0.20;  $q^2=0.64$ ;  $r_{pr}^2=0.62$ ) and the of ER $\beta$ /ER $\alpha$  selectivity ( $r^2=0.96$ ;  $q^2=0.52$ ; SEE=0.15;  $r_{pr}^2=0.70$ ). Both the CoMFA and the molecular docking results consistently suggested that the introduction of an appropriate bulky group into the structures increases the ER $\beta$  inhibitory activity and reduces the ER $\alpha$  inhibitory activity, thus improving the selectivity of the designed ligands.

With the same aim to outline the selectivity features for subtype binding, CoMFA studies were performed on a dataset of 81 ER modulators (6-phenylnaphthalenes and 2-phenylquinolines) for which binding affinity values were collected for both ER $\alpha$  and ER $\beta$  by Salum et al. (2008). The models were developed on a training set of 65 compounds and the remaining 16 compounds were used as a test set. CoMFA models showed

rather similar statistical parameters: (i) for ER $\alpha$ :  $r^2=0.94$ ,  $q^2=0.76$ , and  $r_{pr}^2=0.80$ ; (ii) for ER $\beta$ :  $r^2=0.96$ ,  $q^2=0.73$ , and  $r_{pr}^2=0.88$ . Five ER crystal structures were used in GRID/PCA investigations to generate MIF maps. The several similarities between the GRID/PCA pseudocontour plots and QSAR 3D CoMFA contour maps were detected. Based on the CoMFA model of each receptor, they outline structural regions involved in the ER subtype selectivity.

The same authors used CoMFA and HQSAR (hologram QSAR) analyses on a large set of ER $\alpha$  modulators and developed predictive 3D and HQSAR models (Salum et al., 2007). A structurally diverse dataset was used including three major chemical classes, namely flavanoids, dihydrobenzoxathiins, and dihydrobenzodithiins. A training set of 99 compounds was used to generate binding affinity models, and 28 structures were selected from different chemical classes to form the external test set. The information for the ligand-binding conformations was adopted from the receptor-binding site. For that purpose, docking procedure was applied using GOLD docking program. Based on this, predictive models were obtained ( $r^2=0.93$ ,  $q^2=0.79$ , SEE=0.26). The external predictive ability was expressed as residuals between experimental and calculated values. The maximal residual was 0.42 log units. The models outline higher importance of the steric fields, explained by the fact that ER ligand-binding cavity possesses hydrophobic features which must be fitted correctly by selective ligands. Contour maps were built that were informative for the design of new modulators.

A set of 15 raloxifene analogues was analyzed by Menezes et al. (2006) using GRIND QSARs. For this purpose, MIFs were computed using the GRID program. GRIND descriptors were calculated by ALMOND program (Pastor et al., 2000). They describe the geometry of the interaction in an alignment-independent way. GRIND QSAR models were built and outlined the favorable formation of several HBs as well as favorable  $\pi$ - $\pi$  interactions in the ER pocket. CoMFA and CoMSIA models were developed as well (36 structures in the training set), revealing steric, electrostatic, and HB atom modifications that can cause variations in binding potency.

Wolohan and Reichert (2007) used a genetic algorithm to model a diverse set of novel rhenium-based ER ligands with relative-binding affinities (RBA) to ER $\alpha$  with respect to 17 $\beta$ -estradiol. The binding properties were studied with a combination of CoMSIA and docking. A total of 29 ER

ligands consisting of 11 rhenium complexes and 18 organic ligands were docked inside the LBD of ER $\alpha$  utilizing the program GOLD. The top-ranked poses were used to develop CoMSIA models from a training set of 22 randomly selected compounds. The model combining CoMSIA steric, electrostatic, and hydrophobic indices together with the polar volume showed highest predictive ability among the different runs ( $r^2=0.94$ ;  $q^2=0.68$ ;  $SEE=0.24$ ). Analysis of the scoring functions from GOLD showed particularly poor correlation to RBA ER $\alpha$ . In comparison, the combined CoMSIA and polar volume model ranked correctly the ligands in order of increasing RBA. Thus the utility of this method as a prescreening tool in the development of novel rhenium-based ER ligands was demonstrated.

A study of the structural requirements for ER $\alpha$  and ER $\beta$  selectivity was performed by [Demyttenaere-Kovatcheva et al. \(2005\)](#). CoMFA models were developed for a training set of 72 benzoxazole and benzisoxazole derivatives and validated on a test set of 32 compounds. The models developed for ER $\alpha$  and ER $\beta$  had statistical parameters as follows: for ER $\alpha$ :  $r^2=0.91$  and  $q^2=0.60$ ; for ER $\beta$ :  $r^2=0.95$  and  $q^2=0.40$ .

A number of pharmacophoric hypotheses for estrogen-binding ligands were developed using the [Catalyst software \(2005\)](#). It enables automatic pharmacophore construction by using a collection of structurally diverse molecules with activities ranging over a number of orders of magnitude. The models outline the importance of hydrophobic, hydrophobic aromatic regions as well as HB donors, acceptors for the binding activities. For example, [Taha et al. \(2010\)](#) developed pharmacophore models for a diverse set of 119 ER $\beta$  ligands as collected from the literature. The affinities were expressed as the concentrations of the test compounds that displaced 50% of the bound [3H]17 $\beta$ -estradiol. The structures were carefully selected (i) to have significantly dissimilar affinities to ED $\alpha$  and ED $\beta$  to allow development of selective models for ED $\beta$  ligands; (ii) to have wide structural diversity; and (iii) to span big range of affinity values (over 4.0 orders of magnitude). A total of 210 pharmacophore models emerged from 24 automatic CATALYST modeling rounds. Classical QSAR analysis was performed to search for the best combination of pharmacophore(s) and 2D descriptors (calculated by employing the C2.DESRIPTOR module of CERIUS2 software, Accelrys Inc.) capable of explaining bioactivity variation across the dataset. The models were built for the training set of 96 compounds and tested on the remaining 23 compounds giving  $\eta_{PRESS}^2$

of 0.54–0.56. Four binding hypotheses were outlined in the best QSAR equations suggesting the existence of distinct binding modes accessible to ligands within the ED $\beta$ -binding pocket. Based on the similarity between them, they were merged in two hybrid models. The resulting models and associated QSAR equations were employed to screen the National Cancer Institute list of compounds (238,819 compounds) and an in-house built database of known drugs and agrochemicals (2602 compounds) to search for new ER $\beta$  ligands. After screening with the models and filtering with Lipinski's and Veber's drug-likeness rules, a list of 1176 and, respectively, 409 compounds (for NCI database) and 73 and, respectively, 57 compounds (for the in house database) for both models was extracted.

Brogi et al. (2009) explored the pharmacophore features of a comprehensive set of SERMs derivatives, tested for their inhibitory activity toward MCF7 cell line. The existence of a quantitative correlation between inhibition of MCF7 human breast carcinoma cell line proliferation (measured by IC<sub>50</sub> values) and ER $\alpha$  receptor binding (RBA by competition with 17 $\beta$ -estradiol) has been reported for a series of SERM derivatives structurally related to raloxifene. This suggests that receptor binding is the first step in the pathway that leads to inhibition of tumor cell proliferation. Specifically, a dataset of 53 SERMs belonging to several different structural classes was compiled. These compounds covered four orders of magnitude activity range, centered at the IC<sub>50</sub> median value of 13  $\mu$ M, and included active and less active derivatives reported in the literature as well as metabolites isolated and tested by authors. The selected SERMs were divided into a training set of 24 compounds and a complementary test set of 29 compounds. The generated pharmacophore hypothesis by Catalyst software consisted of five features, namely one hydrophobic, two hydrophobic aromatic, one HB acceptor, and one HB donor. A screening of the Asinex GOLD collection database was performed by coupling pharmacophore hypothesis with a docking filtration, which resulted in a selection set of 12 new scaffolds to be investigated as potential SERMs. The inhibitory activity of these compounds was evaluated *in vitro* using MCF7 human breast adenocarcinoma cell line. Ten of the 12 compounds were found to be active with inhibitory activity between 26 and 188  $\mu$ M thus confirming the utility of the pharmacophore hypothesis generated.

A pharmacophore model was developed by Islam et al. (2008) on a training set of 35 compounds, including flavones, coumestans, isoflavones,



triphenylethylenes, steroids with a phenolic A-ring, etc. It outlined the presence of an aromatic ring and a hydrophobic region at a distance of 3.816 Å and two HB acceptors at a distance of 11.975 Å as crucial for binding affinity to ER. Testing the model on a set of 102 substances gave correlation coefficient  $r^2=0.96$ .

COREPA method was used to derive an hER-binding affinity model for a dataset of 645 chemicals making use of the scaffold of high-affinity synthetic ER ligands (Serafimova et al., 2007). Analysis of reactivity patterns based on the distance between nucleophilic sites resulted in identification of distinct interaction types: a steroid-like (AB) type described by frontier orbital energies and distance between nucleophilic sites with specific charge requirements; an AC type where local hydrophobic effects are combined with electronic interactions to modulate binding; and mixed ABC (AD) type. Chemicals were grouped by type, after which COREPA models were developed within specific RBA ranges. Analysis of the models showed that AB mechanism is probably associated with contribution of stereoelectronic and global HINTs: the functional parameter  $Q\_Distance$  (parameter combining the distance between two nucleophilic atoms and charge of one of these sites) and log Kow are used as COREPA discriminating parameters. Typical for the AC mechanism were found to be HINTs only, where the parameter describing local hydrophobicity and log Kow appears to be discriminative. Interactions underlying the third group of chemicals (ABC type) are described again by stereoelectronic and global hydrophobic parameters, Diameff (effective cross-section diameter), EHOMO (energy of the highest occupied molecular orbital), and log Kow. The performance of the models was illustrated by screening of 232 chemicals tested for rER (rat ER). The screening exercise showed a concordance of 0.60 between the predicted hER and experimentally observed rER data, accounting for the distribution of chemicals across the potency bins. This result is comparable with the interspecies correlation coefficient between hER and rER, which is 0.68.

The VirtualToxLab was used to estimate the binding affinity of small molecules toward the ER based on the X-ray crystal structure of the hER $\alpha$  ligand-binding domain with bound diethylstilbestrol (Vedani et al., 2005). A dataset of 106 diverse ER-binding molecules (comprising six molecular classes, including steroids, biphephyls, and stilbestrols) was used—88 training and 18 test set. The structures were docked to the binding pocket, and each complex was fully optimized using the Yeti software. Up to four

binding modes per molecule were accepted thus composing the input data for multidimensional QSAR, using the Quasar software. The comparison between the constructed receptor surrogate and the binding site at the true biological receptor showed that characteristic properties (H-bond acceptor residues, H-bond donor residue, and the residues forming the larger hydrophobic pocket) are well identified by the receptor surrogate. The resulting multidimensional QSAR model had a  $q^2=0.903$  and a  $r_{pr}=0.89$ .

### B. Docking and Virtual Screening Studies

Several factors make the ER very suitable for docking and VS: (i) there is a large quantity of data in terms of both X-ray data available for ER–ligand complexes and binding affinity data for many ligands; (ii) knowledge of the critical protein–ligand binding motifs in the LBD; and (iii) the binding pocket of ER is well defined in terms of size, shape, and polarity (Knox et al., 2008).

As already discussed, binding of an agonist or antagonist to ER leads to significant conformational changes. Thus, it is important to take into account the flexibility of the receptor during docking by using relevant docking and VS tools.

Based on the principles of molecular mechanics, AMMOS provides five different levels of protein atoms flexibility—from rigid to fully flexible protein, while the ligands are always flexible. Applied to ER, receptor relaxation with AMMOS led to considerable improvement of the enrichment factors, retrieving 83% of the actives after docking in the top 3% of the processed database (Pencheva et al., 2008).

The molecular dynamics (MD) simulations allow including full flexibility of the ligand-binding pocket by generating an ensemble of protein conformations. For instance, an ensemble of 51 energetically favorable structures of the hER $\alpha$ -binding pocket was collected from MD simulations in the study of Sivanesan et al. (2005). Detailed analysis of the MD results showed that nine amino residues were highly flexible and, in turn, influence the ligand binding (Asp351, Ile326, Phe404, Met421, Ile424, Phe425, His524, Met528, and Lys540). It is worth noting that among those are His524 involved in HB interaction with E2 and Met421 shown to differ in ER $\beta$  (Fig. 4). *In silico* screening of 3500 EDCs against

these flexible ligand-binding pockets was performed by docking with the FlexX program. The proposed MD-based *in silico* screening approach had higher hit rates than the rigid crystal structure. Thirty-two compounds were found to bind better to the flexible ligand-binding pockets compared to the crystal structure. These compounds, though belonging to different chemical classes, possess the features typical for the natural hormone 17 $\beta$ -estradiol.

Another important consideration when the ER-mediated effects are investigated is the purpose of the modeling study. Rabinowitz et al. outlined some particularities depending on the purpose of the study—drug discovery versus toxicity screening. For the pharmaceutical industry, the purpose of the initial screen is to limit the number of chemicals that proceed to the next phase of testing, and here, the exclusion of some active chemicals is a reasonable cost. In contrast, minimizing the number of false negatives is critical when screening environmental chemicals because the expectation is that positive chemicals will be tested later in an experimental protocol (Rabinowitz et al., 2008).

Exploring the possible computational approaches for chemical screening and testing prioritization the same authors performed VS of a library of 281 environmentally relevant chemicals into four rat ER targets (Rabinowitz et al., 2009). Ninety-five percent of the chemicals in the library were not active. Two docking protocols were applied—eHiTS (Zsoldos et al., 2006) and FRED (McGann et al., 2003). According to the score-based ranking, all of the active molecules were discovered in the top 16% of the ranked chemicals. When a pharmacophoric filter was applied on the basis of the geometry of binding to the ER, the results were improved, and all of the active molecules were discovered in the top 8% of the chemicals.

Kiss and Allen (2007) applied docking algorithms to predict and rank ligands according to their binding affinities. They docked two ligand subsets, ER agonists (seven structures) and SERMs (five structures), to ER $\alpha$  and ER $\beta$ , utilizing the Lamarckian genetic docking algorithm, as implemented in AutoDock (Morris et al., 1998) and the potentials of mean force scoring function. The ligands were ranked based upon the calculated ligand–receptor interaction energies, as well as experimental RBAs.  $r^2$  values, ranging from 0.55 to 0.93, indicate a good correlation between the virtual and experimental ranking.

A comparative study of two VS approaches, PBVS and DBVS, was performed by Chen et al. (2009) using eight pharmacologically important and

structurally diverse target proteins, including ER $\alpha$ . For each target, the pharmacophore model was constructed based on several X-ray crystal structures of this target protein in complex with ligands (mostly inhibitors), and one high-resolution crystal structure of the ligand–protein complex was used to generate the model for DBVS. LigandScout software was used for pharmacophore generation (Wolber and Langer, 2005). Three docking programs, namely DOCK, GOLD, and Glide, were used in the DBVS. Docking-based methods showed varying performance depending on the nature of the target binding sites. In comparison, for most cases, the pharmacophore-based method outperformed the docking-based methods, and the average PBVS enrichment over the virtual screens against the eight targets was much higher than those of DBVS. For ER $\alpha$ , in particular, using PBVS, 9 of 32 actives were recovered among the top 5% of the database, and this result was superior compared to DBVS where a maximum of two actives were identified.

In the study of Celik et al. (2008), the binding of 43 compounds to the hER $\alpha$  LBD was investigated. The dataset included selected PCBs, plasticizers, and pesticides, which were considered to be potential EDCs. Different conformations of the hER $\alpha$  LBD were used as identified by MD simulations. Glide program was applied for docking. It was found that most suspected EDCs could bind in the steroid-binding cavity, interacting with at least one of the two hydrophilic ends of the binding pocket. The best binders were the pesticides. It was predicted that these compounds can interact with the different conformations of hER $\alpha$  LBD with comparable affinities indicating that they can serve as universal binders to the hER $\alpha$  LBD, regardless of the receptor conformation.

Nose et al. (2009) suggested a new method to discriminate between agonist and antagonist binding to ER. The method was based on the estimated difference between binding energy in the activation conformation and the inactivation receptor conformation. The agonists were found to be more stable in the activation conformation, while antagonists were more stable in the inactivation conformation. A parameter was suggested to reflect these differences. This agonist/antagonist differential-docking screening method was used to evaluate 4-(1-adamantyl)phenol—one of the essential raw materials for nanoporous organosilicate thin films. It was predicted as an agonist of the hER $\alpha$  and this was confirmed by testing in a reporter-gene assay.

Wang et al. (2006) applied a multistep computational approach to a database of 200,000 commercially available organic chemicals in order to

identify potential EDCs. First, the Shape Signatures tool was applied. This is a computational tool that compares molecules on the basis of similarity in shape, polarity, and other bio-relevant properties. 4-Hydroxy tamoxifen and diethylstilbestrol were used as input queries. The hits identified by Shape Signatures method were docked inside the ligand-binding pocket of both the ER $\alpha$ -agonist and the ER $\alpha$ -antagonist X-ray crystal structures. Most of the compounds had better docking scores to the antagonist form of ER $\alpha$ , thus predicting them to be antagonists. On the basis of GOLD scores, eight compounds were predicted to be active and they were subsequently tested. The experimental values versus the calculated one gave  $r^2=0.65$ .

D'Ursi et al. (2005) characterized the molecular interaction of seven EAS with ER and other steroid receptors, using flexible docking protocol. AutoDock and QXP (McMartin and Bohacek, 1997) programs were used. The chemicals were organic polychlorinated compounds, including DDT, its metabolites DDE and DDD, and the hydroxylated PCB derivative (PCB-OH). All ligands docked in the buried hydrophobic pocket. The interaction was characterized by multiple hydrophobic contacts involving a different number of residues facing the binding pocket, depending on the orientation of the ligands. The EAS ligands did not display a unique binding mode; this was explained with their lipophilicity and flexibility, which conferred on them a great adaptability in the large and hydrophobic binding pocket of steroid receptors.

## V. CONCLUSIONS

In the past few years, there has been significant progress in the computational modeling of the ER, and this can be attributed to a number of factors: (i) the ER is involved in a variety of molecular pathways of physiological and toxicological relevance; (ii) various regulatory initiatives to identify potential EDCs have led to intensive application of computational modeling, for both cost and animal welfare reasons; (iii) the various crystallographic studies of the ER make it very suitable object for exploration with different molecular modeling techniques. Historically, the 3D QSAR techniques were the first to be applied. Thus, not surprisingly, a number of 3D QSAR models exist in the scientific literature. They are informative when ligand binding to the ER needs to be predicted and/or new active ligands need to be constructed. Of course one should bear in mind the limitation of these approaches to ligands with the the same

mode of action, which restricts their application to structurally homogeneous series of ligands. Further, when analyzing and applying such models one should carefully judge the quality of the model—many of the existing models still need more precise estimation of their robustness and external predictivity.

In view of the considerable progress in structural studies of the ER and the availability of X-ray structures with sufficiently high resolution, receptor-based techniques such as docking and VS are being increasingly exploited. This allows large chemical inventories to be screened and the number of compounds for experimental testing to be dramatically reduced. There are some issues that should be taken into account: (i) the quality of the test data; (ii) preparation of the ligand library for docking; (iii) the choice of docking protocols and scoring functions; and (iv) consideration of the protein flexibility. These needs should direct the future scientific efforts in the field. Finally, in order to improve our ability to describe and predict the (beneficial and adverse) effects of EAS, the information derived from the use of *in silico* methods will need to be integrated with data generated by *in vitro* methods, including high-throughput screening approaches. This approach will provide a means of linking apical effects and adverse outcomes at the *in vivo* and population levels with key events, such as receptor–ligand binding, in the underlying molecular pathways.

#### ACKNOWLEDGMENTS

The first three authors thank the National Science Fund of Bulgaria (Grant DTK 02-58) for the support.

#### REFERENCES

- Anstead, G. M., Carlson, K. E., Katzenellenbogen, J. A. (1997). The estradiol pharmacophore: ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* **62**, 268–303.
- Brogi, S., Kladi, M., Vagias, C., Papazafiri, P., Roussis, V., Tafi, A. (2009). Pharmacophore modeling for qualitative prediction of antiestrogenic activity. *J. Chem. Inf. Model.* **49**, 2489–2497.
- Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engström, O., et al. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **389**, 753–758.

- Catalyst, Accelrys, Inc., San Diego, CA, 2005.
- Celik, L., Davey, J., Lund, D., Schiøtt, B. (2008). Exploring interactions of endocrine-disrupting compounds with different conformations of the human estrogen receptor  $\alpha$  ligand binding domain: a molecular docking study. *Chem. Res. Toxicol.* **21**, 2195–2206.
- Chen, Z., Li, H. L., Zhang, Q. J., Bao, X. G., Yu, K. Q., Luo, X. M., et al. (2009). Pharmacophore-based virtual screening versus docking-based virtual screening: a benchmark comparison against eight targets. *Acta Pharmacol. Sin.* **30**, 1694–1708.
- Connor, K., Ramamoorthy, K., Moore, M., Mustain, M., Chen, I., Safe, S., et al. (1997). Hydroxylated polychlorinated biphenyls (PCBs) as estrogens and antiestrogens: structure–activity relationships. *Toxicol. Appl. Pharmacol.* **145**, 1111–1123.
- Demyttenaere-Kovatcheva, A., Cronin, M. T., Benfenati, E., Roncaglioni, A., Lopiparo, E. (2005). Identification of the structural requirements of the receptor-binding affinity of diphenolic azoles to estrogen receptors alpha and beta by three-dimensional quantitative structure-activity relationship and structure-activity relationship analysis. *J. Med. Chem.* **48**, 7628–7636.
- Devillers, J., Marchand-Geneste, N., Carpy, A., Porcher, J. M. (2006). SAR and QSAR modeling of endocrine disruptors. *SAR QSAR Environ. Res.* **17**, 393–412.
- D'Ursi, P., Salvi, E., Fossa, P., Milanese, L., Rovida, E. (2005). Modelling the interaction of steroid receptors with endocrine disrupting chemicals. *BMC Bioinf.* **6**(Suppl. 4), S10.
- EC (1996). Report of Proceedings of European Workshop on the Impact of Endocrine Disruptors on Human Health and Wildlife, Weybridge, UK, December 1996. [http://ec.europa.eu/environment/endocrine/documents/reports\\_conclusions\\_en.htm](http://ec.europa.eu/environment/endocrine/documents/reports_conclusions_en.htm)
- Ekins, S., Bravi, G., Binkley, S., Gillespie, J. S., Ring, B. J., Wikel, J. H., et al. (1999). Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics* **9**, 477–489.
- Evans, R. M. (1988). The steroid and thyroid hormone receptor superfamily. *Science* **240**, 889–895.
- Fang, H., Tong, W., Shi, L. M., Blair, R., Perkins, R., Branham, W., et al. (2001). Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.* **14**, 280–294.
- Fang, H., Tong, W., Welsh, W. J., Sheehan, D. M. (2003). QSAR models in receptor-mediated effects: the nuclear receptor superfamily. *J. Mol. Struct. (Theochem)* **622**, 113–125.
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., et al. (2004). Glide: a new approach for rapid, accurate docking and scoring. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749.
- Cruciani, G. (Ed.) (2006). Molecular Interaction Fields Applications in Drug Discovery and ADME Prediction. Wiley-VCH, Weinheim.
- Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857.

- Grese, T. A., Dodge, J. A. (1998). Selective estrogen receptor modulators (SERMs). *Curr. Pharm. Des.* **4**, 71–92.
- Kubinyi, H. (Ed.) (1993). 3D QSAR in Drug Design. Theory, Methods and Applications. ESCOM Science Publishers B.V., Leiden.
- Höltje, H. D., Sippl, W., Rognan, D., Folkers, G. (2008). Virtual screening and docking. In: *Molecular Modeling: Basic Principles and Applications* Wiley-VSH, Weinheim.
- Hopfinger, A. J., Wang, S., Tokarski, J., Jin, B., Albuquerque, M. A. B., Madhav, P. J., et al. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **119**, 10509–10524.
- Islam, M. A., Nagar, S., Das, S., Mukherjee, A., Saha, A. (2008). Molecular design based on receptor-independent pharmacophore: application to estrogen receptor ligands. *Biol. Pharm. Bull.* **31**, 1453–1460.
- Jain, A. N. (2009). Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *J. Comput. Aided Mol. Des.* **23**, 355–374.
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748.
- Katzenellenbogen, B. S. (1996). Estrogen receptors: bioactivities and interactions with cell signaling pathways. *Biol. Reprod.* **54**, 287–293.
- Katzenellenbogen, B. S., Montano, M., Ekena, K., Herman, M., McInerney, E. (1997). Antiestrogens: mechanisms of action and resistance in breast cancer. *Breast Cancer Res. Treat.* **44**, 23–38.
- Katzenellenbogen, J. A., Muthyala, R. (2003). Interactions of exogenous endocrine active substances with nuclear receptors. *Pure Appl. Chem.* **75**, 1797–1817.
- Kavlock, R., Ankley, G., Blancato, J., Breen, M., Conolly, R., Dix, D., et al. (2008). Computational toxicology—a state of the science mini review. *Toxicol. Sci.* **103**, 14–27.
- Kellogg, G., Semus, S., Abraham, D. (1991). HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des.* **5**, 545–552.
- Kiss, G., Allen, N. W. (2007). Automated docking of estrogens and SERMs into an estrogen receptor alpha and beta isoform using the PMF forcefield and the Lamarckian genetic algorithm. *Theor. Chem. Acc.* **117**, 305–314.
- Klebe, G. (1998). Comparative molecular similarity indices analysis: CoMSIA. In: *3D QSAR in Drug Design. Recent Advances*, vol. 3, Kubinyi, H., Folkers, G., and Martin, Y. Eds.), pp. 87–104. Kluwer/ESCOM, Dordrecht.
- Knox, A. J. S., Yang, Y., Lloyd, D. G., Meegan, M. J. (2008). Virtual screening of the estrogen receptor. *Expert Opin. Drug Discov.* **3**, 853–866.
- Kubinyi, H. (1995). Lock and key in the real world: concluding remarks. *Pharm. Acta Helv.* **69**, 259–269.
- Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288.
- Layton, A. C., Sanseverino, J., Gregory, B. W., Easter, J. P., Saylor, G. S., Schultz, T. W. (2002). In vitro estrogen receptor binding of PCBs: measured activity and detection of hydroxylated metabolites in a recombinant yeast assay. *Toxicol. Appl. Pharmacol.* **180**, 157–163.



- Lill, M., Vedani, A. (2007). Computational modeling of receptor mediated toxicity. In: *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*, Ekins, S. (Ed.), pp. 315–352. John Wiley & Sons Inc.
- Lill, M. A., Dobler, M., Vedani, A. (2005). In silico prediction of receptor-mediated environmental toxic phenomena—application to endocrine disruption. *SAR QSAR Environ. Res.* **16**, 149–169.
- Lo Piparo, E., Worth, A. (2010). Review of QSAR Models and Software Tools for predicting Developmental and Reproductive Toxicity. European Commission report EUR 24522 EN.
- Madden, J. C., Cronin, M. T. D. (2010). Three-dimensional molecular modelling of receptor-based mechanisms in toxicology. In: *Silico Toxicology*, Cronin, M. and Madden, J. (Eds.), pp. 210–227. Royal Society of Chemistry, Cambridge.
- McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., Brown, F. K. (2003). Gaussian docking functions. *Biopolymers* **68**, 76–90.
- McKenna, N. J., O'Malley, B. W. (2002). Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell* **108**, 465–474.
- McLachlan, J. A. (2001). Environmental signaling: what embryos and evolution teach us about endocrine disrupting chemicals. *Endocr. Rev.* **22**, 319–341.
- McMartin, C., Bohacek, R. (1997). QXP: powerful, rapid computer algorithms for structure-based drug design. *J. Comput. Aided Mol. Des.* **11**, 333–344.
- Mekenyan, O., Serafimova, R. (2009). Mechanism-based modeling of estrogen receptor binding affinity a common reactivity pattern (COREPA) implementation. In: *Endocrine Disruption Modeling*, Devillers, J. (Ed.), pp. 229–294. CRC Press, Boca Raton.
- Menezes, I. R. A., Leitão, A., Montanari, C. A. (2006). Three-dimensional models of non-steroidal ligands: a comparative molecular field analysis. *Steroids* **71**, 417–428.
- Mocklinghoff, S., Rose, R., Carraz, M., Visser, A., Ottmann, C., Brunsveld, L. (2010). Synthesis and crystal structure of a phosphorylated estrogen receptor ligand binding domain. *Chembiochem* **11**, 2251–2254.
- MOE 2010.10 (Molecular Operating Environment), Chemical Computing Group, 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, Canada H3A 2R7.
- Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., et al. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639–1662.
- Nose, T., Tokunaga, T., Shimohigashi, Y. (2009). Exploration of endocrine-disrupting chemicals on estrogen receptor alpha by the agonist/antagonist differential-docking screening (AADS) method: 4-(1-Adamantyl)phenol as a potent endocrine disruptor candidate. *Toxicol. Lett.* **191**, 33–39.
- Pastor, M., Cruciani, G., McLay, I., Pickett, S., Clementi, S. (2000). Grid independent descriptors (GRIND). A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **43**, 3233–3243.
- Pencheva, T., Lagorce, D., Pajeva, I., Villoutreix, B. O., Miteva, M. (2008). AMMOS: automated molecular mechanics optimization tool for in silico screening. *BMC Bioinformatics* **9**, 438–452.

- Pike, A. C. (2006). Lessons learnt from structural studies of the oestrogen receptor. *Best Pract. Res. Clin. Endocrinol. Metab.* **20**, 1–14.
- Pike, A. C., Brzozowski, A. M., Hubbard, R. E., Bonn, T., Thorsell, A. G., Engström, O., et al. (1999). Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist. *EMBO J.* **18**, 4608–4618.
- Rabinowitz, J. R., Goldsmith, M. R., Little, S. B., Pasquinelli, M. A. (2008). Computational molecular modeling for evaluating the toxicity of environmental chemicals: prioritizing bioassay requirements. *Environ. Health Perspect.* **116**, 573–577.
- Rabinowitz, J. R., Little, S. B., Goldsmith, M. R. (2009). Molecular modeling for screening environmental chemicals for estrogenicity: Use of the toxicant-target approach. *Chem. Res. Toxicol.* **22**, 1594–1602.
- Rarey, M., Kramer, B., Lengauer, T., Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**, 470–489.
- Roncaglioni, A., Benfenati, E. (2008). In silico-aided prediction of biological properties of chemicals: oestrogen receptor-mediated effects. *Chem. Soc. Rev.* **37**, 441–450.
- Salum, L. de B., Polikarpov, I., Andricopulo, A. D. (2007). Structural and chemical basis for enhanced affinity and potency for a large series of estrogen receptor ligands: 2D and 3D QSAR studies. *J. Mol. Graph. Model.* **26**, 434–442.
- Salum, L. de B., Polikarpov, I., Andricopulo, A. D. (2008). Structure-based approach for the study of estrogen receptor binding affinity and subtype selectivity. *J. Chem. Inf. Model.* **48**, 2243–2253.
- Schneider, G., Baringhaus, K. H. (2008). Receptor-ligand interaction. In: *Molecular Design: Concepts and Applications* Wiley-VCH, Weinheim.
- Serafimova, R., Todorov, M., Nedelcheva, D., Pavlov, T., Akahori, Y., Nakai, M., et al. (2007). QSAR and mechanistic interpretation of estrogen receptor binding. *SAR QSAR Environ. Res.* **18**, 389–421.
- Shanle, E. K., Xu, W. (2011). Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chem. Res. Toxicol.* **24**, 6–19.
- Sivanesan, D., Rajnarayanan, R. V., Doherty, J., Pattabiraman, N. (2005). In-silico screening using flexible ligand binding pockets: a molecular dynamics-based approach. *J. Comput. Aided Mol. Des.* **19**, 213–228.
- Soto, A. M., Chung, K. L., Sonnenschein, C. (1994). The pesticides endosulfan, toxaphene, and dieldrin have estrogenic effects on human estrogen-sensitive cells. *Environ. Health Perspect.* **102**, 380–383.
- Sumpter, J. P. (1998). Xenoendocrine disrupters—environmental impacts. *Toxicol. Lett.* **102–103**, 337–342.
- Taha, M., Tarairah, M., Zalloum, H., Abu-Sheikha, G. (2010). Pharmacophore and QSAR modeling of estrogen receptor  $\beta$  ligands and subsequent validation and in silico search for new hits. *J. Mol. Graph. Model.* **28**, 383–400.
- Tong, W., Fang, H., Hong, H., Xie, Q., Perkins, R., Anson, J., et al. (2003). Workshop 1.2: regulatory application of SAR/QSAR for priority setting of endocrine disruptors: a perspective. *Pure Appl. Chem.* **75**, 2375–2388.
- Vedani, A., Smiesko, M. (2009). In silico toxicology in drug discovery—concepts based on three-dimensional models. *ATLA Altern. Lab. Anim.* **37**, 477–496.

- Vedani, A., Briem, H., Dobler, M., Dollinger, H., McMasters, D. (2000). Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J. Med. Chem.* **43**, 4416–4427.
- Vedani, A., Dobler, M., Lill, M. A. (2005a). Combining protein modeling and 6D-QSAR—simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **48**, 3700–3703.
- Vedani, A., Dobler, M., Zbinden, P. (1998). Quasi-atomistic receptor surface models: a bridge between 3D-QSAR and receptor modeling. *J. Am. Chem. Soc.* **120**, 4471–4477.
- Vedani, A., Dobler, M., Lill, M. (2006). The challenge of predicting drug toxicity in silico. *Basic Clin. Pharmacol. Toxicol.* **99**, 195–208.
- Vedani, A., Smiesko, M., Spreafico, M., Peristera, O., Dobler, M. (2009). Virtual ToxLab in silico prediction of the toxic (endocrine-disrupting) potential of drugs, chemicals and natural products. Two years and 2,000 compounds of experience: a progress report. *ALTEX* **26**, 183–193.
- Wang, C. Y., Ai, N., Arora, S., Erenrich, E., Nagarajan, K., Zauhar, R., et al. (2006). Identification of previously unrecognized antiestrogenic chemicals using a novel virtual screening approach. *Chem. Res. Toxicol.* **19**, 1595–1601.
- Wolber, G., Langer, T. (2005). LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **45**, 160–169.
- Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58**, 109–130.
- Wolohan, P., Reichert, D. (2007). CoMSIA and docking study of rhenium based estrogen receptor ligand analogs. *Steroids* **72**, 247–260.
- Wurtz, J. M., Bourguet, W., Renaud, J. P., Vivat, V., Chambon, P., Moras, D., et al. (1996). A canonical structure for the ligand-binding domain of nuclear receptors. *Nat. Struct. Biol.* **3**, 87–94.
- Xiao, A., Zhang, Z., An, L., Xiang, Y. (2008). 3D-QSAR and docking studies of 3-arylquinazolinethione derivatives as selective estrogen receptor modulators. *J. Mol. Model.* **14**, 149–159.
- Yang, W., Liu, X., Liu, H., Wu, Y., Giesy, J. P., Yu, H. (2010). Molecular docking and comparative molecular similarity indices analysis of estrogenicity of polybrominated diphenyl ethers and their analogues. *Environ. Toxicol. Chem.* **29**, 660–668.
- Zsoldos, Z., Reid, D., Simon, A., Sadjad, B. S., Johnson, A. P. (2006). eHiTS: an innovative approach to the docking and scoring function problems. *Curr. Protein Pept. Sci.* **7**, 421–435.

# MULTISCALE COMPUTATIONAL METHODS FOR MAPPING CONFORMATIONAL ENSEMBLES OF G-PROTEIN-COUPLED RECEPTORS

By NAGARAJAN VAIDEHI AND SUPRIYO BHATTACHARYA

Department of Immunology, Beckman Research Institute of the City of Hope,  
Duarte, California, USA

I.	Introduction .....	254
II.	Conformational Flexibility in GPCRs.....	256
	A. Experimental Evidence for Conformational Changes upon Activation of GPCRs.....	256
	B. Dynamics and Conformational State Ensemble for GPCRs.....	258
	C. Insights on Conformational Flexibility from Thermostable Mutants.....	260
III.	Computational Approaches for Studying Conformational Ensembles of GPCRs.....	261
	A. Coarse Grain Simulation Methods for Mapping Conformational Changes upon Activation .....	261
	B. Computational Methods to Calculate Activation Pathways for GPCRs .....	264
	C. Need for Multiscale Methods to Map the Conformational Ensemble of GPCRs .....	267
IV.	Activation Mechanism of GPCRs .....	268
	A. Activation Mechanism of Class A GPCRs .....	268
	B. Role of Water in Receptor Activation .....	271
	C. Insights into the Role of Water from All-Atom MD Simulations.....	271
V.	Concluding Remarks.....	274
	References.....	275

## ABBREVIATIONS

$\beta$ 2-AR  $\beta$ 2-adrenergic receptor  
ENM elastic network model  
GPCR G-protein-coupled receptor

## ABSTRACT

G protein-coupled receptors (GPCRs) belong to a large superfamily of membrane proteins and they mediate many physiological and pathological processes in cell signaling. GPCRs exhibit remarkable structural homology in

spite of large diversity in their amino acid sequence and their function. The efficacy of an agonist depends on the nature of the molecule, as well the receptor and intracellular proteins that the receptor couples to. Many GPCRs show basal activity to various extents even in the absence of any stimulating ligands. They achieve fine modulation in signaling specificity through adapting an ensemble of conformations rather than a two-state model of inactive and active states. There is ample experimental evidence to show that GPCRs exist in an ensemble of conformations and binding of agonists, and the intracellular signaling proteins, such as the trimeric G-proteins, cooperatively activate and stabilize the active state of the receptor. Crystal structures of class A GPCRs have shown that the structure of the active state is different from the inactive state. The signaling specificity achieved by the activation process of GPCRs is determined not only by the lowest energy receptor state as in the crystal structure but also by the range of nearly degenerate conformational states that the receptor explores. Multiscale computational techniques play a key role in integrating the sparse and fragmented data obtained from experiments to map the potential energy landscape of the receptor, as well as the conformational ensemble of states. In this review, we demonstrate the power of the multiscale methods and delineate the need for further development of such multiscale computational methods to study the ensemble of inactive and active states for GPCRs. We review the insights into the receptor activation that emerged from a confluence of biophysical experimental as well as computational data.

