Computer simulations to engineer PDZ-peptide recognition

Francesco Villa

université PARIS-SACLAY

NNT: 2018SACLX076

Thèse de doctorat

Computer simulations to engineer PDZ-peptide recognition

Thèse de doctorat de l'Université Paris Saclay
préparée à l'École Polytechnique

École doctorale nº573
INTERFACES: approches interdisciplinaires, fondements, applications et innovation
Spécialité de doctorat: Biologie

Thèse présentée et soutenue à Palaiseau, le 23 Octobre 2018 par

Francesco Villa

Composition du Jury:

M. Amedeo CAFLISCH Professeur (University of Zurich)	Rapporteur
M. Jean-Philip PIQUEMAL Professeur (Université Pierre-et-Marie-Curie)	Rapporteur
Mme Nathalie BASDEVANT Maître de conférences (Université d'Évry)	Examinatrice
M. Tâp HA DUONG Professeur (Université Paris Sud)	Président
M. GEORGIOS ARCHONTIS Professeur (University of Cyprus)	Examinateur
M. Thomas SIMONSON Professeur (École Polytechnique)	Directeur

Remerciements

Je tiens à remercier les membres du jury *Amedeo Caflisch*, *Jean-Philip Piquemal*, *Nathalie Basdevant*, *Tâp Ha Duong* et *Georgios Archontis* pour avoir accepté d'examiner mon travail de thèse. Je voudrais adresser un remerciement particulier à mon directeur de thèse, *Thomas Simonson*. Merci d'avoir accepté ma candidature il y a trois ans et d'avoir suivi l'avancement de mon travail jour après jour avec une rigueur scientifique extraordinaire. Merci pour les nombreuses discussions, les conseils et les enseignements. Merci enfin pour le temps consacré à lire et commenter ce manuscrit.

Je remercie particulièrement *Ernesto J. Fuentes* pour son importante collaboration dans le projet PDZ, *Alex MacKerell* et *Benoît Roux* pour leur grande disponibilité et soutien avec le développement du champ de force Drude.

Je remercie tout l'équipage du laboratoire BIOC, surtout le directeur *Yves Méchulam* pour son soutien pendant le processus de sélection de l'école doctorale. Merci à *Thomas G.* et *David* pour le support technique et scientifique pendant ces trois ans. Merci à *Alexey* pour les discussions et tous les pauses café, à *Savvas* et *Karen* pour m'avoir aidé à faire les premiers pas dans le monde du CPD avec Proteus. Merci aux camarades *Nicolas P.* et *Vaitea*, pour votre soutien pendant les jours de travail et après les nombreux bières partagé ensemble. Merci à *Maxime* et *Yann*. Je remercie aussi *Rojo*, *Alexandrine*, *Marie-Pierre* et à tous ceux qui ont participé temporairement aux activités de l'équipe de bioinformatique. Merci à *Paula*, et bon courage pour ta thèse. Je remercie *Clara*, *Pierre*, *Marc D.* pour les pauses déjeuner. Un grand merci à *Mélanie* et *Caroline* pour toute votre aide. Je remercie aussi tous les autres membres du labo: *Pierre-Damien*, *Emma*, *Marc G.*, *Alexis*, *Titine*, *Mimi*, *Gabrielle*, *Michou*, *Nathalie* et *Guillaume*. Merci aux doctorants et postdoctorants *Giuliano*, *Nicolas M.*, *Amlam*, *Ditipriya*, *Irène*, *Ramy*, *Nhan* et *Angelina*.

Merci aux amis, particulièrement aux *ritals* pour tous les soirées parisiennes passées ensemble et aussi pour les futures. Ovunque saremo fra tre anni, *avremo sempre Parigi*. Merci *Chiara*, pour tout ce que tu as partagé avec moi jusqu'ici. Merci à ma *belle famille*, à *Roberto* et *Ivana*, pas besoin d'expliquer pourquoi. E infine grazie a Mamma, Papà e Stella, perché senza di voi non sarei mai arrivato fin qui.

All the best.

Table of contents

List of figures

List of tables

Introduction

Protein-protein interactions (PPIs) regulate complex functional networks in cells. They are often mediated by specialized domains such as SH3 and PDZ domains. The specific recognition between such domains controls important biological functions. The understanding of these phenomena is extremely difficult and challenging, even when the interaction partners are relatively small protein domains. Indeed, PPI networks exploit a large number of binding events to transfer information through communication pathways, governing signalling inside and outside cells. Disrupting or altering the equilibrium between PPIs plays an important role in several diseases: one example is the observed altered expression of important signalling proteins found in diseased tissues. The inibition of targeted PPIs is a recognized strategy for the development of new drugs (Böhm & Schneider [2003]).

Understanding and engineering PPIs can be addressed by molecular modeling approaches, including free energy simulations and Computational Protein Design. Computational Protein Design (CPD) allows one to perform large screenings of many protein variants in a limited amount of time. Several protein mutants can be studied starting from a native structure, optimizing a preferred quantity such as stability or binding. CPD applications increased dramatically in the last ten years, but there is a strong need of continuous improvement of both theoretical and computational models to obtain reliable and predictive results that match experiments as closely as possible.

In the present thesis we focus on PDZ domains, which are among the most widespread PPI domains. The human genome has more than 400 PDZ domains belonging to more than 200 different proteins. They are also present in bacteria, yeasts, plants, insects and vertebrates (from the *SMART* database - Schultz [2000] ; Letunic & Bork [2017]). PDZs are small domains, composed of roughly 90 amino acids that specifically recognize the $5-10$ C-terminal amino acids of their binding partners. They are often able to bind the corresponding peptides in isolation. The experimental *screening* of binding of peptide and protein variants is the most common strategy to better characterize their interactions (Songyang [1997] ; Wiedemann *et al.* [2004] ; Stiffler *et al.* [2007]).

PDZ domains share a common secondary structure, composed of 5-6 β strands and two α helices, as shown in figure 1. In their typical bound configuration, C-terminal residues of the binding partner assume an elongated β strand conformation in the binding pocket between the strand β_2 and helix α_2.

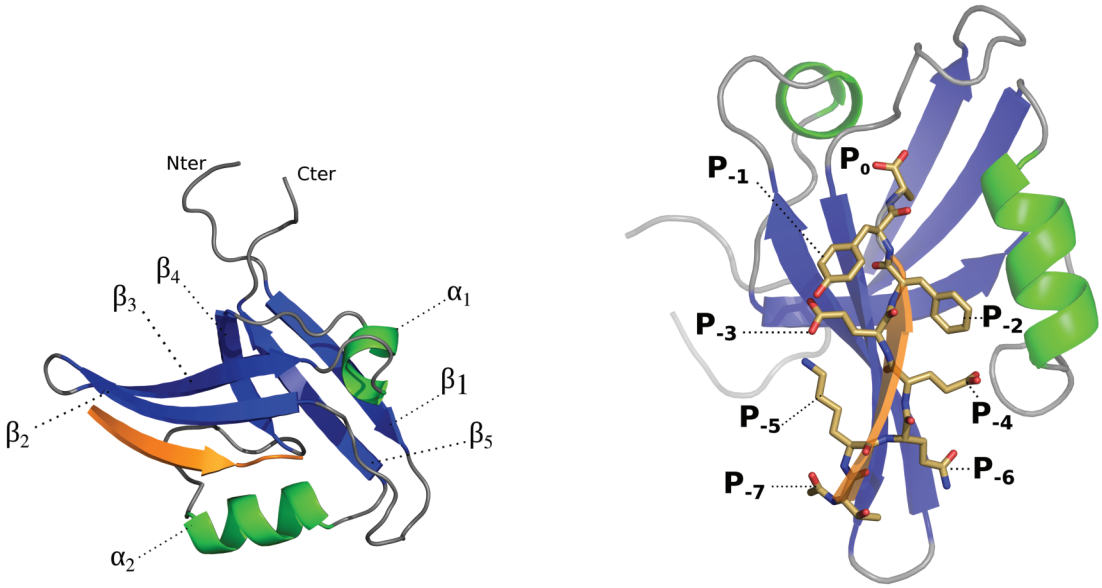


Figure 1 – **PDZ domain structure. Left**: with labels indicating α helices and β strands from Nter to Cter residues, colored in green and blue. The peptide ligand is colored in orange and interacts with residues in the binding pocket situated between β_2 and α_2. **Right**: with labels indicating ligand positions, using negative numbering: position P_0 is the most buried and corresponds to the peptide Cter residue.

Several PDZs of different amino acid sequences show similar affinities for the same ligand (Ernst *et al.* [2010]). This suggests that important hotspot residues for their PPIs are situated in specific domain regions. Crystal structures and combinatorial screening showed that strands β_2, β_3 as well as helix α_2 are important for the PDZ:peptide binding [fig. 1]. PDZs can be classified according to their specificity for different peptide ligands. We number the C-ter residues of the peptide ligand in inverse order where position P_0 is the most buried C-ter residue, P_{-1} is the second-last and so on. Four PDZ classes are defined according to the composition of the last 3 C-terminal positions of the peptide ligand, $P_{0,-1,-2}$. The first class binds the pattern $\mathrm{S/T} - \mathrm{X} - \Phi$ (Serine/Threonine followed by an amino acid X and then a hydrophobic residue). The second class binds $\Phi - \mathrm{X} - \Phi$ (any residue between two hydrophobic amino acids). The third binds $\mathrm{D/E} - \mathrm{X} - \Phi$ and $\mathrm{X} - \Psi - \mathrm{D/E}$ where Ψ is an aromatic residue. Another PDZ classification uses 16 patterns based on the last seven Cter residues (Tonikian *et al.* [2008]). In this study the authors analyzed binding specificity on a large scale. A huge combinatorial peptide library was used to produce more than 3000 peptide ligands able to bind to 54 PDZ domains from the human genome and 28 from the *Caenorhabditis*

elegans worm genome. PDZs showed capability to bind many different peptides, with patterns which are conserved between worm and human.

All these experimental strategies need expression and purification of PDZ domains: the procedure can be quite expensive especially when considering several mutants. Also the structural characterization of variants can be quite difficult and time-demanding. For this reason experiments have been coupled with computational studies, often involving molecular dynamics (MD) simulations which are also able to give insights about thermodynamic properties of the systems of interest. Basdevant *et al.* [2006] performed MD simulation of 12 different PDZ:peptide systems and evaluated binding contributions. Comparing the simulated complexes, they found that non-polar interactions are predominant for specificity, as expected from the observed promiscuity. The authors also found the importance of dynamical behavior and peptide reorganization upon binding. Other studies can be found in Tian *et al.* [2011], Steiner & Caflisch [2012], Blöchliger *et al.* [2015] and Sensoy & Weinstein [2015].

Among the known PDZs, we will focus on the domain belonging to the Tiam1 protein. It is known for modulating signaling activity for cell proliferation and migration, whose dysregulation increases metastatis growth in cancer. Indeed, both overexpression and subexpression of Tiam1 are connected to several tumors. For example, overexpression was found in sick tissues (Minard *et al.* [2006]), suggesting that it can be used as a marker (Zhao *et al.* [2010] ; Ding *et al.* [2009]). Starting from now, we will use "Tiam1" to indicate the PDZ domain of the Tiam1 protein. It is a class 2 PDZ domain, meaning that it has a preference to bind $\Phi - \mathrm{X} - \Phi$ peptides (any residue between two hydrophobic amino acids at positions $P_{0,-2}$). Four X-ray structures of Tiam1 complexes where the domain is bound to different ligands can be found in the PDB with access codes 3KZD, 3KZE, 4GVC and 4GVD (Shepherd *et al.* [2010]). Structures are also known for three quadruple mutants (4NXP, 4NXQ and 4NNR, Liu *et al.* [2013]) and one NMR structure has also been solved for the unbound, *Apo* form of Tiam1 (2D8I, Qin *et al.* [2006]). Tiam1 is composed of 94 residues. An experimental screening of an 8-residue peptide library (Shepherd *et al.* [2011]) showed its capability to bind 70 different peptides. For selected ligand variants, their relative binding free energies were measured. A 2.2 kcal/mol interval was observed between the best and worst binding free energies.

The Tiam1 natural *wildtype* binder is the Sdc1 peptide, composed of the last 8 amino acid residues of the Tiam1 partner Syndecan1, TKQEEFYA. However, Tiam1 is also able to bind to other natural peptides like Caspr4 (Spiegel *et al.* [2002]) of sequence ENQKEYFF. These interactions are involved in cell adesion processes, which are important factors for the metastatic spread of several cancers (Bendas & Borsig [2012]).

In the first part of this manuscript, we begin with a general introduction about Computational Protein Design (CPD), with more detailed information regarding the Proteus software (Simonson *et al.* [2013]) developed at École Polytechnique. Next, we apply the method to design peptide sequences that bind the Tiam1 PDZ domain. Results were published (Villa *et al.* [2018b]) and the paper can be found attached to chapter two.

A more detailed and quantitative study of Tiam1:peptide binding will be described in the second part of this thesis. We performed polarizable free energy simulations to study relative binding free energies for charge mutations in the binding pocket, using the Drude polarizable force field. Results were compared to those obtained with two additive force fields (Amber and Charmm). General information about polarizable models with a particular interest on Drude will be given in chapter three. In chapter four, we describe more technical aspects related to Drude free energy simulations using the a dual topology approach. Our results were published in a larger study (Panel *et al.* [2018]).

In the third part of the thesis, we describe the development of a polarizable Drude model for methyl phosphate and phosphotyrosine using established Drude parametrization procedures. Methyl phosphate parameters were validated computing Mg^{2+} standard binding free energies which showed good agreement with experiments. Results were recently published (Villa *et al.* [2018a]).

Finally, we also performed a CPD study of acid:base equilibrium constants for a collection of proteins (more information can be found in Villa *et al.* [2017]). We compared two approximations used to compute electrostatic interactions in Proteus, which both rely on a Generalized Born (GB) implicit solvent model. The more sophisticated model could be used in future design work.

Chapter 1

Computational design of PDZ-peptide binding

The protein design problem can be defined as looking for the amino acid sequences that fit a specific structure according to a scoring function which depends on the system energy. Random mutagenesis is used to produce a set of candidates that are subjected to selective pressure. Since CPD must be efficient, combinatorial approaches are preferred. A simplified and discretized conformational space, a simple scoring function and a fast exploration algorithm are key ingredients of all CPD software.

The starting point of a typical CPD simulation is the three-dimensional structure of a given protein. The wildtype sequence associated to this native conformation can then be altered by modification of side chain atoms belonging to one or more residues. This is done efficiently if the conformational space is discretized, usually keeping the protein backbone fixed. An energy function is used to score different sequences generated by the sampling algorithm. The conceptual difference between energy function and sequence score is important: the first is the formula used to compute interactions between groups of atoms in a specific configuration. The latter depends on the system energy but can vary according to the quantity to optimize. For example, optimizing protein stability one uses a scoring function which is based on the unfolding free energy difference upon mutation. On the other hand, to optimize ligand binding one has a scoring function which is based on the change in binding affinity. To optimize protonation states, one has to use a scoring function based on protonation free energy differences, and so on.

In this chapter we start by describing the most popular *energy functions* in CPD, how the conformational space is usually discretized and some widely-used exploration routines. Then we will give more detailed information about the *Proteus* software and how sequences are scored according to different quantities, with a special interest in ligand binding.

1 Energy functions for CPD

CPD energy functions are based on molecular mechanics plus an implicit solvent model. The first is borrowed from an existing force field for molecular simulation (for example Amber, Charmm) while the latter is often based on simple dielectric shielding. One widely-used implicit solvent model is *Generalized Born* (GB) combined with solvent accessible surface area (SA), also called GB/SA.

Several CPD softwares use optimized, *statistical* potentials, where the physical-based energy function terms are weighted or scaled by empirical parameters in order to fit experimental data. However in the present discussion we will focus on *physical* potentials, in the spirit of our in-house CPD software Proteus. We start by describing molecular-mechanics for internal degrees of freedom, then nonbonded interaction terms.

1.1 Bonded and nonbonded interactions

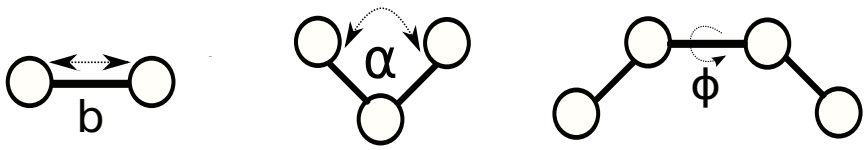


Figure 1.1 – **Bonded interactions** Bond (left), Angle (center) and Dihedral (right).

Bonded energy term Interactions between bonded atoms (up to 1-2, 1-3, 1-4 couples) are computed using a sum of several contributions

$$E_{bonded} = E_{bond} + E_{angl} + E_{Urey-Bradley} + E_{dihedral} + E_{improper} \tag{1.1}$$

Most of the terms share the same, harmonic potential functional form (bonds, angles, Urey-Bradley, Impropers) while the energy associated to dihedral angles is usually periodic

$$\begin{aligned} E_{bond} &= k_b \left(b - b_0\right)^2 \qquad E_{angl} = k_\alpha \left(\alpha - \alpha_0\right)^2 \\ E_{U-B} &= k_u \left(u - u_0\right)^2 \qquad E_{impr} = k_\omega \left(\omega - \omega_0\right)^2 \\ E_{dih} &= k_\phi \left[1 + cos\left(n\phi - \delta\right)\right] \end{aligned} \tag{1.2}$$

In general, bond-angle-dihedral terms are applied to all 1-2, 1-3, 1-4 couples [fig. 1.1], while Urey-Bradley and Impropers are used in more special cases [fig. 1.2]. Urey-Bradley controls angle bending and is applied between a few 1-3 couples which form an angle α with a central

atom. Impropers are generally used to avoid out-of-plane bending and are applied between one central atom and other 3 bonded atoms.

Figure 1.2 – **Bonded interactions** Urey-Bradley (left) and Improper (right).

Bonded parameters (like bond distance b_0 and force constant k_b) usually depend on *atom types*. Their values are optimized in order to reproduce quantum-mechanical (QM) optimized geometries or experimental structures, as well as QM vibrational spectra and QM energy scans. In most of the existing force fields, bonded parameters are quite transferable between different molecules when they refer to the same chemical group.

Nonbonded energy term Nonbonded interactions are not computed for 1-2 and 1-3 pairs (their interactions are completely included in internal terms). 1-4 couples can be excluded or scaled. The interaction between all other not excluded pairs of atoms is then sum of two terms

$$E_{nonbonded} = E_{LJ} + E_{elec} \tag{1.3}$$

Van der Waals interactions between atoms of type i and j at distance r_{ij} are expressed in the attractive part of the Lennard-Jones potential energy

$$E_{LJ}(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.4}$$

where ϵ_{ij} and σ_{ij} are force field constants for a couple of atom types. The first term expresses the strong nuclear repulsion that occurs at short distances because of the Pauli principle. The force field parameters ϵ_{ij} and σ_{ij} are optimized targeting QM interaction energies and experimental consended-phase data. As for the corresponding bonded interaction terms, LJ parameters are usually transferable between different molecules.

Partial atomic charges define the magnitude of Coulomb electrostatic interactions in a dielectric ϵ between atoms i and j at distance r_{ij}

$$E_{elec}(r_{ij}) = \frac{q_i q_j}{\epsilon r_{ij}} \tag{1.5}$$

In contrast to bonded and LJ parameters, atomic charges are less transferable. They depend on the specific molecule and are usually parametrized targeting its QM dipole moment, as well as interactions between water and model compound acceptors/donors.

1.2 Implicit solvent models

Implicit solvent models are based on the decomposition of the system solvation free energy ΔG^{solv} by integration over the solvent degrees of freedom. In general, the total ΔG^{solv} is split into two parts: the first counts electrostatic interactions between charged atoms and the second counts the nonpolar, non-electrostatic interactions:

$$\Delta G^{solv} = \Delta G^{solv}_{elec} + \Delta G^{solv}_{apol} \tag{1.6}$$

The solvation process is decomposed in 3 steps: **1)** formation of a solvent cavity which depends on the solute volume/shape, without solute-solvent interactions **2)** turn on solute-solvent vdW dispersion interactions, **3)** turn on electrostatic interactions. Steps 1 and 2 define ΔG^{solv}_{apol} while step 3 defines ΔG^{solv}_{elec}. We now describe how popular implicit models used in CPD express the different terms of equation 1.6.

Surface-accessible Surface Area In SASA models, the non-polar solvation free energy is written as a linear combination of contributions from solute atoms:

$$\Delta G^{solv}_{surf} = \sum_i \sigma_i S_i \tag{1.7}$$

For each atom i the solvent accessible surface S_i [fig. 1.3] is multiplied by a coefficient σ_i. It should be noted that equation 1.7 is based on a strong assumption of additivity.

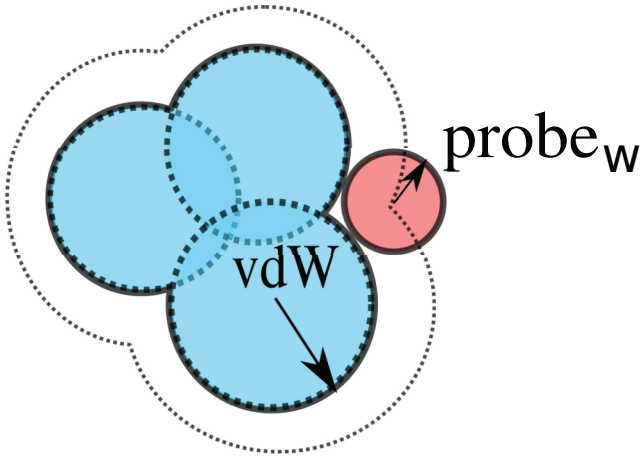


Figure 1.3 – **Solvent accessible surface** obtained rolling a sphere of radius r which represents a rigid water molecule in contact with the solute. For surface calculations in Proteus we use a probe radius $r = 1.5$Å.