## I. INTRODUCTION

G-protein-coupled receptors (GPCRs) are seven helical transmembrane (TM) proteins with remarkable structural homology given their diversity in sequence and function. The ligands that activate the receptor for cell signaling are known as agonists. The size and shape of the agonists for GPCRs vary from photon to small molecules, to proteins. Upon binding of an agonist, the receptor stabilizes a conformational state with high affinity for coupling either with G-proteins or with other proteins involved in G-protein-independent signaling pathways (Lefkowitz and Shenoy, 2005). The conformational state with both agonist and G-protein bound is the “active state” of the receptor. GPCRs are pleiotropic in function, and the G-protein-coupling signaling pathways lead to varied biological effects inside the cell (Hamm, 1998; Johnston and Siderovski, 2007). The non-G-protein-dependent pathways such as those mediated by  $\beta$ -arrestin

family of proteins lead to activation of MAPK signaling pathways inside the cell (Azzi et al., 2003; Shenoy et al., 2006; Violin and Lefkowitz, 2007). A given agonist could activate the same GPCR with varied efficacies for different signaling pathways depending on the protein that the receptor couples to in the intracellular region. This property of the agonist is called “functional selectivity” (Galandrin et al., 2007; Urban et al., 2007; Kenakin, 2008) or “biased agonism” (Kenakin, 2007). A ligand–receptor pair could achieve functional selectivity for different signaling pathways by changing the conformation of the receptor.

Many GPCRs exhibit “basal activity” even in the absence of any agonist leading to the conclusion that they are highly dynamic and adopt many conformations (Seifert and Wenzel-Seifert, 2002; Bond and Ijerman, 2006). The basal activity is suppressed upon binding of an inverse agonist, stabilizing an “inactive” conformation of the receptor. Thus the receptor could, in the absence of any ligand or any intracellular protein coupling, adopt an ensemble of conformational states (Vaidehi and Kenakin, 2010). Subsequent ligand and intracellular protein binding stabilizes a conformational state from this ensemble that leads to triggering a specific signaling pathway and a specific biological effect. Thus mapping the ensemble of receptor active and inactive conformational states is important in understanding the signal transduction mechanisms and in drug discovery.

Recently, the crystal structures of a thermostabilized turkey  $\beta 1$  adrenergic receptor ( $\beta 1$ -AR) with full and partial agonists bound have been solved (Warne et al., 2011). The receptor conformations with agonists bound exhibit marginal conformational changes upon agonist binding. These observations on  $\beta 1$ -AR structure are consistent with another crystal structure of human  $\beta 2$  adrenergic receptor ( $\beta 2$ -AR) bound covalently to a high-affinity agonist (Rosenbaum et al., 2011). However, the crystal structure of  $\beta 2$ -AR bound to an agonist and a G-protein mimic (nanobody) shows larger conformational changes in the intracellular region of the receptor where it couples to the G-protein (Rasmussen et al., 2011). The receptor conformation in this crystal structure could be the closest representation of the active state of a GPCR. Thus both the G-protein and the agonist binding are required to stabilize the receptor in the active state. The mechanism and dynamics of how the receptor goes from the inactive to the active state are still unknown. It is likely that agonist binding causes reshuffling of short-range interhelical contacts that trigger large-scale domain motions in the receptor leading to activation. Thus a combination

of both atomistic and coarse grain computational techniques is needed to delineate the activation mechanism.

Here, we review the current state of our knowledge in understanding the activation mechanism of GPCRs and offer insights into the activation pathways. In particular, we focus on the insights obtained from a combination of biophysical experimental results combined with multiscale computational methods used to understand the conformational ensembles and activation of GPCRs.

## II. CONFORMATIONAL FLEXIBILITY IN GPCRS

### A. *Experimental Evidence for Conformational Changes upon Activation of GPCRS*

There is ample experimental evidence in the literature demonstrating the conformational flexibility of GPCRs in the presence and absence of ligands (Ghanouni et al., 2001; Vaidehi 2010). The agonist-independent basal activity exhibited by many GPCRs (Seifert and Wenzel-Seifert, 2002) along with certain single point mutations (constitutively active mutants) that lead to a marked increase in basal activity (Cotecchia, 2007) clearly demonstrates that the receptor can sample active state conformations even in the absence of agonists. Biophysical studies on the visual receptor rhodopsin, using spin-labeling techniques, solid-state NMR, fluorescent spectroscopy, computational methods, and crystal structure of the partially active state opsin show several interhelical contacts also known as “conformational switches” made and broken upon activation (Krishna et al., 2002; Hubbell et al., 2003; Schertler, 2005; Park et al., 2008; Ahuja and Smith, 2009; Zaitseva et al., 2010). These conformational changes that happen during rhodopsin activation have been documented in previous reviews (Hubbell et al., 2003; Ahuja and Smith, 2009). Briefly, comparison of the rhodopsin and opsin crystal structures shows considerable conformational changes in the intracellular region of TM5 and TM6. TM5 also shows elongation of the helix in this region. Comparatively smaller changes have been observed in the extracellular loop 2 (ECL2) (Ahuja and Smith, 2009). There are substantial rearrangements in all the three intracellular loops in the opsin structure. Thus it is evident that activation is associated with considerable conformational changes in the receptor.

Comparison of the inactive and active state of the human  $\beta 2$ -AR crystal structures (Cherezov et al., 2007; Rosenbaum et al., 2011) shows side-chain rearrangements in the extracellular region especially on the three conserved serines on TM5 (shown in Fig. 1A). There is larger bulge on TM5 around P211<sup>5.50</sup> in the active state (shown in pink in Fig. 1A and B) compared to the inactive state (shown in green in Fig. 1A and B). Here, we have used GPCR-specific residue numbering (Ballesteros and Weinstein, 1995). The first number refers to the TM helix in which the

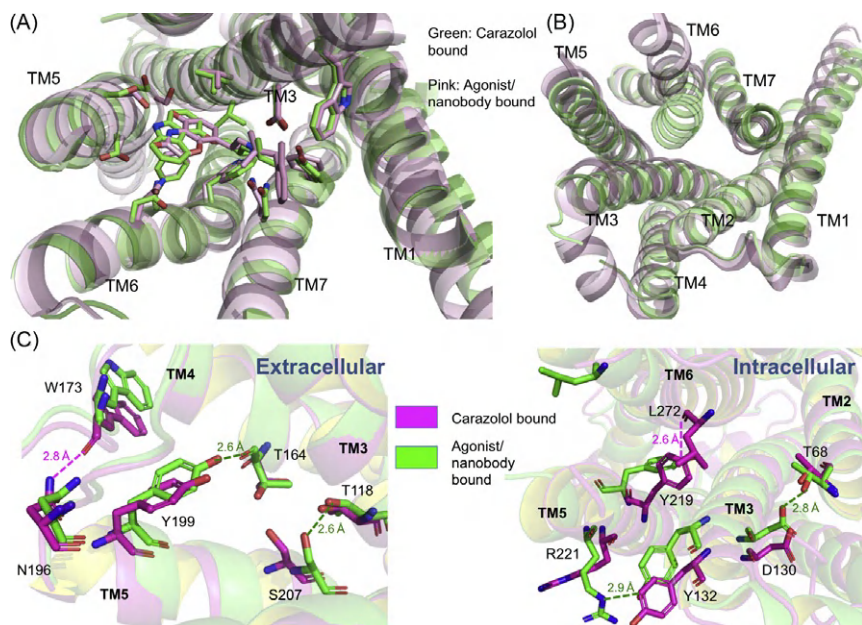


FIG. 1. Comparison of crystal structures of the partial inverse agonist-bound inactive state (pdb ID: 2RH1) and agonist/nanobody-bound active state of human  $\beta 2$ -AR (pdb ID: 3P0G). (A) Extracellular view of the ligand binding site showing small changes in the side chains of the residues in the binding site. The agonist and inverse agonist bind in the same region of the receptor; note the increase in bulge at the Pro on TM5 indicated by an arrow. (B) Intracellular view showing the large movement of transmembrane helix 6 at the G-protein binding site. (C) Breaking and making of hydrogen bond networks in the carazolol-bound and agonist/nanobody-bound crystal structures of human  $\beta 2$ -AR. The hydrogen bonds formed in the carazolol-bound structure are shown in magenta, while the ones in the nanobody-bound structure are shown in green.



residue is located, and the second number indicates the location of this residue with respect to the most conserved residue on that helix being numbered 50. TM1 shows a change in the tilt angle inwards toward the interior of the protein in the active state. [Figure 1B](#) shows conformational changes in the intracellular region of the protein. There is a large outward movement of TM6 similar to that of opsin but substantially larger (13.5 Å is the distance between the C $\alpha$  atoms of K267<sup>6,29</sup> on TM6 in the active and inactive state) movement than in opsin. There is an inward movement of TM3 and TM7. The intracellular loop 2 forms a helix in the active state, and it is unstructured in the inactive state. The change in helical arrangement in the active structure leads to the breaking of a few interhelical hydrogen bonds and formation of new ones ([Fig. 1C](#)). In the inactive crystal structure of  $\beta$ 2-AR, the side chain of Y219<sup>5,58</sup> on TM5 forms a hydrogen bond with the backbone of L272<sup>6,34</sup> on TM6. In the active crystal structure, this hydrogen bond is disrupted as the intracellular end of TM6 moves outward. Additionally, the hydrogen bond between N196<sup>5,35</sup> at the extracellular end of TM5 and W173 (backbone) on ECL2 is broken. The active state  $\beta$ 2-AR structure shows a few new hydrogen bonds that were absent in the inactive structure. S207<sup>5,46</sup> on TM5 forms a hydrogen bond with T118<sup>3,37</sup> on TM3 as well as with the hydroxyl group of the polycyclic moiety of the agonist BI-167107. Mutating S207<sup>5,46</sup> only affects agonist binding and not antagonist binding ([Rosenbaum et al., 2011](#)). Both S207<sup>5,46</sup> and T118<sup>3,37</sup> are conserved in many biogenic amine GPCRs. Thus S207–T118 hydrogen bond could be an important molecular switch for stabilizing the active state in several class A GPCRs. Another new hydrogen bond that is observed in the active  $\beta$ 2-AR structure is between Y199<sup>5,38</sup> on TM5 and backbone of T164<sup>4,56</sup> on TM4. Several new hydrogen bonds are formed near the intracellular interface of TM3, involving the DRY motif. D130<sup>3,49</sup> on TM3 forms a hydrogen bond with T68<sup>2,39</sup> on TM2, and Y132<sup>3,51</sup> forms a hydrogen bond with R221<sup>5,60</sup> on TM5. These hydrogen bonds stabilize the inward tilt of TM3 in the active conformation.

### *B. Dynamics and Conformational State Ensemble for GPCRs*

Kobilka and coworkers have measured the conformational changes upon activation of human  $\beta$ 2-AR purified in detergents, by agonists and partial agonists using bimane-labeled fluorescence spectroscopy ([Swaminath et al., 2005](#); [Yao et al., 2006](#)). Fluorescent tags attached to

select locations on the receptor are excited, and the resulting quenching of fluorescence is monitored. Depending on the local receptor environment, both the lifetime and intensity of the fluorescence can vary and this gives an estimate of the conformational changes in the receptor. Using a fluorophore attached to the intracellular end of TM6 in  $\beta$ 2-AR, it was found that in the absence of any ligand, the receptor shows a single population of fluorescence lifetime (Ghanouni et al., 2001). Antagonists reinforced this single fluorescence peak, whereas agonists showed an additional peak in fluorescence lifetime population distribution. This additional peak, which was absent for antagonists, is indicative of a distinct receptor conformation stabilized by the agonists and is indicative of the active state of the receptor. Moreover, the position of the peak differed between full and partial agonists suggesting that different agonists stabilize distinct receptor conformations. The above results serve as evidence for multiple active states of GPCRs. These analyses have shown that agonists with different efficacies could stabilize different conformations of the receptor.

More recently, Sunahara and coworkers have reconstituted the bimane-labeled monomeric human  $\beta$ 2-AR in high-density lipid nanoparticles that mimic the lipid environment in the cell more closely than the detergent solution and studied the effect of various agonists, inverse agonists, neutral antagonist, and G-proteins on the conformational states of the receptor (Yao et al., 2009). This study shows that the change in fluorescence intensity and frequency that reflect a receptor conformational change is similar upon full agonist or the Gs protein binding. However, the conformational changes upon binding of both full agonist and Gs protein binding are significantly larger than either one of them binding alone. These studies also showed that an inverse agonist prevents the receptor from forming a complex with the G-protein but does not disrupt the preformed receptor–G-protein complexes. Thus it is evident from these measurements that the receptor conformational states are slightly different depending on whether inverse agonist, or full agonist or the G-protein is bound. The receptor conformation is very different when both the G-protein and agonist are bound. This points to the fact that the receptor G-protein coupling is weak in the absence of the agonist (possibly the low-affinity state of the receptor) but gets substantially enhanced in the presence of the agonist (a high-affinity state of the receptor). These results are direct evidence that the receptor takes many conformational states

depending on the ligand. There exists a conformational ensemble of receptor states, a subset of which gets selected upon ligand and/or G-protein binding. Bouvier and coworkers have performed some elegant bioluminescence resonance energy transfer experiments in intact cells, showing that agonist binding is biphasic (Galés et al., 2005; Galandrin et al., 2007; Leduc et al., 2009), and kinetic fluorescence resonance energy transfer studies show a dependence of the efficacy of an agonist on the rate of conformational change in the GPCRs (Lohse et al., 2008).

### C. *Insights on Conformational Flexibility from Thermostable Mutants*

Tate and coworkers derived thermostable mutant GPCRs by mutating most of the residues in the receptor to alanine and selecting for either agonist or antagonist binding at elevated temperatures. Thus these mutants are chosen to stabilize the receptor conformation selectively in an agonist- or antagonist/inverse agonist-bound conformational state (Magnani et al., 2008; Serrano-Vega et al., 2008; Shibata et al., 2009; Tate 2010). An increase in thermal stability of 21 °C was achieved for turkey  $\beta$ 1-AR mutant called m23. This mutant was crystallized with an antagonist cyanopindolol subsequently (Warne et al., 2008). Recently, Warne et al. also crystallized other thermostable mutants with partial agonists, dobutamine and salbutamol, and full agonists, carmoterol and isoprenaline (Warne et al., 2011). However, the structural basis for the thermal stability of the mutants is not known from experiments. Molecular dynamics (MD) simulations of the wild-type and three thermostable mutants of  $\beta$ 1-AR showed that the flexibility of the receptor structures is similar in the wild-type as well as in the thermostable mutants. The stabilizing mutations stabilized the functional microdomains of the GPCR structure in an inactive conformation and hence provided more stability to the inactive conformation (Balaraman et al., 2010). Functional microdomains are conserved regions in the receptor structure that make interhelical contacts which break upon activation (Ballesteros et al., 1998). Thus while the overall receptor conformation of the thermostable mutant may still be flexible, the functional microdomains are constrained in their inactive state. These simulations clearly demonstrate that GPCRs exist in an ensemble of conformational states and the number of degenerate states in the ensemble is lowered upon thermostabilizing mutations, compared to the wild type (Balaraman et al., 2010).

### III. COMPUTATIONAL APPROACHES FOR STUDYING CONFORMATIONAL ENSEMBLES OF GPCRS

Computational methods contribute substantially to the understanding of the activation mechanism of GPCRs, as well as to map the receptor active and inactive conformational ensembles. These processes are not fully accessible by experimental methods. Given that GPCRs are highly dynamic, crystal structures are clearly inadequate to explain their functional behavior emerging from the receptor dynamics. However, crystal structures are an important starting point to studying the dynamics of the receptor using computational methods. Additionally, it is experimentally challenging to obtain the crystal structure or any structural information of the true active conformational state of the receptor. The structural information obtained from NMR and fluorescence spectroscopic techniques are sparse and hence require computational methods to integrate this experimental information to provide a molecular level understanding and rationalization of the receptor activation. In this section, we enumerate the computational methods and their use in bridging between structure and function of GPCRs. Specifically, we describe the computational methods used in predicting the active conformational state from the inactive state crystal structure as well as the activation pathway of GPCRs. We will exemplify the fact that multiscale methods, with a combination of coarse grain and fine grain all-atom methods, are required to understand the conformational ensembles of GPCRs.

#### A. *Coarse Grain Simulation Methods for Mapping Conformational Changes upon Activation*

The conformational changes effected by the agonist and G-protein binding lead to the activation of the receptor. Such processes happen in the range of microseconds, and molecular level insight into the conformational changes and the energetics of activation will greatly aid our understanding of the physiology of GPCRs. All-atom MD simulations in explicit lipid bilayer and water are commonly used to study the dynamics of membrane protein structures (Pitman et al., 2005; Isin et al., 2008; Khelashvili et al., 2008; Dror et al., 2009). For example, the recent microseconds of all-atom MD simulations starting from the inverse agonist-bound  $\beta$ 2-AR show that there are two possible inactive states for the

receptor (Dror et al., 2009). The results of all these simulations show no semblance even to the beginnings of the activation process. This is because all-atom MD simulations are inadequate to map the large-scale conformational changes resulting from activation of the receptor.

### 1. Targeted MD Methods

Targeted MD methods such as metadynamics that drive the receptor structure from the inactive state to the active state have been successful in capturing the gross features of activation (Laio and Parrinello, 2002; Provasi and Filizola, 2010). In metadynamics simulation, a Gaussian term is added to the potential energy to discourage the system from returning to already sampled states. This allows efficient sampling of rugged free-energy surfaces by facilitating barrier crossing. The pitfall of these methods is the bias that is used to drive the inactive conformational state to the active state. Rather than predicting the active state, these methods require prior knowledge of the receptor active state and are thus limited by the paucity of the available information. Also in metadynamics, the results are sensitive to the parameters of the driving force. Finally, a prudent selection of reaction coordinates (statistically independent conformational changes that define the activation events) is critical for identifying the intermediates and major steps toward activation. Using metadynamics and an active state model of rhodopsin based on the crystal structure of opsin, Filizola and coworkers have identified several intermediates along the activation pathway that are characterized by successive tilts of TM6 (Provasi and Filizola, 2010).

### 2. Elastic Network Models

The elastic network model (ENM) method is a coarse grained method for mapping the direction of movement along the low frequency modes without excessive computational cost (Bahar et al., 2010). In ENMs, the protein is modeled as a collection of beads connected by springs: here, beads refer to single or clusters of residues and springs represent the inter-residue contacts. The ENM method has been used to study the key events in the activation of rhodopsin. Using simulated thermal unfolding and ENM analysis, Rader et al. identified the regions in the rhodopsin structure that are responsible for maintaining the stability of the protein core that include domains near the retinal binding site and the conserved

disulfide bond between the cysteines on TM3 and ECL2 (Rader et al., 2004). In a later work, an ENM description of activated rhodopsin was developed, where the connections between the elastic nodes were based on experimentally obtained inter-residue distance constraints (Isin et al., 2006). Such models have been successful in predicting fluorescence decay rates of rhodopsin mutants and have identified residues that could play key roles in the activation process. Recently, Romo et al. compared ENMs to observations from microsecond MD simulations of several GPCRs (rhodopsin,  $\beta$ 2-AR, cannabinoid receptor 2) and showed that the parameters such as force constants and equilibrium distances in ENM could be optimized to improve predictions of the lower-frequency motions and thus reproduce results from very long timescale MD simulations (Romo and Grossfield, 2011).

### 3. *The Discrete Conformational Sampling Method for GPCRs*

The LITiCon computational method samples the receptor conformation in the coarse grain degrees of freedom, thus avoiding the built in bias in targeted MD methods (Bhattacharya et al., 2008a,b; Balaraman et al., 2010; Bhattacharya and Vaidehi, 2010). In the LITiCon method, the seven helices are treated as rigid bodies connected by flexible loops. The seven transmembrane helices are rotated in a desired range of rotation angles (typically the range is  $+40^\circ$  to  $-40^\circ$ ) in increment of  $\pm 5^\circ$  or smaller as desired by the user. The side-chain conformations are optimized for each backbone conformation generated using a rotamer library, and the potential energy is minimized using the all-atom forcefield function. This method generates an energy landscape for the GPCR in the rotational space of the TM helices. The local minima in the resulting energy landscape are identified, clustered, and sorted by total protein interaction energy and also by ligand binding energy. The global minimum state of this energy landscape is chosen as the most stable state of the protein with other nearly degenerate minima representing the ensemble of possible conformational states. Such coarse grain rigid body optimization techniques had been used to predict membrane protein structures (Filizola et al., 1998; Pappu et al., 1999; Vaidehi et al., 2002; Barth et al., 2009). In LITiCon, these coarse grain sampling techniques are used to predict the ensemble of active and inactive states starting from the inactive state crystal structure. Starting from the crystal structure of  $\beta$ 2-AR, the remodeling of the energy landscape of  $\beta$ 2-AR by inverse agonists and agonists of varied

efficacies was calculated using computational method, LITiCon (Bhattacharya and Vaidehi, 2010).

The energy landscape of inverse agonist (carazolol)-bound receptor ensemble shows a broad and deep potential well centered around the inactive state crystal structure, with high barriers to access the well-separated agonist-stabilized states that are located in an energetically unfavorable region outside this potential well. On the contrary, agonist-bound energy landscapes such as that of norepinephrine and epinephrine are highly flexible with a broad potential well of energetically favored states. The agonist-bound states show a favorable energy channel connecting the inverse agonist (carazolol)-bound conformational state to the norepinephrine or to the epinephrine-bound state (Fig. 2). This implies that the receptor is flexible and able to sample the inverse agonist states while bound to norepinephrine. Alternately, the inverse agonist carazolol trapped the receptor in the inactive conformation, making the agonist-bound states inaccessible, thus reducing the basal activity of the receptor. This reduction in basal activity could be due to the reduced affinity of the inactive state of the receptor toward G-protein (Galés et al., 2005).

### *B. Computational Methods to Calculate Activation Pathways for GPCRs*

A detailed mapping of the conformational transition pathway leading to GPCR activation is very difficult to achieve with experiments because all of the intermediate states are short lived. Conventional all-atom MD simulations are limited to short timescale processes as well as limited in the conformational search as shown by some microsecond simulations on GPCRs (Grossfield et al., 2008; Dror et al., 2009; Rasmussen et al., 2011). Therefore, the coarse grain approaches, presented in the section above, although lower in resolution compared to all-atom simulations, would facilitate crossing transition barriers between conformations and sample large conformational changes. Using biased all-atom molecular dynamics methods, Provasi and Filizola calculated the possible activation pathways of the GPCR, bovine rhodopsin starting from the crystal structure of a photoactivated deprotonated conformation of rhodopsin to the structure of partially activated ligand-free opsin in the presence of all-*trans* retinal. The free-energy landscape calculated along the trajectories using a path collective variable approach shows that the inactive and partially active opsin states are connected by at least two different pathways, with

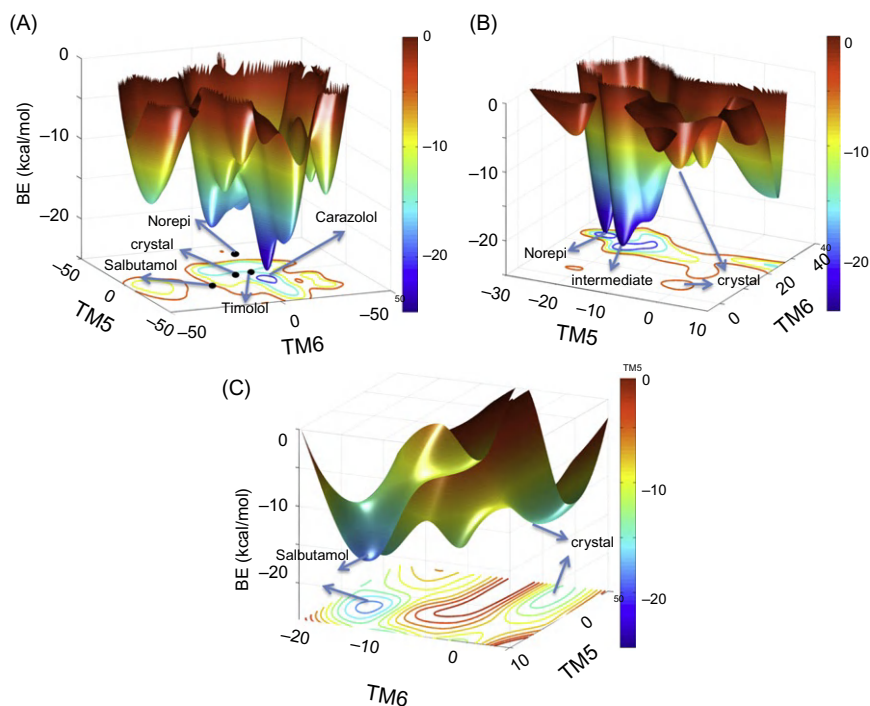


FIG. 2. Binding energy surfaces of human  $\beta_2$ -AR with (A) partial inverse agonist carazolol; (B) full agonist norepinephrine; (C) partial agonist salbutamol. The X and Y axes represent the rotation angles of transmembrane helices 5 (TM5) and 6 (TM6) in degrees. The various predicted ligand-stabilized states are marked on the landscapes. Norepinephrine is abbreviated as norepi.

at least four metastable states characterized by different levels of outward movement of TM6 along the pathway of activation.

Bhattacharya and Vaidehi alternatively developed a Monte Carlo method to calculate the pathway going from inactive state to the agonist-bound state (Bhattacharya et al., 2008a,b; Bhattacharya and Vaidehi, 2010). Starting from the crystal structure of the inactive state of  $\beta_2$ -AR, the ligand-stabilized conformations of various agonists, partial agonists, and inverse agonist were predicted. The Monte Carlo method was used to search for the minimum energy pathway going from the inactive state to the ligand-stabilized states (shown in Fig. 3A). The percentage



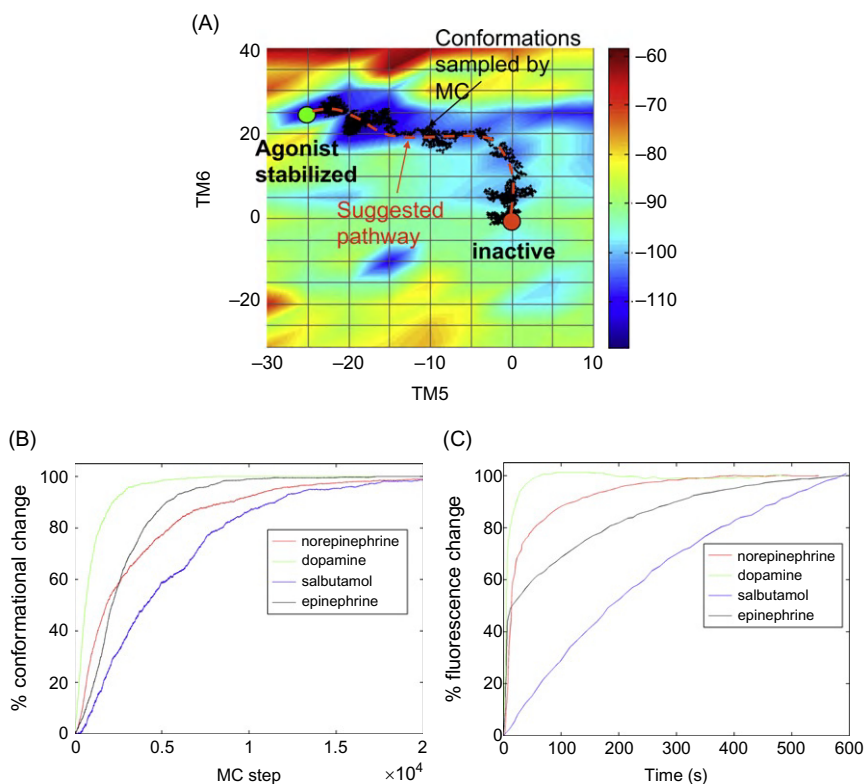


FIG. 3. (A) The minimum energy activation pathway calculated using Monte Carlo method is shown in black dots on the two-dimensional binding energy landscape for the agonist norepinephrine-bound  $\beta 2$ -AR. The energy unit is kcal/mol. The axis coordinates represent the rotation angles of the helices (degrees) relative to the inactive state. The red and the green circles denote the inactive and the agonists stabilized states, respectively. (B) Comparison of the calculated % change in conformation upon ligand binding to the fluorescence intensity measurement over time, for the  $\beta 2$ -AR agonists.

conformational change calculated along the minimum energy pathway correlates well with the fluorescent intensity lifetime measurements on  $\beta 2$ -AR as shown in Fig. 3B. Further, the receptor conformations generated along the minimum energy pathway can be used to map the

conformational ensemble of several kinetic intermediates as well as ligand-stabilized states.

*C. Need for Multiscale Methods to Map the Conformational Ensemble of GPCRs*

Targeted MD approaches sample wider conformational space and can lead to active state conformation starting from the inactive state. These approaches lead to a simulated activation pathway that could provide insight into intermediate structures along the pathway. The caveat is that these simulations are limited by the knowledge of the active state structure. However, the LITiCon method that samples based on coarse grain degrees of freedom avoids any such bias. However, the limitations are that the energy landscape is generated on a coarse grain grid search and could miss significant barriers along the activation pathway. Multiple all-atom MD simulations in explicit lipid and water, starting from various structures along the coarse grain minimum energy pathway from the Monte Carlo-LITiCon procedure, could provide a powerful approach to map the conformational ensemble of GPCRs in the active and inactive states. Roux and coworkers have studied the conformational transition in Src activation by generating a swarm of all-atom MD trajectories starting from targeted MD simulation structures along the activation pathway of Src (Yang et al., 2009).

Starting from various LITiCon generated coarse grain structures along the minimum energy activation pathway of human  $\beta$ 2-AR, Niesen et al. performed multiple all-atom MD simulations to generate a swarm of trajectories to map the conformational ensemble sampled by the receptor without any ligand and the receptor with agonists of varied efficacies. They analyzed the conformational space sampled by the  $\beta$ 2-AR in the absence of any ligand and in the presence of full agonist norepinephrine, partial agonist salbutamol, and partial inverse agonist carazolol. These results were calculated using principal component analysis on the swarm of all-atom MD trajectories starting from various conformations along the minimum energy pathway calculated from coarse grain method LITiCon. The apoprotein (without any ligand bound) samples a wider range of conformational space than any ligand-bound receptor. Inverse agonist-bound receptor conformations are confined to a subset of smaller conformational space, while agonist norepinephrine bound shows wider conformational span. The crystal structures of

the agonist and inverse agonist-bound conformations  $\beta$ 2-AR are located in the most densely populated regions in this conformational ensemble. This is clear evidence that the binding of an agonist leads to conformational selection of the various low-energy minima that are sampled by the aporeceptor (Niesen et al., 2011).

#### IV. ACTIVATION MECHANISM OF GPCRS

##### A. *Activation Mechanism of Class A GPCRS*

The activation of GPCRS proceeds through a discrete set of conformational changes and intermediates. One of the major goals is to understand the activation pathway by mapping all these intermediates using both experimental techniques and computational methods. A detailed mapping of the conformational transition pathway and the conformational substates leading to GPCR activation is very difficult to achieve with experiments because many of the intermediate states are short lived and not accessible in the experimental timescales. In the previous section, we introduced the computational methods such as targeted MD and directed Monte Carlo methods that have been used to calculate the minimum energy pathway of activation.

Using site-directed spin labeling, the change of mobility of different residues on TM6 upon rhodopsin photoactivation were measured (Dunham and Farrens, 1999; Hubbell et al., 2003). They identified that a well-conserved interhelical salt bridge between TM3 and TM6 (popularly known as the “ionic lock” in the GPCR community) breaks upon activation of rhodopsin (Farrens et al., 1996). The breaking of the ionic lock in rhodopsin leads to the flexibility of helix 6, and hence, it moves away from helix 3 in the intracellular part of the receptor. Overall, these changes can be interpreted as a complex motion of TM6, combining both rotation and outward tilt, as discussed by the authors. The spin-labeling results also suggest a lack of deformation in the helices, since none of the spin labels on the outer surfaces of the helices showed any change in mobility. This indicates that the conformational changes in rhodopsin involve mainly rigid body helical movements. Results from other spin-labeled residues indicate a movement of TM2 toward TM4. The conformational change in TM6 is further corroborated by the interhelical cross-linking

experiments using disulfide and zinc binding (Schwartz et al., 2006). Cross-links involving the intracellular end of TM6 blocked transducin activation, whereas cross-links toward the extracellular end retained activity.

The inter-residue distance constraints obtained from the above experiments can be used in MD simulations in driving the conformational changes involved in activation. These simulation results provide a molecular level visualization of the activation process. For example, distance measurements from NMR on rhodopsin have been used as constraints in MD simulations by Smith and coworkers to uncover the mechanism of activation of rhodopsin (Hornak et al., 2010). These studies showed a displacement of ECL2 from all-*trans* retinal upon photoactivation (Ahuja and Smith, 2009). It was suggested that steric interaction with the C19 methyl group of retinal triggered the conformational change in ECL2, since removing the C19 methyl group prevented activation. NMR data also showed a rearrangement of the hydrogen bond network near TM3–TM5 interface, where the hydrogen bond between H211<sup>5,46</sup> backbone and E122<sup>3,37</sup> is disrupted and a new hydrogen bond is formed between H211<sup>5,46</sup> side chain and E122<sup>3,37</sup>. Recently, Ye et al. used Fourier transform IR spectroscopy on azido-Phe incorporated rhodopsin to study the early movements of TM6 in going through the various substates of rhodopsin activation (Ye et al., 2010). They observed that helix 6 undergoes an anticlockwise rotation (when viewed from the extracellular region) in the early stages of activation much before the fully active state is reached. Thus there are several receptor substates that form an ensemble in the pathway to activation. The ionic lock is one of the features of the inactive state of rhodopsin that is broken upon activation. Although the residues that make the ionic lock between helices 3 and 6 are highly conserved across all class A GPCRs, the ionic lock is not present in any of the other class A GPCR crystal structures of the inactive state. This could be because 11-*cis*-retinal, the inverse agonist for rhodopsin, is covalently linked to the receptor, thus completely annulling the receptor basal activity. All other class A GPCRs bind to diffusible ligands and show some level of basal activity. In these receptors, the inactive state is again an ensemble of states, as evidenced from all-atom MD simulations of  $\beta$ 2-AR (Dror et al., 2009; Balaraman et al., 2010). Thus the tight ionic lock could be a feature of completely quiet receptor state.

While information on rhodopsin activation was obtained using spin labeling, NMR, and disulfide and zinc cross-linking, the activation dynamics of another class A GPCR, the human  $\beta 2$ -AR, was studied using fluorescence lifetime measurements pioneered by Kobilka and coworkers. Using a fluorophore attached to the intracellular end of TM6 in  $\beta 2$ -AR, time-resolved fluorescence spectroscopy was used to record the change in fluorescence intensity with time upon agonist binding (Swaminath et al., 2004, 2005). While full agonists epinephrine and norepinephrine showed a biphasic change in fluorescence (fast change followed by a slower change), partial agonists, dopamine and salbutamol, showed monophasic fast and slow changes, respectively. These results indicate that the full and partial agonists stabilize distinct active states and follow different pathways toward activation.