CASA model In the Coulombic Accessible Surface Area model, the total (polar+apolar) solvation free energy is expressed using the simple function

$$\Delta G^{solv} = \left(\frac{1}{\epsilon} - 1\right) E_{elec} + \Delta G^{solv}_{surf} \tag{1.8}$$

The apolar contribution to the solvation free energy is obtained using the surface energy term (SASA model), while the electrostatic contribution is obtained scaling Coulomb interactions. This is one of the first implicit solvent models used in CPD. Despite its simplicity, successful applications can be found in the literature, for example in the work of Dahiyat & Mayo [1997].

Generalized Born model To describe continuum electrostatics in the context of GB, we follow the analytical treatment of Schaefer & Karplus [1996]. Their integral formula for the system electrostatic energy

$$E^{el} = \frac{1}{8\pi\epsilon_w}\int_w \bar{D}^2(\bar{x})d^3x + \frac{1}{8\pi\epsilon_p}\int_p \bar{D}^2(\bar{x})d^3x \tag{1.9}$$

is based on the assumption that the system is split into two regions (solvent and solute) characterized by two distinct dielectric constants ϵ_w and ϵ_p. Only solute atoms i contribute to the additive dielectric displacement $\bar{D} = \sum_{i=1}^{N} \bar{D}_i$.

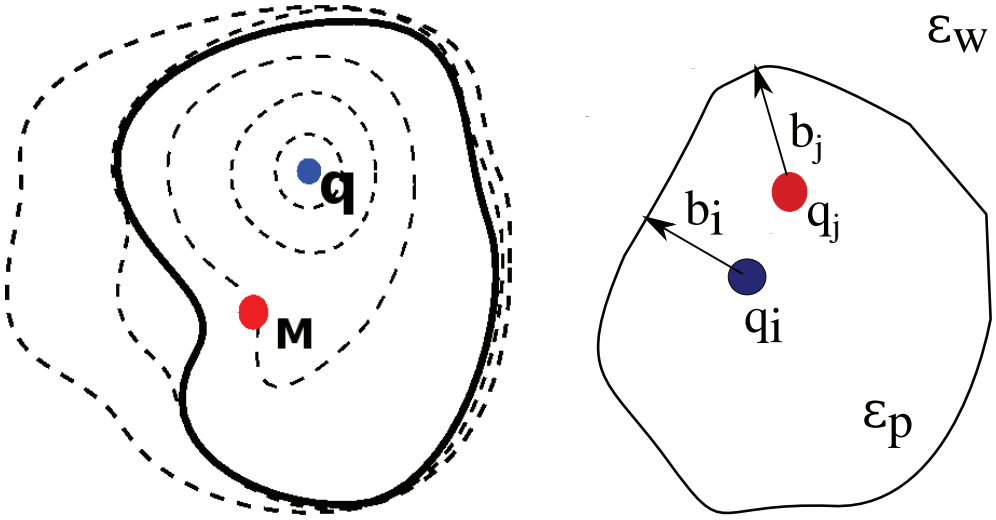


Figure 1.4 – **Interaction between solute charges Left**: potential at point M generated by a charge q buried in the solute interior. **Right**: Solvation radii of two charges q_i and q_j, interacting in the solute ϵ_p and screened by the solvent ϵ_w.

One can write

$$E^{el} = \frac{1}{8\pi\epsilon_w}\int_{\mathbf{R}^3} \bar{D}^2(\bar{x})d^3x + \frac{\tau}{8\pi}\int_p \bar{D}^2(\bar{x})d^3x \tag{1.10}$$

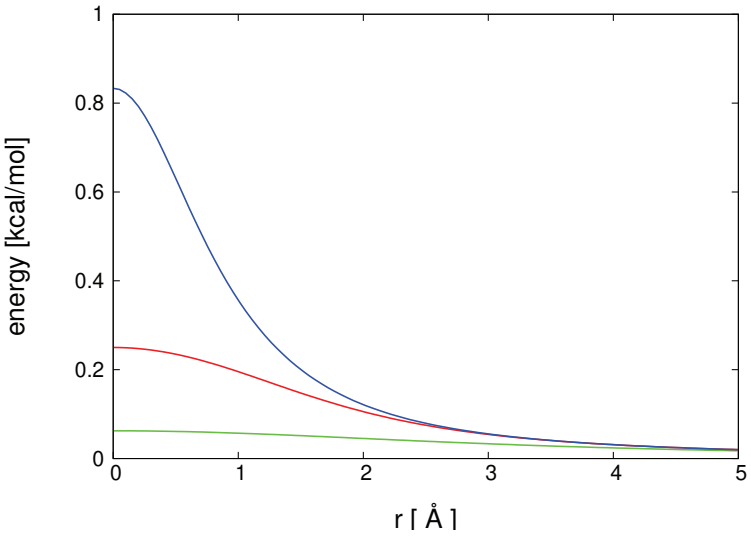


Figure 1.5 – **GB interaction energy** as function of the distance between a couple of charges r. **Blue** line: interaction for a couple of exposed charges, low $b_i b_j$. **Green** line: interaction for a couple of buried charges, high $b_i b_j$. **Red** line: interaction for intermediate value of $b_i b_j$

where

$$\tau = \frac{1}{\epsilon_p} - \frac{1}{\epsilon_w} \tag{1.11}$$

It is possible to manipulate the $\bar{D}^2$ integrals of equation 1.10. The full calculation is shown in Schaefer & Karplus [1996]. One can split diagonal $\bar{D}_i^2$ and off-diagonal $\bar{D}_i \bar{D}_j$ elements and then sum up results in *self-energy* and *interaction-energy* terms. Separating the standard, well-known Coulomb interaction between charges in the solute environment and the contribution from the solvent one obtains

$$E^{el} = \sum_{i<j} \frac{q_i q_j}{\epsilon_p r_{ij}} + \Delta E^{solv} \tag{1.12}$$

$$\Delta E^{solv} = \sum_i \Delta E_i^{self} + \sum_{i<j} \Delta E_{ij}^{int} \tag{1.13}$$

Using the approximation of Still *et al.* [1990], the solvent contribution is computed using

$$\Delta E_{ij}^{int} = -\frac{\tau q_i q_j}{\left(r_{ij}^2 + b_i b_j \exp\left[-r_{ij}^2 / 4 b_i b_j\right]\right)^{1/2}} \tag{1.14}$$

where b_i is called solvation radius of the solute charge i. It has a physical interpretation, as the measure of the distance between charge i and the solute surface. It is an analytical function of all other solute atoms positions

$$b_i = b_i(r_1, r_2, ..., r_n) \tag{1.15}$$

Figure 1.5 shows the limiting behaviour of this interaction.

The accuracy of all GB models relies on the accuracy of solvation radii calculation. In several applications, they are obtained from the definition of the self-energy potential

$$\Delta E_i^{self} \stackrel{def}{=} -\tau \frac{q_i^2}{2b_i} \tag{1.16}$$

Using this definition, solvation radii are obtained from the calculation of associated self-energies. To calculate this quantity, it is necessary to approximate the integral of the squared dielectric displacement using a continuos density function $P_k(x)$ to describe the solute volume

$$\Delta E_i^{self} = -\tau \frac{q_i^2}{2\epsilon_w R_i} + \tau q_i^2 \sum_{k \neq i} E_{ik}^{self} = -\tau \frac{q_i^2}{2\epsilon_w R_i} + \sum_{k \neq i} \frac{\tau}{8\pi} \int_{\mathbf{R}^3} D_i^2 P_k(x) dx \tag{1.17}$$

where R_i is an atomic radius, part of the force field parameters.

We now describe two GB variants that differ in the method used to approximate equation 1.17. In **GB/ACE**, the integral is calculated using a density function $P_k(x)$ where volume occupied by charges is expressed with gaussians

$$E_{ik}^{self} = \frac{1}{\omega_{ik}} exp(-r_{ik}^2 \sigma_{ik}^2) + \frac{V_k}{8\pi} \left(\frac{r_{ik}^3}{r_{ik}^4 + \mu_{ik}^4} \right) \tag{1.18}$$

ω_{ik} and μ_{ik} are functions of the volume V_k, of the atomic radii $R_k = (3V_k/4\pi)$ and a smoothing parameter λ. Another approximation of equation 1.17 is **GB/HCT**, where

$$4E_{ik}^{self} = \frac{1}{L_{ik}} - \frac{1}{U_{ik}} + \frac{r_{ik}}{4} \left(\frac{1}{U_{ik}^2} - \frac{1}{L_{ik}^2} \right) + \frac{1}{2r_{ik}} ln \frac{L_{ik}}{U_{ik}} + \frac{R_{ik}^2}{4r_{ik}} \left(\frac{1}{L_{ik}^2} - \frac{1}{U_{ik}^2} \right) \tag{1.19}$$

and

$$L_{ik} = \begin{cases} 1 & r_{ik} + R_k \leq R_i \\ R_i & r_{ik} - R_k \leq R_k \leq r_{ik} + R_k \\ r_{ij} - R_k & R_i \leq R_k \leq r_{ik} - R_k \end{cases}$$

$$U_{ik} = \begin{cases} 1 & r_{ik} + R_k \leq R_i \\ r_{ik} - R_k & R_i \leq r_{ik} + R_k \end{cases} \tag{1.20}$$

Onufriev proposed a slightly modified version of equation 1.16 to compute b_i. The atomic radius R_i of each atom is reduced by a constant and the self-energy is scaled by a factor λ,

then the solvation radius is reduced by δ

$$b_i = \left[(R_i - \rho_0)^{-1} - \lambda \sum_{k \neq i} E_{ik}^{self} \right]^{-1} - \delta \tag{1.21}$$

In the Proteus GB implementation, $\lambda = 1.4$, $\rho_0 = 0.09$ Å and $\delta = 0.15$ Å.

LK model In the Lazaridis-Karplus model (Lazaridis & Karplus [1999]), the free energy of solvation is a sum of atomic terms ΔG_i^{solv}

$$\Delta G^{solv} = \sum_i \Delta G_i^{solv} \tag{1.22}$$

where the sum runs over atoms and

$$\Delta G_i^{solv} = G_i^{ref} - \sum_j \int_{V_j} f_i(r_{ij}) dV \simeq G_i^{ref} - \sum_{j \neq i} f_i(r_{ij}) V_j \tag{1.23}$$

The "reference" solvation free energies G_i^{ref} measure the change in solvation free energy for a process where the atom i is transferred from a fully-solvated, *unfolded* state to a partially solvated or buried *folded* state. Each integral is approximated by volumes V_j occupied by the other atoms j. The f_i terms are gaussian energy density functions

$$4\pi r_{ij}^2 f_i(x_{ij}) = \alpha_i \exp(-x_{ij}^2) \qquad x_{ij} = \frac{r_{ij} - R_i}{\lambda_i} \tag{1.24}$$

where R_i is the Van der Waals radius of atom i, r_{ij} is the distance between i and j and λ_i is a parameter called *gaussian correlation length* such that when i is fully solvated, the total ΔG_i^{solv} equals zero.

Dispersion term In most popular implicit solvation models, the apolar contribution to the solvation free energy in the three steps process [eqn. 1.6] has a cavity-formation term calculated with the accessible surface area model. Moreover, the attractive part of the LJ potential [eqn. 1.4] can be integrated for each solute atom, in order to obtain total solute-solvent vdW dispersion

$$\Delta G_{disp}^{solv} = \sum_i U_i^{vdW} = \sum_i \left(-\rho_w \int_{V_{solv}} \frac{4\epsilon_{iw}\sigma_{iw}^6}{(r - r_i)^6} dV \right) \tag{1.25}$$

In the formula, ϵ_{iw} and σ_{iw} are the nonbonded parameters of eqn. 1.4 between solute atom i and the water oxygen. The integral is performed over the solvent volume, considering the number density of water molecules ρ_w as a constant. It can be split in two parts

$$U_i^{vdW} = -\frac{4\pi}{3} \rho_w \left(4\epsilon_{iw}\sigma_{iw}^6 \right) \left(\frac{1}{R_i^3} - \frac{3}{4\pi} \int_{V_i}^{solute} \frac{4\epsilon_{iw}\sigma_{iw}^6}{(r - r_i)^6} dV \right) \tag{1.26}$$

where V_i is the solute volume excluding atom i. Onufriev and coworkes proposed an approximation of the integral: the solution consists in two terms

$$I_i^{disp} = \frac{3}{4\pi} \int_{V_i} \frac{dV}{(r - r_i)^6} \simeq I_i^{main} + I_i^{neck} \tag{1.27}$$

where *main* and *neck* are respectively contributions relative to the volume region occupied by atomic spheres and the volume region between atomic spheres. The two integrals are obtained analytically (Aguilar *et al.* [2010]).

1.3 Structural models for CPD

Discretization of the protein conformational space Using an implicit solvent model, the cost to compute interactions is aggressively reduced. Protein degrees of freedom must also be simplified in order to explore large ensembles of protein variants in an acceptable computing time. Most CPD implementations use a fixed backbone while side chains assume a discrete set of conformations, or rotamers. These are defined by a rotamer library, which can be backbone dependent (Dunbrack & Karplus [1993]) or not (Tuffery *et al.* [1991]). The use of rotamer libraries is justified by the fact that amino acid side chains in proteins assume a limited number of configurations, as confirmed by several studies (Finkelstein & Ptitsyn [1977] ; Janin *et al.* [1978] ; Ponder & Richards [1987]). CPD models can also use *native rotamers*, extending existing libraries by introduction of side chain configurations extracted from the crystallographic structure of the system of interest.

Pairwise-additive energy function To explore vast ensembles of structures and sequences, CPD software needs to rapidly compute interactions between couples of atoms in a discretized conformational space. In many situations it is convenient to follow a combinatorial approach. The protocol is based on a pre-calculated energy matrix, where pairwise-additive interactions between groups of atoms (in general between couples of residues) are stored and can be accessed during the sequence/structure exploration (Dahiyat & Mayo [1997] ; Gaillard & Simonson [2014]). Additional information will be given below, with more details about our in-house software Proteus.

We describe now the general case, where such an energy function would be

$$E_{tot} = \sum_i E_{ii} + \sum_{i<j} E_{ij} \tag{1.28}$$

The sum runs over residues i and j which occupy one of their possible rotamers. Using equation 1.28, a problem arises when one wants to define a pairwise-additive energy function based on molecular-mechanics, where the solvent is described implicitly with one of the previously

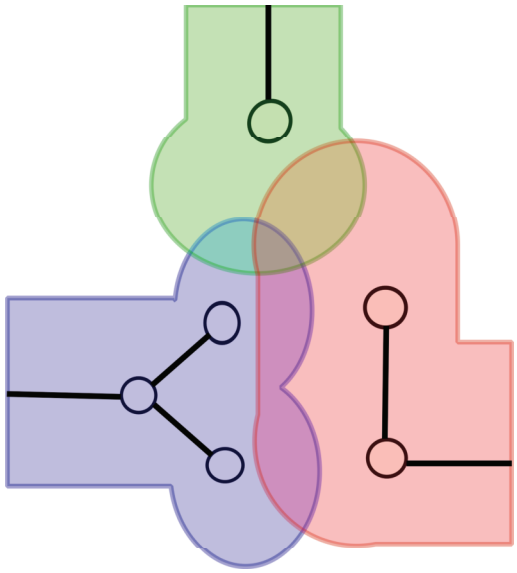


Figure 1.6 – **Surface overlap** between three buried residues. Individual surfaces are colored in green, red, blue.

discussed models, for example GB. Indeed, pairwise diagonal and off-diagonal terms E_{ii} and E_{ij} should depend only on atomic coordinates of residues i and j. However, the GB interaction between couples of atoms belonging to two residues [eqn. 1.14] has an implicit dependency on all solute atom positions through the atomic solvation radii b_i [eqn. 1.15]. Further approximations are needed to make E_{tot} residue-pairwise. One possible solution is to calculate GB interactions between pairs of residues using solvation radii obtained in the protein *native* conformation. This solution is called the *Native Environment Approximation* (NEA). Another strategy (Archontis & Simonson [2005]) was recently implemented in Proteus and makes use of a *Fluctuating Dielectric Boundary* (FDB, Villa *et al.* [2017]). Another problem arises considering the calculation of the solvent accessible surface (important when using CASA or GBSA). Indeed, considering a couple of residues i and j the intersection between their surfaces can also include their intersection with a third residue, as well as all other residues in the system [fig. 1.6]. To avoid double counting of buried surfaces, Street & Mayo [1998] proposed to apply a correction factor to exclude overlapping volume defined by surfaces of nearby buried residues.

2 Exploration of the structure-sequence space

Most CPD programs generate a limited number of protein variants, solving problems of reduced complexity in an acceptable computing time. Indeed, one could be interested in the lowest energy state or a group of sequences close to this *Global Minimum Energy Configuration* (GMEC). Several approaches can be found in literature: some of them are heuristic,

others are stochastic.

Heuristic methods Heuristic methods are useful to determine a group of low energy configurations starting from one or more random states. A popular method was introduced by Wernisch *et al.* [2000]. They proposed a simple sampling method, where starting from a random protein sequence, a single position was picked and and its rotamer optimized. The procedure was repeated for many iterations, until a local minimum was found. The method successfully reproduced side chain conformations for surface residues and stability changes for mutations applied in the protein core or at its surface. However, this method generally produces only a few variants. For this reason it is not particularly adapted to perform high-throughput design, where one wants to generate a distribution of many protein variants with a single simulation.

Monte-Carlo methods Monte Carlo (MC) sampling can be used to obtain a set of protein sequences generated from a stochastic process. One popular Monte Carlo method suitable for protein design is based on the Metropolis algorithm. With MC, it is possible to generate a *distribution* of protein sequences that are distribuited according to a particular probability density function. The system energy is used to accept or refuse new system configurations that populate the desired distribution. For MC, convergence is assured only for a very long simulation, and the sampling can be stuck in local minima. However, several advanced sampling techniques can be employed to avoid too long simulations and to jump free energy barriers. More technical information will be given below, with a detailed description about implementation in our software Proteus.

The main advance of Monte Carlo is that is usually simple to implement and can be easily adapted to many different problems. Sampling Boltzmann probability, is also possible to extract statistical-mechanics properties (for example free energy differences) which can be easily related to experimental data or to results of molecular dynamics simulations.

3 CPD softwares

Before introducing our in-house software Proteus, we briefly describe two CPD programs used in the lab during the last few years.

Rosetta is a collection of programs suitable for molecular modelling developed by a large community of researchers (RosettaCommons is an organization that counts 150 developers around the world). *RosettaDesign* is a utility used for CPD of protein stability; other tools like *RosettaDock* or *RosettaAbInitio* are used to predict conformations of protein-ligand complexes or de novo protein structure prediction. Several Rosetta energy functions are inspired

by molecular mechanics but make extensive use of statistical terms used to fit experimental data. The first versions were based on a fixed backbone and a backbone-dependent rotamer library, while models with limited backbone flexibility were introduced more recently. Despite its remarkable success in the protein design field, the Rosetta energies are usually expressed in unphysical, *rosetta units*. Even if some effort has been made to translate them to kcal/mol (Alford *et al.* [2017]), results are still difficult to interpret, especially when compared to those obtained with experiments or state-of-the-art molecular simulations. Moreover, for the design procedure the developers preferred fast algorithms (for example Monte Carlo with Simulated Annealing) which are able to generate a few variants in a limited computer time, rather than generating many protein variants with a single run. The user interested in generating an ensemble of sequences often needs to run several independent Rosetta simulations and then discard repeated variants.

Toulbar2 is a C++ solver for cost function networks (CFN) developed at INRA in Toulouse (Allouche *et al.* [2014]). The program uses a cost function to find the GMEC from a matrix containing interactions between couples of residues (Traoré *et al.* [2013]). These interactions, however, must be computed using an external program like *Proteus* or *Rosetta-fixbb* (fixed backbone design).

4 The Proteus CPD framework

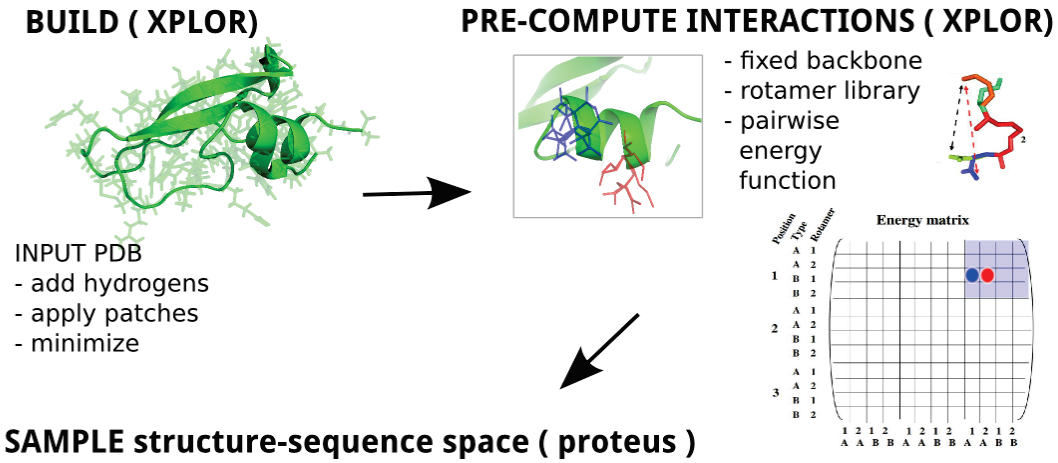


Figure 1.7 – **General Proteus CPD protocol** of 3 steps, as described in the text

Proteus (Simonson *et al.* [2013]) is a complete software distribution for Computational Protein Design. It is composed of several programs and scripts: structural data and energy calculations are managed using an in-house version of XPLOR, while the exploration of structure-sequence space is performed with a C code called *proteus* (small p) which implements several sampling

algorithms.