Besides fluorescence, NMR spectroscopy has been used to probe conformational changes to the ECLs of  $\beta 2$ -AR, near the salt bridge between K305<sup>7.33</sup> on TM7 and D192 on ECL2 (Bokoch et al., 2010). This salt bridge is observed in the crystal structure of inverse agonist-bound  $\beta 2$ -AR. NMR spectroscopy using <sup>13</sup>C-labeled methylated lysine side chains showed a change in chemical environment of the K305<sup>7.33</sup> side chain (in comparison to apoprotein) upon inverse agonist binding. Antagonist binding, however, showed a minimal change, while agonist binding leads to a weakening of the salt bridge. These results imply that agonists bound to the orthosteric site can modulate the ECL conformations and vice versa. The crystal structures of the inactive state of the GPCRs show the conserved water molecules near the highly conserved functional microdomains of the receptor: D2.50 on TM2, the conserved WxP motif in TM6 and NPxxY motif on TM7. Since these residues are conserved among all class A GPCRs, the water molecules near them could be important to the structural stability of the receptors and possibly the activation process as well. Notably, none of these water molecules are observed in the crystal structure of the active state of  $\beta 2$ -AR. The conserved waters in this structure are either absent or not well resolved. The water molecules are proposed to serve as bridges in communicating the activation signal from the ligand binding site to the G-protein binding site at the intracellular surface of the receptors (Angel et al., 2009b). Recently, experiments and computational studies have uncovered the role of water in the activation of rhodopsin, as discussed in the next section. Future experiments using methods such as radiolytic protein footprinting will delineate the

role of water molecules in the activation of  $\beta$ 2-AR and other GPCRs (Angel et al., 2009b; Orban et al., 2010).

### B. *Role of Water in Receptor Activation*

There are emerging studies highlighting the role of water molecules on activation of rhodopsin and other class A GPCRs. Angel et al. (2009a) analyzed the water molecules observed in the available GPCR crystal structures and showed that there is increased disorder as well as rearrangement of water molecules upon activation. Angel et al. also labeled selected residues in the TM region of rhodopsin using radiolytic footprinting method to identify the differences in the position of tightly bound waters in the inactive and active state of rhodopsin (Angel et al., 2009b). This analysis showed that activation of the receptor is accompanied by rearrangements in tightly bound waters and these waters could act as allosteric communicators of signals from outside the TM region to the cytoplasmic region. Changes in rates of oxidation observed when comparing ground state and activated receptor reflect local structural changes upon formation of both Meta II and opsin. Using a combination of MD simulations and magic angle spinning NMR, Grossfield et al. showed that water molecules interact with several conserved residues showing that they play an important role in activation of rhodopsin.

### C. *Insights into the Role of Water from All-Atom MD Simulations*

In rhodopsin and other GPCRs, modulation of the helical kink caused by conserved proline residue in TM6 is proposed to play a role in the activation of the receptor (Shi et al., 2002). Based on previous Monte Carlo simulations (Shi et al., 2002), it has been speculated that the modulation of the proline kink in TM6 could be the result of the change in the side-chain rotamer of the nearby highly conserved W265<sup>6,48</sup>. This movement of the side chain of W265<sup>6,48</sup> upon activation has not been observed in any of the GPCR crystal structures of the active or partially active states so far but has been observed in the NMR measurements of rhodopsin activation (Ahuja and Smith, 2009). It is possible that this effect is dynamic and can be observed only in solution. In the MD simulation of the *trans*-retinal-bound activated rhodopsin, Bhattacharya et al. observed a strong correlation between the rotamer angle of W265<sup>6,48</sup> and the kink

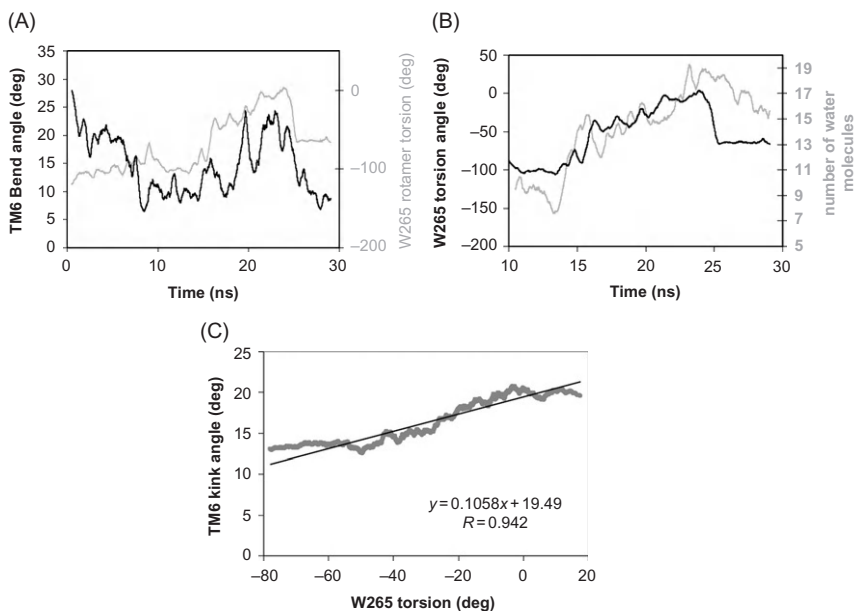


FIG. 4. Dynamics of residues are important for activation of GPCRs. Methods such as NMR and MD simulations are required to observe this dynamics. Shown in this figure is the dynamics of the side chain of W265<sup>6,48</sup> that modulates the helical kink in TM6, as seen in the MD simulations. Water plays a critical role in assisting this dynamics. (A) The torsional angle (torsion is measured between the C<sub>β</sub> and C<sub>γ</sub> atoms of W265<sup>6,48</sup>) of W265<sup>6,48</sup> and the helical kink angle of TM6 as function of time; (B) the torsion angle of the side chain of W265<sup>6,48</sup> and number of water molecules in the vicinity of the conserved proline that causes the helical kink, as function of time; (C) TM6 kink angle as function of the torsion angle of the side chain of W265<sup>6,48</sup>. The gray curve represents the actual data (block averaged over 0.5 ns), whereas the dark line represents the best fit trend line. The equation of the fitted trend line and the correlation coefficient are shown in the box. This shows a correlation between the dynamics of the side chain of W265<sup>6,48</sup> and the helical kink.

angle of TM6 hinged on the P267<sup>6,50</sup> as shown in Fig. 4. The change in the side-chain rotamer of W265<sup>6,48</sup> is assisted by a number of water molecules that have penetrated the region as shown in Fig. 5. The diffusion of water into the protein cavity weakens the hydrogen bond between H211<sup>5,46</sup> and W265<sup>6,48</sup> and assists in the movement of the side chain of W265<sup>6,48</sup>. The energy required for breaking this hydrogen bond comes from the formation of new water mediated hydrogen bonds between the neighboring

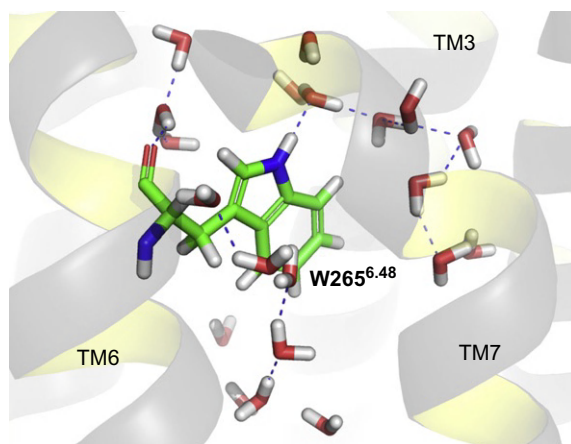


FIG. 5. The dynamics of the side-chain conformation of  $W265^{6.48}$  is assisted by water. Shown in this figure are water molecules that form a “water wire” with  $W265^{6.48}$  and TM7. Formation of this water wire facilitates the side-chain movement of  $W265^{6.48}$ .

water molecules and both the residues  $H211^{5.46}$  and  $W265^{6.48}$ . During the course of the dynamics, the side chain of  $W265^{6.48}$  traverses through several intermediate states such as the ones shown in the snapshots in Fig. 5. These intermediate side-chain conformations are stabilized by the formation of hydrogen bonds with the neighboring waters. When the  $W265^{6.48}$  side chain moves from one intermediate state to the next, a few water-mediated hydrogen bonds are broken and new ones are formed to stabilize the rotamer in the next intermediate state. Therefore, water directly assists in the movement of the side chain of  $W265^{6.48}$  by providing stabilizing contacts to all the intermediate rotamer conformations and in lowering the energy barrier involved in the side-chain movement of  $W265^{6.48}$  upon activation.

Several MD studies have highlighted the role of water in the activation of rhodopsin and other GPCRs. During 1.5  $\mu$ s of MD simulations involving all-*trans* retinal-bound rhodopsin, there was a large influx of water into the transmembrane region (Grossfield et al., 2008). Using NMR, this increase in internal hydration was attributed to the Meta-I intermediate. MD simulations on squid rhodopsin showed a continuous water channel to extend from the retinal binding site to the helical proline kink of TM6 (Jardón-Valadez et al., 2008). Additionally, the network of water-mediated



hydrogen bonds between highly conserved residues Y306<sup>7.53</sup> of NPxxY motif on TM7, D83<sup>2.50</sup> on TM2, and W265<sup>6.48</sup> of the WxP motif on TM6 was stable throughout the simulation. The water channel and the water-mediated hydrogen bond network have been hypothesized to relay the signal from the retinal isomerization site to the intracellular interface of the receptor. Microsecond MD simulations on the cannabinoid CB2 receptor have shown that agonist entry into the binding site triggers a conformational change to the TM6–TM7 interface followed by an influx of water into the transmembrane cavity connecting the ligand binding domain to the intracellular surface of the receptor (Hurst et al., 2010). The role of water in modulating the movement of the side chain of W386<sup>6.48</sup> was studied using microsecond scale MD and metadynamics (Selent et al., 2010). Hydrated sodium ions were shown to bind to an allosteric site between the conserved aspartate D80<sup>2.50</sup> on TM2, the NPxxY motif on TM7, an S121<sup>3.39</sup> on TM3, and W386<sup>6.48</sup> on TM6 in the human dopamine D2 receptor. During the course of MD, water molecules clustered around the positively charged sodium ion modulated the side-chain movement of the W386<sup>6.48</sup> residue. The above results collectively suggest that transmembrane water molecules help in transducing the activation signal from the ligand binding site to the functional microdomains and to the G-protein binding interface of the receptor. Additionally, water can aid the diffusion of soluble ligands into the binding cavity.

## V. CONCLUDING REMARKS

GPCRs are highly dynamic and exist in an ensemble of conformations even in the absence of any ligand. Ligand binding leads to conformational selection and selective stabilization of certain conformations, thus causing a population shift. There is further shift in population of various states upon binding of intracellular proteins such as the G-proteins or  $\beta$ -arrestin. Existence of the ensemble of receptor conformations is consistent with the fact that the same ligand shows varied efficacies to different downstream effectors or different signaling pathways. This has been termed as “functional selectivity” or “biased agonism.” The crystal structures are only one of the low-energy conformations in the ensemble, and the signaling specificity achieved by the activation process of GPCRs is determined not only by the lowest energy receptor state as in the crystal structure but also by the range of nearly degenerate conformational states that the

receptor explores. However, crystal structures serve as an excellent starting conformation point for mapping the ensemble of conformations describing the dynamics and efficacy.

The receptor gets stabilized in the active state only in the presence of both the agonist and the G-protein as revealed in the recent crystal structure of  $\beta$ 2-AR with agonist and nanobody (mimic of the G-protein) bound. This crystal structure in conjunction with the results from other biophysical experiments on purified proteins shows that TM6 undergoes large outward movement upon activation. Water plays an important role in lowering the energy barriers for breaking of interhelical hydrogen bonds and making of new ones.

Computational methods have been used extensively in modeling GPCR structure and dynamics. It is amply clear that long timescale all-atom MD simulations starting from any single state, such as the inactive state of the receptor, do not map the entire ensemble of conformational states that the receptor would sample, including the active state. We have clearly demonstrated that coarse grain methods are useful in mapping multiple conformational states of the receptors although these conformations are of low resolution. Combining the coarse grain methods, with all-atom MD simulations, gives a better sampling of the conformational ensemble of the receptor as well as a description of the activation process. The multiscale computational methods, in conjunction with the crystal structures of GPCRs, can be used to successfully address the problem of receptor flexibility in functionally specific drug design.

#### REFERENCES

- Ahuja, S., Smith, S. O. (2009). Multiple switches in G protein-coupled receptor activation. *Trends Pharmacol. Sci.* **30**, 494–502.
- Angel, T. E., Chance, M. R., Palczewski, K. (2009a). Conserved waters mediate structural and functional activation of family A (rhodopsin-like) G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **106**, 8555–8560.
- Angel, T. E., Gupta, S., Jastrzebska, B., Palczewski, K., Chance, M. R. (2009b). Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proc. Natl. Acad. Sci. USA* **106**, 14367–14372.
- Azzi, M., Charest, P. G., Angers, S., Rousseau, G., Kohout, T., Bouvier, M., et al. (2003). Beta-arrestin-mediated activation of MAPK by inverse agonists reveals distinct active conformations for G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **100**, 11406–11411.

- Bahar, I., Lezon, T. R., Bakan, A., Shrivastava, I. H. (2010). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497.
- Balaraman, G. S., Bhattacharya, S., Vaidehi, N. (2010). Structural insights into conformational stability of wild-type and mutant  $\beta_1$ -adrenergic receptor. *Biophys. J.* **99**, 568–577.
- Ballesteros, J. A., Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25**, 366–428.
- Ballesteros, J., Kitanovic, S., Guarnieri, F., Davies, P., Fromme, B. J., Konvicka, K., et al. (1998). Functional microdomains in G-protein-coupled receptors: the conserved arginine-cage motif in the gonadotropin-releasing hormone receptor *J. Biol. Chem.* **273**, 10445–10453.
- Barth, P., Wallner, B., Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. *Proc. Natl. Acad. Sci. USA* **106**, 1409–1414.
- Bhattacharya, S., Vaidehi, N. (2010). Computational mapping of the conformational transitions in agonist selective pathways of a G-protein coupled receptor. *J. Am. Chem. Soc.* **132**, 5205–5214.
- Bhattacharya, S., Hall, S. E., Li, H., Vaidehi, N. (2008a). Ligand-stabilized conformational states of human  $\beta_2$  adrenergic receptor: insight into G-protein-coupled receptor activation. *Biophys. J.* **94**, 2027–2042.
- Bhattacharya, S., Hall, S. E., Vaidehi, N. (2008b). Agonist induced conformational changes in bovine rhodopsin: insight into activation of G-protein coupled receptors. *J. Mol. Biol.* **382**, 539–555.
- Bokoch, M. P., Zou, Y., Rasmussen, S. G. F., Liu, C. W., Nygaard, R., Rosenbaum, D. M., et al. (2010). Ligand-specific regulation of the extracellular surface of a G protein coupled receptor. *Nature* **463**, 108–112.
- Bond, R. A., Ijerman, A. P. (2006). Recent developments in constitutive receptor activity and inverse agonism, and their potential for GPCR drug discovery. *Trends Pharmacol. Sci.* **27**, 92–96.
- Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, G. F., Thian, F. S., Kobilka, T. S., et al. (2007). High-resolution crystal structure of an engineered human  $\beta_2$ -adrenergic G protein-coupled receptor. *Science* **318**, 1258–1265.
- Cotecchia, S. (2007). Constitutive activity and inverse agonism at the  $\alpha_1$  adrenoreceptors. *Biochem. Pharmacol.* **73**, 1076–1083.
- Dror, R. O., Arlow, D. H., Borhani, D. W., Jensen, M.Ø., Piana, S., Shaw, D. E. (2009). Identification of two distinct inactive conformations of the  $\beta_2$ -adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl. Acad. Sci. USA* **106**, 4689–4694.
- Dunham, T. D., Farrens, D. L. (1999). Conformational changes in rhodopsin. Movement of helix F detected by site-specific chemical labeling and fluorescence spectroscopy. *J. Biol. Chem.* **274**, 1683–1690.
- Farrens, D. L., Altenbach, C., Yang, K., Hubbell, W. L., Khorana, H. G. (1996). Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science* **274**, 768–770.

- Filizola, M., Perez, J. J., Carteni-Farina, M. (1998). BUNDLE: a program for building transmembrane domains of G-protein coupled receptors. *J. Comput. Aid. Mol. Des.* **12**, 111–118.
- Galandrin, S., Oligny-Longpr, G., Bouvier, M. (2007). The evasive nature of drug efficacy: implications for drug discovery. *Trends Pharmacol. Sci.* **28**, 423–430.
- Galés, C., Rebols, R. V., Hogue, M., Trieu, P., Brelt, A., Hébert, T. E., et al. (2005). Real-time monitoring of receptor and G-protein interactions in living cells. *Nat. Methods* **2**, 177–184.
- Ghanouni, P., Gryczynski, Z., Steenhuis, J. J., Lee, T. W., Farrens, D. L., Lakowicz, J. R., et al. (2001). Functionally different agonists induce distinct conformations in the G protein coupling domain of the  $\beta_2$  adrenergic receptor. *J. Biol. Chem.* **276**, 24433–24436.
- Grossfield, A., Pitman, M. C., Feller, S. E., Soubias, O., Gawrisch, K. (2008). Internal hydration increases during activation of the G-protein-coupled receptor rhodopsin. *J. Mol. Biol.* **381**, 478–486.
- Hamm, H. E. (1998). The many faces of G protein signaling. *J. Biol. Chem.* **273**, 669–672.
- Hornak, V., Ahuja, S., Eilers, M., Reeves, P. J., Sheves, M., Smith, S. O. (2010). A view of the activated state of rhodopsin from guided molecular dynamics simulations. *J. Mol. Biol.* **396**, 510–527.
- Hubbell, W. L., Altenbach, C., Hubbell, C. M., Khorana, H. G. (2003). Rhodopsin structure, dynamics, and activation: a perspective from crystallography, site directed spin labeling, sulfhydryl reactivity, and disulfide cross-linking. *Adv. Protein Chem.* **63**, 243–290.
- Hurst, D. P., Grossfield, A., Lynch, D. L., Feller, S., Romo, T. D., Gawrisch, K., et al. (2010). A lipid pathway for ligand binding is necessary for a cannabinoid g protein-coupled receptor. *J. Biol. Chem.* **285**, 17954–17964.
- Isin, B., Rader, A. J., Dhiman, H. K., Klein-Seetharaman, J., Bahar, I. (2006). Predisposition of the dark state of rhodopsin to functional changes in structure. *Proteins* **65**, 970–983.
- Isin, B., Schulten, K., Tajkhorshid, E., Bahar, I. (2008). Mechanism of signal propagation upon retinal isomerization: insights from molecular dynamics simulations of rhodopsin restrained by normal modes. *Biophys. J.* **95**, 789–803.
- Jardón-Valadez, E., Bondar, A.-N., Tobias, D. J. (2008). Dynamics of the internal water molecules in squid rhodopsin. *Biophys. J.* **96**, 2572–2576.
- Johnston, C. A., Siderovski, D. P. (2007). Receptor-mediated activation of heterotrimeric G-proteins: current structural insights. *Mol. Pharmacol.* **72**, 219–230.
- Kenakin, T. (2007). Functional selectivity through protean and biased agonism: who steers the ship? *Mol. Pharmacol.* **72**, 1393–1401.
- Kenakin, T. (2008). Functional selectivity in GPCR modulator screening. *Comb. Chem. High Throughput Screen.* **11**, 337–343.
- Khelashvili, G., Grossfield, A., Feller, S. E., Pitman, M. C., Weinstein, H. (2008). Structural and dynamic effects of cholesterol at preferred sites of interaction with rhodopsin identified from microsecond length molecular dynamics simulations. *Proteins* **76**, 403–417.

- Krishna, A., Menon, S. T., Terry, T. J., Sakmar, T. P. (2002). Evidence that helix 8 of rhodopsin acts as a membrane-dependent conformational switch. *Biochemistry* **41**, 8298–8309.
- Laio, A., Parrinello, M. (2002). Escaping free energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566.
- Leduc, M., Breton, B., Gals, C., Le Gouill, C., Bouvier, M., Chemtob, S., et al. (2009). Functional selectivity of natural and synthetic prostaglandin EP4 receptor ligands. *Pharmacol. Exp. Ther.* **331**, 297–307.
- Lefkowitz, R. J., Shenoy, S. K. (2005). Transduction of receptor signals by beta-arrestins. *Science* **308**, 512–517.
- Lohse, M. J., Nikolaev, V. O., Hein, P., Hoffmann, C., Vilardaga, J.-P., Bünemann, M. (2008). Optical techniques to analyze real-time activation and signaling of G-protein-coupled receptors. *Trends Pharmacol. Sci.* **29**, 159–165.
- Magnani, F., Shibata, Y., Serrano-Vega, M. J., Tate, C. G. (2008). Co-evolving stability and conformational homogeneity of the human adenosine A<sub>2a</sub> receptor. *Proc. Natl. Acad. Sci. USA* **105**, 10744–10749.
- Niesen, M., Bhattacharya, S., Vaidehi, N. (2011). Conformational selection upon ligand binding in G-protein coupled receptors. *J. Am. Chem. Soc.* accepted.
- Orban, T., Gupta, S., Palczewski, K., Chance, M. R. (2010). Visualizing water molecules in transmembrane proteins. *Biochemistry* **49**, 827–834.
- Pappu, R. V., Marshall, G. R., Ponder, J. W. (1999). A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Biol.* **6**, 50–55.
- Park, J. H., Scheerer, P., Hoffman, K. P., Choe, H.-W., Ernst, O. P. (2008). Crystal structure of the ligand-free G-protein coupled receptor opsin. *Nature* **454**, 183–187.
- Pitman, M. C., Grossfield, A., Suits, F., Feller, S. E. (2005). Role of cholesterol and polyunsaturated chains in lipid–protein interactions: molecular dynamics simulation of rhodopsin in a realistic membrane environment. *J. Am. Chem. Soc.* **127**, 4576–4577.
- Provasi, D., Filizola, M. (2010). Putative active states of a prototypic G-protein-coupled receptor from biased molecular dynamics. *Biophys. J.* **98**, 2347–2355.
- Rader, A. J., Anderson, G., Isin, B., Gobind Khorana, H., Bahar, I., Klein-Seetharaman, J. (2004). Identification of core amino acids stabilizing rhodopsin. *Proc. Natl. Acad. Sci. USA* **101**, 7246–7251.
- Rasmussen, S. G., Choi, H. J., Fung, J. J., Pardon, E., Casarosa, P., Chae, P. S., et al. (2011). Structure of a nanobody-stabilized active state of the  $\beta_2$  adrenoceptor. *Nature* **469**, 175–180.
- Romo, T. D., Grossfield, A. (2011). Validating and improving elastic network models with molecular dynamics simulations. *Proteins* **79**, 23–24.
- Rosenbaum, D. M., Zhang, Z., Lyons, J. A., Holl, R., Aragao, D., Arlow, D. H., et al. (2011). Structure and function of an irreversible agonist– $\beta_2$  adrenoceptor complex. *Nature* **469**, 236–240.
- Schertler, G. F. X. (2005). Structure of rhodopsin and the metarhodopsin I photo-intermediate. *Curr. Opin. Struct. Biol.* **15**, 408–415.

- Schwartz, T. W., Frimurer, T. M., Holst, B., Rosenkilde, M. M., Elling, C. E. (2006). Molecular mechanism of 7TM receptor activation—a global toggle switch model. *Annu. Rev. Pharmacol. Toxicol.* **2006**(46), 481–519.
- Seifert, R., Wenzel-Seifert, K. (2002). Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **366**, 381–416.
- Selent, J., Sanz, F., Pastor, M., De Fabritiis, G. (2010). Induced effects of sodium ions on dopaminergic G-protein coupled receptors. *PLoS Comput. Biol.* **6**, e1000884.
- Serrano-Vega, M. J., Magnani, F., Shibata, Y., Tate, C. G. (2008). Conformational thermostabilization of  $\beta_1$ -adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci. USA* **105**, 877–882.
- Shenoy, S. K., Drake, M. T., Nelson, C. D., Houtz, D. A., Xiao, K., Madabushi, S., et al. (2006). Beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. *J. Biol. Chem.* **281**, 1261–1273.
- Shi, L., Liapakis, G., Xu, R., Guarnieri, F., Ballesteros, J. A., Javitch, J. A. (2002).  $\beta_2$  adrenergic receptor activation. Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. *J. Biol. Chem.* **277**, 40989–40996.
- Shibata, Y., White, J. F., Serrano-Vega, M. J., Magnani, F., Aloia, A. L., Grisshammer, R., et al. (2009). Thermostabilization of the neurotensin receptor NTS1. *J. Mol. Biol.* **390**, 262–277.
- Swaminath, G., Xiang, Y., Lee, T. W., Steenhuis, J., Parnot, C., Kobilka, B. K. (2004). Sequential binding of agonists to the beta<sub>2</sub> adrenoceptor: kinetic evidence for intermediate conformational states. *J. Biol. Chem.* **279**, 686–691.
- Swaminath, G., Deupi, X., Lee, T. W., Zhu, W., Thian, F. S., Kobilka, T. S., et al. (2005). Probing the beta<sub>2</sub> adrenoceptor binding site with catechol reveals differences in binding and activation by agonists and partial agonists. *J. Biol. Chem.* **280**, 22165–22171.
- Tate, C. G. (2010). Practical considerations of membrane protein instability during purification and crystallization. *Methods Mol. Biol.* **601**, 187–203.
- Urban, J. D., Clarke, W. P., von Zastrow, M., Nichols, D. E., Kobilka, B., Weinstein, H., et al. (2007). Functional selectivity and classical concepts of quantitative pharmacology. *J. Pharmacol. Exp. Ther.* **320**, 1–13.
- Vaidehi, N. (2010). Dynamics and flexibility of G-protein coupled receptor conformations and their relevance in drug design. *Drug Discov. Today* **15**, 951–957.
- Vaidehi, N., Kenakin, T. (2010). The role of conformational ensembles of seven transmembrane receptors in functional selectivity. *Curr. Opin. Pharmacol.* **10**, 775–781.
- Vaidehi, N., Floriano, W. B., Trabanino, R., Hall, S. E., Freddolino, P., Choi, E. J., et al. (2002). Prediction of structure and function of G-protein coupled receptors. *Proc. Natl. Acad. Sci. USA* **99**, 12622–12627.
- Violin, J. D., Lefkowitz, R. J. (2007). Beta-arrestin-biased ligands at seven-transmembrane receptors. *Trends Pharmacol. Sci.* **28**, 416–422.
- Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., et al. (2008). Structure of a  $\beta_1$ -adrenergic G-protein-coupled receptor. *Nature* **454**, 486–491.

- Warne, T., Moukhametzianov, R., Baker, J. G., Nehmé, R., Edwards, P. C., Leslie, A. G., et al. (2011). The structural basis for agonist and partial agonist action on a  $\beta_1$ -adrenergic receptor. *Nature* **469**, 241–244.
- Yang, S., Banavali, N. K., Roux, B. (2009). Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc. Natl. Acad. Sci. USA* **106**, 3776–3781.
- Yao, X., Parnot, C., Deupi, X., Ratnala, V. R. P., Swaminath, G., Farrens, D., et al. (2006). Coupling ligand structure to specific conformational switches in the  $\beta_2$ -adrenoceptor. *Nat. Chem. Biol.* **2**, 417–422.
- Yao, X. J., Ruiz, G., Whorton, M. R., Rasmussen, S. G., De Vree, B. T., Deupi, X., et al. (2009). The effect of ligand efficacy on the formation and stability of a GPCR-G protein complex. *Proc. Natl. Acad. Sci. USA* **106**, 9501–9506.
- Ye, S., Zaitseva, E., Caltabiano, G., Schertler, G. F. X., Sakmar, T. P., Deupi, X., et al. (2010). Tracking G-protein-coupled receptor activation using genetically encoded infrared probes. *Nature* **464**, 1386–1389.
- Zaitseva, E., Brown, M. F., Vogel, R. (2010). Sequential rearrangement of interhelical networks upon rhodopsin activation in membranes: the Meta II(a) conformational substate. *J. Am. Chem. Soc.* **132**, 4815–4821.

# ADVANCES IN IMPLICIT MODELS OF WATER SOLVENT TO COMPUTE CONFORMATIONAL FREE ENERGY AND MOLECULAR DYNAMICS OF PROTEINS AT CONSTANT PH

By YURY N. VOROBYEV

Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Science, Novosibirsk, Russia

I.	Introduction .....	282
II.	Formulation of General Implicit Solvent Model for Calculating Conformational Free Energy .....	283
	A. Transport of a Protein from Gas Phase into Water-Proton Bath .....	283
III.	Continuum Solvent Models .....	286
	A. Free Energy of Nonpolar Interactions .....	286
	B. Free Energy of a Solvent Cavity .....	287
	C. The Solute-Solvent van der Waals Interactions .....	289
	D. Solvent Polarization Free Energy .....	290
	E. Continuum Electrostatic Poisson Model .....	291
	F. A Smooth Molecular Surface .....	292
	G. Fast Adaptive Multigrid Boundary Element Method .....	294
	H. Generalized Born Model .....	296
IV.	Protein Ionization .....	302
	A. PMF of Equilibrium Titration .....	302
	B. Practical Calculation of PMF of Implicit Titration .....	304
	C. Method FAMBEpH-GB .....	306
V.	Examples of Simulations with Implicit Solvent Models .....	308
	A. Practical Advantages of an Implicit Solvent Models .....	308
	B. Free Energy of Protein Decoys and Protein Decoy Discrimination .....	309
	C. Protein Folding .....	310
	D. Constant pH MD Simulations .....	311
	E. Limitations of Current Implicit Continuum Solvent Models and Further Direction .....	313
	References .....	314

## ABSTRACT

Modern implicit solvent models for macromolecular simulations in water-proton bath are considered. The fundamental quantity that implicit models approximate is the solute potential of mean force, which is obtained by averaging over solvent degrees of freedom. The implicit solvent models suggest practical ways to calculate free energies of



macromolecular conformations taking into account equilibrium interactions with water solvent and proton bath, while the explicit solvent approach is unable to do that due to the need to account for a large number of solvent degrees of freedom. The most advanced realizations of the implicit continuum models by different research groups are discussed, their accuracy are examined, and some applications of the implicit solvent models to macromolecular modeling, such as free energy calculations, protein folding, and constant pH molecular dynamics are highlighted.

## I. INTRODUCTION

Computer simulations in which solvent molecules are treated explicitly represent one of the most detailed approach to study the influence of solvation on biomolecules (Brooks et al., 1988). However, an accurate description of the aqueous environment for realistic biomolecular simulations, that is, via method of molecular dynamics (MD), requires a large number of solvent molecules to be placed around it (Karplus and McCammon, 2002; McDowell et al., 2007). In practical simulations, a large fraction of computer time is spent calculating a detailed trajectory of the solvent molecules, while it is the solute behavior that is primarily of interest. Despite their cost, computer simulations with explicit solvent molecules are not exempt from approximations, for example, difficulties arise in calculations involving charged molecules when long-range electrostatic interactions are truncated or summed over periodic array of simulation boxes using Ewald techniques (Hünneberg and McCammon, 1999). While free energy perturbation methods, based on microscopic simulation of a macromolecule with explicit solvent, may in principle be suitable for free energy calculations (Kollman, 1993; Radmer and Kollman, 1997), this in practice meets with tremendous difficulties due to the large molecular size and the need to sample adequately over large number of solvent and solute conformations and properly evaluate long-range electrostatic interactions (Bogusz et al., 1998; Kollman et al., 2000). An accurate calculation of the free energy of a macromolecule in an aqueous solution requires sampling over the whole volume of accessible phase space. While it is possible for structurally highly organized macromolecule from the results of a simulation, the same is difficult task for a solvent. Partly due to these difficulties, approximate schemes treating the solvent implicitly have been developed in past decades; some

of them are reviewed (Roux and Simonson, 1999; Bashford and Case, 2000; Chen and Brooks, 2008; Onufriev, 2008). Elaboration of implicit models of water and proton bath as a solvent media is important task for reliable simulation of proteins with many titratable groups at a given solvent pH. This chapter reviews modern explicit solvent models applicable for calculating the free energy of a macromolecule in aqueous solution and simulation via method of MD at constant pH.

## II. FORMULATION OF GENERAL IMPLICIT SOLVENT MODEL FOR CALCULATING CONFORMATIONAL FREE ENERGY

### A. *Transport of a Protein from Gas Phase into Water-Proton Bath*

To avoid the difficult problem of properly sampling solvent configuration, an implicit description of solvation can be adopted, and thereby obtained a partition function of the solute in which the interactions with the solvent are represented through a solvation potential or potential of mean force (PMF) which depends explicitly on the solute's coordinates (Hill, 1986; Vorobjev et al., 1998; Roux and Simonson, 1999). The partition function,  $Z$  of a solute molecule (atomic coordinates  $\mathbf{x}$ ) in a solvent (coordinates  $\mathbf{y}$ ) can be written as the ratio of the partition functions for solution and pure solvent systems (containing identical numbers of solvent molecules)

$$Z = \frac{\int d\mathbf{x} \int d\mathbf{y} \exp\{-\beta[U_m(\mathbf{x}) + U_{ms}(\mathbf{x}, \mathbf{y}) + U_{ss}(\mathbf{y})]\}}{\int d\mathbf{y} \exp[-\beta U_{ss}(\mathbf{y})]} \quad (1)$$

Here,  $U_m(\mathbf{x})$  is the intramolecular potential energy,  $U_{ms}(\mathbf{x}, \mathbf{y})$  is the potential energy of the solute-solvent interactions, and  $U_{ss}(\mathbf{y})$  is the potential energy of the solvent-solvent interactions. This can be rewritten as a partition function with solvent-mediated interactions between atoms of the solute molecule:

$$Z = \int d\mathbf{x} \exp\{-\beta[U_m(\mathbf{x}) + W(\mathbf{x})]\} \quad (2)$$

where the sum  $[U_m(\mathbf{x}) + W(\mathbf{x})]$  presents an effective energy, and

$$\exp[-\beta W(\mathbf{x})] = \frac{\int d\mathbf{y} \exp\{-\beta[U_{ms}(\mathbf{x}, \mathbf{y}) + U_{ss}(\mathbf{y})]\}}{\int d\mathbf{y} \exp[-\beta U_{ss}(\mathbf{y})]} \quad (3)$$

Here,  $W(\mathbf{x})$  is the free energy of solvation or PMF of the solute molecule with conformation  $\mathbf{x}$ . Considering scaled molecule–solvent interaction with coupling parameter  $\lambda$ , the solvation free energy,  $W(\mathbf{x})$ , can be written in the framework of the free energy perturbation method:

$$W(\mathbf{x}) = \int_0^1 d\lambda \frac{\int U_{\text{ms}}(\mathbf{x}, y) dy \exp\{-\beta[\lambda U_{\text{ms}}(\mathbf{x}, y) + U_{\text{ss}}(y)]\}}{\int dy \exp\{-\beta[\lambda U_{\text{ms}}(\mathbf{x}, y) + U_{\text{ss}}(y)]\}} \quad (4)$$

This provides an expression suitable for a microscopic simulation. Considering a multistep sequential “turning on” of different types of solute–solvent interactions in Eq. (4), one can see that the process of dissolving a gas-phase protein in water in the presence of hydrogen ions can be modeled as a four-stage thermodynamic process (Honig et al., 1993; Ripoll et al., 1996; Vorobjev et al., 2008): (stage 1) creation of a solute-sized cavity in water; (stage 2) insertion of the zero-charged protein (with all atoms having zero partial charges) into the cavity in water; (stage 3) charging of the protein to the gas-phase partial atomic charges  $\mathbf{q}^0 = (q^0_1, \dots, q^0_N)$  in which all ionizable groups are maintained neutral; and (stage 4) an equilibrium titration of the protein at a given pH (Fig. 1). The first three stages of this partition describe the solvation free energy of a protein with fixed gas-phase partial charges on all atoms  $\mathbf{q}^0$

$$W(\mathbf{x}, \mathbf{q}^0) = G_{\text{cav}}(\mathbf{x}) + G_{\text{vdw}}(\mathbf{x}) + G_{\text{pol}}(\mathbf{x}, \mathbf{q}^0) \quad (5)$$

where  $G_{\text{cav}}(\mathbf{x})$  is the free energy for creation of the molecular cavity in water (stage 1),  $G_{\text{vdw}}(\mathbf{x})$  is the free energy of van der Waals interactions between the solute and the water solvent (stage 2),  $G_{\text{pol}}(\mathbf{x}, \mathbf{q}^0)$  is the free energy of polarization of the water solvent by the protein with gas-phase partial charges on all atoms (stage 3),  $\Delta G_{\text{inz}}(\mathbf{x}, \text{pH})$  is the free energy of equilibrium titration of protein for a given pH and conformation  $\mathbf{x}$  which leads to a change of the protein gas-phase partial atomic charges  $\mathbf{q}^0$  of the neutral ionization microstate  $\mathbf{z}^0 = (z^0_1, \dots, z^0_\zeta)$ , all  $z^0_i = 0$ , where  $\zeta$  is the total number of titratable protons (or groups), to a new values  $\mathbf{q}_{\text{inz}}$  for equilibrium ionization state  $\langle \mathbf{z} \rangle$  which is coupled with conformation  $\mathbf{x}$  and pH value. The whole thermodynamic cycle defines the free energy  $G_t(\mathbf{x}, \text{pH})$  of transport of a single protein molecule into water at a given pH in an instantaneous microscopic conformation  $\mathbf{x}$ :

$$G_t(\mathbf{x}, \text{pH}) = W(\mathbf{x}, \mathbf{q}^0) + \Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) \quad (6)$$

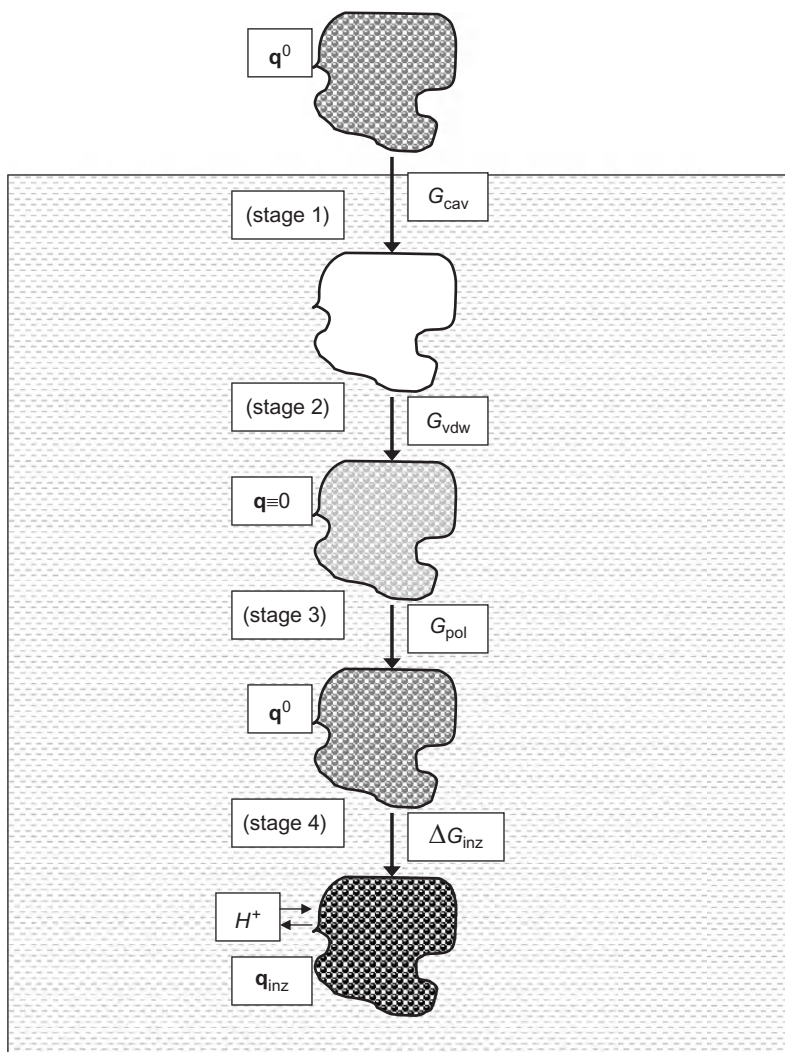


FIG. 1. Thermodynamic cycle of transport of molecule from gas phase into water-proton bath;  $\mathbf{q}^0$  is partial atomic charges in the gas phase in which all ionizable groups are neutral; (stage 1) creation of a solute-sized cavity in water; (stage 2) insertion of the zero-charged protein (with all atoms having zero partial charge) into the cavity in water; (stage 3) charging of the protein to the gas-phase partial atomic charges  $\mathbf{q}^0 = (q^0_1, \dots, q^0_N)$ ; and (stage 4) an equilibrium titration of the protein at a given pH value,  $\mathbf{q}_{inz}$  is partial atomic charges of titrated protein.

It should be noted that a transport of the neutral protein molecule from gas phase into water solvent at a given pH is not accompanied by the transfer of a net charge. The protein molecule becomes charged in water-proton bath due to equilibrium proton binding and releasing into solvent, that is, by means of equilibrium redistribution of protons between the solvent and the solute in a given conformation  $\mathbf{x}$ . The total free energy of protein for a given conformation  $\mathbf{x}$  in the solvent at given pH is equal to

$$G(x, \text{pH}) = U_m(x; q^0) + W(x, \mathbf{q}^0) + \Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) \quad (7)$$

where  $U_m^0(\mathbf{x}; \mathbf{q}^0)$  is the intramolecular conformational potential energy of the protein computed in the gas-phase with gas-phase atomic charges ( $\mathbf{q}^0$ ). A prediction of conformational preference of proteins in water-proton bath based on Eq. (7) makes it more reliable (Arnautova et al., 2009).

Considering all phase space of a solute molecule as a sum of subspaces A, B, . . . , each of which describes a distinct macroscopic solute conformation, it follows from Eq. (2) that the free energy  $G_A$  of a solute molecule in a macroscopic conformation A can generally be presented in terms of average configurational energy and entropy over the molecular degrees of freedom

$$G_A = \langle U_m(\mathbf{x}; \mathbf{q}^0) \rangle_A + \langle W(\mathbf{x}, \mathbf{q}^0) + \Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) \rangle_A - TS_c \quad (8)$$

where  $\langle \rangle_A$  denotes an average over micro-configurations of the conformation A,  $S_A$  is the entropy of the conformation A, which can be estimated over MD trajectory in quasi-harmonic approximation (Srinivasan et al., 1998; Vorobjev et al., 1998; Kollman et al., 2000).

### III. CONTINUUM SOLVENT MODELS

#### A. Free Energy of Nonpolar Interactions

The sum of free energy of solvent cavity formation and solute-solvent van der Waals interactions presents a free energy of nonpolar solvation (or hydration)  $G_{\text{np}}$ :

$$G_{\text{np}} = G_{\text{cav}} + G_{\text{vdw}} \quad (9)$$

The nonpolar solvation has a complex physical nature, and the associated energy generally has a small amplitude than the polar counterpart; however, it is well appreciated that hydrophobic association is one of the

principal interaction that determines biomolecular structures (Scheraga, 1998). Recently, it was well understood that the nonpolar solvation should include two terms, that is, the free energy of solvent cavity  $G_{\text{cav}}$  formation and solute–solvent van der Waals free energy  $G_{\text{vdw}}$ . These two terms are balanced each other depending differently on structure and conformation of interacting chemical groups (Chen and Brooks, 2007). Their sum describes a length-scale dependence of the free energy of solvating hydrophobic solutes and free energy of hydrophobic association (Chen and Brooks, 2008). Further, we consider recent advanced implicit models describing the cavity and solvent–solute van der Waals free energies.

### B. Free Energy of a Solvent Cavity

Analysis of the nonpolar free energy in the integral form of thermodynamic perturbation theory (Roux and Simonson, 1999), experimental data (Hermann, 1972; Ben-Naim and Marcus, 1984; Ben-Naim, 1990), microscopic simulations on small systems (Hummer et al., 1995, 1996; Wallqvist and Berne, 1995a,b), and scaled particle theory (Reiss et al., 1960; Pierotti, 1976; Postma et al., 1982; Jackson and Sternberg, 1994, 1995) shows consistently that the cavity free energy changes linearly with the surface  $S$  of the cavity

$$G_{\text{cav}} \approx \gamma_{\text{micro}} S \quad (10)$$

where the cavity surface is defined as a smooth molecular surface (MS) confining the molecular solvent-excluded volume (SEV) (Connolly, 1983a; Vorobjev and Hermans, 1997) or in some applications as a solvent-accessible surface (SAS) (Chothia, 1974; Richards, 1977). The SAS is generated by the center of water solvent probe molecule, modeled as a rigid sphere of radius  $R_w = 1.4 \text{ \AA}$ , when this rolls about external van der Waals surface of protein atoms, each represented by a spherical ball of atomic van der Waals radius  $R_{\text{vdw},i}$ . It is common practice to assume that the atomic van der Waals radii are independent of atomic charges. The MS is generated by the inward-facing surface of the solvent probe molecule, when it rolls about the van der Waals surface of the molecule. The proportionality factor,  $\gamma_{\text{micro}}$ , is a microscopic surface tension. An optimum choice for the proportionality factor,  $\gamma_{\text{micro}}$ , between surface area and cavity free energy is coupled with the choice of a type of surface, the MS or the SAS and remains to be determined. Simulations with an explicit water model show the free energy of creating an uncharged “bubble” in an aqueous solution to be proportional to the

macroscopic surface of the cavity, with an interfacial surface tension similar to the experimental gas-solvent surface tension,  $\gamma_{\text{macro}}$ , for bubbles well exceeding a water molecule in size (Jackson and Sternberg, 1994). The value of the microscopic surface free energy,  $\gamma_{\text{micro}}$ , used to compute  $G_{\text{cav}}$  is smaller because, on a molecular scale, the microscopic surface of an interface is much more irregular and somewhat larger than the corresponding macroscopic surface. For planar atomic arrays of densely packed van der Waals atomic spheres either in contact or interpenetrating up to 30%, one finds the smooth macroscopic surface to be smaller than the irregular microscopic surface by a factor of about 0.66 (Vorobjev et al., 1998; Vorobjev and Hermans, 1999). Correspondingly, the microscopic surface free energy should be smaller than the macroscopic surface tension of water by the same factor. With experimental  $\gamma_{\text{macro}}$  equal to 102 cal/(mol Å<sup>2</sup>), this gives a value of 67 cal/(mol Å<sup>2</sup>) for  $\gamma_{\text{micro}}$ , in good agreement with the estimate of 70 cal/(mol Å<sup>2</sup>) that has been found to optimize the correlation between the results of free energy estimates and experimental data for protein stability and protein-protein binding of mutant proteins (Jackson and Sternberg, 1995; Novotny et al., 1997).

Recently, the analytical generalized Born plus nonpolar (AGBNP) implicit solvent model has been presented, which includes estimators for solute cavity formation work and solute-solvent van der Waals energy (Gallicchio et al., 2002; Levy et al., 2003; Gallicchio and Levy, 2004). The cavity formation free energy term is described by Eq. (11), where cavity term  $G_{\text{cav}}$  is presented as a sum over partial atomic surfaces  $s_i$  with atom-dependent scaling factors  $\gamma_i$

$$G_{\text{cav}} = \sum_i \gamma_i s_i \quad (11)$$

A set of atomic factors  $\gamma_i$  are adjusted empirically on a training set of small molecules, a uniform value  $\gamma_i = 117$  cal/(mol Å<sup>2</sup>) independent on atom type. Solvent-accessible atomic surfaces  $s_i$  have been calculated as van der Waals surface of atoms with increased atomic radii  $R_i = \sigma_i/2 + 0.5$  Å, where  $\sigma_i$  is OPLS forcefield van der Waals parameter (Jorgensen et al., 1996). The improved implicit solvation model AGBNP2 (Gallicchio et al., 2009) describes the cavity formation free energy by Eq. (11) with various  $\gamma_i$  which are obtained from fitting Eq. (11) to the hydration energies of alkane cavities. The atomic parameters  $\gamma_i$  are in the range of 117–129 cal/(mol/Å<sup>2</sup>), for details see Table I in the chapter (Gallicchio et al., 2009).

### C. *The Solute–Solvent van der Waals Interactions*

The free energy  $G_{\text{vdw}}(\mathbf{x})$  of solute–solvent van der Waals interactions can be accurately estimated by averaging the potential energy of solute–solvent van der Waals energy over MD trajectory  $\sim 100$  ps with explicit solvent for a solute frozen at a given conformation  $\mathbf{x}$  (Vorobjev et al., 1998). This is a good approximation because the free energy of solvent reorganization due to the weak attractive van der Waals interaction with the solute is small (Tomasi and Persico, 1994). Due to a short-range nature of van der Waals potential, the energy  $G_{\text{vdw}}$  can be approximated by the linear expression over area of molecular surface  $S$ ,

$$G_{\text{vdw}}(\mathbf{x}) = -\gamma_{\text{vdw}}S \quad (12)$$

The average proportionality factor  $\gamma_{\text{vdw}} = +30 (\pm 17)$  cal/(mol  $\text{\AA}^2$ ) has been found from MD simulations of the solute–solvent van der Waals energy for a set of medium-sized proteins in an explicit SPC water (Vorobjev et al., 1998). The consistency of the implicit model in reproducing the cavity term and solute–solvent van der Waals energy is demonstrated by the agreement between the distance dependence of the implicit solvent PMF of nonpolar interactions between two methane molecules on the distance  $r$  in water with the PMF calculated by microscopic simulations via Monte-Carlo and MD (Vorobjev and Hermans, 1999). A recent study (Sobolevski et al., 2007) confirmed the observation that MS area in the Eqs. (10) and (12) provides a reasonable description of hydrophobic association of hydrocarbons and reproduces desolvation maximum of the rigorous PMF of hydrophobic association calculated by the MD free energy simulation in an explicit water solvent.

The nonpolar hydration free energy of Eq. (9) has been modeled by SAS area empirical models (Ooi et al., 1987; Langlet et al., 1988; Honig et al., 1993; Simonson and Brünger, 1994; Sitkoff et al., 1994; Tomasi and Persico, 1994; Juffer et al., 1995) which are still widely employed (Lee et al., 2000; Fogolary et al., 2001; Pellegrini and Field, 2002; Curutchet et al., 2003; Jorgensen and Tirado-Rives, 2004; Wagoner and Baker, 2006; Chen and Brooks, 2007, 2008).

The AGBNP model describes the solute–solvent van der Waals free energy by expression which is obtained as integral of van der Waals solute–solvent interactions over the solvent volume modeled as a uniform continuum (Levy et al., 2003).



$$G_{\text{vdw}} = \sum_i \alpha_i \frac{a_i}{(B_i + R_w)^3} \quad (13)$$

where

$$a_i = -\frac{16}{3} \pi \rho_w \varepsilon_w \sigma_{iw}^6 \quad (14)$$

where  $\rho_w = 0.033428 \text{ \AA}^{-3}$  and  $\sigma_{iw}$  and  $\varepsilon_{iw}$  are the OPLS force field parameters (Jorgensen and Madura, 1985) for van der Waals potential between atom  $i$  and water oxygen,  $B_i$  is the Born radius of atom  $i$  in the molecule of given conformation and  $R_w = 1.4 \text{ \AA}$  is radius of water molecule. The values of parameters  $\alpha_i \sim 1$  have been set so as to reproduce as best as possible the solute-solvent van der Waals energies of individual atoms of a large set of proteins and small molecules obtained from the results of explicit solvent simulations with TIP4P3 (Jorgensen and Madura, 1985; Levy et al., 2003; Gallicchio et al., 2009). It should be noted that the description of the nonpolar hydrations via Eqs. (11) and (13) with atomic scaling factors  $\alpha_i$  and  $\gamma_i$  empirically accounts for a dependence of atomic van der Waals radii  $R_{\text{vdw},i}$  on atomic charges.

#### D. Solvent Polarization Free Energy

The protein's charges (charge  $q_i$  is at position  $\mathbf{r}_i$ ) for conformation  $\mathbf{x} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$  induce in the solvent a polarization charge density,  $\langle \rho_{\text{pol}}(\mathbf{r}) \rangle$ , which produces an electrostatic potential,  $V_{\text{pol}}(\mathbf{r})$ , at the point  $\mathbf{r}_i$

$$\langle V_{\text{pol}}(r_i) \rangle = \int \frac{\langle \rho_{\text{pol}}(r) \rangle}{|r - r_i|} d\mathbf{r} \quad (15)$$

The polarization free energy is a work done in a charging process in which the charges of the protein are gradually “turned on” by factor  $\lambda$

$$G_{\text{pol}} = \int_0^1 d\lambda \sum_i [q_i \langle V_{\text{pol}}(r_i) \rangle_\lambda] \quad (16)$$

With a linear response approximation for solvent polarization,  $V_{\text{pol}}$  and  $\rho_{\text{pol}}$  both are proportional to  $\lambda$ , and this gives

$$G_{\text{pol}} = \frac{1}{2} \sum_i q_i \int \frac{\langle \rho_{\text{pol}}(r) \rangle}{|r - r_i|} d\mathbf{r} \quad (17)$$

In a dynamics simulation with explicit solvent,  $\rho_{\text{pol}}$  is identical with the distribution of the average charges of the solvent atoms, and a common approach is to use Eq. (16) to compute  $G_{\text{pol}}$  with thermodynamic integration or perturbation (Kollman, 1993).

The validity of the linear response approximation for the solvent reaction potential of an aqueous solvent has been examined by direct simulations of its dependence on  $\lambda$  in MD free energy simulations (Jayaram et al., 1989; Roux et al., 1990; Levy et al., 1991; Rick and Berne, 1994; Hummer et al., 1995; Aqvist and Hansson, 1996; Vorobjev and Hermans, 1999). In a majority of simulations of polar and charged molecules, a nearly linear response has been observed for a moderately charged solute, that is, one whose partial atomic charges do not exceed  $\sim 1 e$  and whose electrostatic field near the solute surface does not exceed  $50 \text{ kT}/(e \text{ \AA})$ .

### E. Continuum Electrostatic Poisson Model

The validity of linear response approximation assumes that the calculation of the average-induced polarization charge density,  $\langle \rho_{\text{pol}}(\mathbf{r}) \rangle$ , can be done also in the framework of macroscopic electrostatics, that is, with an implicit continuum solvent description. The average electrostatic potential  $\Phi(\mathbf{r})$  contains contributions from the fixed charges of the protein and the induced polarization charges in the solvent, according to the Poisson equation,

$$\nabla^2 \Phi(r) = -4\pi \sum_i q_i \delta(r - r_i) - 4\pi \langle \rho_{\text{pol}}(r) \rangle \quad (18)$$

and with use of standard relations connecting the average-induced charge density  $\langle \rho_{\text{pol}}(\mathbf{r}) \rangle$  with the average polarization, and the polarization with the electric field (Jackson, 1975; Landau and Lifshitz, 1988), one obtains Poisson equation with a position-dependent dielectric constant  $D(\mathbf{r})$

$$\nabla D(r) \nabla \Phi(r) = -4\pi \sum_i q_i \delta(r - r_i) \quad (19)$$

If the position-dependent dielectric constant  $D(\mathbf{r})$  is known, Eqs. (18) and (19) define the distribution of  $\langle \rho_{\text{pol}}(\mathbf{r}) \rangle$  for a given conformation of the protein so that  $G_{\text{pol}}$  can be calculated with Eq. (17).

A fundamental question is how to model the distribution of the dielectric constant,  $D(\mathbf{r})$ . Inside the protein molecule's SEV, the dielectric constant  $D_{\text{I}}=1$  because the solvation free energy has to be calculated for

a fixed internal degrees of freedom and nonpolarizable charge distribution, in a single conformation (Roux et al., 1999). In the solvent space, it is common practice to use the bulk water solvent dielectric constant  $D_0=80$ . Near the water–solute interface, the density of water drops sharply, over a distance of about 0.5 Å, from the bulk density to zero, as it has been shown by extensive MD simulation of solvent density around proteins (Lounnas et al., 1994). Therefore, a model with a sharp stepwise approximation to the solvent density is reasonable. Based on integral equations of liquids (Beglov and Roux, 1996, 1997), it was shown that the position-dependent dielectric constant  $D(\mathbf{r})$  can be modeled by the equation

$$D(\mathbf{r}) = D_i + \theta(\mathbf{r})(D_0 - D_i) \quad (20)$$

where  $\theta(\mathbf{r})$  is a sharp switching function equal to zero inside the SEV. The exact choice of where to locate the solute–solvent dielectric boundary is empirical and compensate for deviations of the actual dependence of the dielectric constant from the assumed step function near the protein surface. An optimal set of atomic radii defining dielectric interface MS has been calculated by fitting the implicit model polarization free energy to a set of experimental data (Sitkoff et al., 1994) and/or data obtained by calculations via free energy perturbation method with explicit solvent for a training set of small molecules (Nina et al., 1997, 1999; Vorobjev et al., 2008).

#### F. A Smooth Molecular Surface

The method used to compute the dielectric interface in Eqs. (19) and (20) must be defined with precision because it is crucial component for an accurate prediction in macromolecular electrostatic applications. It is assumed that the MS is a good approximation of a dielectric surface border between high dielectric polar solvent and low dielectric interior of solute molecule in continuum dielectric methods on the base of numerical solution of the Poisson equation (19) (Zauhar and Morgan, 1988; Rashin, 1990; Sharp and Honig, 1990; Juffer et al., 1991; Vorobjev et al., 1992; Rashin et al., 1994; Bharadwaj et al., 1995; Vorobjev and Hermans, 1997; Vorobjev and Scheraga, 1997; Vorobjev et al., 1998, 2008). Calculation of molecular properties on the MS and integration of a function over the MS require a numerical representation of the MS as a manifold  $S(\mathbf{s}_i, \mathbf{n}_i, \Delta s_i)$  of boundary elements (BEs), where  $\mathbf{s}_i$ ,  $\mathbf{n}_i$ , and  $\Delta s_i$  are coordinates, normal vector in outward direction, and area of a small

surface element. The MS contains three types of components or faces, which are termed “contact,” “saddle,” and “concave reentrant,” according to whether the solvent probe sphere simultaneously touches one, two, or three atoms, respectively (Connolly, 1983a,b, 1985). The true Connolly’s MS of a protein may contain hundreds of singular regions with singularities in the direction of the normal vector. The direction of the normal vector  $\mathbf{n}$  is not continuous in the vicinity of a singular point of the MS (Vorobjev and Hermans, 1997). Singularities called cusps and holes appear when the probe can almost, but not quite, pass through a group of two or three atoms of the protein (Connolly, 1985; Zauhar, 1995; Vorobjev and Hermans, 1997). It has been shown (Vorobjev and Hermans, 1997; Vorobjev and Scheraga, 1997; Vorobjev et al., 1998) that accurate solution of Poisson equation via BE method needs MS with smoothed singularities.

None of programs, MSROLL (Connolly, 1985), MSED (Perrot et al., 1992), MS (Varshney et al., 1994), and MSMS (Sanner et al., 1996), were specifically designed for the BE method application and provide a dot MS of poor quality as was tested by (Vorobjev and Hermans, 1997) to be used with BE method. The Connolly’s method of MS calculation (Connolly, 1983a,b, 1985) has been revised, and a new method generating smooth invariant molecular surface (SIMS) (Vorobjev and Hermans, 1997) has been developed. The SIMS method (i) produces a near-homogeneous dot distribution, (ii) is invariant to molecular rotation and translation, and (iii) recognizes all types of singularities of the MS and smoothed them with specified minimal radius of curvature. An optimal practical choice of the radius of the smoothing sphere is  $\sim 0.4 \text{ \AA}$ . The SIMS method generates a dot MS of good numerical quality, which can be used in a variety of implicit continuum models for calculating solvation free energy and for molecular electrostatics with the BE method in dielectric continuum models. The influence of a choice and composition of BEs on convergence of the solution of the Poisson equation by numerical methods has been investigated in details using Connolly’s MSROLL (Connolly, 1985) and SIMS programs to generate BE on the solute–solvent dielectric surface (Kar et al., 2007). It has been found that the SIMS program generates the BEs of better quality and achieves convergence faster using smaller number of the surface elements than the MSROLL program, by a factor of  $\sim 1.5$ – $2.0$ , in the test on a set of 35 medium-sized proteins. A complete description of the SIMS method can be found elsewhere (Vorobjev and Hermans, 1997). The timing of the SIMS method is somewhat better than

the timing of Connolly's method, the CPU time scales as the number of atoms in the molecule (Vorobjev et al., 1998). The SIMS program is available from the authors on request (ynvorob@niboch.nsc.ru).

### *G. Fast Adaptive Multigrid Boundary Element Method*

The evaluation of the solvent polarization charge density for proteins for a complicated atomic charge distribution is done numerically with Eqs. (19) and (20) using finite-element methods in 3D space or on the dielectric surface boundary. The finite difference (FD) method solves Poisson (or Poisson–Boltzmann) equation in differential form Eq. (19) using multigrid elements in 3D space of rectangular box which includes the solute and a volume of solvent around it (Honig et al., 1993; Madura et al., 1994; Simonson and Brünger, 1994; Sitkoff et al., 1994; Holst and Saied, 1995; Holst et al., 1994, 2000; Rocchia et al., 2002; Zhou et al., 2008). The alternative is a BE method which is used for numerical solution of an integral equation over the dielectric boundary, to which the original Poisson equation (19) can be analytically converted (Bharadwaj et al., 1995). The BE method finds a solution in terms of electrostatic potential and/or solvent polarization charge density induced on BEs tessellated the solute–solvent dielectric surface (Yoon and Lenhoff, 1990; Juffer et al., 1991; Vorobjev et al., 1992, 1998; Bharadwaj et al., 1995; Vorobjev and Scheraga, 1997; Lu et al., 2006; Lu and McCammon, 2007; Vorobjev et al., 2008). The BE method shows its invariance to rotation and translation of the solute molecule, and a comparison of multigrid BE and FD methods shows (Bharadwaj et al., 1995) that the BE method exhibits a higher degree of consistency. Improved methods of solving the Poisson equation for inhomogeneous dielectric media using multigrid and multilevel techniques have been developed (Holst et al., 1994; Goel et al., 1995; Holst and Saied, 1995; McKenney and Greengard, 1995; Douglas, 1996; Zhou et al., 1996; Rocchia et al., 2002). Multilevel and multisized BE techniques have been applied to the iterative BE method (Rashin, 1990; Vorobjev et al., 1992; Rashin et al., 1994; Zauhar and Varnek, 1996). The iterative BE methods suffer from slow convergence and are more time consuming than multigrid FD methods.

Recently, a new efficient implementations of the BE method have been developed (Lu et al., 2006; Lu and McCammon, 2007; Vorobjev et al., 2008). The BE integral equation, to which the Poisson equation (19) is analytically converted (Bharadwaj et al., 1995), is solved by the fast

adaptive multigrid boundary element (FAMBE) method (Vorobjev and Scheraga, 1997; Vorobjev et al., 2008) for the induced surface polarization charge density  $\sigma(\mathbf{t})$

$$\sigma(t) = f \int_S \frac{\sigma(s)(t-s)n(t)ds}{|t-s|^3} + \frac{f}{D_1} \sum_i n_i E_i(t) \tag{21}$$

where  $f = (1/2\pi)(D_1 - D_0)/(D_1 + D_0)$  and  $\mathbf{n}(\mathbf{t})$  is the outward normal vector to the MS at point  $\mathbf{t}$ ,  $D_1$  and  $D_0$  are the dielectric constants inside and outside the surface, respectively, and  $\mathbf{E}_i(\mathbf{t})$  is electrostatic field generated by the charge  $i$  at the surface point  $\mathbf{t}$ . The induced charge density  $\sigma(\mathbf{t})$  approximates the average solvent-induced charge density, in Eq. (15). The solvent polarization free energy  $G_{\text{pol}}$  can be found with Eq. (17), replacing volume integral and volume charge density with surface integral and surface charge density  $\sigma(\mathbf{t})$

$$G_{\text{pol}} = \frac{1}{2} \sum_i q_i \int_S \frac{\sigma(s)}{|r - r_i|} ds \tag{22}$$

Since the term  $\mathbf{E}_i(\mathbf{t})$  is linear in the charges  $q_i$ , it is possible to split  $\sigma(\mathbf{t})$  given by Eq. (21) into a sum of terms, each one of which represents the induced polarization charge density,  $\sigma_i(\mathbf{t})$ , generated by a single group of charges. The FAMBE method splits Eq. (21) into set of independent minor BE equations, one each for the induced polarization charge density generated by a single charge (or small compact group of charges)

$$\sigma_i(t) = f \int_S \frac{\sigma_i(s)(t-s)n(t)ds}{|t-s|^3} + \frac{f}{D_1} n_i E_i(t), \quad i = 1, 2, \dots \tag{23}$$

the total surface charge,  $\sigma(\mathbf{t})$  is the sum of the components  $\sigma_i(\mathbf{t})$ . The reason for such decomposition is that the integral equation, Eq. (23), for each component,  $\sigma_i(\mathbf{t})$ , can be converted into a discrete linear equation of low dimensionality of a matrix  $\mathbf{M}_i$  over the set  $i$  of adaptive multisized BEs

$$\boldsymbol{\sigma}_i = \mathbf{M}_i \boldsymbol{\sigma}_i + \mathbf{E}_i \tag{24}$$

For each charge,  $q_i$  the size of the BEs steadily increases with distance  $R$  from the source of the molecular electrostatic field. Thereby the MS is tessellated by the unique set of multisized BEs, thusway, for any given single charge  $q_i$  the dimension of the vector of surface charge densities  $\boldsymbol{\sigma}_i$  and of the matrix  $\mathbf{M}_i$  is significantly lower than the total number of surface elements that would be

encountered if the surface were tessellated by the finest uniform BEs in Eq. (21). The number of multisized BEs  $N_{\text{MBE}}$ , that is, the matrix  $\mathbf{M}_i$  size for any single charge  $q_i$ , which tessellates an MS with area  $A_S$  scales as:

$$\text{NMBE} \approx n_{\text{loc}} \ln \left( \frac{A_S}{A_{\text{loc}}} \right) \quad (25)$$

where  $n_{\text{loc}}$  and  $A_{\text{loc}}$  are the average number of BEs and the size of the local area with the finest tessellation. Each minor matrix (24) is solved by the preconditioned biconjugate gradient method (Press et al., 1988). Only a few iterations (five or six) are needed to find a solution of linear equation (24) with a relative accuracy of  $10^{-4}$ – $10^{-5}$ . The estimated computational complexity of the FAMBE method scales as:

$$\text{Complexity} \approx N_z \left[ n_{\text{loc}} \ln \left( \frac{A_S}{A_{\text{loc}}} \right) \right]^2 \quad (26)$$

where  $n_{\text{loc}}$  and  $A_{\text{loc}}$  are the average number of BEs and the size of the local area with the finest tessellation, and  $N_z$  is the number of charges (or charged groups) in the solute molecule. The further details of the FAMBE method can be found elsewhere (Vorobjev et al., 2008). Test calculations for several proteins show that the CPU time of the FAMBE method scales approximately linearly with the number of atoms of the molecule. The FAMBE method (Vorobjev et al., 2008) shows a high degree of internal self-consistency and higher accuracy and speed of calculations in comparison with one of the latest realization of BE method by other authors (Lu et al., 2006; Lu and McCammon, 2007). The free energy calculated with the FAMBE method includes dependence on salt effects implicitly (Vorobjev et al., 2008). A good numerical quality and a high speed suggest the FAMBE method as ideal tool for a post-processing of MD trajectories for free energy estimations via Eq. (8) with important applications for systems undergoing a large conformational changes. The FAMBE program is available from the authors on request (ynvorob@niboch.nsc.ru).

#### H. Generalized Born Model

However, solving the Poisson equation by the fastest available methods is still too time consuming to be used for calculation of solvation energy and atomic forces on the fly as it is required in the method of MD. Therefore,

other simplified and significantly faster approaches like the generalized Born (GB) method has received considerable recent attention (Bashford and Case, 2000; Onufriev, 2008). In this model, the electrostatic contribution to the free energy of solvent polarization is defined analytically

$$G_{\text{pol}} = -\frac{1}{2} \left( \frac{1}{D_1} - \frac{1}{D_0} \right) \sum_{i,j} \frac{q_i q_j}{f_{\text{GB}}(r_{ij})} \quad (27)$$

where  $r_{ij}$  is distance between protein charges  $q_i$ ,  $q_j$ ,  $D_1$  and  $D_0$  are internal and external molecular volume dielectric constants, and  $f_{\text{GB}}(r)$  is a function that interpolates between “effective Born radius”  $B_{ij}$ , of atoms  $i, j$  when the distance between atoms  $r_{ij}$  is short, and  $r_{ij}$  itself at the large distances  $r_{ij}$  (Still et al., 1990)

$$f_{\text{GB}}(r_{ij}) = \left[ r_{ij}^2 + B_i B_j \exp \left( -\frac{r_{ij}^2}{4B_i B_j} \right) \right]^{1/2} \quad (28)$$

where  $B_i$ ,  $B_j$  are effective Born radii of atoms  $i$  and  $j$ . The basic idea of the GB approach can be viewed as an interpolation formula between analytical solutions for a single sphere and for widely separated spheres. Considering protein with one charged atom  $q_i$ , the self-polarization free energy of the charge  $q_i$  via the Poisson equation method equation is

$$G_{i,\text{pol}}^{\text{P}} = \frac{1}{2} q_i \int_S \frac{\langle \sigma_i(\mathbf{s}) \rangle}{|\mathbf{s} - \mathbf{r}_i|} d\mathbf{s} \quad (29)$$

where  $\sigma_i(\mathbf{s})$  is a polarization charge density induced by the protein charge  $q_i$  over the molecular surface  $S$  of the protein. On the other hand, GB, Eq. (27), defines that self-polarization energy as

$$G_{i,\text{pol}}^{\text{GB}} = -\frac{q_i^2}{2B_i} \left( \frac{1}{D_1} - \frac{1}{D_0} \right) \quad (30)$$

Comparing Eqs. (27) and (30), one obtains a formal way to define “ideal” effective Born radius  $B_i$  of atom  $i$  of the protein in particular conformation

$$B_i = -\frac{q_i^2}{2G_{i,\text{pol}}^{\text{P}}} \left( \frac{1}{D_1} - \frac{1}{D_0} \right) \quad (31)$$

Incorporation of salt effects in the GB model is achieved by the simple substitution



$$\left(\frac{1}{D_1} - \frac{1}{D_0}\right) \rightarrow \left(\frac{1}{D_1} - \frac{\exp(-\kappa f(r_{ij}))}{D_0}\right) \quad (32)$$

where  $\kappa$  is the Debye–Hückel screening parameter. The goal of the GB model can be thought of as an effort to find a relatively simple analytical formula, which for real molecular conformations will reproduce, as much as possible, the results of the Poisson equation. The GB model using the “ideal” Born atomic radii  $B_i$ , which are defined by Eq. (31), provides the accurate approximation of the Poisson polarization free energy of proteins (Onufriev et al., 2002; Feig et al., 2004) with errors within a few percent  $\sim 1$ –3%. It should be noted that calculation of the “ideal” Born radii set on the base of Eq. (29), that is, by solving Poisson equation, thought to be impractical (Bashford and Case, 2000); therefore, much effort have been done to find a more rapid and still reasonable approximations for the effective Born radii to its “ideal” values. Assuming that effective Born radii can be computed efficiently for each atom of molecule, computational advantage of the GB model relative to numerical FD or BE solution becomes apparent, the GB formula is simple: its analytical derivatives with respect to atomic positions provide electrostatic atomic forces required in the MD simulation method.

Usually, the effective Born radii  $B_i$  are estimated by expression using Coulomb field approximation (CFA) (Still et al., 1990) for electrostatic field completely neglecting a solvent reaction field in a solvent and protein volume. The CFA self-polarization free energy  $G_i^{\text{CFA}}$  of a charge  $q_i$

$$G_i^{\text{CFA}} = \frac{q_i^2}{2 \times 4\pi} \left(\frac{1}{D_0} - \frac{1}{D_1}\right) \int_{r>\text{SEV}} \frac{dV}{|r - r_i|^4} \quad (33)$$

where SEV is the solvent-excluded volume. The effective Born radii in the CFA approximation are defined as

$$B_i^{-1} = R_{i,\text{vdw}}^{-1} - \frac{1}{4\pi} \int_{r>R_{i,\text{vdw}}}^{\text{SEV}} \frac{dV}{|r - r_i|^4} \quad (34)$$

where  $R_{\text{vdw},i}$  is van der Waals radius of atom  $i$ . The volume integral of Coulomb field energy density (Eq. (34)) is evaluated by numerical integration (Still et al., 1990) over the volume of the van der Waals spheres of the solute atoms instead of the SEV volume, that is

$$B_i^{-1} = R_{i,\text{vdw}}^{-1} - \frac{1}{4\pi} \int_{r>R_{\text{vdw},i}}^{\text{vdw}} \frac{dV}{|r - r_i|^4} \quad (35)$$

It implies a definition of a solute volume in terms of a set of van der Waals atomic spheres, rather than as the SEV confined by a complex MS commonly used in the Poisson calculations (Connolly, 1983a; Vorobjev and Hermans, 1997). A closed form of analytical expressions for two overlapping spheres from which expression for volume integral has been derived in the pairwise approximation is given by (Schaefer and Froemmel, 1990; Hawkins et al., 1996). Originally, the GB model with HTC (Hawkins et al., 1996) Born radii formula (35) has been developed for small molecules, where it was found to reproduce solvation energies and individual charge–charge interactions quite well (Hawkins et al., 1996; Curutchet et al., 2003) if a slightly reduced values for atomic van der Waals radii  $R_{i,\text{vdw}} = R_{i,\text{vdw}} - 0.09 \text{ \AA}$  are used. For macromolecules, the approach based on Eq. (35) with integral over the van der Waals volume tends to underestimate the values of Born radii for buried atoms (Onufriev et al., 2002) because the integration procedure for Eq. (35) treats small vacuum-filled crevices between the van der Waals spheres of protein atoms as being filled with water. The HTC formula assigns Born radii for medium-sized proteins in narrow interval  $\sim 1.5\text{--}4.0 \text{ \AA}$ , while the range of values for the “ideal” Born radii is much large  $\sim 1.5\text{--}10 \text{ \AA}$ . Two approximations are used by the HTC-Born radii model (Eq. (35)): (i) the CFA, which neglects a solvent reaction field; (ii) van der Waals volume of integration is a crude approximation of the excluded solvent volume ESV. The OBC-Born radii model (Onufriev et al., 2004) defines Born radii by an empirical function of volume integral equation (35) with empirical parameters. The OBC-Born radii model improves accuracy of Born radii estimation for proteins so that the OBC distribution of  $B_i(\text{OBC})$  covers interval  $1.5\text{--}6 \text{ \AA}$ . However, the deviation of the OBC-Born radii for buried atoms from the “ideal” Born radii is still large so that, for many buried atoms, the OBC-Born radii are lower by the factor of 2–3, compared to its “ideal” values.