A typical Proteus simulation is composed of three steps: **1)** model construction, **2)** pre-calculation of residue-residue interactions using a pairwise-additive energy function and then **3)** fast exploration of structure-sequence space [fig. 1.7]. The first steps are performed starting from an input structure, given in PDB format. The initial model usually needs to be refined by adding missing hydrogens and performing few energy minimization steps. The system energy is expressed with an energy function which uses classical molecular mechanics plus an implicit solvent based on GB/SA as described later.

A given protein composed of n amino acids has sequence $S = \{t_1,t_2,...,t_n\}$ where t_i is the amino acid type at site i. In a Proteus simulation, selected residues i called *active positions* are able to mutate. They can vary their type t_i within a *mutation space* T_i which is defined a priori by the user. The only exception is for glycines and prolines, usually not mutable and modelled as part of the fixed protein backbone. Considering all active positions with their mutation space, the protein is then able to explore a *sequence space* which is defined by the ensemble of all possible protein sequences. This is a usually a very large number: as an example, for a mutation space of 20 possible types for N active positions one has 20^N possible sequences.

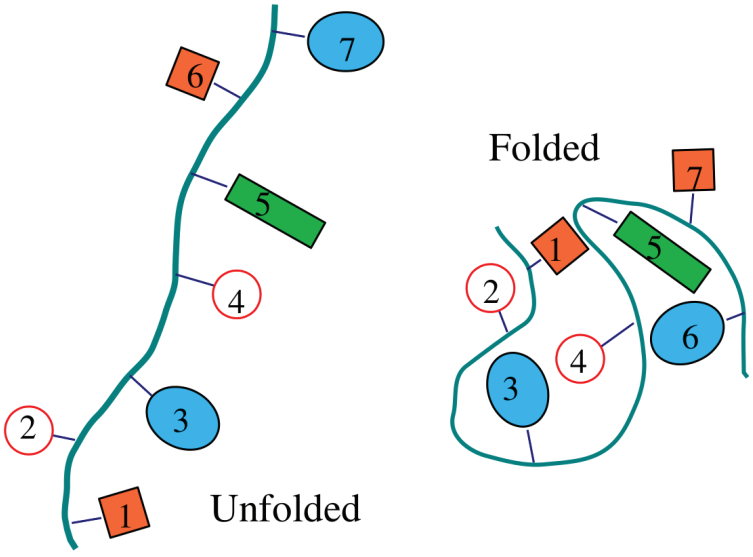


Figure 1.8 – **Unfolded and Folded states** as described in the text. The unfolded state is modeled as an extended polymer where side chains belonging to different residues do not interact with pairwise terms.

4.1 Folded and unfolded states

For each visited sequence we consider its *folded* and *unfolded* states, such that for two sequences s and s' the associated folding energy difference is

$$\Delta E = \Delta E_f(s \to s') - \Delta E_{uf}(s \to s') \tag{1.29}$$

The backbone conformation is the one of the initial model, given in input. We refer to it as *native* backbone. Side chains can assume a limited number of configurations belonging to a library of *rotamers* described by different torsional angles. By default Proteus uses the rotamer library from Tuffery (1995 or 2003), but additional rotamers can easily be included. C_β atoms of each site are fixed together with the backbone, while the rest of the side chain can vary. For a given amino acid type t_i (for example ASP or GLU shown in figure 1.9) rotamers are defined by a group of different torsional angles χ that belong to the *rotamer space* R_i. Active positions can assume rotamers of different amino acid types, while the other positions, called *inactive*, can vary their conformation within rotamers of their fixed, *native* amino acid type. Another capability of Proteus is to include ligand rotamers, specified by the user, or multiple backbone conformations for desired protein segments. These *backbone rotamers* are usually obtained from molecular dynamics simulation snapshots of the wildtype protein.

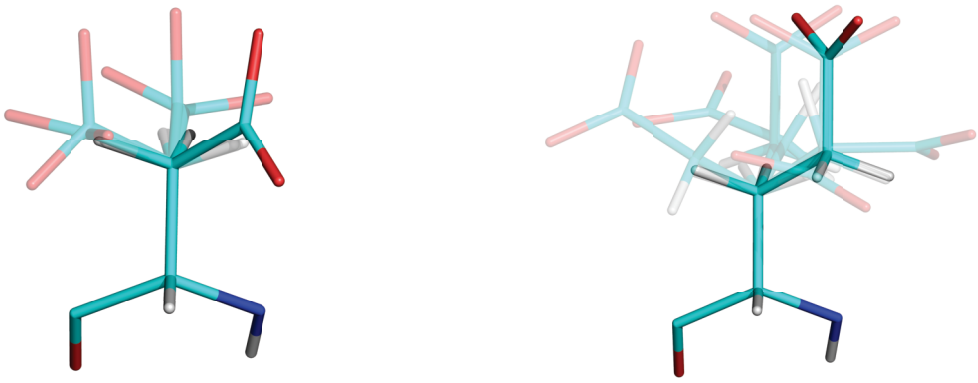


Figure 1.9 – **Selected rotamers** for side chains of ASP and GLU (left, right).

Folded state interactions The system energy in the folded state is described by a pairwise-additive function

$$E_f(s,\{r_i\}) = \sum_{i}^{n} E_{ii}(t_i,r_i) + \sum_{i<j}^{n} E_{ij}(t_i,r_i,t_j,r_j) \tag{1.30}$$

the sum runs over all the residues $i = 1, ..., n$. Diagonal terms include the interactions between atoms belonging to side chain i (of type t_i and rotamer r_i) with itself and with the fixed backbone. Off-diagonal terms contain the interactions between atoms belonging to two different side chains i and j.

Unfolded state interactions On the other hand, the *unfolded state* energy is a sum of diagonal terms:

$$E_{uf}(S) = \sum_i E_i^{uf}(t_i) \tag{1.31}$$

It corresponds to a simple physical model of non-interacting sites. Each $E_i^{uf}(t_i)$ term is called the *unfolded energy* of amino acid type t_i at position i. Unfolded energy values depend on the specific problem of interest, where they assume a different physical meaning. For example, to study protein stability, unfolded energies of different amino acid types are calculated using fully solvent-exposed amino acid model compounds. Moreover, the unfolded energy of an amino acid type can be the same for all protein positions $i = 1, ..., n$ or it can vary for different sites. This flexibility allows to define particular models for the unfolded state. For example, one can use a set of unfolded energies for exposed positions and a different set for buried positions. Using equation 1.29, this choice allows to better represent the energetic cost in transferring side chains from pure solvent to different dielectric environments.

4.2 Proteus Energy Function

Interactions are computed using molecular-mechanics and an implicit solvent

$$E = E_{MM} + E_{solv} \tag{1.32}$$

where

$$E_{MM} = E_{bond} + E_{angl} + E_{dih} + E_{impr} + E_{vdw} + E_{Coul} \tag{1.33}$$

Interactions between solute atoms are computed borrowing parameters from the Amber ff99SB force field

$$E_{vdw}(i,j) + E_{Coul}(i,j) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon_p r_{ij}} \tag{1.34}$$

where ϵ_p is the solute (protein) dielectric constant.

With the pairwise-additive energy function, interactions between pairs of residues (and between residues and the protein backbone) in presence of the implicit solvent are computed before exploring the sequence-structure space, for each couple of possible amino acid types and each couple of rotamers. Values are stored in an *energy matrix* [fig. 1.10]. This calculation can be accomplished in parallel on a small cluster or on a multi-core desktop machine. During the exploration, energy updates are calculated by picking the necessary elements from the pre-computed matrix.

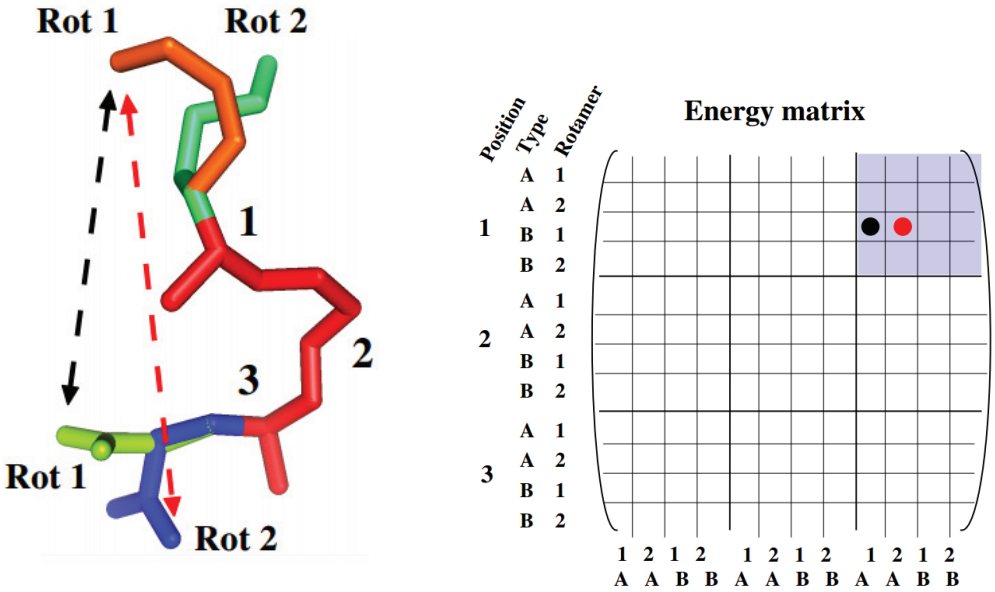


Figure 1.10 – **Proteus energy matrix calculation** as described in the text. Matrix elements contain the interaction energy between couples of *positions* i and j when occupy two rotamers Rot_i and Rot_j of two types A and B

4.3 Pairwise residue GB interaction

We consider briefly a method to compute the many-body GB interaction accurately yet efficiently, so that the NEA can be avoided. Using the Still formula to approximate the GB interaction between a pair of charges, we compute the interaction between a couple of residues R and R' as

$$\Delta E^{int}_{RR'} = \sum_{i\in R}\sum_{j\in R'} \frac{\tau q_i q_j}{\left(r^2_{ij} + B_R B_{R'} \exp\left[-r^2_{ij}/4B_R B_{R'}\right]\right)^{1/2}} \tag{1.35}$$

The right side of the equation has *residue solvation radii* B_R and $B_{R'}$ instead of the usual atomic radii b_i, b_j. Radii B_R can be calculated as a harmonic average over $b_i \in R$,

$$\Delta E^{self}_R = \sum_{i\in R} \Delta E^{self}_i = -\frac{\tau}{2}\sum_{i\in R}\frac{q_i^2}{b_i} \stackrel{def}{=} -\frac{\tau}{2}\frac{(\sum_{i\in R} q_i^2)}{B_R} \tag{1.36}$$

This formulation is called *Residue-GB*. As in atomic-GB, the interaction between two group of charges depends on the configuration of the whole solute through the *residue solvation radii* which are average of the atomic radii $b_i = b_i(r_1,...,r_N)$.