Other attempts to improve GB model (Im et al., 2003; Lee et al., 2003) named as GBSV/MS model use (i) more realistic definition of a protein volume as a union of smoothed solvent exclusion functions centered on atoms, to approximate the rigorous SEV more accurately, but still computationally effectively, and (ii) corrected CFA is used for definition of self-polarization free energy of charged atoms. The self-polarization energy

$G_{\text{pol},i}$  has been expressed as a sum of empirical correction terms to the CFA (Lee et al., 2003; Feig et al., 2004) and demonstrated great improvement over the CFA for the calculated effective Born radii. The last corrected GB models (Lee et al., 2003; Feig et al., 2004; Mongan et al., 2007a,b) have a good agreement for polarization free energy with calculations by the Poisson equation method, showing relative errors of about 3–5%. Currently, a variety of the corrected GB models are implemented in a modern simulation packages, for example, AMBER and CHARMM. Because of their algorithmic simplicity and reasonable accuracy, they are commonly used in many applications (Onufrev, 2008). A recent study (Chen, 2010) continues to search for a more effective and accurate GB models for Born radii as a series of empirical correction terms to the CFA (Lee et al., 2003; Feig et al., 2004). The GBSV/MS2 model suggests empirical expression for Born radii with three parameters, instead of two in the original GBSV/MS model of Lee et al. (2003). The GBSV/MS2 model has been parameterized by minimizing the root-mean-square deviation (RMSD) error between GB and Poisson results for effective Born radii and self-polarization free energy of all atoms for 22 small proteins. It was found that the average relative unsigned errors for GBSV/MS2 Born radii  $\Delta B = \langle |B_i(\text{GBSV/MS2}) - B_i(\text{ideal})| / B_i(\text{ideal}) \rangle \sim 0.25$ , for buried atoms with  $B_i > 4 \text{ \AA}$ . Many buried atoms still have much lower effective Born radii in the GBSV/MS and GBSV/MS2 models up to factor 2.0, compared to the values of the respective “ideal” Born radii (Chen, 2010). This discrepancy leads to errors in estimation of pair atom–atom electrostatic interactions and respective forces. This observation shows a limitation of approaches developing an accurate GB model based on an empirical corrections for self-polarization free energy.

Levy’s group has been developing analytical version of GB model during past decade (Gallicchio et al., 2000, 2002; Levy et al., 2003; Gallicchio and Levy, 2004; Gallicchio et al., 2009). The most elaborated AGBNP2 (analytical GB nonpolar) model (Gallicchio et al., 2009) introduces two key innovations to the nonpolar (has been discussed earlier) and electrostatic components. The electrostatic solvation model in the AGBNP is based on the HTC (Hawkins et al., 1996) pairwise descreening GB scheme, whereby the Born radius of each atom is obtained by summing an appropriate descreening function over its neighbors. The main distinction between the AGBNP method and conventional HTC pairwise descreening volume integral implementation is that in the AGBNP method, the solute volume

is modeled as a set of overlapping atomic spheres which in turn are approximated by the Gaussian density functions proposed by (Grant and Pickup, 1995). The solute volume is computed by the inclusion–exclusion formula. The model defines analytically the self-volume and van der Waals surface of atom  $i$  with a set of empirically adjusted switching functions. The Born radii of the AGBNP model are obtained by analytical evaluation of the integral equation (35) over the volume occupied by the solute atoms (Gallicchio and Levy, 2004). The AGBNP2 model (Gallicchio et al., 2009) introduces method to approximate the SEV by the van der Waals integration volume of Eq. (35) using empirically augmented van der Waals radii and volume rescaling factors, while keeping the analytical expressions obtained for van der Waals intersecting spheres. Validation of the AGBNP2 method of integration over approximately defined SEV is done by a direct comparison of Born radii  $B_i(\text{AGBNP2})$  with calculation of the Born radii  $B_i(\text{SEV})$  over accurate numerically defined SEV. The comparison shows that the AGBNP2 model improves estimation of Born radii of the AGBNP model which integrates Eq. (35) over the volume occupied by the van der Waals spheres of solute atoms. The average ratio  $B_i(\text{AGBNP2})/B_i(\text{SEV}) \sim 1.2\text{--}2.0$ , while the ratio  $B_i(\text{AGBNP})/B_i(\text{SEV}) \sim 1.4\text{--}3.0$  for buried atoms with Born radii  $B_i(\text{SEV}) > 5 \text{ \AA}$ . In spite of that deficiency, the AGBNP2 model is implemented in the MD packages and shows a reasonable performance on a large set of test proteins (Gallicchio et al., 2009).

A different expression to compute the effective Born radii was proposed in the study (Grycuk, 2003), the “ $R6$  radii”, as an alternative to the CFA approximation (34)

$$B_i^{-1} = \left( R_{\text{vdw},i}^{-3} - \frac{3}{4\pi} \int_{r>R_{\text{vdw},i}}^{\text{SEV}} \frac{dV}{|r - r_i|^6} \right)^{1/3} \quad (36)$$

Unlike the CFA radii in Eq. (34), the “ $R6$  radii” formula is exact for any location of a charged atom within a perfect spherical solute in the limit  $D_0/D_1 \gg 1$  (Morgan et al., 2007b; Aguilar et al., 2010). It has been shown that when “ $R6$  radii” are computed by accurate numerical integration over exact MS or SEV (Morgan et al., 2007b), the resulting effective Born radii are in very close agreement with “ideal” Born radii. The study of Aguilar et al. (2010) suggests a new analytical method (AR6) to compute the effective Born radii as empirical function based on  $R6$  integral

equation (36) with pairwise van der Waals approximation of the SEV molecular volume and several molecular volume correction terms to approximate more exactly the “true” molecular volume in a vicinity of the atom in question. Finally, the AR6 effective Born radii are defined by empirical function with several parameters which are defined by parametrization. The RMSD between the inverse effective AR6 and the “ideal” Born radii for medium-sized protein lysozyme is about 0.064. It actually means that Born radii of buried atoms with Born radii  $B_i > 3 \text{ \AA}$ , that is,  $B_i^{-1} < 0.3$ , are estimated by the AR6 model with errors  $> 20\%$  and the error is increased up to 50% for deeply buried atom with Born radii  $B_i > 6 \text{ \AA}$ . This observation suggests that accurate and numerically fast analytical or numerical approximation of the exact SEV molecular volume is the modern research frontier for improvement in the GB approximation. It should be noted that for the small drug-like molecules the AR6 model with cavity term, of Eq. (11), and van der Waals solvation term, of Eq. (13), reproduces the experimental solvation free energies with good accuracy, the RMSD error is equal to 1.73 kcal/mol.

#### IV. PROTEIN IONIZATION

##### A. PMF of Equilibrium Titration

Transport of protein molecule from gas phase into a water–proton bath is accompanied by an ionization of titratable residues. The work required for that is the free energy of ionization  $\Delta G_{\text{inz}}$  (Eq. (6)). This free energy is the implicit titration potential of mean force (IT-PMF) for a protein in water–proton bath. A rigorous statistical mechanical formulation of IT-PMF has been considered by Baptista et al. (1997) in terms that eliminate the explicit reference to a variable number of protons. The IT-PMF free energy  $\Delta G_{\text{inz}}(\mathbf{x}, \text{pH})$  of protein at a given pH in water solvent is defined as

$$\Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) = -kT \ln \sum_{n, \mathbf{z}} \exp \left[ \frac{(n\mu - \Delta G(\mathbf{x}, \mathbf{z}))}{kT} \right] \quad (37)$$

where  $\Delta G(\mathbf{z}, \mathbf{x})$  is a free energy of a protein at ionization microstate  $\mathbf{z} = (z_1, \dots, z_c)$  relative to the neutral state  $\mathbf{z}^0$  in water, for the conformation  $\mathbf{x}$ ,

$$\Delta G(\mathbf{x}, \mathbf{z}) = G(\mathbf{x}, \mathbf{z}) - G(\mathbf{x}, \mathbf{z}^0) \quad (38)$$

$n$  is a total number of bound protons for the ionization microstate  $\mathbf{z}$ ,  $\mu$  is a chemical potential of protons, that is,  $\mu = -kT \cdot (\ln 10) \text{pH}$ . A canonical MD simulation of a protein with free energy described by Eq. (37) at constant temperature is the constant pH MD (CpHMD) simulation of the titratable system in the IT-PMF. To perform such simulation, one has to express  $\Delta G_{\text{inz}}(\mathbf{x}, \text{pH})$  in terms of quantities that can be computed. The implementation of the implicit titration potential  $\Delta G(\mathbf{x}, \text{pH})$  for CpHMD method developed by [Baptista et al. \(1997\)](#) was too simplified because it was based on the mean field approximation, that is, the pair correlation of ionization degrees  $\langle z_i z_j \rangle$  of sites  $i, j$  has been approximated as a product  $\langle z_i z_j \rangle \approx \langle z_i \rangle \cdot \langle z_j \rangle$ , and the average values of  $\langle z_i \rangle$  have been calculated by modified Tanford–Kirkwood method ([Tanford and Roxby, 1972](#)), assuming a spherical shape for the protein.

An accurate practical implementation of the IT-PMF addresses two questions: (i) What is the optimal algorithm to compute the multisite ionization equilibrium and related free energy and atomic forces? (ii) What is the optimal protocol to produce fast and accurate CpHMD simulations? A practical solution of the first problem is provided by the new method FAMBEpH ([Vorobjev et al., 2008](#)) which generalizes FAMBE method ([Vorobjev and Scheraga, 1997](#)) for calculating the free energies of solvent polarization  $G_{\text{pol}}(\mathbf{x})$  and ionization  $\Delta G_{\text{inz}}(\mathbf{x}, \text{pH})$ , in Eqs. (5) and (6), of a protein at a given pH. The FAMBE method is used for a fast evaluation of the “ideal” Born atomic radii calculating the self-polarization free energy of each charged atom of the protein. The GB method with “ideal” Born radii allows one to perform analytical calculation of all electrostatic atomic forces for MD simulation. Thereby the FAMBEpH–GB method provides one with (i) the solvation free energies of the ionizable residues in water, (ii) an accurate estimation of an average ionization degrees  $\langle z_i \rangle$ , their pair correlations  $\langle z_i z_j \rangle$ , and (iii) the free energy of ionization and respective atomic forces due to the IT-PMF. The IT-PMF gives an instant equilibrium response of the proton bath at given pH; therefore, the CpHMD with the IT-PMF can be more effective, then the commonly used today’s approaches which are based on an explicit stochastic titration method considering explicitly a vast number of ionization microstates which are generated randomly ([Mongan et al., 2004](#); [Machuqueiro and Baptista, 2006](#); [Williams et al., 2010](#)).

### B. Practical Calculation of PMF of Implicit Titration

The ionization free energy,  $\Delta G_{\text{inz}}(\mathbf{x}, \text{pH})$ , can be calculated by thermodynamic integration method as a titration process from zero hydrogen-ion concentration to a given value of pH via the Tanford–Schellman integral (Tanford, 1970; Schellman, 1975). From Eq. (37), it follows

$$\frac{\partial \Delta G_{\text{inz}}(\mathbf{x}, \text{pH})}{\partial \text{pH}} = kT(\ln 10) \sum_{i=1}^{2^{\xi}} \theta_i \langle z_i(\mathbf{x}, \text{pH}) \rangle \quad (39)$$

where  $\langle z_i(\mathbf{x}, \text{pH}) \rangle$  represents the average ionization degree of site  $i$  in the protein in conformation  $\mathbf{x}$  and parameter  $\theta_i$  is equal to 1 or  $-1$  if the ionizing group is a base or an acid, respectively. Integrating over pH one obtains practically treatable expression (Yang and Honig, 1993; Vorobjev et al., 2008) to calculate the free energy of ionization

$$\Delta \Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) - \Delta \Delta G_{\text{inz}}(\mathbf{x}, \infty) = kT(\ln 10) \sum_{i=1}^N \theta_i \int_{\infty}^{\text{pH}} \left[ \langle z_i(\mathbf{x}, \text{pH}) \rangle^i - \langle z_{i,\text{mod}}(\text{pH}) \rangle \right] d\text{pH} \quad (40)$$

where the functions  $\langle z_i(\mathbf{x}, \text{pH}) \rangle$  and  $\langle z_{i,\text{mod}}(\mathbf{x}, \text{pH}) \rangle$  represent the average ionization degree of site  $i$  in the protein in conformation  $\mathbf{x}$ , and in the isolated model compound, respectively and

$$\Delta \Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) = \Delta G_{\text{inz}}(\mathbf{x}, \text{pH}) - \Delta G_{\text{inz,mod}}(\mathbf{x}, \text{pH}) \quad (41)$$

is the free energy of ionization of protein relative to the total free energy of ionization of the all titratable residues in model compounds. For site  $i$  in protein conformation  $\mathbf{x}$  at a given pH, the average ionization degrees  $\langle z_i(\mathbf{x}, \text{pH}) \rangle$  can be calculated by a Monte-Carlo random walk in the space of the ionization microstates  $\mathbf{z}$

$$\langle z_i(\mathbf{x}, \text{pH}) \rangle^i = \frac{1}{Z_{\text{inz}}} \sum_{\mathbf{z}}^{2^{\xi}} z_i \exp \left( -\Delta G \left( \frac{\mathbf{x}, \mathbf{z}, \text{pH}}{kT} \right) \right) \quad (42)$$

where  $Z_{\text{inz}}$  is the partition function over all ionization microstates. It is shown (Vorobjev et al., 2008) that a direct calculation of the free energy from partition function

$$\Delta G_{\text{inz}}(\mathbf{x}; \text{pH}) = -kT \ln \left\{ \sum_{\mathbf{z}}^{2^{\xi}} \exp \left[ \frac{-\Delta G(\mathbf{x}, \mathbf{z}, \text{pH})}{kT} \right] \right\} \quad (43)$$

and calculation by the integral, Eq. (40), give well-coincided numerical values for protein BPTI. The free energy  $\Delta G(\mathbf{x}, \mathbf{z}, \text{pH})$  has an electrostatic nature and can be represented as the sum of energies of ionization of individual titratable residues and energies of their pair electrostatic interactions in solution

$$\Delta G(\mathbf{x}, \mathbf{z}, \text{pH}) = \sum_{i=1}^{\xi} z_i [\theta_i k_B T \ln 10 (\text{pH} - \text{pK}_{\text{mod},i}) + (\Delta g_i(\mathbf{x}) - \Delta g_{\text{mod},i})] + \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}}^N z_i z_j \Delta w_{ij}(\mathbf{x}) \quad (44)$$

where  $\Delta g_i(\mathbf{x})$  is an increment of the total electrostatic energy in a solvent due to ionization of one titratable group  $i$  of the protein with all other titratable groups kept in the zero ionization state;  $\Delta g_{\text{mod},i}$  is an increment of the total electrostatic energy of the model compound  $i$  due to its ionization in a solvent;  $\text{pK}_{\text{mod},i}$  is an ionization constant of the model compound  $i$ ; and  $\Delta w_{ij}(\mathbf{x})$  is the excess electrostatic potential between ionized sites  $i, j$  with respect to the nonionized sites.

In general, the total energy  $\Delta G_{\text{inz}}(\mathbf{x}, \text{pH})$  of Eq. (43) can be presented relative to any reference ionization microstate  $\mathbf{z}^r$ . Assuming that the  $\Delta G_{\text{inz}}^r(\mathbf{x}, \text{pH})$  is the free energy of ionization of the protein at a given pH with respect to the reference ionization microstate  $\mathbf{z}^r$ , from Eq. (43), one obtains

$$\Delta G_{\text{inz}}^r(\mathbf{x}, \text{pH}) + G(\mathbf{x}, \mathbf{z}^r, \text{pH}) = \Delta G_{\text{inz}}^0(\mathbf{x}, \text{pH}) + G(\mathbf{x}, \mathbf{z}^0, \text{pH}) \quad (45)$$

From Eq. (45), it follows that the energy  $\Delta G_{\text{inz}}^r(\mathbf{x}, \text{pH})$  has a minimal absolute value if the reference ionization microstate  $\mathbf{z}^r$  is equal to the most probable ionization microstate  $\mathbf{z}^p$  with minimal energy  $G(\mathbf{x}, \mathbf{z}^p, \text{pH})$ . The expression for the free energy  $\Delta G_{\text{inz}}^p(\mathbf{x}, \text{pH})$  follows from Eqs. (44) to (45)

$$\Delta G_{\text{inz}}^p(\mathbf{x}, \text{pH}) = \Delta G_{\text{inz}}^0(\mathbf{x}, \text{pH}) - \sum_{i=1}^{\xi} z_i^p \theta_i \left[ k_B T \ln 10 (\text{pH} - \text{pK}_{\text{mod},i}) + (\Delta g_i(\mathbf{x}) - \Delta g_{\text{mod},i}) \right] - \frac{1}{2} \sum_{\substack{i,j \\ i \neq j}}^{\xi} z_i^p z_j^p \Delta w_{ij}(\mathbf{x}) \quad (46)$$

where  $z_i^p$  is ionization degree of the titratable site  $i$  in the most probable ionization microstate  $\mathbf{z}^p$ . Finally, the total free energy  $G(\mathbf{x}, \text{pH})$  of a protein



in water–proton bath can be presented relative to the most probable ionization microstate  $\mathbf{z}^P$

$$G(\mathbf{x}, \text{pH}) = U^P_{\text{mol}}(\mathbf{x}) + [G^P_{\text{cav}}(\mathbf{x}) + G^P_{\text{pol}}(\mathbf{x})] + \Delta G^P_{\text{inz}}(\mathbf{x}, \text{pH}) \quad (47)$$

The first three terms of this equation describe physically real protein structure in the ionization microstate  $\mathbf{z}^P$ ; the IT-PMF  $\Delta G^P_{\text{inz}}(\mathbf{z}, \text{pH})$  describes correction due to deviations of the equilibrium ensemble of ionization microstates from the ionization microstate  $\mathbf{z}^P$ . The PMF  $\Delta G^P_{\text{inz}}(\mathbf{z}, \text{pH})$  of implicit titration has a minimal amplitude for the optimal ionization microstate  $\mathbf{z}^P$ .

Atomic forces produced by the IT-PMF  $\Delta G^P_{\text{inz}}(\mathbf{x}, \text{pH})$ , Eq. (46), are defined by expression

$$\begin{aligned} \frac{\partial \Delta G^P_{\text{inz}}(\mathbf{x}; \text{pH})}{\partial \mathbf{r}_l} &= \left\langle \frac{\partial \Delta G(\mathbf{x}, \mathbf{z}, \text{pH})}{\partial \mathbf{r}_l} \right\rangle_{\mathbf{z}} - \sum_{i=1}^{\xi} z_i^P \frac{\partial}{\partial \mathbf{r}_l} \Delta g_i(\mathbf{x}) \\ &\quad - \frac{1}{2} \sum_{i \neq j}^{\xi} z_i^P z_j^P \frac{\partial}{\partial \mathbf{r}_l} \Delta w_{ij}(\mathbf{x}) = \sum_{i=1}^{\xi} [\langle z_i \rangle - z_i^P] \frac{\partial}{\partial \mathbf{r}_l} \Delta g_i(\mathbf{x}) \\ &\quad + \frac{1}{2} \sum_{i \neq j}^{\xi} [\langle z_i z_j \rangle - z_i^P z_j^P] \frac{\partial}{\partial \mathbf{r}_l} \Delta w_{ij}(\mathbf{x}) \end{aligned} \quad (48)$$

Electrostatic energy  $\Delta g_i(\mathbf{x})$  of ionization of the titratable group  $i$ , energies of pair interactions  $\Delta w_{ij}(\mathbf{x})$  of titratable groups  $i, j$ , the optimal ionization microstate  $\mathbf{z}^P = (z^P_1, \dots, z^P_\xi)$  and average ionization degrees  $\langle z_i \rangle$ , and pair correlations  $\langle z_i z_j \rangle$  of ionization degrees of titratable groups  $i, j$  are calculated by the method FAMBEpH (Vorobjev, et al., 2008). An effective calculation of the gradients  $\Delta g_i(\mathbf{x})$  and  $\Delta w_{ij}(\mathbf{x})$  over coordinate of atom  $\mathbf{r}_i$  is done in the framework of the GB method with “ideal” FAMBE defined Born radii.

### C. Method FAMBEpH–GB

Method FAMBEpH–GB is the method FAMBEpH conjugated with the GB model which uses the FAMBE defined “ideal” Born radii. Method FAMBEpH solves the Poisson equation by the FAMBE method and calculates a set of partial polarization densities  $\sigma_i(\mathbf{s})$  generated by each atom  $i$  with atomic charge  $q_i$  on the MS of protein molecule. The energy of

solvent polarization  $G_{\text{pol}}(\mathbf{x})$  of the FAMBE method depends on surface integral over the protein MS

$$\begin{aligned}
 G_{\text{pol}}(\mathbf{x}) &= \frac{1}{2} \sum_i q_i \int_S \frac{\sigma_i(\mathbf{s}) d\mathbf{s}}{|\mathbf{x}_i - \mathbf{s}|} + \frac{1}{2} \sum_{i \neq j} q_i \int_S \frac{\sigma_j(\mathbf{s}) d\mathbf{s}}{|\mathbf{x}_i - \mathbf{s}|} \\
 &= \sum_i g_i(\mathbf{x}) + \frac{1}{2} \sum_{i \neq j} w_{ij}(\mathbf{x})
 \end{aligned}
 \tag{49}$$

where  $g_i$  is the energy of solvent self-polarization by atom  $i$ , and  $w_{ij}$  is the pair interaction of atoms  $i, j$  due to the solvent polarization. The FAMBE method is very effective to calculate a full set of partial atomic polarization densities  $\sigma_i(\mathbf{s})$ , polarization energy, and atomic forces for a given protein conformation  $\mathbf{x}$ . However, a multiple calculation of the forces on the fly by the FAMBE method for MD simulations is still time consuming. To accelerate calculation of electrostatic interactions and atomic forces due to solvent polarization, the method FAMBEpH is conjugated with the GB method. The FAMBEpH-GB method calculates a full set of “ideal” atomic Born radii  $B_i$  for a given conformation  $\mathbf{x}$  of the protein calculating the set of self-polarization energies  $g_i(\mathbf{x})$  for all protein atoms by the method FAMBE

$$g_i(\mathbf{x}) = \frac{q_i}{2} \int_S \frac{\sigma_i(\mathbf{s}) d\mathbf{s}}{|\mathbf{r}_i - \mathbf{s}|}
 \tag{50}$$

where  $S$  is the protein MS. The “ideal” Born atomic radii  $B_i$  of the atom  $i$  is the radius which gives the GB self-polarization energy equal to the one defined by the FAMBE method, that is

$$B_i = \left( \frac{1}{D_0} - \frac{1}{D_1} \right) \frac{q_i^2}{g_i(\mathbf{x})}
 \tag{51}$$

The total energy of solvent polarization of the GB method is a sum of atomic self-polarization energies,  $g_i^{\text{GB}}$  and the PMF  $w_{ij}^{\text{GB}}$  of pair interactions of atoms  $i, j$

$$\begin{aligned}
 G_{\text{pol}}^{\text{GB}}(r) &= \left( \frac{1}{D_0} - \frac{1}{D_1} \right) \sum_i \frac{q_i^2}{B_i} + \frac{1}{2} \sum_{i \neq j} \frac{q_i q_j}{f_{\text{GB}}(r_{ij}, B_i, B_j)} \left( \frac{1}{D_0} - \frac{1}{D_1} \right) \\
 &= \sum_i g_i^{\text{GB}} + \frac{1}{2} \sum_{i \neq j} w_{ij}^{\text{GB}}
 \end{aligned}
 \tag{52}$$

By definition, the GB method with “ideal” Born atomic radii reproduces exactly the values of atomic self-polarization energies  $g_i^{\text{FM}}$  calculated by the FAMBE method. It is shown by us that the pair atom–atom potentials  $w_{ij}(\text{FAMBE})$  calculated by the FAMBE method and potentials  $w_{ij}(\text{GB})$  calculated by the GB method with “ideal” Born radii are coincided with average unsigned error  $\sim 1.5\%$ . Therefore, a calculation of electrostatic energies and atomic forces for a given conformation  $\mathbf{x}$  of a protein can be done by the GB method without loss of accuracy. A reliable protocol calculating the CpHMD trajectory consists of (i) periodic update of optimal ionization microstate, (ii) calculation of MD trajectory of a protein in the “frozen” optimal ionization microstate  $\mathbf{z}_p$  during time  $t_{\text{MD}}(\text{frozen-}\mathbf{z}) \sim 1\text{--}2$  ps, (iii) periodic update of the set of “ideal” atomic Born radii  $B_p$ , which are dependent on the protein conformation, with update time  $\tau_B$ . It is found that a reasonable values of  $\tau_B \sim 0.01\text{--}0.03$  ps or  $\tau_B$  is equal to 10–30 elementary MD time steps of typical length of 0.001 ps for simulations at temperature  $T \sim 300$  K.

## V. EXAMPLES OF SIMULATIONS WITH IMPLICIT SOLVENT MODELS

### A. *Practical Advantages of an Implicit Solvent Models*

The implicit solvent models have several advantages over the explicit molecular water representation in MD simulation (Onufriev et al., 2004): (i) the computational cost associated with the use of implicit models is considerably smaller than the cost of representing water explicitly, (ii) the implicit models describe an instantaneous solvent dielectric response, which eliminate the need for the lengthy equilibration of water that is necessary in explicit water simulations, (iii) the solute molecule can more quickly explore the available conformational space due to absence of viscosity and “solvent reorganization energy barriers” associated with explicit water environment, (iv) the implicit continuum model corresponds to solvation in an infinite volume of solvent avoiding possible artifacts of solute replica electrostatic interactions in the periodic systems typically used with explicit solvent models, (v) estimating free energies of solvated structures is much more straightforward than with explicit water models. The same is true for the implicit titration methods which describe an instant response of a proton bath and eliminate the need for a vast

number of ionization microstates to model equilibrium ionization state. Therefore, a real implicit solvent model finds a wide application in biomolecular simulations. A new implicit solvent model should be carefully optimized in conjunction with particular force field to reproduce the experimental solvation energies for representative set of small molecules, the PMF of interactions between pairs of protein side chains in explicit solvent and the conformational equilibrium for peptides (Chen and Brooks, 2008; Gallicchio et al., 2009; Chen, 2010).

### *B. Free Energy of Protein Decoys and Protein Decoy Discrimination*

During the past decade, genome sequencing has revealed a vast number of new unknown sequences. The growing gap between the solved structures by the X-ray or the NMR methods increases the usefulness and interest in the development of reliable computational methods to predict unknown structures. All-atom force fields and implicit solvation models represent a very important tool for scoring and refining protein models produced by coarse grain and heuristic methods such as ROSETTA (Bradley et al., 2005), TASSER (Zhang and Skolnick, 2004), 3D-SHOTGAN (Fisher, 2003), and so forth. These methods were shown to produce sets of models which contain relatively accurate native-like models, but these methods are usually not able to identify the native-like conformations reliably among a set of other nonnative conformations. A necessary requirement for free energy prediction method to produce accurate protein structure models is that they must recognize the native state of the protein or a set of similar native-like conformations as models having lowest free energies. Scoring of a large set of protein models to discriminate native or near-native conformations from nonnative structures is a quality test which is carried out for free energy models.

Tests on a set of misfolded proteins have shown that solvation term is important part of the total free energy of protein in a solvent and improves success rate of discrimination native structure from decoys (Lazaridis and Karplus, 1998). The CHARMM 19 force field with GB solvent model was able to identify the misfolded structures with more than 90% accuracy (Dominy and Brooks, 2002). A high success rate has been reported for discrimination test of a set of protein decoys performed by Felts et al. (2002) using a local energy minimization with OPLS all-atom force field

and GBNP implicit solvent model (Gallicchio et al., 2002). Native structures have a lowest free energy for almost 90% of proteins considered (Felts et al., 2002). Later, Wroblewska and Skolnick (2007) found that a long MD relaxation of protein decoys with AMBER/GB force field led to significant deterioration of discriminative ability of the force field. The lowest energy structures were obtained from the short  $\sim 5$  ps native MD trajectories for 70% proteins, while a longer relaxation up to  $\sim 2$  ns decreases the success rate of discrimination of the native structures up to 20%.

The FAMBE method in conjunction with CHARMM19 force field was used by Vorobjev et al. (1998) and Vorobjev and Hermans (2001) for estimation of an average protein solvation energy  $\langle W(\mathbf{x}, \mathbf{q}) \rangle$ , Eq. (5), over an equilibrated MD trajectories of  $\sim 50$ – $100$  ps obtained for protein decoy structures in explicit water. It was found that for all proteins of Park and Lewit decoy set and for a set of the CASP3 protein models the native structures were correctly found to be more stable than decoy structures for all proteins considered Vorobjev and Hermans (2001).

It was recognized that discriminative ability of a force field and solvation model depends on quality of protein decoy set and on the protocol used to compute free energies of protein decoys (Vila et al., 2005; Arnautova et al., 2009). When local energy minimization or a short MD trajectory is substituted by a long MD trajectory of nanosecond timescale, decoy conformations become to be well relaxed within a given force field and solvation model, unfavorable atom–atom contacts disappear, and discrimination of native-like structure from a set of competing decoys becomes being a real challenge. It was shown (Arnaudova et al., 2009) that discriminative accuracy on a high quality independently generated decoy set of the ECEPP05 force field (Arnaudova et al., 2006) combined with FAMBEpH solvation–ionization model, Eq. (8) (Vorobjev et al., 2008), and structure relaxation is superior with success rate  $\sim 89\%$ , compared to other less elaborated solvation models.

### C. Protein Folding

All-atom MD simulations with implicit GB and nonpolar solvation of *ab initio* protein folding have been reported for small proteins, 20-residue trp-cage protein (Simmerling et al., 2002; Lee and Olson, 2010), 36-residue

villin headpiece, and 46-residue helix bundle (Onufriev, 2008). The average helicity and conformational equilibria of  $\beta$ -hairpin of several model peptides versus temperature have been modeled by Chen (2010) using the improved GBSW/MS2 implicit solvation model. The MD folding simulations of trp-cage protein “NLYIQWLKDGGPSSGRPPS” were structurally determined by NMR (PDB ID: 1L2Y). Lee and Olson (2010) have used GBMV2 implicit solvation model and replica exchange script *aarex.pl* of the CHARMM simulation package of version c33b2 in conjunction with the self-guided Langevin dynamics. Several simulations each of length 100–200 ns were performed starting from extended model of protein structure or experimental NMR-defined structure. These simulations allowed one to reconstruct free energy folding landscape, estimate the melting temperature of the native structure with a reasonable accuracy of about 10° and gain insights into folding mechanism.

#### D. Constant pH MD Simulations

The charges of atoms of titratable groups are not being fixed during conformational changes of protein in a solvent at fixed pH value, as it is assumed in commonly used MD simulations. The protein atomic charges and conformation are strongly coupled, and this coupling can affect protein folding pathway and stability of different conformations. Over past decade, several methods have been proposed which enable MD to be carried out at constant pH with changing protonation states. These CpHMD methods can be classified into two categories: methods of explicit stochastic titrations, which consider physical ionization microstates and methods of implicit titrations without explicit references to titratable protons.

The explicit stochastic titration methods (Mongan et al., 2004; Baptista et al., 1999; Machuqueiro and Baptista, 2006; Williams et al., 2010) operate directly in physical ionization phase space; ionization microstates are randomly generated and accepted or rejected on the fly according to Metropolis criterion of the Monte-Carlo method during the course of the MD simulation. The GB model was used to calculate (i) relative free energies of ionization microstates and (ii) atomic forces to propagate MD simulation of the accepted ionization microstate. Due to a large number of ionization

microstates  $\sim 10^9$  for a medium-sized protein with  $\sim 30$  titratable residues, a convergence of that hybrid MC–MD method is slow (Williams et al., 2010).

The recent works (Lee et al., 2004; Khandogin and Brooks, 2005; Khandogin and Brooks, 2006; Khandogin et al., 2006) represent an advanced realization of the explicit  $\lambda$ -titration method using the method of  $\lambda$ -dynamic (Kong and Brooks, 1996) to simulate proton binding/release by a set of titratable sites. The method of Lee et al. (2004) describes the protonation/deprotonation process at titrating residues by a set of continuous variables  $\lambda_i$ , the end points of which define deprotonated ( $\lambda_i=1$ ) and protonated ( $\lambda_i=0$ ) states. The total phase space of a titratable protein is extended by the set of titration variables  $\lambda_i$ . The  $\lambda$ -titration free energy profile for each titratable proton is predefined by additional simulations with GBSW solvent model. To produce movement along the titration variables  $\lambda_i$ , the Nose–Hoover method (Nose, 1984; Hoover, 1985) is used. The titration variables  $\lambda_i$  are coupled with effective masses which move in the effective  $\lambda_i$ -dependent potential by conventional MD method. The protonation dependence of the electrostatic energies is incorporated through the atomic partial charges on atoms of the titrating residue, which are interpolated between the values in the protonated and deprotonated states linearly over  $\lambda_i$ . Further development of the  $\lambda$ -titration method (Khandogin and Brooks, 2006) includes a coupling of continuous titration with replica exchange method and an improved GBSW solvent model with salt-screening Debye function for energy and force calculation. The REX protocol enables to enhance sampling of protonation and conformational states; the accuracy of the REX CpHMD method is demonstrated by 1 ns titration simulation of 10 proteins. The experimental  $pK_a$  values of these proteins are reproduced with RMSD of 0.6–1.2 with maximum errors of 1.0–4.2 pK units for buried residues.

The CpHMD simulation method with implicit titration PMFs, Eqs. (43) and (44), without explicit reference to ionization microstates, and titratable protons are ongoing research which is not explored in details. The CpHMD with implicit titration PMF in general potentially is more effective approach than available explicit titration methods because the implicit titration PMF describes an instantaneous response of proton bath, which eliminate the need to run a large ensemble of explicit ionization microstates for each protein conformation to model average equilibrium ionization degree of titratable groups at given pH.

*E. Limitations of Current Implicit Continuum Solvent Models and Further Direction*

The fundamental approximation of finite-sized molecular solvent by continuum solvent eliminates a number of effects that depend on the finite size of water molecules, such that (i) water–water correlations and inhomogeneous solvent density and its dielectric response, in a vicinity of protein MS and inside a deep pockets or grooves; (ii) tightly bound water molecules and water bridges, which may be important for stability of macromolecular structure and protein folding; (iii) coupling between values of atomic charges and radii defining SEV and respective MS. These limitations of modern implicit solvent models are well known and can be partially corrected introducing new features to the available models. An important limitation for transferability of implicit solvation models is that their optimized parameters are tightly coupled with the specific molecular force field used for modeling intramolecular potential energy, and therefore, they are generally are not transferable. Use of the MS surface in the modern implicit models GBSV/MS2 (Chen, 2010) and AGBNP2 model (Gallicchio, et al., 2009) instead of van der Waals atomic surface implicitly accounts for the finite size of water molecule and considerably improves quality of that models in reproducing of a desolvation maximum of the accurate PMF between charged groups calculated with explicit solvent via free energy perturbation simulations. The most recent AR6 GB model (Aguilar et al., 2010) looks very promising if a better and still fast approximation for the SEV or the MS will be developed. The AGBNP2 model (Gallicchio et al., 2009) introduces implicit short-range hydrogen bonding correction function, which describes an additional hydrogen bonding interaction of solute donor (acceptor) with virtual solvent water molecule. Water bridge hydrogen bonding between two solute atoms can be modeled also by a similar method. Dependence of atomic radii defining solute–solvent dielectric interface on atomic charges has been recently studied. It was found that linear function of atomic radii on atomic charges describes this dependence reasonably and has been parametrized for small molecules (Hou et al., 2010). In general, two independent atomic radii sets can be considered for calculating the solute–solvent dielectric surface interface for continuum dielectric model, the first one is for neutral and the second one is for ionized residues of a protein, respectively. Thereby the modern implicit solvent



models demonstrate a number of options for self-improvements to become more accurate and fast in approximations of the most detailed explicit solvent model. It is likely that improvements in the implicit solvent models accompanied by careful optimizations of implicit model empirical parameters with accumulation of practical experience will make the implicit solvent models a standard well-defined powerful option of a modern simulation packages for computational structural biology.

#### ACKNOWLEDGMENTS

This work was supported by a grant from the Russian Fund of Basic Research #09-04-00136, by grants (#26-2009 and #119-2009) from the Siberian Branch of Russian Academy of Science, by the Grant for Scientific Schools (#3185.2010.4), by Ministry of Education and Science of Russian Federation # 02.740.11.0079.

#### REFERENCES

- Aguilar, B., Shadrach, R., Onufriev, A. V. (2010). Reducing the secondary structure bias in the generalized Born model via R6 effective Radii. *J. Chem. Theory Comput.* **6**, 3613–3630.
- Aqvist, J., Hansson, T. (1996). On the validity of electrostatic linear response in polar solvent. *J. Phys. Chem.* **100**, 9512–9521.
- Arnautova, E. Y., Jagielska, A., Scheraga, H. A. (2006). A new force field ECEPP05 for peptides, proteins and organic molecules. *J Phys. Chem. B* **110**, 5025–5044.
- Arnautova, E. Y., Vorobjev, Y. N., Vila, J. A., Scheraga, H. A. (2009). Identifying native-like protein structures with scoring functions based on all-atom ECEPP force fields, implicit solvent models and structure relaxation. *Proteins* **77**, 38–51.
- Baptista, M., Martel, P. J., Petersen, S. B. (1997). Simulation of protein conformation freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins* **27**, 523–544.
- Baptista, M., Martel, P. J., Soares, C. M. (1999). Simulation of electron-proton coupling with a Monte Carlo method: application to cytochrome c(3) using continuum electrostatics. *Biophys. J.* **76**, 2978–2998.
- Bashford, D., Case, A. D. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129–152.
- Beglov, D., Roux, B. (1996). An integral equation to describe the solvation of polar molecules in liquid water. *J. Chem. Phys.* **104**, 8678–8689.
- Beglov, D., Roux, B. (1997). Solvation of complex molecules in a polar liquid: an integral equation theory. *J. Phys. Chem.* **101**, 7821–7826.
- Ben-Naim, A. (1990). Solvent effects on protein association and protein folding. *Biopolymers* **29**, 567–596.

- Ben-Naim, A., Marcus, Y. (1984). Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **81**, 2016–2027.
- Bharadwaj, R., Windemuth, A., Sridharan, S., Honig, B., Nicholls, A. (1995). The fast multipole boundary element method for molecular electrostatics: an optimal approach for large systems. *J. Comput. Chem.* **16**, 898–913.
- Bogusz, S., Cheatham, T. E., III., Brooks, R. R. (1998). Removal of pressure and free energy artifacts in charged periodic system via net charge corrections to the Ewald potential. *J. Chem. Phys.* **108**, 7070–7084.
- Bradley, P., Misura, K. M., Baker, D. (2005). Towards high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871.
- Brooks, C. L., III., Karplus, M., Pettitt, B. M. (1988). Proteins a theoretical perspectives of dynamics, structure and thermodynamics. In: *Advance in Chemical Physics*, vol. LXXI, Prigogine, I. and Rice, S. A. (Eds.). John Wiley and Sons, New York.
- Chen, J. (2010). Effective approximation of molecular volume using atom-centered dielectric functions in generalized Born models. *J. Chem. Theory Comput.* **6**, 2790–2803.
- Chen, J., Brooks, C. (2007). Critical importance of length-scale dependence in implicit modeling of hydrophobic interactions. *J. Am. Chem. Soc.* **129**, 2444–2445.
- Chen, J., Brooks, C. (2008). Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.* **10**, 471–481.
- Chothia, C. H. (1974). Hydrophobic bonding and accessible area in proteins. *Nature* **248**, 338–339.
- Connolly, M. L. (1983a). Analytical molecular surface calculation. *J. Appl. Crystallogr.* **16**, 548–558.
- Connolly, M. L. (1983b). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713.
- Connolly, M. L. (1985). Computation of molecular volume. *J. Am. Chem. Soc.* **107**, 1118–1124, <http://www.netsci.org/Science/Compchem/feature14e.html>.
- Curutchet, C., Cramer, C. J., Truhlar, D. G., Ruiz-Lopez, M. F., Rinaldi, D., Orozco, M., et al. (2003). Electrostatic component of solvation: comparison of SCRF continuum models. *J. Comput. Chem.* **24**, 284–297.
- Douglas, C. C. (1996). Multigrid methods in science and engineering. *Comput. Sci. Eng.* **3**, 55–68.
- Dominy, B. N., Brooks, C. L. (2002) Identifying native-like protein structures using physics-based potentials. *J. Comput. Chem.* **23**, 147–160
- Feig, M., Onufriev, A., Lee, M., Im, W. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **25**, 265–284.
- Felts, A. K., Gallicchio, E., Wallqvist, A., Levy, R. M. (2002). Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized Born solvent model. *Proteins* **48**, 404–422.
- Fisher, D. (2003). 3D-SHORTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins* **51**, 434–444.

- Fogolari, F., Esposito, G., Viglino, P., Molinari, H. (2001). Molecular mechanics and dynamics of biomolecules using a solvent continuum model. *J. Comput. Chem.* **22**, 1830–1842.
- Gallicchio, E., Levy, R. (2004). AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.* **25**, 479–499.
- Gallicchio, E., Kubo, M. M., Levy, R. M. (2000). Enthalpy-entropy and cavity decomposition of alkane hydration free energies: Numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B* **104**, 6271–6285.
- Gallicchio, E., Zhang, L. Y., Levy, R. M. (2002). The SGB/NP hydration free energy model based on the surface generalized Born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem.* **23**, 517–529.
- Gallicchio, E., Paris, K., Levy, R. (2009). The AGBNP2 implicit solvation model. *J. Chem. Theory Comput.* **5**, 2544–2564.
- Goel, N. S., Gang, F., Ko, Z. (1995). Electrostatic field in inhomogeneous dielectric media. Indirect boundary element method. *J. Comput. Phys.* **118**, 172–179.
- Grant, A., Pickup, B., Nicholls, A. (2001). A smooth permittivity function for Poisson-Boltzmann solvation methods. *J. Comput. Chem.* **22**, 608–640.
- Grycuk, T. (2003). Deficiency of the Coulomb-field approximation in the Generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.* **119**, 4817–4826.
- Hawkins, G. D., Cramer, C. J., Truhlar, D. G. (1996). Parametrized models of aqueous free energies of solvation based pairwise solute descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **100**, 19824–19836.
- Hermann, R. B. (1972). Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **76**, 2754–2759.
- Hill, T. L. (1986). *An Introduction to Statistical Thermodynamics*. Dover, New York.
- Holst, M., Saied, F. (1995). Numerical solution of the nonlinear Poisson-Boltzmann equation: developing more robust and efficient methods. *J. Comput. Chem.* **16**, 337–364.
- Holst, M., Kozack, R. E., Saied, F., Subramaniam, S. (1994). Treatment of electrostatic effects in proteins: multigrid-based Newton iterative method for solution of the full nonlinear Poisson-Boltzmann equation. *Proteins* **18**, 231–245.
- Holst, M., Baker, N., Wang, M. (2000). Adaptive multilevel finite element solution of the Poisson-Boltzmann equation. I. Algorithms and examples. *J. Comput. Chem.* **21**, 1319–1342.
- Honig, B., Sharp, K., Yang, A. S. (1993). Macroscopic models of aqueous solutions: biological and chemical applications. *J. Phys. Chem.* **97**, 1101–1109.
- Hoover, W. G. (1985). Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695–1697.
- Hou, G., Zhu, X., Cui, Q. (2010). An implicit solvent model for SCC-DFTB with charge-dependent radii. *J. Chem. Theory Comput.* **6**, 2303–2314.
- Hummer, G., Pratt, L. R., Garcia, A. E. (1995). Hydration free energy of water. *J. Phys. Chem.* **99**, 14188–14194.

- Hummer, G., Pratt, L. R., Garcia, A. E. (1996). Free energy of ionic hydration. *J. Phys. Chem.* **100**, 1206–1215.
- Hünneberg, P. H., McCammon, J. A. (1999). Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatic study. *Biophys. Chem.* **78**, 69–88.
- Im, W., Lee, M. S., Brooks, C. L., III. (2003). Generalized Born model with a simple smoothing function. *J. Comput. Chem.* **24**, 1691–1702.
- Jackson, J. D. (1975). *Classical Electrodynamics*. Wiley, New York.
- Jackson, R. M., Sternberg, J. E. (1994). Application of scaled particle theory to model the hydrophobic effect: implications for molecular association and protein stability. *Protein Eng.* **7**, 371–383.
- Jackson, R. M., Sternberg, J. E. (1995). A continuum model for protein-protein interactions: applications to the docking problem. *J. Mol. Biol.* **250**, 258–275.
- Jayaram, B., Fine, R., Sharp, K., Honig, B. (1989). Free energy calculations of ion hydration: an analysis of the Born model in terms of microscopic simulations. *J. Phys. Chem.* **93**, 4320–4327.
- Jorgensen, W. L., Madura, J. D. (1985). Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Mol. Phys.* **56**, 1381–1392.
- Jorgensen, W., Tirado-Rives, J. (2004). Free energies of hydration from a generalized born model and an all-atom force field. *J. Phys. Chem. B.* **108**, 16264–16270.
- Jorgensen, W. L., Maxwell, D. S., Tirado-Rives, J. J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225–11236.
- Juffer, A. H., Botta, E. F. F., Bert, A. M., van Keulen, B. A. M., van der Ploeg, A., Berendsen, H. J. C. (1991). The electric potential of a macromolecule in a solvent: a fundamental approach. *J. Comput. Phys.* **97**, 144–171.
- Juffer, A. H., Eisenbaher, S. J., Hubbard, S. J., Walter, D., Argos, P. (1995). Comparison of atomic solvation parametric sets: applicability and limitations in protein folding and binding. *Protein Sci.* **4**, 2499–2509.
- Kar, P., Wei, Y., Hansmann, U. E., Höfinger, S. (2007). Systematic study of the boundary composition in Poisson Boltzmann calculations. *J. Comput. Chem.* **28**, 2538–2544.
- Karplus, M., McCammon, A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652.
- Khandogin, J., Brooks, C. L., III. (2005). Constant pH Molecular Dynamics with Proton Tautomerism. *Biophys. J.* **89**, 141–157.
- Khandogin, J., Brooks, C. L., III. (2006). Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochemistry* **2006**(45), 9363.
- Khandogin, J., Chen, J., Brooks, C. L., III. (2006). Exploring atomistic details of pH-dependent peptide folding. *PNAS* **103**, 18546–18550.
- Kollman, P. (1993). Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.* **93**, 2395–2417.
- Kollman, P., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**, 889–897.

- Kong, X., Brooks, C. L., III. (1996).  $\lambda$ -dynamics: a new approach to free energy calculations. *J. Chem. Phys.* **105**, 2414–2423.
- Landau, L. D., Lifshitz, E. M. (1988). *Electrodynamics of Continuous Media*. Volume 8 of Course of Theoretical Physics. Translated from the Russian. Pergamon Press, Oxford.
- Langlet, J., Claverie, P., Caillet, J., Pullman, A. (1988). Improvements of the continuum model. Application to the calculation of the vaporization thermodynamic quantities of non-associated liquids. *J. Phys. Chem.* **92**, 1617–1631.
- Lazaridis, T., Karplus, M. (1998). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* **288**, 477–487.
- Lee, M. R., Duan, Y., Kollman, P. A. (2000). Use of MM-PB/SA in estimating the free energies of proteins: application to native, intermediates, and unfolded villin headpiece. *Proteins* **39**, 309–316.
- Lee, M. S., Feig, M., Salsbury, F. R., Jr., Brooks, C. L., III. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **24**, 1348–1356.
- Lee, M. S., Salsbury, F. R., Jr., Brooks, C. L., III. (2004). Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738–752.
- Lee, M. S., Olson, M. A. (2010). Protein Folding Simulations Combining Self-Guided Langevin Dynamics and Temperature-Based Replica Exchange. *J. Chem. Theory Comput.* **2010**, 2477–2487.
- Levy, R. M., Belhadj, M., Kitchen, D. B. (1991). Gaussian fluctuation formula for electrostatic free energy changes in solution. *J. Chem. Phys.* **95**, 3627–3633.
- Levy, R. M., Zhanh, L. Y., Gallicchio, E., Felts, A. (2003). On the non polar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. *J. Am. Chem. Soc.* **25**, 9523–9530.
- Lounnas, V., Pettitt, B. M., Phillips, G. N., Jr. (1994). A Global model of protein solvent interface. *Biophys. J.* **66**, 601–614.
- Lu, B., McCammon, A. (2007). Improved boundary element method for Poisson-Boltzmann electrostatic potential and force calculations. *J. Chem. Theory Comput.* **3**, 1134–1142.
- Lu, B., Cheng, X. L., Hang, J. F., McCammon, A. (2006). Order N algorithm for computation of electrostatic interactions in biomolecular systems. *Proc. Natl. Acad. Sci. USA* **103**, 19314–19319.
- Machuqueiro, M., Baptista, A. M. (2006). Constant-pH molecular dynamics with ionic strength effects: protonation-conformation coupling in decalysine. *J. Phys. Chem.* **110**, 2927–2933.
- Madura, J. D., Davis, M. E., Gilson, M. K., Wade, R. C., Luty, B. A., McCammon, J. A. (1994). Biological application of electrostatic calculations and Brownian dynamics simulations. *Rev. Comput. Chem.* **5**, 229–267.
- McDowell, S. C., Špackova, N., Šponer, J., Walter, N. G. (2007). Molecular dynamics simulations of RNA: an in silico single molecule approach. *Biopolymers* **85**, 169–184.
- McKenney, A., Greengard, L. (1995). A fast Poisson solver for complex geometries. *J. Comput. Phys.* **118**, 348–355.

- Mongan, J., Case, D. A., McCammon, J. A. (2004). Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **25**, 2038–2064.
- Mongan, J., Simmerling, C., McCammon, J., Case, D., Onufriev, A. (2007a). A generalized Born model with a simple, robust molecular volume correction. *J. Chem. Theory Comput.* **3**, 156–159.
- Mongan, J., Svrcek-Seiler, W. A., Onufriev, A. (2007b). Analysis of integral expressions for effective Born radii. *J. Chem. Phys.* **127**, 18510–18521.
- Nina, M., Beglov, D., Roux, B. (1997). Atomic radii for continuum electrostatic calculations based on molecular dynamics free energy simulations. *J. Phys. Chem.* **101**, 5239–5248.
- Nina, M., Im, W., Roux, B. (1999). Optimized atomic radii for protein continuum electrostatic solvation forces. *Biophys. Chem.* **78**, 89–96.
- Nose, S. J. (1984). A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511–519.
- Novotny, J., Brucooleri, R. E., Davis, M., Sharp, K. A. (1997). Empirical free energy calculations: a blind test and further improvements of the method. *J. Mol. Biol.* **268**, 401–411.
- Onufriev, A. (2008). Implicit solvent models in molecular dynamics simulations: a brief overview. *Annu. Rep. Comput. Chem.* **4**, 125–137.
- Onufriev, A., Case, D., Bashford, D. (2002). Effective Born radii the generalized Born approximation: the importance of being perfect. *J. Comput. Chem.* **23**, 1297–1304.
- Onufriev, A., Bashford, D., Case, D. (2004). Exploring protein native states and large scale conformational changes with modified generalized Born model. *Proteins* **55**, 383–394.
- Ooi, T., Oobatake, M., Nemethy, G., Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA* **84**, 3086–3090.
- Pellegrini, E., Field, M. J. (2002). A generalized-born solvation model for macromolecular hybrid-potential calculations. *J. Phys. Chem. A* **106**, 1316–1326.
- Perrot, G. B., Cheng, B., Gibson, K. D., Vila, J., Palmer, K. A., Nayeem, A., et al. (1992). MSEED: a program for rapid analytical determination of accessible surface areas and their derivatives. *J. Comput. Chem.* **13**, 1–11.
- Pierotti, R. A. (1976). A scaled particle theory of aqueous and non-aqueous solutions. *Chem. Rev.* **76**, 717–726.
- Postma, J. P. M., Berendsen, H. J. C., Haak, J. R. (1982). Thermodynamics of cavity formation in water. *Faraday Symp. Chem. Soc.* **17**, 55–67.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. (1988). *Numerical Recipes in C*. Cambridge University Press, Cambridge.
- Radmer, R. J., Kollman, P. A. (1997). Free energy calculation methods: a theoretical and empirical comparison of numerical errors and a new method for qualitative estimates of free energy changes. *J. Comput. Chem.* **18**, 902–919.
- Rashin, A. A. (1990). Hydration phenomena, classical electrostatics, and the boundary element method. *J. Phys. Chem.* **94**, 1725–17358.
- Rashin, A. A., Young, L., Topol, I. A. (1994). Quantitative evaluation of hydration thermodynamics with continuum model. *Biophys. Chem.* **51**, 359–374.

- Reiss, H., Frisch, H. L., Helfand, E., Lebovitz, J. L. (1960). Aspects of the statistical thermodynamic of real fluids. *J. Chem. Phys.* **32**, 119–124.
- Richards, F. M. (1977). Areas, volume, packing and protein structures. *Annu. Rev. Biophys. Biophys. Chem.* **19**, 301–332.
- Rick, S. W., Berne, B. J. (1994). The aqueous solvation of water: a comparison of continuum methods with molecular dynamics. *J. Am. Chem. Soc.* **116**, 3949–3954.
- Ripoll, D. R., Vorobjev, Y. N., Liwo, A., Vila, J. A., Scheraga, H. A. (1996). Coupling between folding and ionization equilibria: effect of pH on the conformational preferences of polypeptides. *J. Mol. Biol.* **264**, 770–783.
- Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., Honig, B. (2002). Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.* **23**, 128–137.
- Roux, B., Simonson, T. (1999). Implicit solvent models. *Biophys. Chem.* **78**, 1–20.
- Roux, B., Yu, H. A., Karplus, M. (1990). Molecular basis for the Born model of ion solvation. *J. Phys. Chem.* **94**, 4683–4688.
- Sanner, M. F., Olson, A. J., Spehner, J. C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320.
- Schaefer, M., Froemmel, C. (1990). A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.* **216**, 1045–1066.
- Schellman, J. A. (1975). Macromolecular binding. *Biopolymers* **14**, 999–1018.
- Scheraga, H. A. (1998). Theory of hydrophobic interactions. *J. Biomol. Struct. Dynamics* **16**, 447–460.
- Sharp, K. A., Honig, B. (1990). Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Chem.* **19**, 301–332.
- Simonson, T., Brünger, A. (1994). Solvation free energies estimated from macroscopic continuum theory: an accuracy assessment. *J. Phys. Chem.* **98**, 4683–4694.
- Sitkoff, D., Sharp, K. A., Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978–1988.
- Simmerling, C., Strockbine, B., Roitberg, A. E. (2002). All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259.
- Sobolevski, E., Makowski, M., Czaplewski, C., Liwo, A., Oldziej, S., Scheraga, H. A. (2007). Potential of mean force of hydrophobic association: dependence on solute size. *J. Phys. Chem. B* **111**, 10765–10774.
- Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., Case, D. A. (1998). Continuum solvent studies of stability of DNA, RNA and phosphoramidate DNA helicases. *J. Am. Chem. Soc.* **120**, 9401–9409.
- Still, W. C., Tempezyk, A., Hawley, R. C., Hendricson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129.
- Tanford, C. (1970). Protein denaturation: part C. Theoretical models for denaturation. *Adv. Protein Chem.* **24**, 1–95.
- Tanford, C., Roxby, R. (1972). The interpretation of protein titration curves. Application to lysozyme. *Biochemistry* **1972**(11), 2192–2198.

- Tomasi, J., Persico, M. (1994). Molecular interactions in solution: overview of methods based on continuum distribution of the solvent. *Chem. Rev.* **94**, 2027–2094.
- Varshney, A., Brooks, F. P., Wright, W. V. (1994). Computing smooth molecular surface. *IEEE Comput. Graph. Appl.* **14**, 19–25.
- Vila, J., Ripoll, D. R., Arnautova, Y. A., Vorobjev, Y. N., Scheraga, H. A. (2005). Coupling between conformation and proton binding in proteins. *Proteins* **61**, 56–68.
- Vorobjev, Y. N., Grant, J. A., Scheraga, H. A. (1992). A Combined Iterative and Boundary Element Approach for Solution of the Nonlinear Poisson-Boltzmann Equation. *J. Am. Chem. Soc.* **114**, 3189–3196.
- Vorobjev, Y. N., Hermans, J. (1997). SIMS, computation of a smooth invariant molecular surface. *Biophys. J.* **73**, 722–732.
- Vorobjev, Y. N., Hermans, J. (1999). ES/IS: estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. *Biophys. Chem.* **78**, 195–205.
- Vorobjev, Y. N., Hermans, J. (2001). Free energies of protein decoys provide insight into determinant of protein stability. *Protein Sci.* **10**, 2498–2506.
- Vorobjev, Y. N., Scheraga, H. A. (1997). A fast adaptive multigrid boundary element method for macromolecular electrostatics in a solvent. *J. Comput. Chem.* **18**, 569–583.
- Vorobjev, Y. N., Almagro, J. C., Hermans, J. (1998). Discrimination between native and intentionally misfolded conformation of proteins: ES/IS, new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and implicit solvent continuum model. *Proteins* **32**, 399–413.
- Vorobjev, Y. N., Vila, J., Scheraga, H. A. (2008). FAMBE-pH: a fast and accurate method to compute the total solvation free energies of proteins. *J. Phys. Chem. B* **112**, 11122–11136.
- Wagoner, J., Baker, N. (2006). Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. *Prot. Natl. Acad. Sci. USA* **103**, 8331–8336.
- Wallqvist, W., Berne, B. J. (1995a). Molecular dynamics study of the dependence of water solvation free energy on solute curvature and surface area. *J. Phys. Chem.* **99**, 2885–2892.
- Wallqvist, W., Berne, B. J. (1995b). Computer simulation of hydrophobic hydration forces on stacked plates at short range. *J. Phys. Chem.* **99**, 2893–2899.
- Williams, S. L., Oliveira, C. A. F., McCammon, J. A. (2010). Coupling constant pH molecular dynamics with accelerated molecular dynamics. *J. Chem. Theory Comput.* **6**, 560–568.
- Wroblewska, L., Skolnick, J. (2007). Can a physics-based, all-atom potential find a protein's native structure among misfolded structures?. I. Largescale AMBER benchmarking. *J. Comput. Chem.* **28**, 2059–2066.
- Yang, S. A., Honig, B. (1993). On the pH dependence of protein stability. *J. Mol. Biol.* **231**, 459–474.
- Yoon, B. J., Lenhoff, A. M. (1990). A boundary element method for molecular electrostatics with electrolyte effects. *J. Comput. Chem.* **11**, 1080–1086.



- Zauhar, R. J., Morgan, R. S. (1988). The rigorous computation of the molecular electric potential. *J. Comput. Chem.* **9**, 171–187.
- Zauhar, R. J. (1995). SMATR: a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. Comput. Aid. Mol. Des.* **9**, 149–159.
- Zauhar, R. J., Varnek, A. A. (1996). Fast and space-efficient boundary element method for computing electrostatics and hydration effects in large molecules. *J. Comput. Chem.* **17**, 864–877.
- Zhang, Y., Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* **101**, 7594–7599.
- Zhou, Z., Payne, P., Vasquez, M., Kuhn, N., Levitt, M. (1996). Finite-difference solution of the Poisson-Boltzmann equation: complete elimination of self-energy. *J. Comput. Chem.* **17**, 1344–1351.
- Zhou, Y. C., Feig, M., Wei, G. W. (2008). Highly accurate biomolecular electrostatics in continuum dielectric environments. *J. Comput. Chem.* **29**, 87–97.

## AUTHOR INDEX

### A

Abagyan, R., 192–193, 202  
Abraham, D. J., 4–5, 229–230  
Abseher, R., 202  
Abu-Sheikha, G., 231, 239–240  
Acevedo, O., 90, 92–93, 98  
Aemissegger, A., 103  
Aguilar, B., 301–302, 313–314  
Aharoni, A., 113  
Ahn, K., 3  
Ahuja, S., 256, 269, 271–273  
Ai, N., 244–245  
Akahori, Y., 241  
Albuquerque, M. A. B., 231–232  
Alexander, J. P., 3  
Alexandrova, A. N., 85, 86  
Alexandrov, V., 202  
Alexov, E., 294  
Allen, N. W., 243  
Almagro, J. C., 283, 286, 289, 292–294, 310  
Almond, H. R., 243  
Aloia, A. L., 260  
Alonso, H., 87  
Alov, P., 217–252  
Altenbach, C., 256, 268–269  
Althoff, E. A., 88–90  
Amadei, A., 97–98  
Amadio, D., 3  
Andersen, H. C., 152  
Anderson, G., 262–263  
Andrec, M., 57–58, 62  
Andreozzi, C., 88–89  
Andres, J., 90, 98–101, 102, 103, 104–105,  
107, 108–109, 112, 115,  
116–117, 119–120  
Andrews, P. R., 98–101, 102, 119–120  
Andricopulo, A. D., 5, 237–238  
Angel, T. E., 270–271  
Angers, S., 254–255

Ankley, G., 218  
An, L., 237  
Anson, J., 219–220  
Anstead, G. M., 224  
Antonietti, F., 5  
Aqvist, J., 12, 291  
Aragao, D., 255–256, 257–258  
Arendse, A., 3  
Argos, P., 289  
Arhipov, A., 198  
Arlow, D. H., 255–256, 257–258, 261–262,  
264–265, 269  
Arnautova, E. Y., 286, 310  
Arnautova, Y. A., 310  
Arnold, E., 71  
Arnold, F. H., 114–115  
Arnold, G. F., 71  
Arora, S., 244–245  
Asif-Ullah, M., 98–101, 102, 110, 119–120  
Atilgan, A. R., 191–192, 202  
Azzi, M., 254–255

### B

Baase, W. A., 59–60  
Babbitt, P. C., 88  
Babtie, A., 113–114  
Backes, A. C., 112–113  
Bahar, I., 191–192, 202, 261–263  
Bajorath, J., 4–5  
Bakan, A., 262–263  
Baker, D., 88–89, 263–264, 309  
Baker, J. G., 255–256, 260  
Baker, N. A., 53, 69–70, 289, 294  
Bakowies, D., 92–93  
Balaraman, G. S., 260, 263–264, 269  
Baldrige, K. K., 115–116  
Ballesteros, J. A., 260, 271–273  
Bambico, F. R., 2–3

- Banavali, N. K., 267  
 Bandura, A. V., 91–92  
 Banks, J. L., 233  
 Bao, X. G., 243–244  
 Baptista, A. M., 302–303, 311–312  
 Barbany, M., 98–101  
 Barbieri, C. M., 43  
 Baringhaus, K. H., 233  
 Barth, P., 263–264  
 Bartlett, P. A., 105, 106–107  
 Bashford, D., 12–13, 53, 69–70, 282–283,  
     296–298, 299, 308–309  
 Bash, P. A., 90  
 Basso, E., 11  
 Bastolla, U., 202  
 Bayly, C. I., 53  
 Beauchamp, K., 186–188  
 Beck, T. L., 38–39  
 Beglov, D., 291–292  
 Belew, R., 243  
 Belhadj, M., 291  
 Bellissent-Funel, M., 192–193, 202  
 Bellott, M., 12–13, 53  
 Benfenati, E., 229, 234–237, 239  
 Benfield, A. P., 71–72  
 Benkovic, S. J., 83–84, 107–108  
 Ben-Naim, A., 287–288  
 Bennett, C. H., 56–57  
 Benson, T. E., 15  
 Berendsen, H. J. C., 151–152, 287–288, 289,  
     292–293, 294  
 Berend Smit, B., 167, 168–169  
 Berglund, P., 113–114  
 Berman, H. M., 184–185  
 Berne, B. J., 66–67, 287–288, 291  
 Bershtein, S., 113  
 Bert, A. M., 289, 292–293, 294  
 Bertekap, R., 3  
 Bertrán, J., 81–142  
 Besold, G., 157  
 Beusen, D. D., 3–5  
 Bharadwaj, R., 292–293, 294–295  
 Bhattacharya, S., 260, 263–264, 265–268, 269,  
     271–273  
 Bhat, T. N., 184–185  
 Binder, H., 162  
 Bini, R., 128  
 Binkley, S., 231–232  
 Blair, R., 219  
 Blancato, J., 218  
 Blaney, J. M., 233  
 Bligaard, T., 125–126  
 Bliznyuk, A. A., 87  
 Boehr, D. D., 87  
 Bogusz, S., 282–283  
 Bohacek, R., 245  
 Bohr, H., 126–127  
 Bokoch, M. P., 270–271  
 Bolduc, J. M., 88–89  
 Bolon, D. N., 86  
 Bondar, A.-N., 273–274  
 Bond, R. A., 255  
 Bonn, T., 222–224  
 Bordi, F., 3–4, 5  
 Boresch, S., 30–32, 60–61  
 Borhani, D. W., 261–262, 264–265, 269  
 Bornscheuer, U. T., 87–88, 113–114  
 Botta, E. F. F., 289, 292–293, 294  
 Bottegoni, G., 202  
 Boudon, S., 60–61  
 Bourguet, W., 222  
 Bourin, C., 3  
 Bouvier, M., 254–255, 259–260  
 Bouzida, D., 57–58  
 Bowman, G. R., 186–188  
 Boyce, S. E., 51–52, 59–60, 66, 73  
 Bracey, M. H., 11  
 Bradley, P., 309  
 Branduardi, D., 4–5  
 Branham, W., 219  
 Branson, K., 28–29, 66  
 Braun, R., 186–188, 198  
 Bravi, G., 231–232  
 Breen, M., 218  
 Breitenbucher, J. G., 3  
 Brelt, A., 259–260, 264  
 Brenk, R., 59–60  
 Brent, R., 85  
 Breton, B., 259–260  
 Brezovsky, J., 88  
 Briem, H., 231–232  
 Brogi, S., 240  
 Bromley, S. T., 91–92  
 Brooijmans, N., 28

- Brooks, B. R., 155  
 Brooks, C. L., I. I., 4–5, 53, 65, 69–70,  
 185–186, 282–283, 286–287, 299–300,  
 308–309, 312  
 Brooks, F. P., 293–294  
 Brooks, R. R., 282–283  
 Brown, F. K., 243  
 Brown, F. L. H., 145–146  
 Brown, M. F., 256  
 Brown, S. P., 28–29, 66, 70  
 Bruck, J., 85  
 Brucooleri, R. E., 287–288  
 Bruce, T. C., 98–101, 104–105, 107–108  
 Brünger, A., 289, 294  
 Brunsveld, L., 225  
 Brzozowski, A. M., 222–224  
 Bucher, D., 91–92  
 Buckner, J. K., 60–61  
 Buldyrev, S. V., 194–195, 210  
 Bünemann, M., 259–260  
 Burris, K., 3  
 Bursulaya, B. D., 87–88  
 Bush, B. L., 29, 30–32, 35–36, 40–41, 59, 60  
 Bussi, G., 66–67  
 Busto, E., 113–114
- C**
- Caillet, J., 289  
 Caltabiano, G., 269  
 Camps, J., 186–188, 193–194, 208  
 Canepa, C., 124–125  
 Cannon, W. R., 83–84  
 Cantor, R. S., 164  
 Capoferri, L., 21–22  
 Carlson, H. A., 16, 98–101  
 Carlson, K. E., 224  
 Carmi, C., 5  
 Carpenter, S. H., 107–108, 109  
 Carpy, A., 228–229, 234–237  
 Carraz, M., 225  
 Carrillo, O., 186–188, 193–195, 208, 210  
 Carteni-Farina, M., 263–264  
 Carter, E. A., 85–86, 120–121  
 Casarosa, P., 255–256, 264–265  
 Case, A. D., 282–283, 296–298  
 Case, D. A., 53, 61–62, 69–70, 286, 297–298,  
 299–300, 303, 308–309, 311–312  
 Catlow, R., 129  
 Cavalli, A., 4–5  
 Cavazzoni, A., 5  
 Celik, L., 244  
 Chacon, P., 192–194, 202, 204–205  
 Chae, P. S., 255–256, 264–265  
 Chambon, P., 222  
 Chance, M. R., 270–271  
 Chang, C.-E.A., 45–46, 67–69, 70, 72  
 Charest, P. G., 254–255  
 Charifson, P. S., 72  
 Cheatham, T. E., I. I., 282–283, 286  
 Chemtob, S., 259–260  
 Cheng, B., 293–294  
 Cheng, X. L., 294–296  
 Chen, I., 227  
 Chen, I.-J., 67–68  
 Chen, J., 53, 69–70, 72, 282–283, 286–287, 299–  
 300, 308–309, 310–311, 312, 313–314  
 Chennubhotla, C., 202  
 Chen, W., 45–46, 67–69, 70  
 Chen, Z., 243–244  
 Cherezov, V., 257–258  
 Chiabrera, A., 294  
 Chipot, C., 28–29, 54–56, 65  
 Chodera, J. D., 28–29, 48, 51–52, 58–60, 66, 73  
 Choi, E. J., 263–264  
 Choi, H. J., 255–256, 264–265  
 Cho, K.-B., 125–126  
 Chong, L. T., 69–70, 282–283, 286  
 Chook, Y. M., 107–108, 109  
 Chothia, C. H., 287–288  
 Christensen, C. H., 125–126  
 Christov, C., 11, 12–13  
 Chuang, Y., 94  
 Chudyk, E., 21–22  
 Chung, K. L., 227  
 Cieplak, P., 53, 286  
 Clapper, J. R., 21–22  
 Clark, D. E., 86–87  
 Clarke, W. P., 254–255  
 Clark, T., 86  
 Claverie, P., 289  
 Cleland, W. W., 83–84  
 Clementi, C., 190–191

- Clementi, S., 238  
 Clements, J. H., 71–72  
 Coitiño, E. L., 94  
 Colizzi, F., 4–5  
 Colombo, M. C., 91–92  
 Compton, T. R., 3  
 Connolly, M. L., 287–288, 292–294, 299  
 Connor, K., 227  
 Conolly, R., 218  
 Conway, C., 3  
 Copley, S. D., 98–101, 113  
 Corchado, J. C., 94  
 Cornell, W. D., 53  
 Cornils, B., 83–84  
 Cotecchia, S., 256  
 Cramer, C. J., 93, 289, 299, 300–301  
 Cravatt, B. F., 2, 3, 11, 15  
 Cronin, M. T. D., 231, 239  
 Cruciani, G., 229, 238  
 Cui, Q., 90, 98–101, 193–194, 202, 313–314  
 Curutchet, C., 289, 299  
 Czaplewski, C., 198–199, 289
- D**
- D'Abramo, M., 186, 193–194  
 Dahiyat, B. I., 88–89  
 Dahlke, E. E., 91–92  
 Dallavalle, S., 3  
 Dal Peraro, M., 91–92  
 Damborsky, J., 88  
 Danyliv, O., 91–92  
 Darden, T. A., 59–60  
 Das, R., 88–89  
 Das, S., 240–241  
 Daura, X., 53  
 Dauter, Z., 222–224  
 Davey, J., 244  
 Davies, P., 260  
 Davis, M. E., 287–288, 294  
 Da-Wei, Li, 166–167  
 DeClue, M. S., 115–116  
 Decornez, H., 4–5  
 Deeth, R. J., 91–92  
 De Fabritiis, G., 186–188, 273–274  
 DeGrado, W. F., 88  
 de Groot, B. L., 146  
 de Kruiff, B., 164  
 Delbruck, M., 175  
 Dellerue, S., 192–193, 202  
 DeLorbe, J. E., 71–72  
 de Meyer, F., 155, 167, 168–169, 176, 177, 178  
 Demirel, M. C., 191–192, 202  
 Demyttenaere-Kovatcheva, A., 239  
 Deng, Y., 4–5, 28, 33–34, 59–61, 63, 64  
 den Heeten, R., 126–127  
 De Simone, C., 3  
 Deupi, X., 258–260, 269, 270  
 Deuss, P. J., 126–127  
 Devillers, J., 228–229, 234–237  
 De Vivo, M., 15  
 De Vree, B. T., 259–260  
 de Vries, A. H., 91–92, 146, 197  
 Dewar, M. J. S., 93  
 Dhiman, H. K., 262–263  
 Difley, S., 125–126  
 Dill, K. A., 28–29, 33–34, 48, 51–52, 59–60, 66, 73  
 Di Marzo, V., 2–3  
 Ding, F., 194–195, 196, 210  
 Dinola, A., 151–152  
 Dix, D., 218  
 Dixon, R. W., 28–29, 66  
 Dobbins, S. E., 202  
 Dobler, M., 231–233, 241–242  
 Dobson, C. M., 127–128  
 Dodge, J. A., 227  
 Doherty, J., 242–243  
 Dokholyan, N. V., 194–195, 196, 210  
 Dollinger, H., 231–232  
 Douglas, C. C., 294  
 Drake, M. T., 254–255  
 Dror, R. O., 261–262, 264–265, 269  
 Duan, Y., 289  
 Duke, R. E., 59–60  
 Dumas, R., 94  
 Dunbrack, R. L., 12–13, 53  
 Duranti, A., 2–3, 5, 6, 7–8, 11, 12, 15, 16–22  
 Durell, S. R., 191–192, 202  
 D'Ursi, P., 245  
 Dwyer, M. A., 115  
 Dyguda-Kazimierowicz, E., 21–22