Even if residue-GB is not pairwise in the interaction term, an alternative functional form was proposed by Archontis and Simonson [2005] and can be used to extrapolate the pairwise "information" about a couple of residues. Indeed, equation 1.35 can be fitted for a large range

of B values using the formula

$$\Delta E^{int}_{RR'}(\{r_i,q_i\}; B = B_R B_{R'}) \simeq c_1(\{r_i,q_i\}) + c_2(\{r_i,q_i\})B + \\ + c_3(\{r_i,q_i\})B^2 + c_4(\{r_i,q_i\})B^{-1/2} + c_5(\{r_i,q_i\})B^{-3/2}$$

For all pairs of residues in all posible rotamers whose coordinates and charges $\{r_i,q_i\}$ are known, 5 coefficients $c_i(\{r_i,q_i\})$ can be extracted and then inserted in the matrix. Then ΔE^{int} can be directly computed during the MC simulation, keeping updated residues solvation radii from pairwise self-energies. More information can be found in (Villa *et al.* [2017]).

4.4 Monte Carlo exploration

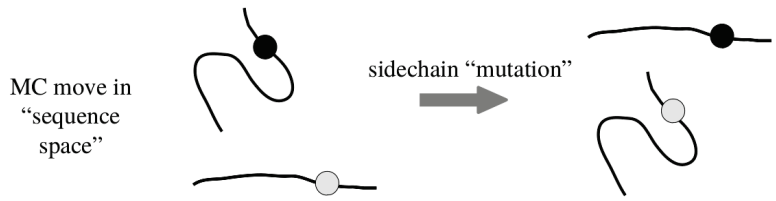


Figure 1.11 – **Amino acid mutation in MC**. Two different types are represented using black and white circles. For each mutation in the folded state, the inverse mutation is performed in the unfolded protein, as described in the text.

Without knowing the system partition function, we can compute *a priori* the relative probability to visit points in the configurational and sequence space. We generate a Markov chain of states using the well-known Metropolis acceptance method (Metropolis *et al.* [1953]). In principle, the algorithm allows to sample any kind of probability density function. In our case, we look for states that are populated according to a Boltzmann distribution. Starting from an initial state (defined by a protein sequence and the set of occupied rotamers) we perform N iterations to generate N states of the system. At each iteration we perform *trial moves* to change conformation. These moves are accepted or rejected according to *acceptance rules* that exploit the *detailed balance* property of long, well-behaved Markov chains.

One possible Proteus *trial move* is a *mutation*. At a given simulation timestep, we modify the side chain at a random amino acid position i in the folded protein, mutating it to a random amino acid type $t_i \rightarrow t_i^{'}$ picked from mutation space T_i. A random rotamer is then selected from the rotamer space R_i of the extracted type. At the same time, the *reverse mutation* is performed in an unfolded copy of the system, $t_i^{'} \rightarrow t_i$. The total energy change for the mutation

which counts both folded and unfolded copies is calculated using equation 1.29. Another trial move changes the rotamer of a randomly picked residue $r_i \rightarrow r_i^{'}$ without modification of its amino acid type.

Mutation and rotamer moves can be combined together to obtain a *trial move* that involves multiple positions in a single attempted transition. Other possible moves are the so-called *mut-mut* or *rot-rot* moves where we change two side chains at two different positions. To increase the efficiency of these moves, we define lists of neighbours that are pre-computed before the simulation looking at pairwise interactions between residues. If two residues interact with an energy that is lower than a threshold value, they cannot be selected together in one of these "double" moves.

With the transitions and energy changes defined above, the desider Boltzmann distribution of states is determined by accepting or rejecting moves. Since moves are selected randomly they have an associaged "generation" probability $\alpha(o \rightarrow n)$ where o and n indicates the old (current) and the new, generated states. We call $acc(o \rightarrow n)$ the corresponding move *acceptance probability*. The overall probability associated with the move is the product $\pi = \alpha(o \rightarrow n)acc(o \rightarrow n)$. At thermal equilibrium, the average number of leaving moves from a state o must be equal to the number of moves going from all other states n to the state o. In other words, the flux of configurations in one direction is canceled by the flux in the opposite direction. This is the detailed balance condition, which holds in the limit of a very long simulation. On average, the probability of finding the system in the two configurations corresponds to the equilibrium populations $N(o)$ and $N(n)$ which obey Boltzmann distributions. The *detailed balance* reads

$$N(o)\pi(o \rightarrow n) = N(n)\pi(n \rightarrow o) \tag{1.37}$$

so

$$\frac{N(n)}{N(o)} = \frac{\pi(o \rightarrow n)}{\pi(n \rightarrow o)} = \frac{\alpha(o \rightarrow n)acc(o \rightarrow n)}{\alpha(n \rightarrow o)acc(n \rightarrow o)} \tag{1.38}$$

Since

$$\frac{N(n)}{N(o)} = \exp\left[-\beta\Delta E(o \rightarrow n)\right] \tag{1.39}$$

the acceptance probability must obey

$$\frac{acc(o \rightarrow n)}{acc(n \rightarrow o)} = \exp\left[-\beta\Delta E(o \rightarrow n)\right] \frac{\alpha(n \rightarrow o)}{\alpha(o \rightarrow n)} \tag{1.40}$$

This is verified for the Metropolis scheme

$$acc(o \rightarrow n) = \exp\left[-\beta\Delta E(o \rightarrow n)\right] \frac{\alpha(n \rightarrow o)}{\alpha(o \rightarrow n)} \qquad if \quad \Delta E(o \rightarrow n) > 0; \quad 1 \quad otherwise \tag{1.41}$$

5 Constant-activity and constant-pH Monte Carlo

The main goal of CPD is to produce one or more protein variants with an associated *sequence score* in order to optimize a desired property. For example, one can use a scoring function based on the protein unfolding free energy in order to study protein stability. In general, one is not directly interested in the *quantitative* estimation of free energy differences but in producing a relatively small set of protein variants, which will be eventually studied with more sophisticated methods. For this reason, several scoring functions do not directly represent a physical quantity and are expressed in arbitrary units. However, the CPD model along with the Monte Carlo sampling described above can be used to estimate equilibrium thermodynamic quantities like binding free energy differences or protonation probabilities at constant pH. Proteus is particularly suitable for this kind of calculations: the fact that it is based on well-established physical models allows to define sampling methods which target different equilibrium properties.

Constant-activity Monte Carlo and ligand titration A special case is the design of ligand binding. We consider the simple case of a protein of fixed sequence S wich can bind two different ligands, represented by one mutating site of mutation space L and $L^{'}$.

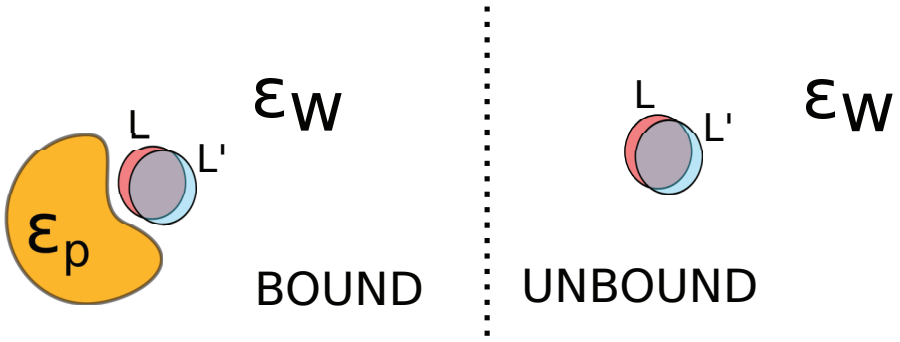


Figure 1.12 – **Bound and unbound states** for two ligands L and $L^{'}$

Following the Proteus design protocol, an energy matrix is computed from a structure representing the *bound state*. Interactions between the ligand of type L or $L^{'}$ with protein residues are pre-calculated for all possible couples of protein rotamers. In this special case, the bound state is represented using the *folded state* energy function [eqn. 1.30] while the unbound state is represented using the unfolded state energy function [eqn. 1.31]. The ligand unfolded energies (in this case we call them unbound energies) have a logarithmic dependence on the concentration

$$E^{ub}(X) = kT \ln [X] + e^{ub}(X) \tag{1.42}$$

where X is equal to L or $L^{'}$ and kT is the thermal energy. The second term $e^{ub}(X)$ is the energy of the unbound ligand calculated ahead of time using the molecular mechanics energy function,

including the implicit solvent. During Monte Carlo, mutations are performed swapping the ligand type between L and L'.

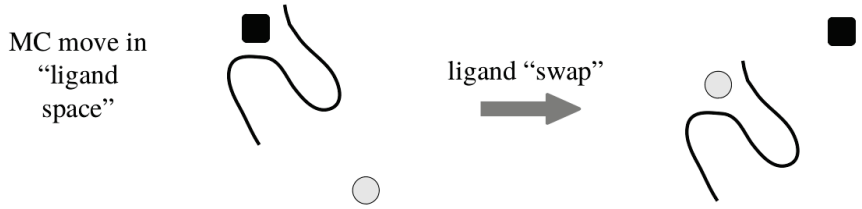


Figure 1.13 – **Ligand mutation** where two different ligands are represented with black and white circles. For a ligand mutation in the bound state, the inverse mutation is performed in the unbound state.