**E**

- Easter, J. P., 227  
 Echols, L., 202  
 Edwards, P. C., 255–256, 260  
 Eilers, M., 269  
 Eisenbaher, S. J., 289  
 Ekena, K., 219  
 Ekins, S., 231–232  
 Elling, C. E., 268–269  
 Elstner, M., 86  
 Emperador, A., 186–188, 194–196, 208, 210  
 Engstrom, O., 222–224  
 Erenrich, E., 244–245  
 Eriksson, A. E., 59–60  
 Eriksson, L., 229–230  
 Erion, M. D., 65  
 Español, P., 147–148, 150–151, 154  
 Esposito, G., 289  
 Essex, J. W., 4–5, 56, 61–62, 65–67  
 Eudes, R., 5  
 Evanseck, J. D., 12–13, 53  
 Evans, R. M., 220–221  
 Eyal, E., 202

**F**

- Faiella, M., 88–89  
 Fang, H., 219–220  
 Farrens, D. L., 256, 258–259, 268–269  
 Favia, A. D., 3–4, 113–114  
 Fegley, D., 3  
 Feibelman, P., 125–126  
 Feig, M., 4–5, 297–298, 299–300  
 Feller, S. E., 155, 261–262, 264–265, 273–274  
 Felts, A. K., 28, 57–58, 62, 288, 289–290, 300–301, 309–310  
 Feng, Y. P., 166–167  
 Feng, Z., 184–185  
 Fenollosa, C., 186, 193–194  
 Ferguson, D. M., 53  
 Ferreira, K. N., 83–84  
 Ferrenberg, A. M., 57–58  
 Ferrer-Costa, C., 186, 193–194, 202  
 Ferrer, S., 116–117, 119–120  
 Ferrin, T. E., 233

- Feynman, R., 123  
 Fezza, F., 3  
 Field, M. J., 12–13, 53, 90, 289  
 Filizola, M., 262, 263–265  
 Fine, R., 291  
 Fioni, A., 21–22  
 Fiser, A., 4–5  
 Fisher, D., 309  
 Fitzgerald, L. R., 3  
 Flannery, B. P., 295–296  
 Floriano, W. B., 263–264  
 Fogolary, F., 289  
 Foguel, D., 128  
 Folkers, G., 233  
 Foloppe, N., 67–68, 70  
 Fossa, P., 245  
 Frank, I., 91–92  
 Freddolino, P. L., 198, 263–264  
 Freindorf, M., 90  
 Frey, P. A., 83–84  
 Friesner, R. A., 16, 28, 66–67, 233  
 Frimurer, T. M., 268–269  
 Froemmel, C., 299  
 Fromme, B. J., 260  
 Fuhrmans, M., 197, 198  
 Fujitani, H., 60–61  
 Fung, J. J., 255–256, 264–265  
 Furr, J. R., 4–5

**G**

- Gadea, F. X., 193–194, 202  
 Gaetani, S., 3  
 Gaidukov, L., 113  
 Gaille, C., 115–116  
 Galandrin, S., 254–255, 259–260  
 Galés, C., 259–260, 264  
 Gallicchio, E., 37–38, 38, 40, 41, 42, 45,  
 51–52, 53, 57–58, 62, 63–64, 66–67,  
 69–70, 71, 72, 288, 289–290, 300–301,  
 308–310, 313–314  
 Gals, C., 259–260  
 Gang, F., 294  
 Gao, C., 72  
 Gao, J., 83–84, 85, 86, 90, 107–108  
 Gao, L., 171

- Gao, Y. Q., 87  
 Garcia, A. E., 191, 287–288, 291  
 Garcia-Viloca, M., 83–84, 85, 86, 107–108  
 Gardiner, C. W., 207–208  
 Gatti, G., 5, 15  
 Gattinoni, S., 3  
 Gawrisch, K., 264–265, 273–274  
 Geerlings, P., 124–125  
 Gelin, B. R., 185–186  
 Gelpi, J. L., 186, 193–194  
 Gerlt, J. A., 88, 113  
 Gerstein, M., 202  
 Gervasio, F. L., 4–5, 66–67  
 Ge, X., 59–60  
 Ghanouni, P., 256, 258–259  
 Ghosh, A., 16  
 Gibson, K. D., 293–294  
 Giesy, J. P., 237  
 Gillespie, J. S., 231–232  
 Gilliland, G., 184–185  
 Gilson, M. K., 28, 29, 30–32, 35–36, 40–41, 43, 44–46, 59, 60, 67–69, 70, 72, 294  
 Giupponi, G., 186–188  
 Given, J. A., 29, 30–32, 35–36, 40–41, 59, 60, 67–68  
 Glasner, M. E., 113  
 Glen, R. C., 233  
 Gobbi, G., 2–3  
 Gobind Khorana, H., 262–263  
 Goel, N. S., 294  
 Goetz, R., 164  
 Goldsmith, M. R., 243  
 Golubkov, P. A., 59–60  
 Golynskiy, M. V., 89–90  
 Go, N., 190–191  
 González-Lafont, A., 95–97  
 Goodford, P. J., 229  
 Goodsell, D., 243  
 Gopala Krishna, A., 256  
 Gotor-Fernandez, V., 113–114  
 Gouda, H., 69–70  
 Gould, I. R., 53  
 Grafmüller, A., 171, 172  
 Grant, J. A., 243  
 Graves, A. P., 51–52, 59–60, 66, 73  
 Greengard, L., 294  
 Gregory, B. W., 227  
 Grese, T. A., 227  
 Grest, G. S., 152  
 Grishammer, R., 260  
 Gronenborn, A., 202  
 Groot, R. D., 43, 151, 153, 154, 157, 160–161, 162, 163, 178  
 Grossfield, A., 261–263, 264–265, 273–274  
 Grubmuller, H., 146  
 Gryczynski, Z., 256, 258–259  
 Guarnieri, F., 260, 271–273  
 Guidoni, L., 91–92  
 Guigas, G., 162, 173–174, 175–177  
 Guijarro, A., 22  
 Gumbart, J., 186–188, 198  
 Guo, H., 98–101  
 Gupta, S., 270–271  
 Gustin, D. J., 98–101  
 Gutiérrez-de-Terán, H., 98–101  
 Guvench, O., 28
- ## H
- Haak, J. R., 151–152, 287–288  
 Halgren, T. A., 233  
 Halliday, R., 243  
 Hall, R. J., 98–101  
 Hall, S. E., 263–264, 265–267  
 Hamelberg, D., 97–98  
 Hammes-Schiffer, S., 83–84, 107–108  
 Hamm, H. E., 254–255  
 Hang, J. F., 294–296  
 Hannongbua, S., 4–5  
 Hansch, C., 3–4  
 Hansmann, U. E., 293–294  
 Hanson, M. A., 11, 106–107, 257–258  
 Hansson, T., 291  
 Haranczyk, M., 90, 92–93  
 Hardy, L. W., 4–5  
 Hart, W., 243  
 Harvey, J. N., 124–125  
 Harvey, M. J., 186–188  
 Haslam, E., 98–101  
 Hawkins, G. D., 299, 300–301  
 Hawley, R. C., 298–299  
 Hayashi, R., 128  
 Haynes, M., 107–108

Head-Gordon, T., 199, 201  
 Head, M. S., 67–68  
 Hébert, T. E., 259–260, 264  
 Heerklotz, H., 162  
 Hehre, W. J., 86  
 Hein, P., 259–260  
 Helfrich, W., 146  
 Hellinga, H. W., 88–89  
 Hemley, R. J., 128  
 Henchman, R. H., 54  
 Henderson, R., 260  
 Hendricson, T., 298–299  
 Heremans, K., 128  
 Herman, M., 219  
 Hermann, J. C., 2, 11  
 Hermann, R. B., 287–288  
 Hermans, J., 283, 286, 287–288, 289, 292–294  
     303, 310  
 Herrmann, W. A., 83–84  
 Herschlag, D., 113–115, 122  
 Hess, B., 186–188, 198  
 Hilbers, C. W., 202  
 Hill, T. L., 283  
 Hilvert, D., 98–101, 107–108, 109  
 Himo, F., 91–92  
 Hindle, S. A., 98–101  
 Hinsen, K., 192–193, 202  
 Hodoscek, M., 66–67  
 Hoffmann, C., 259–260  
 Höfinger, S., 293–294  
 Hogue, M., 259–260, 264  
 Hollfelder, F., 113–114  
 Holl, R., 255–256, 257–258  
 Holst, B., 268–269  
 Holst, M., 294  
 Holt, D. A., 59–60  
 Holthausen, M. C., 86  
 Höltje, H. D., 233  
 Hong, B., 170–171  
 Hong, H., 219–220  
 Honig, B., 284, 289, 291–293, 294–295, 304  
 Hoogerbrugge, P., 147–148  
 Hoover, W. G., 151–152, 312  
 Hopfinger, A. J., 231–232  
 Hornak, V., 269  
 Horstink, L., 202  
 Hospital, A., 186–188, 193–194, 208

Hotta, K., 112–113  
 Hou, G., 313–314  
 Houtz, D. A., 254–255  
 Hritz, J., 66–67  
 Hünneberg, P. H., 282–283  
 Hubbard, R. E., 70, 222–224, 223  
 Hubbard, S. J., 289  
 Hubbell, C. M., 256, 268–269  
 Hubbell, W. L., 256, 268–269  
 Huey, R., 243  
 Hughes, B. D., 175–176  
 Hu, H., 90, 92–93, 98  
 Hult, K., 113–114  
 Hummer, G., 287–288, 291  
 Hung, L., 120–121  
 Huo, S., 69–70, 282–283, 286  
 Hur, S., 98–101  
 Hurst, D. P., 273–274  
 Hyodo, S., 165, 166, 169–170

## I

Ijerman, A. P., 255  
 Illya, G., 164–165  
 Imada, K., 198  
 Im, W., 297–298, 299–300  
 Inoue, Y., 67–68  
 Isin, B., 261–263  
 Islam, M. A., 240–241  
 Ito, M., 60–61  
 Iverson, T. M., 83–84

## J

Jackson, J. D., 291  
 Jackson, R. M., 287–288  
 Jagielska, A., 310  
 Jain, A. N., 234  
 Jakobsen, A. F., 155–156, 157, 179  
 Jang, Y., 193–194, 202  
 Jardón-Valadez, E., 273–274  
 Jarzynski, C., 54–55, 56  
 Jastrzebska, B., 270–271  
 Jaun, B., 103  
 Javitch, J. A., 271–273



- Jayachandran, G., 50–52, 59–61  
 Jayaram, B., 291  
 Jee, J. G., 202  
 Jensen, M. O., 261–262, 264–265, 269  
 Jensen, R. A., 113  
 Jeong, J. I., 193–194, 202  
 Jernigan, R. L., 191–193, 202  
 Jiang, L., 88–90, 113  
 Jiang, W., 66–67, 73  
 Jiao, D., 59–60  
 Jin, B., 231–232  
 Johnson, A. P., 243  
 Johnson, D. S., 3  
 Johnston, C. A., 254–255  
 Jonas, S., 113–114  
 Jones, G., 233  
 Jonsson, A. L., 186  
 Jorgensen, W. L., 5, 16, 28, 34–35, 53, 60–61, 64, 65, 86–87, 90, 92–93, 98–101, 203, 288, 289–290  
 Juffer, A. H., 289, 292–293, 294
- K**
- Kalinowski, S., 198–199  
 Kamerlin, S. C., 90, 92–93  
 Kamtekar, S., 15  
 Kangas, E., 98–101  
 Kantorovich, L., 91–92  
 Kaplan, J., 88  
 Kar, P., 293–294  
 Karplus, M., 30–32, 53, 60–61, 70, 87–88, 90, 185–186, 194–195, 198, 202, 210, 282–283, 309–310  
 Karplus, P. A., 107–108, 109  
 Karttunen, M., 150–151, 153–154  
 Kastner, J., 91–92, 98  
 Kast, P., 98–101, 102, 110, 115–116, 119–120  
 Kathuria, S., 3  
 Kato, M., 128  
 Katzenellenbogen, B. S., 219  
 Katzenellenbogen, J. A., 223, 224, 226, 228  
 Kaul, M., 43  
 Kavlock, R., 218  
 Kazlauskas, R. J., 113–115  
 Keefe, J., 115, 118–119, 120  
 Ke, H., 107–108, 109  
 Kellogg, G., 229–230  
 Kenakin, T., 254–255  
 Keskin, O., 191–192, 202  
 Khalili, M., 198–199  
 Khandogin, J., 53, 69–70, 312  
 Khanjin, N. A., 109–110  
 Khavrutskii, I. V., 66–67  
 Khelashvili, G., 261–262  
 Khersonsky, O., 88–89, 113–114  
 Khorana, H. G., 256, 268–269  
 Kienhöfer, A., 110  
 Kim, B., 66–67  
 Kim, H., 67–68  
 Kimmel, A. V., 91–92  
 Kim, M. K., 193–194, 202  
 King, A. R., 21–22  
 King, M. A., 66–67  
 Kiss, G., 243  
 Kitanovic, S., 260  
 Kitchen, D. B., 4–5, 291  
 Kladi, M., 240  
 Klebe, G., 4–5, 229–230, 233  
 Klein-Seetharaman, J., 262–263  
 Klicic, J. J., 233  
 Kneller, G., 192–193, 202  
 Knight, J. L., 65  
 Knowles, J. R., 98–101, 113  
 Knox, A. J. S., 242  
 Kobilka, B. K., 254–255, 270  
 Kobilka, T. S., 257–259, 270  
 Koch, W., 86  
 Koder, R. L., 86, 89–90  
 Koelman, J., 147–148  
 Kofke, D. A., 39, 55, 56–57  
 Kohout, T., 254–255  
 Kollman, P. A., 57–58, 60–61, 69–70, 93, 282–283, 286, 289, 291  
 Kolossvary, I., 72  
 Kondrashov, D. A., 193–194, 202  
 Kong, X., 312  
 Kong, Y., 202  
 Konialian, A. L., 59–60  
 Kontoyianni, M., 87–88  
 Konvicka, K., 260  
 Koopman, E. A., 152  
 Kovacs, J. A., 192–193, 202

- Ko, Y. H., 67–68  
 Ko, Z., 294  
 Kozack, R. E., 294  
 Kramer, B., 233  
 Kranenburg, M., 167–168  
 Krebs, W. G., 202  
 Kremer, K., 152  
 Kroemer, R. T., 4–5  
 Kubinyi, H., 228, 229–230  
 Kubo, M. M., 300–301  
 Kubo, R., 205–206  
 Kuhn, B., 69–70, 93, 282–283, 286  
 Kuhn, N., 294  
 Kumar, S., 57–58, 97–98  
 Kuntz, I. D., 28, 69–70, 233  
 Künzler, D. E., 115–116  
 Kushick, J., 70
- L**
- Lagorce, D., 234, 242  
 Laio, A., 66–67, 91–92, 262  
 Lai, W., 125–126  
 Lakowicz, J. R., 256, 258–259  
 Landau, L. D., 291  
 Langer, T., 243–244  
 Langlet, J., 289  
 Langridge, R., 233  
 Lantsch, G., 162  
 Lapelosa, M., 37–38, 38, 40, 41, 42, 51–52, 58,  
   62, 63, 66–67, 70, 71  
 Laradji, M., 165, 169–170  
 Lassila, J. K., 115, 118–119, 120  
 Layton, A. C., 227  
 Lazaridis, T., 53, 309–310  
 Leach, A. R., 4–5, 233  
 Leduc, M., 259–260  
 Lee, A. Y., 107–108, 109  
 Lee, M., 297–298, 299–300  
 Lee, M. R., 289  
 Lee, M. S., 34–35, 64, 66–67, 70, 299–300, 312  
 Lee, T. W., 256, 258–259, 270  
 Lee, Y. S., 103, 110  
 Lefkowitz, R. J., 254–255  
 Le Gouill, C., 259–260  
 Leitao, A., 238  
 Leitgeb, M., 30–32, 60–61  
 Lemons, D. S., 205–206  
 Lengauer, T., 233  
 Lenhoff, A. M., 294  
 Leo, A., 3–4  
 Lerner, R. A., 105  
 Lesk, V. I., 202  
 Leslie, A. G., 255–256, 260  
 Levitt, M., 90, 115, 294  
 Levy, R. M., 28, 37–38, 40, 41, 42, 45, 51–52,  
   53, 57–58, 62, 63–64, 66–67, 69–70, 71,  
   72, 288, 289–290, 291, 300–301,  
   308–310, 313–314  
 Lezon, T. R., 262–263  
 Liapakis, G., 271–273  
 Lifshitz, E. M., 291  
 Li, G., 59–60  
 Li, H. L., 243–244, 263–264, 265–267  
 Liimatta, M., 3  
 Lill, M. A., 232–233, 234–237, 241–242  
 Lindahl, E., 146–147, 186–188, 198  
 Lin, H., 91–93  
 Lin, L. C. L., 145–146  
 Linssen, A. B. M., 97–98  
 Lipowsky, R., 148–149, 157, 163, 164–165,  
   166–167, 171, 172  
 Lipscomb, W. N., 202  
 Little, S. B., 243  
 Liu, B. K., 113–114  
 Liu, C. W., 270–271  
 Liu, H., 237  
 Liu, P., 66–67  
 Liu, X. Y., 166–167, 237  
 Liwo, A., 198–199, 284, 289  
 Li, Z. X., 162  
 Llarrull, L. I., 91–92  
 Lloyd, D. G., 242  
 Lodola, A., 2, 3–4, 5, 8, 11, 12–13, 16–22, 86  
 Lohse, M. J., 259–260  
 Lonsdale, R., 12–13, 85, 86, 124–125  
 Looger, L. L., 115  
 Lopes, D., 127–128  
 López-Blanco, J. R., 193–194, 202  
 Lopiparo, E., 239  
 Lorenzi, S., 3–4  
 Louhivouri, M., 146–147  
 Lowe, C. P., 152

Lu, B., 294–296  
 Ludwig, H., 128  
 Luengo, J. I., 59–60  
 Lu, J. R., 162  
 Lu, N., 39, 55, 56–57  
 Lund, D., 244  
 Luo, H., 34–35, 40–41  
 Luo, X. M., 243–244  
 Luty, B. A., 294  
 Lutz, S., 87–88  
 Lu, Y., 86  
 Lynch, D. L., 273–274  
 Lyne, P. D., 98–101  
 Lyons, J. A., 255–256, 257–258

### M

Ma, B., 87  
 Machuqueiro, M., 303, 311–312  
 MacKerell, A. D., 12–13, 28, 53  
 Madabushi, S., 254–255  
 Madden, J. C., 231  
 Mader, M. M., 105, 106–107  
 Madhav, P. J., 231–232  
 Madura, J. D., 288, 289–290, 294  
 Magistrato, A., 91–92  
 Maglio, O., 86  
 Magnani, F., 260  
 Mainz, D. T., 233  
 Ma, J., 202  
 Makowski, M., 289  
 Malaisree, M., 4–5  
 Mandal, A., 98–101  
 Marchand-Geneste, N., 228–229, 234–237  
 Marcus, Y., 287–288  
 Marelius, J., 12  
 Mark, A. E., 146, 161  
 Marques, O., 193–194, 202  
 Marrink, S. J., 146–147, 161, 197, 198  
 Marshall, G. R., 3–5, 263–264  
 Martel, P. J., 302–303  
 Martínez, J. I., 125–126  
 Martín, M. E., 90, 98–101  
 Martín, R. M., 120–121  
 Martí, S., 90, 95–97, 98–101, 102, 103,  
 104–105, 107, 108–109, 110, 112, 115,  
 116–117, 119–120  
 Maruyama, Y., 166  
 Massova, I., 69–70, 282–283, 286  
 Masuda, K. R., 11  
 Mata, R. A., 90  
 Mattei, P., 98–101  
 Matthews, B. W., 59–60  
 Matubayasi, N., 64  
 Maxwell, D. S., 53, 288, 289  
 Mayo, S. L., 86, 88–89  
 McCammon, J. A., 29, 30–32, 35–36, 40–41,  
 54, 59, 60, 65, 87–88, 185–186, 282–283,  
 294–296, 299–300, 303, 311–312  
 McClellan, L. M., 87–88  
 McDowell, S. C., 282–283  
 McGann, M. R., 243  
 McInerney, E., 219  
 McInnes, C., 28  
 McKenna, N. J., 218–219  
 McKenney, A., 294  
 McKerell, A., Jr., 198  
 McKinney, M. K., 2  
 McLachlan, J. A., 227  
 McLay, I., 238  
 McMartin, C., 245  
 McMasters, D., 231–232  
 McReynolds, A. C., 59–60, 73  
 Meegan, M. J., 242  
 Meersman, F., 127–128  
 Mekenyan, O., 231  
 Mendez, R., 202  
 Menezes, I. R. A., 238  
 Meng, Y., 66–67  
 Menon, S. T., 256  
 Merz, K. M., 53, 90  
 Meyer, T., 186, 193–194, 195–196, 210  
 Michel, J., 4–5, 56, 61–62, 65–66  
 Mihailescu, M., 40–41, 43  
 Milanese, L., 245  
 Mileni, M., 15  
 Millspaugh, L. E., 71–72  
 Minkkilä, A., 3  
 Mishra, R., 127–128

- Misura, K. M., 309  
 Miteva, M. A., 5, 234, 242  
 Mitsutake, A., 66–67  
 Miyamoto, S., 60–61  
 Mobley, D. L., 28–29, 33–34, 48, 51–52,  
 59–60, 61–62, 66, 73  
 Mocklinghoff, S., 225  
 Moghaddam, S., 67–68  
 Molinari, H., 289, 309–310  
 Moliner, V., 95–97, 98–101, 119–120  
 Monard, G., 90, 95–97  
 Mongan, J., 97–98  
 Montanari, C. A., 238  
 Montano, M., 219  
 Moore, M., 227  
 Moras, D., 222  
 Moreno-Sanz, G., 22  
 Morgan, J., 299–300, 301–302, 303,  
 311–312  
 Mori, T., 67–68  
 Moritsugu, K., 193–194, 201, 202  
 Mori, Y., 66–67  
 Morley, K. L., 114–115  
 Mor, M., 2, 3–4, 5, 6, 7–8, 11, 12–13, 15,  
 16–21, 86  
 Morokuma, K., 90, 92–93, 95–97  
 Morozova, D., 176, 178  
 Morris, G., 243  
 Morton, A., 59–60  
 Moukhametzianov, R., 255–256, 260  
 Mouritsen, O. G., 157, 179  
 Mowbray, D. J., 125–126  
 Muccioli, G. G., 2  
 Muchmore, S. W., 70  
 Mukherjee, A., 240–241  
 Mulholland, A. J., 2, 4–5, 11, 12–13, 85, 86,  
 98–101, 110  
 Mundorff, E. C., 106–107  
 Murata, K., 66–67  
 Murphy, P. M., 88–89  
 Murphy, R. B., 233  
 Mustain, M., 227  
 Muthyala, R., 223, 226, 228  
 Mysovsky, A. S., 91–92
- N**
- Naganathan, A. N., 183–216  
 Nagarajan, K., 244–245  
 Nagar, S., 240–241  
 Nakai, M., 241  
 Namba, K., 198  
 Nanda, V., 86, 89–90  
 Nannei, R., 3  
 Nardelli, M. B., 126  
 Nared, K. D., 107–108, 109  
 Nastro, F., 86  
 Nayeem, A., 293–294  
 Neale, C., 66–67  
 Nedelcheva, D., 241  
 Neese, F., 91–92  
 Neet, K. E., 83–84  
 Nehmé, R., 255–256, 260  
 Nelson, C. D., 254–255  
 Nemethy, G., 289  
 Nevalainen, T., 3  
 Nguyen, K. A., 94  
 Nicholls, A., 243, 292–293, 294–295  
 Nichols, D. E., 254–255  
 Niesen, M., 267–268, 272  
 Nikolaev, V. O., 259–260  
 Nikolovska-Coleska, Z., 72  
 Nikunen, P., 150–151, 153–154  
 Nilges, M., 202  
 Nina, M., 291–292  
 Nobeli, L., 113–114  
 Norskov, J. K., 125–126  
 Nose, S. J., 151–152, 312  
 Nose, T., 244  
 Novotny, J., 287–288  
 Nussinov, R., 87  
 Nygaard, R., 270–271  
 Nymeyer, H., 190–191
- O**
- Oatley, S. J., 233  
 O'Brien, P. J., 113–115

- Oh, H. J., 59–60  
 Okamoto, Y., 66–67  
 Okumura, H., 72  
 Okur, A., 66–67  
 Oldziej, S., 198–199, 289  
 Oligny-Longpr, G., 254–255, 259–260  
 Oliveira, C. A. F., 303, 311–312  
 Ollila, O. H. S., 146–147  
 Olson, A. J., 293–294  
 Olson, M. A., 34–35, 64, 66–67, 70  
 O'Malley, B. W., 218–219  
 Onuchic, J. N., 190–191  
 Onufriev, A., 282–283, 296–298, 299–300, 308–309, 310–311  
 Onufriev, A. V., 301–302, 313–314  
 Oobatake, M., 289  
 Ooi, T., 289  
 Oostenbrink, C., 65, 66–67  
 Orban, T., 270–271  
 Orellana, L., 186–188, 193–194, 202, 208  
 Orozco, M., 186, 193–196, 202, 204–205, 208, 210, 289, 299  
 Ortega-Carrasco, E., 91–92  
 Ostlund, N. S., 86  
 Ottmann, C., 225
- P**
- Pailthorpe, B. A., 175–176  
 Pajeva, I., 234, 242  
 Palczewski, K., 270–271  
 Palmer, K. A., 293–294  
 Pande, V. S., 28–29, 50–52, 59–60, 66, 69–70, 186–188  
 Papazafiri, P., 240  
 Pappu, R. V., 263–264  
 Pardon, E., 255–256, 264–265  
 Paris, K., 53, 63–64, 288, 289–290, 300–301, 308–309, 313–314  
 Park, M.-S., 72  
 Park, S., 50–52, 59–60  
 Parnot, C., 258–259, 270  
 Parrinello, M., 4–5, 66–67, 91–92, 262  
 Parr, R. G., 86  
 Pasquinelli, M. A., 243  
 Pastor, M., 238, 273–274  
 Pastor, R. W., 155  
 Pattabiraman, N., 242–243  
 Paulaitis, M. E., 38–39  
 Pauling, L., 84, 88, 105  
 Paul, S., 126  
 Pavlov, T., 241  
 Payne, P., 294  
 Payne, V. A., 64  
 Peishoff, C. E., 4–5  
 Pellegrini, E., 289  
 Pencheva, T., 234, 242  
 Penfold, J., 162  
 Perahia, D., 70  
 Pérez, A., 186, 193–194  
 Perez, J. J., 263–264  
 Periole, X., 197, 198  
 Peristera, O., 232–233  
 Perkins, R., 219–220  
 Perola, E., 72  
 Perozzo, R., 4–5  
 Perrot, G. B., 293–294  
 Persico, M., 93, 289  
 Petersen, F., 126–127  
 Petersen, S. B., 302–303  
 Petrescu, A., 192–193, 202  
 Petrosino, S., 2–3  
 Pettitt, B. M., 282–283  
 Philips, G. N., 193–194, 202  
 Phillips, G. N., Jr., 202  
 Phillips, J. C., 186–188, 198  
 Piana, S., 66–67, 91–92, 261–262, 264–265, 269  
 Pickett, S., 238  
 Pierotti, R. A., 287  
 Piersanti, G., 5, 6, 7–8, 15  
 Pike, A. C., 219, 221–224  
 Pilch, D. S., 43  
 Piomelli, D., 2–3, 11, 12–13  
 Piparo, E., 234–237  
 Pitera, J. W., 69–70  
 Pitman, M. C., 261–262, 264–265, 273–274  
 Pjjeva, I., 217–252  
 Plake, H. R., 71–72  
 Plazzi, P. V., 3–4, 5, 8, 11  
 Pohorille, A., 28–29, 38–39, 54–56, 65  
 Polikarpov, I., 237–238  
 Pomès, R., 66–67

- Ponder, J. W., 263–264  
 Porcher, J. M., 228–229, 234–237  
 Porezag, D., 86  
 Postma, J. P. M., 151–152, 287–288  
 Potter, M. J., 68  
 Prat-Resina, X., 95–97  
 Pratt, L. R., 38–39, 287–288, 291  
 Press, W. H., 295–296  
 Proust-De Martín, F., 94  
 Provasi, D., 262, 264–265  
 Pullman, A., 289
- Q**
- Qiu, F., 170–171
- R**
- Rabinowitz, J. R., 243  
 Rabone, K. L., 162, 163, 178  
 Rader, A. J., 262–263  
 Radmer, R. J., 282  
 Radom, L., 86  
 Rajnarayanan, R. V., 242–243  
 Ramamoorthy, K., 227  
 Ramos, M. J., 90  
 Ranaghan, K. E., 12–13, 85, 86  
 Rarey, M., 233  
 Rashin, A. A., 292–293, 294  
 Rasmussen, G. F., 257–258  
 Rasmussen, S. G. F., 255–256, 259–260,  
 264–265, 270–271  
 Raso, V., 105  
 Ratnala, V. R. P., 258–259  
 Rebols, R. V., 259–260, 264  
 Recanatini, M., 4–5  
 Reddy, M. R., 65  
 Reed Murphy, L., 64  
 Reeves, P. J., 269  
 Reichert, D., 238–239  
 Reid, D., 243  
 Reiss, H., 287  
 Rekharsky, M. V., 67–68  
 Renaud, J. P., 222  
 Renka, R., 94  
 Ren, P., 59–60  
 Repchevsky, D., 186, 193–194  
 Reyes, C., 69–70, 282–283, 286  
 Riccardi, D., 90  
 Richards, F. M., 88–89, 287–288  
 Rick, S. W., 291  
 Rinaldi, D., 93, 289, 299  
 Ring, B. J., 231–232  
 Ripoll, D. R., 284, 310  
 Risselada, H. J., 146–147, 197, 198  
 Rivail, J. L., 93  
 Rivara, S., 3–4, 5, 6, 7–8, 11, 12–13, 15, 16–22  
 Robles, V. M., 91–92  
 Roca, M., 110  
 Rocchia, W., 294  
 Rocklin, G. J., 51–52, 59–60, 66, 73  
 Rodgers, J. M., 176, 177, 178  
 Rodinger, T., 66–67  
 Rognan, D., 233  
 Rohrig, U. F., 91–92  
 Roitberg, A. E., 66–67  
 Romo, T. D., 262–263, 273–274  
 Roncaglioni, A., 229, 234–237, 239  
 Roodveldt, C., 113–114  
 Roos, G., 124–125  
 Rosenbaum, D. M., 255–256, 257–258,  
 270–271  
 Rosenberg, J. M., 57–58, 97–98  
 Rosenblatt, M. M., 86  
 Rosenkilde, M. M., 268–269  
 Rose, R., 225  
 Rossi, I., 94  
 Rosky, P., 85–86  
 Rothenberg, G., 129  
 Rothlisberger, D., 85, 86, 88–89  
 Rothlisberger, U., 91–92  
 Rousseau, G., 254–255  
 Roussis, V., 240  
 Roux, B., 4–5, 28, 33–36, 57–58, 59–61, 63,  
 64, 66–67, 73, 97–98, 267, 282–283,  
 287–288, 291–292  
 Rovida, E., 245  
 Roxby, R., 303  
 Rueda, M., 186–188, 193–195, 202, 204–205,  
 208, 210  
 Ruiz, G., 259–260  
 Ruiz-Lopez, M. F., 289, 299

Ruiz-Pernia, J. J., 94–97  
 Russo, R., 22  
 Ruud, K., 127–128

## S

- Saario, S., 3  
 Sadjad, B. S., 243  
 Safe, S., 227  
 Saffmann, P. G., 175  
 Safo, M. K., 4–5  
 Saha, A., 240–241  
 Saied, F., 294  
 Sakmar, T. P., 256, 269  
 Sali, A., 4–5  
 Salsbury, F.R. Jr., 299–300, 312  
 Salum, L. de B., 3, 5, 237–238  
 Salvi, E., 245  
 Sanchez, M. L., 90, 98–101  
 Sanejouand, Y. H., 193–194, 202  
 Sanner, M. F., 293–294  
 Sanseverino, J., 227  
 Sanz, F., 273–274  
 Sasso, S., 115–116  
 Sayler, G. S., 227  
 Scapozza, L., 4–5  
 Schaefer, M., 299  
 Schaefer, P., 90  
 Schaeffer, R. D., 186  
 Schellman, J. A., 304  
 Scheraga, H. A., 198–199, 284, 286–288, 289,  
 291–296, 303, 304, 306, 310  
 Schertler, G. F. X., 256, 269  
 Schettino, V., 128  
 Schiott, B., 244  
 Schmidt, U., 162, 173–174, 175, 176–177, 178  
 Schneider, G., 233  
 Schnieders, M. J., 59–60  
 Schramm, V. L., 105  
 Schuler, L. D., 53  
 Schulten, K., 198, 261–262  
 Schultz, P. G., 105  
 Schultz, T. W., 227  
 Schwartz, T. W., 268–269  
 Scouras, A. D., 186  
 Sebastiani, D., 91–92  
 Seelig, B., 89–90  
 Seierstad, M., 3  
 Seifert, R., 255, 256  
 Selassie, C. D., 3–4  
 Selent, J., 273–274  
 Selvapalam, N., 67–68  
 Semus, S., 229–230  
 Sendrovic Rapp, C., 16  
 Senn, H. M., 12, 90, 91–93, 98  
 Sen, T. Z., 192–193, 202  
 Serafimova, R., 231, 241  
 Serdyuk, I. N., 45  
 Serrano-Vega, M. J., 260  
 Shadrach, R., 301–302, 313–314  
 Shanle, E. K., 220–222, 228  
 Sharma, P. K., 85, 86  
 Sharp, K. A., 34–35, 40–41, 284, 287–288,  
 289, 291–293, 294, 304  
 Shaw, D. E., 261–262, 264–265, 269  
 Sheehan, D. M., 219–220  
 Shenoy, S. K., 254–255  
 Sherwood, P., 91–92  
 Sheves, M., 269  
 Shibata, Y., 260  
 Shih, A. Y., 198  
 Shi, L., 271–273  
 Shillcock, J. C., 147, 148–149, 157, 163,  
 164–165, 166–167, 171, 172  
 Shi, L. M., 219  
 Shimohigashi, Y., 244  
 Shirts, M. R., 28–29, 50–52, 58–61, 66  
 Shluger, A. L., 91–92  
 Shoichet, B. K., 4–5, 28, 51–52, 59–60, 66, 73  
 Shrivastava, I. H., 262–263  
 Shurki, A., 102–103, 104–105  
 Siderovski, D. P., 254–255  
 Siegbahn, P. E. M., 83–84, 91–92  
 Siegel, J. B., 88–89, 115–116  
 Silla, E., 94–97  
 Silva, C. A., 5, 21–22  
 Silva, J. L., 128  
 Simmerling, C., 66–67, 299–300  
 Simms, A. M., 186  
 Simon, A., 243  
 Simonson, T., 35–36, 282–283, 287–288, 289,  
 291–292, 294  
 Singh, J. K., 39, 56–57

- Špackova, N., 282–283  
 Šponer, J., 282–283  
 Sippl, W., 233  
 Sirirak, J., 11, 21–22  
 Sitkoff, D., 289, 291–292, 294  
 Sit, S. Y., 3  
 Sivanesan, D., 242–243  
 Sjöström, M., 229–230  
 Skolnick, J., 309  
 Smeller, L., 128  
 Smiesko, M., 232–233  
 Smit, B., 157, 163, 164, 167–168, 173–174,  
 176, 177, 178  
 Smith, G. D., 98–101, 102, 119–120  
 Smith, J. C., 193–194, 201, 202  
 Smith, S. O., 256, 269, 271–273  
 Smit, M., 155, 176, 177  
 Snow, C. D., 60–61  
 Snyder, J. P., 109–110  
 Sobolevski, E., 289  
 Sokol, A. A., 91–92  
 Solorzano, C., 5  
 Sonnenschein, C., 227  
 Sorensen, M., 199, 201  
 Soto, A. M., 227  
 Soubias, O., 264–265, 273–274  
 Sousa, S. F., 90  
 Spohner, J. C., 293–294  
 Sperotto, M. M., 173–174  
 Spreafico, M., 232–233  
 Sridharan, S., 292–293, 294–295  
 Srinivasan, J., 286  
 Staples, E. J., 162  
 Steenhuis, J., 270  
 Steenhuis, J. J., 256, 258–259  
 Stefanovich, E. V., 91–92  
 Steinbrecher, T., 61–62  
 Steindal, A. H., 127–128  
 Sternberg, J. E., 287–288  
 Sternberg, M. J. E., 202  
 Stern, H. A., 72  
 Stevenson, T., 3  
 Stevens, R. C., 11, 15  
 Still, W. C., 298–299  
 Stollar, B. D., 105  
 Strajbl, M., 102–103, 104–105  
 Stura, E., 107–108  
 Subramaniam, S., 294  
 Sugita, Y., 66–67  
 Suits, F., 261–262  
 Sulimov, V. B., 91–92  
 Summa, C. M., 86  
 Sumpter, J. P., 227  
 Sun, H., 72  
 Sunil Kumar, P. B., 165, 169–170  
 Sushko, M. L., 125–126  
 Sushko, P. V., 91–92, 125–126  
 Swaminath, G., 258–259, 270  
 Swanson, J. M. J., 54  
 Swendsen, R. H., 57–58, 98  
 Swope, W. C., 69–70  
 Sykes, D. G., 91–92  
 Szabo, A., 86  
 Szefczyk, B., 110
- T**
- Tafi, A., 240  
 Taha, M., 231, 239–240  
 Tajkhorshid, E., 186–188, 198, 261–262  
 Taketomi, H., 190–191  
 Tama, F., 193–194, 202  
 Tanford, C., 303, 304  
 Tanida, Y., 60–61  
 Tan, Z., 58–59  
 Tarairah, M., 231, 239–240  
 Tarzia, G., 2–3, 5, 6, 7–8, 11, 12–13, 15  
 Tate, C. G., 260  
 Tawfik, D. S., 113–114  
 Taylor, R., 233  
 Taylor, S. V., 110  
 Tembe, B. L., 65  
 Tempezyk, A., 298–299  
 Teresk, M. G., 71–72  
 Terry, T. J., 256  
 Tettinger, F., 30–32, 60–61  
 Teukolsky, S. A., 295–296  
 Thian, F. S., 257–259, 270  
 Thiel, S., 91–92  
 Thiel, W., 12, 86, 90, 91–93, 98  
 Thomas, G., 86–87  
 Thomas, L. L., 65  
 Thomas, R. K., 162



- Thompson, M. A., 90  
 Thorsell, A. G., 222–224  
 Tidor, B., 98–101  
 Tieleman, D. P., 197  
 Tirado-Rives, J., 53, 60–61, 203  
 Tirado-Rives, J. J., 288, 289  
 Tirion, M. M., 191–192, 202  
 Tobias, D. J., 273–274  
 Todorov, M., 241  
 Tokarski, J., 231–232  
 Tokunaga, T., 244  
 Tokuriki, N., 113–114  
 Tomasi, J., 93, 289  
 Toney, M. D., 116  
 Tong, W., 219–220  
 Tontini, A., 2–3, 5, 6, 7–8, 11, 12, 15,  
 16–22  
 Toofanny, R. D., 186  
 Topol, I. A., 292–293, 294  
 Torrie, G. M., 97–98  
 Toscano, M. D., 113  
 Trovov, M., 87–88  
 Trabanino, R., 263–264  
 Trieu, P., 259–260, 264  
 Tropsha, A., 3–4  
 Truhlar, D. G., 90, 91–93, 120–121, 289, 299,  
 300–301  
 Truong, T., 91–92  
 Tsakovska, I., 217–252  
 Tucker, I., 162  
 Tuñón, I., 81–142  
 Turner, A. J., 95–97  
 Tye, J. W., 83–84
- U**
- Ueda, Y., 190–191  
 Ulmschneider, J. P., 203  
 Ulrich, H. D., 106  
 Urban, J. D., 254–255
- V**
- Vacondio, F., 5, 12, 16–22  
 Vagias, C., 240  
 Vaidehi, N., 255, 256, 260, 263–264, 265–268,  
 269, 272  
 Valiño, F., 3  
 Valleau, J. P., 97–98  
 Van der Kamp, M. W., 186  
 van der Ploeg, A., 289, 292–293, 294  
 van der Spoel, D., 186–188, 198  
 VandeVondele, J., 91–92  
 van Gunsteren, W. F., 53, 65, 151–152  
 Van Kampen, N. G., 207–208  
 van Keulen, B. A. M., 289, 292–293, 294  
 Varnek, A. A., 294  
 Varshney, A., 293–294  
 Vasquez, M., 294  
 Vattulainen, I., 146–147, 150–151, 153–154  
 Vedani, A., 231–233, 234–237, 241–242  
 Venturoli, M., 155, 157, 163, 164, 167–168,  
 173–174, 176, 177  
 Vetterling, W. T., 295–296  
 Vezzosi, S., 5  
 Viglino, P., 289  
 Vila, J. A., 284, 286, 291–296, 303, 304,  
 306, 310  
 Villardaga, J.-P., 259–260  
 Villa, E., 186–188, 198  
 Villà, J., 104–105, 107–108  
 Villoutreix, B. O., 5, 234, 242  
 Violin, J. D., 254–255  
 Visser, A., 225  
 Vivat, V., 222  
 Vlaar, M., 167, 168  
 Voelz, V. A., 186–188  
 Vogel, R., 256  
 von Zastrow, M., 254–255  
 Vorobjev, Y. N., 283, 284, 286, 287–288, 289,  
 291–296, 303, 304, 306, 310  
 Vreven, T., 90, 92–93, 95–97
- W**
- Wachucik, K., 198–199  
 Wade, R. C., 294  
 Wagoner, J., 53, 289  
 Wallner, B., 263–264  
 Wallqvist, A., 66–67, 309–310  
 Wallqvist, W., 287–288

- Walter, D., 289  
 Walter, N. G., 282–283  
 Wang, C. Y., 244–245  
 Wang, J., 59–60  
 Wang, L.-P., 125–126  
 Wang, M., 294  
 Wang, S., 72, 123, 231–232  
 Wang, W., 186–188, 198  
 Warne, T., 255–256, 260  
 Warren, P., 147–148, 150–151, 154,  
 157, 160–161  
 Warshel, A., 83–84, 85, 86, 90, 92–93,  
 102–103, 104–105, 107–108, 115  
 Weaver, L. H., 59–60  
 Wei, B. Q., 59–60  
 Weinstein, H., 254–255, 261–262  
 Weissig, H., 184–185  
 Weiss, M., 162, 173–174, 175–177, 178, 179  
 Wei, Y., 293–294  
 Welsh, W. J., 219–220  
 Wenzel-Seifert, K., 255, 256  
 Westbrook, J., 184–185  
 White, J. F., 260  
 White, L. R., 175–176  
 Whorton, M. R., 259–260  
 Widom, B., 38–39  
 Wikel, J. H., 231–232  
 Willems, T. F., 176, 177, 178  
 Willett, P., 233  
 Williams, I. H., 84  
 Williams, S. L., 303, 311–312  
 Wilson, C. A., 202  
 Windemuth, A., 292–293, 294–295  
 Winter, R., 127–128  
 Wiorkiewicz-Kuczera, J., 198  
 Wolber, G., 243–244  
 Wold, S., 229–230  
 Wolfenden, R., 105  
 Wolohan, P., 238–239  
 Wood, D. C., 15  
 Woods, C. J., 4–5, 66–67  
 Woo, H.-J., 34–35, 57–58, 64  
 Woo, T. K., 91–92  
 Worth, A., 234–237  
 Worthington, S. E., 103, 110  
 Woycechowsky, K. J., 113  
 Wozniak, J. A., 59–60  
 Wright, W. V., 293–294  
 Wu, Q., 113–114  
 Wurtz, J. M., 222  
 Wu, Y., 237
- X**
- Xiang, Y., 237, 270  
 Xiao, A., 237  
 Xiao, K., 254–255  
 Xia, X., 90  
 Xie, K., 3  
 Xie, Q., 219–220  
 Xu, R., 271–273  
 Xu, W., 220–222, 228
- Y**
- Yamamoto, S., 165, 166, 169–170  
 Yamashita, D. S., 59–60  
 Yang, A. S., 284, 289, 294, 304  
 Yang, C., 67–68  
 Yang, C.-Y., 72  
 Yang, K., 268–269  
 Yang, L., 202  
 Yang, S. A., 267, 304  
 Yang, W., 86, 90, 92–93, 98, 237  
 Yang, Y., 170–171, 242  
 Yao, X. J., 258–260  
 Yefimov, S., 197  
 Yeh, I.-C., 66–67  
 Yen, H. K., 59–60  
 Ye, S., 269  
 Yeung, N., 86  
 Yoon, B. J., 294  
 Young, L., 292–293, 294  
 Ytreberg, F. M., 98  
 Yu, H., 202, 237  
 Yu, K. Q., 243–244
- Z**
- Zaccai, G., 45  
 Zaccai, N. R., 45

- Zaera, F., 128–129  
Zaitseva, E., 256, 269  
Zalatan, J. G., 122  
Zalloum, H., 231, 239–240  
Zanghellini, A., 88–89, 113, 115–116  
Zauhar, R., 244–245  
Zauhar, R. J., 292–293, 294  
Zbinden, P., 232  
Zhang, H., 170–171  
Zhang, J., 59–60  
Zhang, L. Y., 288, 289–290, 300–301, 309–310  
Zhang, Q. J., 243–244  
Zhang, Y., 90, 91–92, 123, 155, 309  
Zhang, Z., 237, 255–256, 257–258  
Zhou, H.-X., 28, 44–45, 46, 68–69  
Zhou, Y., 123  
Zhou, Y. C., 294  
Zhou, Y. Q., 194–195, 210  
Zhou, Z., 28, 294  
Zhu, W., 258–259, 270  
Zhu, X., 313–314  
Ziebart, K. T., 116  
Zoebisch, E. G., 93  
Zou, Y., 270–271  
Zsoldos, Z., 243  
Zwanzig, R. W., 54–55

## SUBJECT INDEX

Note: The letters ‘*f*’ and ‘*t*’ following the locators refer to figures and tables respectively.

### A

Automated molecular mechanics  
optimization tool for *in silico* screening  
(AMMOS), 242

### B

$\beta$ 1-Adrenergic receptor ( $\beta$ 1-AR)  
mutant, 260  
thermostabilized turkey crystal  
structures, 255–256

$\beta$ 2-Adrenergic receptor ( $\beta$ 2-AR)  
active and inactive state comparison,  
257–258  
ECLs, 270–271  
inverse agonist bound, 261–262  
structure, 255–256

BAR. *See* Bennet acceptance ratio

BEDAM. *See* Binding energy distribution  
analysis method

Bennet acceptance ratio (BAR)  
estimators, 58–59  
formula, 56–57

Biased agonism, 254–255, 274–275

Binding energy distribution analysis method  
(BEDAM)  
challenges, 63–64  
described, 62  
double decoupling, 63  
*vs.* energy-only estimators, 63  
implicit and explicit solvation, 62

### C

CAs. *See* Catalytic antibodies

Catalytic antibodies (CAs)  
designing, 105–113

1F7, 111–112

TSA

affinity, 106–107  
structure, 111

Catalytic promiscuity, 113

Chorismate mutase (CM)

activation energies, 101

advantages, 98–101

(–)-chorismate conversion, 98–101, 99  
cope rearrangement, carbachorismate  
to carbaprephenate, 103, 103

electrostatic interaction and repulsion,  
103–104, 104*f*

free-energy profiles, 98–101, 101*f*

Glu78 residue, 103

potential-energy barrier and  
electrostatic effects, 102–103

QM and MM regions, 98–101, 100*f*

transition and reactant structures,  
104–105

transition state stabilization, 104–105  
water solution and BsCM, 101, 102*t*

CM. *See* Chorismate mutase

Coarse grain simulation methods

all-atom MD, 261–262

ENM, 262–263

systematic sampling

human  $\beta$ 2-AR binding energy  
surfaces, 264, 265*f*

LITiCon, 263–264

targeted MD, 262

CoMFA. *See* Comparative molecular field  
analysis

Common reactivity pattern (COREPA)  
method, 231, 241

Comparative molecular field analysis  
(CoMFA)

CoMSIA studies, 234–237

docking, 237

Comparative molecular field analysis  
(CoMFA) (*continued*)  
ER  
  modeling, 229–230  
  subtype selectivity, 237–238  
Comparative molecular similarity indices  
  analysis (CoMSIA)  
  CoMFA studies, 234–237  
  docking, 238–239  
  ER modeling, 229–230  
CoMSIA. *See* Comparative molecular  
  similarity indices analysis  
Continuum solvent models  
  cavity free energy  
    atomic factors, 288  
    SAS, 287–288  
  electrostatic Poisson  
    dielectric constant, 291–292  
    induced polarization charge density,  
      291  
  GB, 296–302  
  MS, 292–294

## D

DBVS. *See* Docking-based virtual screening  
Density functional theory (DFT) methods,  
  125–126  
Dissipative particle dynamics (DPD)  
  simulation, membrane and protein  
  barostat  
    dissipation-fluctuation theorem,  
      155–156  
    equilibration time, 156  
    piston force, 155  
    relaxation method, 155  
    zero surface tension, 155  
  bead motion and total force, 148  
  coarse-grained/mesoscopic  
    simulations, 145  
  coarse graining, 146–147, 147*f*  
  described, 144–145, 147–148  
  equations of motion, integration  
    barostat, 155  
  Euler method *vs.*  
    VV approaches, 153  
    loop over steps, 154  
    VV algorithm, 153–154  
  harmonic potential, 149–150

  initial and boundary conditions  
    bilayer fluctuations, 156  
    predefined membrane setting,  
      156–157  
  light microscopy methods, 144  
  molecular dynamics, 146  
  Newton's equations of motion, 148  
  parameters, 157  
  processor speed, 145  
  repulsive conservative force, 148–149,  
    149*f*  
  rigidity, hydrocarbon chain  
    model, 150  
  structure and dynamics investigation  
    budding and fission, 169–171  
    described, 162  
    exogenous factors, 178–179  
    fusion, 171–172  
    lipid aggregates, 166–167  
    lipid bilayers, 163–165, 167–169  
    membrane proteins, 172–178  
    multicomponent membranes, 165  
  testing and calibration  
    barostat, 160–161  
    bead velocity, 158–159, 159*f*  
    conversion, SI units, 162  
    density profiles, 159–160, 160*f*  
    physical quantities, 157  
    temperature, 158, 158*f*  
  thermostat  
    Andersen and Langevin thermostats,  
      152  
    exchange frequency, 152  
    fluctuation-dissipation theorem,  
      150–151  
    Nosé–Hoover thermostat, 151–152  
    random and dissipative forces, 150  
    repulsion parameter, 151  
    Schmidt number, 153  
    time scales and solvent, 145–146  
Docking-based virtual screening (DBVS),  
  243–244

## E

EAS. *See* Endocrine active substances  
EDC. *See* Endocrine-disrupting chemical  
Elastic network model (ENM)

- ed-ENM model, formulation, 160*f*
- proteins
  - deformation movements, 202
  - near-equilibrium dynamics
    - properties, 191–192, 194, 196
  - rhodopsin activation, 262–263
- Endocrine active substances (EAS)
  - adverse effects, 218
  - described, 218
  - estrogenic
    - exogenous compounds, 224, 226*f*
    - industrial chemicals and pesticides, 227
    - pharmaceuticals, 226–227
    - phytoestrogens, 227–228
  - NR binding, 218–219
- Endocrine-disrupting chemical (EDC)
  - binding, 244
  - defined, 218
  - in silico* screening, 242–243
- Estrogen receptor (ER) mediated toxicity,
  - molecular modeling
    - combined studies
      - catalyst software, 239–240
      - CoMFA and CoMSIA, 234–237
      - CoMFA and docking, 237
      - CoMSIA and docking, 238–239
      - COREPA method, 241
      - 3D QSAR, 237
      - GRIND QSARs, 238
      - hologram QSAR models, 237–238
      - SERMs, 240
      - subtype selectivity, 237–238
      - VirtualToxLab, 241–242
  - 3D approaches, 220
  - docking and virtual screening studies
    - agonist *vs.* antagonist binding, 244
    - algorithms and ligand ranking, 243
    - AMMOS, 242
    - chemical screening and testing, 243
    - indication factors, 242
    - MD simulations and *in silico* screening, 242–243
    - PBVS and DBVS, 243–244
    - polychlorinated compounds, 245
    - Shape Signatures tool, 244–245
- EAS
  - adverse effects, 218
  - described, 218
  - NR binding, 218–219
- EDC, 218
- estradiol, 219
- in silico* models, 219–220
- ligand- and receptor-based models, 234, 235*t*
- ligands, 220, 221*f*
- QSARs
  - assays, 228–229
  - binding affinity and risk assessment, 228
  - event simulation, 229
  - multidimensional, 231–233
  - three-dimensional (3D), 229–231
- receptor-based approaches
  - docking procedures, 234
  - placement, molecules, 233
  - Protein Data Bank, 233
  - spatial orientation, 233
  - virtual screening, 234
- structural characterization and ligands
  - activation functions, 221–222
  - binding pocket, 222–224, 223*f*
  - ER $\alpha$  and ER $\beta$ , 220–221, 222*f*
  - estradiol (E2), 224, 225*f*
  - estrogenic EAS, 224–228
  - LBDs, 222, 223*f*
  - signaling, 222–224

## F

- FAAH. *See* Fatty acid amide hydrolase
- FAMBE. *See* Fast adaptive multigrid boundary element
- Fast adaptive multigrid boundary element (FAMBE)
  - CPU time, 295–296
  - FAMBEpH, 303
  - FAMBEpH–GB, 303, 306–308
  - programme, 295–296
  - protein solvation energy estimation, 310
- Fatty acid amide hydrolase (FAAH)
  - catalytic mechanism, 2

## Fatty acid amide hydrolase (FAAH)

*(continued)*

computational approaches

LBDD, 3–4

QSAR, 3–4

SBDD, 4–5

description, 2

development, 5

electron-donor groups, 21–22

enzyme activity, 3

genetic/pharmacological inactivation,  
2–3

LBDD, 5–10

N-alkylcarbamate acid aryl, 3

SBDD, 11–21

therapeutic option, 22

time requirement, 5

**G**

## Generalized Born (GB) method

CFA, 298–299

GBSVMS, 299–301

GPCRs. *See* G-protein-coupled receptors

## G-protein-coupled receptors (GPCRs)

activation

crystal structures comparison,  
257–258, 257*f*

hydrogen bond, 257–258

rhodopsin, 256

activation mechanism, 261

agonists, 254–255

all-atom MD simulations, water

residues dynamics, activation,  
271–273, 273*f*side-chain conformation dynamics,  
W265<sup>6,48</sup>, 271–273, 272*f*

transmembrane molecules, 273–274

 $\beta$ 1-AR and  $\beta$ 2-AR, 255–256

basal/constitutive activity, 255

class A activation mechanism

goal, 268

“ionic lock”, 268–269

NMR spectroscopy, 270–271

rhodopsin, 269

site-directed spin labeling, 268–269

coarse grain simulation methods

ENM, 262–263

systematic sampling, 263–264

targeted MD, 262

computational methods, activation

pathways calculation

conformational transition mapping,  
264–265

Monte Carlo method, 265–267

conformational flexibility, thermostable  
mutants, 260

dynamics and conformational state

ensemble

bimane-labeled monomeric human  
 $\beta$ 2-AR, 259–260

fluorescence, 258–259

functional selectivity/biased agonism,  
274–275

multiscale methods, conformational

ensemble

conformations sampled, human  
 $\beta$ 2-AR, 267–268, 272*f*

targeted MD approaches, 267

water role, receptor activation,  
271–274

## Grid independent descriptors (GRIND)

QSARs, 238

**H**

## Hybrid QM/MM simulations

advantages, 90

approaches

described, 91–92

development and application, 92–93

dual level strategy, 95–97

gas phase reactivity, 93

Hamiltonian, 92–93

interpolated corrections, 94

micro–macro iteration optimization

algorithm, 95–97, 96*f*

minimum energy paths and transition

structures, 94–95, 95*f*

reactant and transition states, 93

semiempirical Hamiltonians

and PES, 93

- atomic resolution predictions, 90–91
- CAs and TSA
- activation free energy, 106
  - affinity and somatic mutations, 106–107
  - Bacillus subtilis* and *Escherichia coli*, 107–108
  - barriers, free-energy, 108–109, 109*t*
  - catalytic efficiency and maturation process, 107
  - design, 105
  - IF7 interactions, 109–110
  - formation, substrate–catalyst complex, 105
  - free-energy profiles, 108, 108*f*
  - in silico* mutations, 111–112
  - immune system structures, 112–113
  - individual amino acid residues, 109–110, 110*f*
  - substrate–protein interactions, 111–112, 112
  - X-ray diffraction study, 111
- catalysis
- described, 83
  - research, 124–125
- catalyst design, 123–124
- catalytic materials, 128–129
- CM, 98–105
- complex chemical processes, 85–86
- computational
- chemistry, 120–121
  - modeling, 85
- design, catalytic functions, 91
- development and application, 122
- DFT methods, 125–126
- directed evolution, 87–88
- docking programs, 86–87
- electronic structure calculations, 120–121
- enzymes
- de novo* engineering, 88
  - described, 83–84
  - rate enhancements, 84
- excited state dynamics, 126–127
- in silico* mutations, 124–125
- MD techniques and docking protocols, 87–88
- metal–organic interfaces, 125–126
- model systems, 125–126
- molecular modeling techniques, 85–86, 129
- multidisciplinary approach, 124–125
- observables, 121
- PMF/free-energy calculations
- biasing potentials and umbrella sampling technique, 97–98
  - described, 97–98
  - molecular simulation, 97–98
- promiscuity, enzyme catalysis
- behavior, PchB, 117–118
  - chemical transformations, 115
  - computational enzyme design, 120
  - condition and substrate, 113
  - directed evolution and rational design, 114–115
  - free-energy profiles, 116–117, 118*f*
  - I207F mutation, 119
  - IPL, 115–116, 116
  - kinetic analyses, 113–114
  - O7–Arg405 interaction, 119–120
  - protein engineering, 113–114
  - Val35Ile and Ala35Ile mutations, 118–119
  - wild-type MbtI *in vitro*, 116, 117
- protein structure and stability, 89–90
- quantum chemistry, 86
- rational/*de novo* protein design, 123
- reaction rate prediction, 126
- Rosetta methodology, 88–89
- static pressure, 127–128
- substrate specificity, 122
- temperature and pressure, 128
- transition state
- chemical reactivity, 84
  - stabilization and active-site structures, 85
- umbrella sampling, 123
- virtual screening methods, 86–87

## I

- Implicit solvent models
- advantages, 308–309



- Implicit solvent models (*continued*)  
 computer simulations, 282–283  
 continuum, 286–302  
 limitations, 313–314  
 pH MD simulations  
 CpHMD, 312  
 explicit stochastic titrations methods,  
 311–312  
 I-titration method, 312  
 protein  
 folding, 310–311  
 ionization, 302–308  
 protein decoy discrimination  
 coarse grain and heuristic methods,  
 309  
 FAMBE method, 310  
 solvation, 310  
 protein transport  
 gas phase, water, 286  
 partition function, 283  
 thermodynamic cycle, 284, 285*f*
- Implicit titration potential of mean force  
 (IT-PMF)  
 atomic forces, 306  
 implementation, 303  
 protein free energy, 302–303
- IPL. *See* Isochorismate pyruvate lyase
- Isochorismate pyruvate lyase (IPL),  
 115–116, 116
- IT-PMF. *See* Implicit titration potential of  
 mean force
- L**
- Langevin dynamics  
 Brownian motion, 205–206  
 Gaussian process, 205–206  
 Newton's equation, 206  
 NMA, 208  
 noise function, 207–208
- LBDD. *See* Ligand-based drug design
- LBDs. *See* Ligand-binding domains
- LIE. *See* Linear interaction energy
- Ligand-based drug design (LBDD)  
 compounds, 6, 6*f*  
 3D-QSAR methods, 5  
*gauche* and *anti* conformation, 6, 7*f*  
 inhibitory potency, 7–8, 7*f*  
 substituents, 8–10
- Ligand-binding domains (LBDs)  
 activation functions, 221–222  
 ER $\alpha$  structure, 222, 223*f*  
 flexibility, 224  
 hER $\alpha$ , 244
- Linear interaction energy (LIE)  
 chemical reactivity, 15  
 ligand binding, 16  
*N*-alkylcarbamic acid biphenyl-3-yl esters,  
 16–20, 17*t*  
 SGB-LIE equation, 16  
 URB880 two-dimension, 20–21, 21*f*
- Lipid bilayers  
 perturbations  
 decay length, 174–175  
 hydrophobic mismatch effects, 173  
 tilt angle, 173–174  
 phase diagrams  
 alcohol-induced interdigitation, 168  
 cholesterol, 168–169  
 composition and temperature, 167  
 head–head repulsion, 168  
 single-chain and double-chain lipids,  
 167–168  
 physical properties  
 bending stiffness, 143–183, 163  
 chain length and asymmetry,  
 164–165  
 lateral pressure profile, 164  
 single and double-chain models, 163,  
 163*f*
- M**
- MD. *See* Molecular dynamics
- Membrane structure and dynamics  
 investigation, DPD method  
 budding and fission  
 coalescence dynamics, 170  
 intracellular protein trafficking, 169  
 paternal vesicle, 169–170  
 vesiculation and periodic boundary  
 conditions, 170–171

- described, 162
  - exogenous factors
    - nonionic surfactants, 178
    - phospholipase, 179
  - fusion
    - energy barriers, 172
    - time and tension, 172
    - vesicle and planar membrane, 171
  - lipid aggregates
    - inverted hexagonal and cubic phase, 166–167
    - number fraction, 166
    - vesicle formation, 166
  - lipid bilayers
    - phase diagrams, 167–169
    - physical properties, 163–165
  - membrane proteins
    - acylation, 178
    - cluster formation, 174*f*, 176
    - described, 172
    - diffusion coefficient determination, 175–176
    - entropy, 177
    - homo-oligomers, 177
    - hydrophobic matching, 172–173
    - hydrophobic shielding, 177
    - model, 173, 174*f*
    - PMF, 174*f*, 176–177
  - multicomponent membranes, 165
  - Metropolis Monte Carlo (mMC) simulation
    - average acceptance rate, 203
    - Cartesian movements, 203
    - essential deformation modes, 204–205
    - flowchart, 204*f*
    - Metropolis algorithm, 203
  - MIFs. *See* Molecular interaction fields
  - Mining minima (MM) binding free energy methods
    - advantage and limitation, 67–68
    - configurational partition function, 68
    - enthalpic and entropic components, 68–69
  - Molecular dynamics (MD)
    - PMF, 97–98
    - simulations
      - all-atom, 271–274
      - Berendsen thermostat, 151–152
      - calculation, 102–103
      - constant pH MD (CpHMD), 311–312
      - I207F mutation, 119
      - Langevin piston barostat, 155
      - ligand-binding pocket, 242–243
      - MARTINI force field, 198
      - Val35Ile variant, 118–119
    - targeted methods, 262
    - water role, all-atom simulations, 271–274
  - Molecular interaction fields (MIFs)
    - 3D QSAR approaches, 229–230
    - GRID method, 229
  - Molecular mechanics/Poisson–Boltzmann surface area (MM/PBSA) method
    - configurational entropies, 70
    - enthalpy/entropy decomposition, 69–70
    - single-trajectory approaches, 70
  - Molecular surface (MS)
    - components, 292–293
    - GBSV, 299–300
    - requirement, 292–293
    - SIMS method, 293–294
    - usage, 313–314
  - MS. *See* Molecular surface
  - Multigrid BE
    - FAMBE, 294–295
    - and FD, 294
  - Multistate Bennett acceptance ratio (MBAR) method, 58–59
- N**
- NMA. *See* Normal mode analysis
  - Noncovalent protein–ligand binding
    - alchemical formulation
      - binding energy, 32
      - decoupled states, 33
      - interaction free energy, 33
      - partition functions, 32
      - thermodynamic cycle, 33–34, 34*f*
    - bound state
      - indicator function, 40–41
      - spectroscopic reporting and exclusion zone, 43
      - surface sites and binding constant, 43
      - T4-lysozyme complex, 41, 42*f*

- Noncovalent protein-ligand binding  
*(continued)*  
 conformational decomposition  
 binding constant, 52–53  
 integration over parts approaches,  
 51–52  
 joint and conditional distributions,  
 50–51  
 macrostate-specific binding constant,  
 50–51  
 modes and free energy, 50  
 relative contribution, macrostate, 52  
 enthalpy/entropy decomposition  
 configurational entropy, 46–47  
 effective potential energy and binding  
 enthalpy, 45–46  
 implicit solvation, 46–47  
 standard binding, 44–45  
 translational and interaction  
 entropy, 45  
 implicit solvent representation  
 binding constant, 35–36  
*vs.* explicit solvent, 37  
 interaction free energy, 35–37  
 PDT, 37–40  
 standard binding free energy, 37  
 molecular association equilibria  
 binding free energy and constant, 30  
 configurational partition functions,  
 30–32  
 indicator function, 30–32  
 PMF formulation, 34–35  
 receptor–ligand interactions, 44  
 reorganization free energy  
 binding process steps, 47  
 defined, 47–48  
 macrostate restraints and population,  
 48–49  
 restrain-and-release decomposition,  
 48, 48*f*
- Nonpolar interactions free energy, 286–287
- Normal mode analysis (NMA)  
 deformation  
 large movements, 202  
 modes, 202  
 distribution, 200  
 frequency eigenvectors, 201  
 practical use, 202  
 Taylor series, 201
- Nosé–Hoover thermostat, 151–152
- NRs. *See* Nuclear hormone receptors
- Nuclear hormone receptors (NRs)  
 described, 218–219  
 ER family, 219
- ## P
- PBVS. *See* Pharmacophore-based virtual  
 screening
- PDT. *See* Potential distribution theorem
- PES. *See* Potential-energy surface
- Pharmacophore-based virtual screening  
 (PBVS), 243–244
- PMF. *See* Potential of mean force
- polarization free energy, 290–291
- Potential distribution theorem (PDT)  
 binding affinity density, 40  
 binding energy distribution, 38, 38*f*  
 effective binding energy, 37–38  
 particle insertion method, 38–39  
 solute–solvent interaction, 39
- Potential-energy surface (PES), 93
- Potential of mean force (PMF)  
 binding  
 constant formulation, 34–35  
 free energy calculation, 57–58, 64  
 calculations, 97–98  
 defined, 176  
 protein–protein interaction, 174*f*,  
 176–177  
 solvent, 35–40
- Protein, coarse graining  
 approximate coarse-grained models,  
 186–188  
 atomistic molecular dynamics (MD)  
 simulations, 185–186  
 bonded and nonbonded interactions,  
 185–186  
 crystal structure, 186–188  
 experimental structure, 184–185  
 flat potentials  
 long-range attractive effects, 195–196  
 MD sampling algorithms, 194–195

- particle–particle distance vibration, 194–195
- protein–protein interactions, 196
- pseudo-physical strategies, 195–196
- fundamental problems, 186
- Gō-like potentials
  - atomistic physical models, 190–191
  - $C\alpha$  coarse-graining, 190–191
  - formalism, complexity, 190–191
  - Langevin dynamics sampling
    - algorithms, 191
  - nearest-neighbor energy, 190–191
- harmonic potentials
  - covalent and non-covalent contacts, 193–194
  - distance-dependent function, 193–194
  - ed-ENM model, 193–194, 194*f*
  - ENM, 191–192
  - inter-residue distance, 192–193
  - Kirchhoff topology matrix, 191–192
  - near-equilibrium dynamics
    - properties, 194
  - pairwise Hookean potential, 191–192
  - protein-fitted scaling factor, 192–193
  - remote residues interactions, 191–192
  - sampling technique, 194
  - topology matrix, 193–194
- MD codes, 186
- MDWeb Server, 189*f*
- Newton's equations, 185–186
- physical and pseudo-physical potentials
  - CHARMM-like, 198
  - coarse-grained model, 198–199
  - four-to-one mapping, 197
  - GROMACS simulation package, 198
  - Lennard–Jones parameters, 197–198
  - MARTINI force field, 197
  - nonbonded residue interactions, 199
  - non-neighboring beads, 197–198
  - quantum mechanical and atomistic
    - dynamics simulations, 198–199
- Protein Data Bank (PDB), 184–185, 184*f*
- sampling techniques
  - discrete molecular dynamics, 208–210
  - Langevin dynamics, 205–208
  - metropolis Monte Carlo (mMC)
    - simulation, 203–205
  - normal mode analysis, 200–202
  - structural variation, 200
- simplification, atoms, 188–190
- 2010 version
  - MoDEL, 187*f*
  - simulations, 187*f*
- Protein ionization
  - equilibrium titration
    - CpHMD, 303
    - IT-PMF, 302–303
  - FAMBEpH-GB method
    - CpHMD trajectory, 308
    - solvent polarization, 307
- implicit titration
  - electrostatic nature, 305
  - structure, 306
  - thermodynamic integration
    - method, 304
- Protein–ligand binding affinities, modeling
  - BEDAM, 62–64
  - docking and empirical scoring
    - approaches, 28
  - double decoupling
    - described, 59
    - indicator function, 60
    - L99A and L99A/M102Q mutants, 59–60
    - ligand restraints, 60–61
    - simulations, 60
    - soft-core hybrid potentials, 61–62
  - force fields, 53
  - free energy estimators
    - BAR formula, 56–57
    - binding PMF approach, 57–58
    - bound ensemble, 55
    - $\lambda$ -dependent hybrid potential, 55–56
    - exponential average, 56–57
    - implicit solvation, 54
    - MBAR method, 58–59
    - perturbation and distribution, 54–55
    - stratification technique, 55–56
    - TI formula, 56
    - umbrella sampling, 57–58
    - WHAM, 57–58
- ligand and receptor reorganization

- Protein–ligand binding affinities, modeling  
(*continued*)  
 chemical rigidification, 71–72  
 entropic model, 72  
 favorable and unfavorable work, 71  
 HIV epitopes, 71  
 MM/GBSA model, 72  
 protein side-chain motion, 73  
 MM binding free energy methods, 67–69  
 MM/PBSA and MM/GBSA approaches  
 configurational entropies, 70  
 enthalpy/entropy decomposition,  
 69–70  
 single-trajectory approaches, 70  
 noncovalent binding theory  
 alchemical formulation, 32–34  
 bound state, 40–43  
 conformational decomposition,  
 50–53  
 enthalpy/entropy decomposition,  
 44–47  
 implicit solvent representation, 35–40  
 molecular association equilibria,  
 30–32  
 PMF formulation, 34–35  
 receptor–ligand interactions, 44  
 reorganization free energy, 47–49  
 physics-based models, 28–29  
 PMF approach, 64  
 RE conformational sampling, 66–67  
 relative binding free energies  
*vs.* absolute binding free energies, 66  
 described, 65–66  
 pharmaceutical applications, 65  
 statistical mechanics theory, 29  
 thermodynamic path and end point  
 methods, 54
- Q**
- QM/MM. *See* Quantum mechanics/  
 molecular mechanics  
 QSARs. *See* Quantitative structure-activity  
 relationships  
 Quantitative structure-activity relationships  
 (QSARs)  
 assays, 228–229  
 binding affinity and risk  
 assessment, 228  
 event simulation, 229  
 GRIND, 238  
 hologram, 238  
 multidimensional  
 4D, 231–232  
 5D and 6D, 232  
 Quasar software, 232  
 VirtualToxLab and docking protocol,  
 232–233  
 three-dimensional (3D)  
*vs.* classical approaches, 229  
 CoMFA and CoMSIA method,  
 229–230  
 COREPA, 231  
 crystallization process, 231  
 estrogenicity prediction, 237  
 geometry optimization and energy  
 minimization, 230  
 MIFs, 229  
 pharmacophore modeling, 231  
 Quantum mechanics/molecular mechanics  
 (QM/MM) mechanistic model  
 application, 12  
 binding orientation, 15  
 energy profile, 13–15, 14*f*  
 Ser241 steps, 12–13, 13*f*  
 URB597, FAAH, 11*f*, 12–13
- S**
- Sampling techniques, protein  
 discrete molecular dynamics  
 covalent bonds, 209  
 DMD calculation, 209  
 Newton's equations, 208  
 Langevin dynamics, 205–208  
 metropolis Monte Carlo (mMC)  
 simulation, 203–205  
 NMA, 200–202  
 structural variation, 200  
 SAS. *See* Solvent-accessible surface  
 Selective estrogen receptor modulators  
 (SERMs), 240

SERMs. *See* Selective estrogen receptor modulators  
SGB. *See* Surface generalized born  
SIMS. *See* Sooth invariant molecular surface  
Solvent-accessible surface (SAS), 287–288, 289  
Sooth invariant molecular surface (SIMS), 293–294  
Structure-based drug design (SBDD)  
  inhibitory potency *vs.* lipophilicity, 10*f*, 11  
  LIE calculations, 15–21  
  *N*-alkylcarbamic acid, 12, 17*t*  
  QM/MM mechanistic model, 12–15  
  URB597, 11, 11*f*  
Surface generalized born (SGB)  
  approach, 16–20  
  continuum model, 16  
  SGB-LIE equation, 16

**T**

Thermodynamic integration (TI)  
  formula, 56  
Transition state analogues (TSA)  
  activation free energy, 106  
  CA affinity, 106–107  
  design, 105

  protein structures, 112–113  
Transmembrane (TM)  
  helix, 257–258  
  proteins, 254–255  
  region, 271  
TSA. *See* Transition state analogues

**V**

Van der Waals interactions, solute–solvent  
  AGBNP, 289–290  
  free energy, 289  
Velocity-Verlet (VV) algorithm  
  DPD  
    dissipative force, 153–154  
    integration, 154  
  *vs.* Euler method, 153

**W**

Weighted histogram analysis method (WHAM)  
  binding free energy estimators, 58–59  
  umbrella sampling, 57–58  
WHAM. *See* Weighted histogram analysis method