For two ligand states X and Y, moves are accepted according to the energy difference $\Delta E^{bound}(X \to Y) - \Delta E^{ub}(X \to Y)$ where

$$\Delta E^{ub}(X \to Y) = kT \ln\left(\frac{[Y]}{[X]}\right) + e^{ub}(Y) - e^{ub}(X) \tag{1.43}$$

Choosing $[X]$, $[Y]$ ahead of time, we then perform a "constant activity" MC. Constant activity MC can be used to estimate binding free energy differences $\Delta\Delta G^0(X \to Y)$. This method has been recently applied to compare the binding affinity between tyrosyl-tRNA synthetase (TyrRS) and several substrates (Druart *et al.* [2015]). It consists in performing a series of simulations, each one at a different $\Delta E^{ref}(X \to Y)$. From equation 1.43, this corresponds to varying the ratio of ligand concentrations $[Y]\,/\,[X]$. From each simulation, one extracts the MC populations of X and Y, then standard binding free energy difference is computed from the titration midpoint [fig. 1.14] where X and Y are equally populated

$$\Delta\Delta G^0_{bind}(X \to Y) = kT \ln\left(\frac{[Y]}{[X]}\right)_{mid} = \Delta E^{ref}_{mid} - \Delta e^{ub}(X \to Y) \tag{1.44}$$

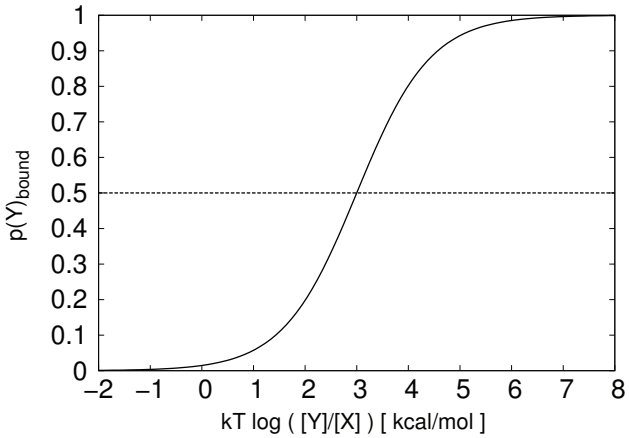


Figure 1.14 – **Titration of two ligands** by variation of their relative concentrations. The binding free energy is calculated extrapolating the midpoint of the sigmoidal function (dashed line).

Constant-pH Monte Carlo As in constant-activity MC, the titration protocol can be applied to simulate *protonation-deprotonation* events at constant pH (Polydorides & Simonson [2013]). This allows to calculate pK_a values of side chains of proteins, a quantity of great interest as it is connected to their stability, binding and many other biochemical properties. Several precise and well-established methods to obtain protein pK_as are based on explicit and implicit solvent free energy calculations (Simonson *et al.* [2004]). However, CPD allows to screen larger proteins in a limited amount of time.

Following the Proteus protocol, titrable protein amino acid side chains are able to mutate in their mutation space. An energy matrix is computed, where pairwise interactions are obtained using protonated and deprotonated rotamers at all titrable positions. We indicate their generic deprotonated and protonated forms X and X^+. For each couple of types X and X^+, their relative population in solution (model compounds) can be controlled using the unfolded energy difference

$$\Delta E^{uf}(X \to X^+) = 2.303kT(pH - pK_a^{model}) + e^{ub}(X^+) - e^{ub}(X) \tag{1.45}$$

where pK_a^{model} is a tabulated quantity, known from experiments. The standard protonation free energy ΔG^0 depends on the dissociation constant at equilibrium through $\Delta G^0 = kTlnK_a = -2.303kTpK_a$. In equation 1.45, the difference $pH - pK_a$ is used to compute the protonation free energy at a different pH value. The pK_a shift ΔpK_a is then obtained from the ratio of protonated/deprotonated populations in the protein extracted from a MC simulation:

$$\Delta\Delta G^0 = \Delta G^0_{protein} - \Delta G^0_{model} = -2.303kT\Delta pK_a \tag{1.46}$$

Chapter 2

High throughput design of protein-ligand binding

The computational design of ligand binding is the search for preferred sequences using a scoring function based on the binding free energy. However, most of CPD methods usually identify few candidates, looking for low energy states close to the global minimum energy configuration (GMEC). They often use empirical energy functions, without giving binding free energies to be compared with experimental affinities. As a consequence, most of CPD programs cannot be used for the more challenging high-throughput design, where one wants to optimize the binding between a receptor and many ligand variants in a short cpu time, obtaining formally exact binding free energies.

Here we present a general high-throughput design method that allows to overcome these limitations. It was implemented in our in-house program Proteus, but can be easily adapted to other CPD programs. Our protocol adaptively flattens the energy landscape in sequence space, and then extracts formally exact binding free energy differences from *biased* Monte Carlo (MC) simulations. Our test system was the Tiam1 PDZ domain (Tiam1), which binds to its natural peptide ligand Syndecan1 (Sdc1) composed of 8 amino acids. We used the method to design peptides that bind to Tiam1 and could serve as inhibitors of its activity. We targeted four of the last five C-terminal amino acids of Sdc1. Considering 18 possible amino acid types at the four "active" positions (we allow all amino types except Gly and Pro), there are 18^4=104976 competiting ligands. Our protocol is efficient, since we obtained binding free energies for almost 75000 peptide variants using just one CPU hour. Method, simulations and results are presented in the attached paper (Villa *et al.* [2018b]).

Selected peptide variants were finally studied using a semi-empirical PB/LIE method, previously optimized to predict binding specificity of the Tiam1 PDZ domain. It was recently developed in our laboratory (Panel *et al.* [2017]). Details about PB/LIE calculations and results are given below, after the attached paper. At the end of this chapter, we also describe some theoretical aspects of the new design protocol, and how biased MC sampling techniques can be used to design sequences either by binding affinity or by specificity.

THE JOURNAL OF CHEMICAL PHYSICS **149**, 072302 (2018)

Adaptive landscape flattening in amino acid sequence space for the computational design of protein:peptide binding

Francesco Villa, Nicolas Panel, Xingyu Chen, and Thomas Simonson[a)]
Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France

(Received 12 January 2018; accepted 16 March 2018; published online 4 May 2018)

For the high throughput design of protein:peptide binding, one must explore a vast space of amino acid sequences in search of low binding free energies. This complex problem is usually addressed with either simple heuristic scoring or expensive sequence enumeration schemes. Far more efficient than enumeration is a recent Monte Carlo approach that adaptively flattens the energy landscape in sequence space of the unbound peptide and provides formally exact binding free energy differences. The method allows the binding free energy to be used directly as the design criterion. We propose several improvements that allow still more efficient sampling and can address larger design problems. They include the use of Replica Exchange Monte Carlo and landscape flattening for both the unbound and bound peptides. We used the method to design peptides that bind to the PDZ domain of the Tiam1 signaling protein and could serve as inhibitors of its activity. Four peptide positions were allowed to mutate freely. Almost 75 000 peptide variants were processed in two simulations of 10^9 steps each that used 1 CPU hour on a desktop machine. 96% of the theoretical sequence space was sampled. The relative binding free energies agreed qualitatively with values from experiment. The sampled sequences agreed qualitatively with an experimental library of Tiam1-binding peptides. The main assumption limiting accuracy is the fixed backbone approximation, which could be alleviated in future work by using increased computational resources and multi-backbone designs. *Published by AIP Publishing.* https://doi.org/10.1063/1.5022249

I. INTRODUCTION

Computational protein design (CPD) is an emerging technique for engineering and understanding protein structure and function.[1–4] CPD explores a large space of amino acid sequences and conformations to identify protein variants that have predefined properties, such as ligand binding. Conformational space is usually defined by a discrete library of side chain rotamers and a small set of allowed backbone conformations. The energy function usually combines physical and empirical terms;[5–7] the solvent and the unfolded state of the protein are described implicitly. The space of sequences and conformations is often explored with a simulated annealing Monte Carlo (MC) approach.[1,8,9] Other approaches search for the lowest energy state, the "Global Minimum Energy Conformation" or GMEC, using combinatorial optimization.[10–14]

To design ligand binding, the quantity to optimize is the binding free energy, which is not a property of the GMEC but a Boltzmann average over many states. This is a complex problem, and most CPD applications have simply tried to identify one or a few low energy states. Methods have been developed to systematically enumerate low energy states within a certain interval above the GMEC and to compute the configuration integral of the molecule(s) up to a predefined accuracy. They are referred to as "partition function" methods. Enumeration methods are applicable to problems with limited numbers of degrees of freedom. They exploit the discrete nature of sequence and rotamer space and use efficient methods from computer science to explore them. Thus, cost function network and dead end elimination approaches have been developed that provide deterministic guarantees on configuration integral accuracy.[15–17] The binding free energy is then obtained as the ratio of bound and unbound configuration integrals. Related methods have been developed to enumerate secondary structures of simplified RNA molecules[18,19] and transmembrane proteins.[20]

Of course, partition functions are not needed to obtain binding free energies. Molecular dynamics (MD) and Monte Carlo (MC) simulation methods have been developed for decades that focus directly on partition function ratios, instead. One explores states that are "intermediate" between the bound and unbound states, through importance sampling, using a bias energy to increase their statistical weights.[21–23] To compute binding free energy *differences* between two ligands, one alchemically transforms one ligand into the other.[24–26] These methods carefully avoid calculating partition functions yet provide excellent precision when comparing pairs of ligands if enough sampling is performed. They can be accelerated by employing an implicit model of the solvent.[27]

MD and MC simulations have also been applied to more complex problems. For the competitive binding of many ligands to the same protein,[28–30] an adaptive, "lambda dynamics" method was proposed. The ligand chemical potentials are

[a)]Electronic mail: thomas.simonson@polytechnique.fr

adaptively adjusted during MD until all possible complexes are similarly populated. Related adaptive MD methods have been used to sample conformational basins of proteins or peptides. An energy penalty is added to basins that have already been explored, leading to a gradual flattening of the conformational free energy surface. Metadynamics,[31–34] Wang-Landau methods,[35] and multicanonical simulations[36–38] are three main examples.

In protein design, the space to explore is especially complex since it includes amino acid sequence variations at multiple positions in a protein or peptide. Thus, adaptive simulations are an attractive perspective. A Wang-Landau MC approach was applied to RNA folding[39] and to protein backbone conformations.[40] A simpler, constant-activity MC protocol was used to study the competition between pairs of ligands binding to the tyrosyl-tRNA synthetase enzyme;[41,42] the relative binding free energies obtained for four small ligands were in qualitative agreement with experiment. Finally, an adaptive, Wang-Landau MC approach was applied to the binding of peptides to several PDZ protein domains.[43] A simulation of the unbound peptide was performed with several positions allowed to mutate freely; a bias potential was adaptively constructed so that all available sequences were equally populated. The same bias potential was then employed in a simulation of the protein:peptide complex. In this simulation, sequences are then populated according to their relative binding free energies, an elegant solution to the free energy design problem.

Here, we propose three improvements to the adaptive Wang-Landau MC approach[43] and apply it to a PDZ:peptide system of biological interest. The first improvement is the use of Replica Exchange MC (REMC) for improved sampling.[44,45] This allows us to explore a much larger mutation space than previously. Below, we simulate four mutating positions and over 10^5 competing ligand sequences using one CPU hour. The second improvement is the use of separate adaptive simulations and bias potentials for the bound and unbound states. By adaptively flattening sequence space for the protein:peptide complex, we make the method more robust and allow improved sampling of weak-binding peptides. Third, we have implemented the method within the established, Proteus CPD software.[41,46] This gives us access to improved, Generalized Born (GB) implicit solvent models,[47–49] efficient exploration using a precalculated energy matrix, a user-friendly interface, and additional Proteus features not applied here, such as multi-backbone design.[40]

We consider the PDZ domain of the Tiam1 protein. The peptide ligand corresponds to the eight C-terminal residues of the Syndecan-1 protein. Tiam1 interacts functionally with Syndecan-1 by binding to its C-terminus through its own PDZ domain, helping to regulate downstream signaling pathways in several cell types.[50,51] From now on, we refer to the peptide as Sdc1 and to the PDZ domain as Tiam1. Our goal is to design peptides analogous to Sdc1 that can bind and inhibit Tiam1 either *in vitro* or in a cellular context. Such peptides could modulate Tiam1 function and have applications as reagents in cell biology or as inhibitors to downregulate Tiam1 in tumor cells. We redesign the Sdc1 peptide by allowing four of its last five residues to mutate (out of eight). This corresponds to a pool of $18^4 = 104\,976$ competing ligands (we allow all amino types except Gly and Pro, 18 in all). The protein and peptide backbone are held fixed, while side chains mutate and/or explore rotamers. The fixed backbone allows precalculation of an energy matrix, which makes the MC exploration very efficient.[41,52] The ligand pool is simulated twice, in the bound and unbound states, respectively. The ligand populations are gradually equalized by an adaptive Wang-Landau bias potential. They can then be reranked based on their binding free energies. Thus, our protocol directly designs peptides for their binding free energies.

Computed binding free energies were obtained for almost 75 000 peptide variants using one simulation of the peptide and one of the PDZ:peptide complex, each included one billion MC steps per REMC replica and required 1/2 h of CPU time on a desktop computer. The free energies were in fair agreement with experiment for seven variants, with a mean unsigned error of 0.8 kcal/mol (excluding one outlier). A sequence logo based on the MC sequences was qualitatively similar to the one based on a combinatorial library of peptide sequences selected experimentally for Tiam1 binding.[53] In future applications, longer simulations would allow larger numbers of mutating positions to be explored and more accurate variants of the Proteus solvent model to be used. We expect that using multiple conformations for the protein and/or peptide backbone, with the help of a hybrid MC method,[40] would also give improved accuracy. The ability to design sequences directly for binding affinity, without the need for expensive partition function calculations, should facilitate a number of biotechnology applications.

II. COMPUTATIONAL METHODS

A. MC exploration of side chain rotamers and types

We consider a polypeptide of L amino acids. Below, it will be either a folded protein:peptide complex or an unbound, extended peptide. We assume that each amino acid i can adopt a few different types s, s', ..., leading to different possible sequences S. The polypeptide backbone has a fixed geometry. The side chains can each explore a few discrete conformations r, r', ... called rotamers (around 10 per side chain type s). The energy of any particular conformation depends on the sequence and the particular set of rotamers. We perform a Monte Carlo exploration, whose goal is to generate a Markov chain of states,[55–57] such that the states are populated according to a Boltzmann distribution. The simulation system explicitly includes one copy of the polypeptide, whose sequence and rotamers can fluctuate. One possible MC move is to change a rotamer r_i at one particular position; the energy change is $\Delta E_{on} = E(\ldots s_i, r'_i \ldots) - E(\ldots s_i, r_i \ldots)$, where the subscripts o and n refer to the old and new rotamer states. Another possible elementary move is a mutation: we modify the side chain type $s_i \to s'_i$ at a chosen position i in the folded protein, assigning a particular rotamer r'_i to the new side chain. Let $\alpha(o \to n)$ be the probability to select a move between two states o and n; let $acc(o \to n)$ be the probability to accept it. The overall probability of the move is $\pi(o \to n) = \alpha(o \to n)\, acc(o \to n)$. The Metropolis-Hastings scheme[55–57] chooses the acceptance

probability as

$$acc(o \to n) = \min(1, \exp(-\beta \Delta E_{on})) \frac{\alpha(o \to n)}{\alpha(n \to o)}, \quad (1)$$

where β is the usual inverse temperature. This, along with the assumption of detailed balance[45,56,57] leads to Boltzmann statistics,

$$\frac{\mathcal{N}(n)}{\mathcal{N}(o)} = e^{-\beta \Delta E_{on}}, \quad (2)$$

where $\mathcal{N}(o)$ is the equilibrium population of state o (respectively, n).

B. Adaptive bias energy

The adaptive simulation method is closely analogous to the Wang-Landau and metadynamics approaches.[32,35] For each position i among a selection of positions to be designed, we perform a long MC simulation where only i can mutate, and we gradually increase a bias potential until all the side chain types at i have roughly equal populations. The bias potential, $E_i^B(r_i, t)$, thus depends on time t and side chain type s_i. In practice, an increment $e_i^b(s_i; t)$ is added at a particular time t; then, we run the simulation for an interval T before adding a new bias increment. At the end of each interval T, the bias increment is only added to one side chain type s: the one populated at the end of the interval. The magnitude of the bias increment depends exponentially on the current bias value, which leads to a gradually decreasing increment. Specifically, the bias potential is defined as follows:

$$E_i^b(s; t) = \sum_{n; nT < t} e_i^b(s_i(nT); nT) \delta_{s_i(nT), s}, \quad (3)$$

$$e_i^b(s(t); t) = e_0 \exp\left(-E_i^b(s(t); t)/E^0\right), \quad (4)$$

where e_0 and E^0 are constant energies and $\delta_{s,s'}$ is the Kronecker delta. The above bias increment scheme is adapted from the well-tempered metadynamics methodology.[32–34] Once the single-position bias potentials E_i^b have been optimized separately, we are ready to perform simulations where all the selected positions can mutate simultaneously. We number these positions arbitrarily from 1 to p. We apply a bias that is the sum of the single-position biases,

$$E^B(s_1, s_2, \ldots, s_p; t) = \sum_{i=1}^{p} E_i^b(s_i; t). \quad (5)$$

C. Relative binding free energy

The protein:peptide complex and the unbound peptide are simulated separately, each with its own bias potential. In the biased simulations, the populations of a particular sequence S in the bound and unbound states are denoted $p_b(S)$ and $p_{ub}(S)$. The binding free energy relative to a reference sequence S_r, in the presence of the biases, has the form

$$\Delta\Delta G_{\text{biased}}(S) = -kT \ln \frac{p_b(S)}{p_b(S_r)} + kT \ln \frac{p_{ub}(S)}{p_{ub}(S_r)}. \quad (6)$$

The relative binding free energy in the absence of bias has the form

$$\Delta\Delta G(S) = \Delta\Delta G_{\text{biased}}(S) - \left(E_b^B(S) - E_{ub}^B(S) - E_b^B(S_r) + E_{ub}^B(S_r)\right). \quad (7)$$

Subscripts b and ub indicate the bias energy in the bound or unbound state, respectively. In practice, Eq. (7) is applied to sequences that have been sufficiently sampled in both the bound and unbound states.

D. Protein design model and test system

The PDZ:Sdc1 complex is modeled using an experimental X-ray structure [Protein Data Bank (PDB) code 4GVD[58]]. The backbone geometries of the protein and peptide are held fixed. Side chains explore a discrete rotamer library.[41,59] Selected peptide positions can mutate freely. The unbound peptide is modeled as a single, extended conformation, with a fixed backbone geometry. A molecular mechanics energy function is used along with a Generalized Born (GB) implicit solvent model. In principle, GB is a many-body model, such that the interaction energy between two amino acids I, J depends on the positions of the other amino acids. Here, a "Native Environment Approximation (NEA)" is made such that the $I - J$ interaction is computed assuming the rest of the system occupies its native sequence and structure.[41,47] With this approximation, all inter-residue interactions can be computed ahead of time and stored in an energy matrix. Later, they are used for sequence exploration. More details are given below.

We considered the X-ray structure of the Tiam1 PDZ domain bound to an octapeptide derived from the C-terminus of the Syndecan-1 (Sdc1) target protein (PDB entry 4GVD). We refer to the peptide as Sdc1. Amino acids in the peptide are numbered backwards from the C-terminus, as usual for PDZ ligands. The positions allowed to mutate here are positions −4 to −1. Position 0 is the C-terminal position. The N-terminus is acetylated and the C-terminus carboxylated. The peptide wildtype (WT) sequence is TKQEEFYA. Protonation states of histidines in Tiam1 were assigned to be neutral, based on visual inspection of hydrogen-bonding patterns in the 3D structure.

E. Effective energy function and energy matrix for MC

The energy calculations are done with a modified version of the Xplor program,[41,60] using the following effective energy function:

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihe}} + E_{\text{impr}} + E_{\text{vdw}} + E_{\text{Coul}} + E_{\text{solv}}. \quad (8)$$

The first six terms in (8) are taken from the Amber ff99SB molecular mechanics energy function.[61] The last term on the right, E_{solv}, represents the contribution of a solvent. We use a "Generalized Born + Surface Area," or GBSA implicit solvent model,[62,63]

$$\begin{aligned} E_{\text{solv}} &= E_{\text{GB}} + E_{\text{surf}} \\ &= \frac{1}{2}\left(\frac{1}{\epsilon_W} - \frac{1}{\epsilon_P}\right) \sum_{ij} q_i q_j \left(r_{ij}^2 + a_i a_j e^{-r_{ij}^2/4 a_i a_j}\right)^{-1/2} \\ &\quad + \sum_i \sigma_i A_i. \end{aligned} \quad (9)$$

Here, $\epsilon_W = 80$ and $\epsilon_P = 4$ are the solvent and protein dielectric constants; r_{ij} is the distance between atoms i, j and a_i is the "solvation radius" of atom i.[62,63] A_i is the exposed solvent accessible surface area of atom i; σ_i is a parameter

that reflects each atom's preference to be exposed or hidden from the solvent. The solute atoms are divided into 4 groups with the following σ_i values [cal/(mol/Å^2)]: unpolar (−5), aromatic (−12), polar (−8), and ionic (−9). Hydrogen atoms have a surface coefficient of zero. In the GB term, the solvation radius a_i is a function of the coordinates of all the protein atoms.[62,63] We use a "Native Environment Approximation," where the solvation radii a_i of a particular group are computed with the rest of the system having its native sequence and conformation.[41,47]

The first stage in our CPD procedure[41,52,64] is to calculate the interaction energies between all side chain pairs and between side chains and the backbone, considering all allowed side chain types and the discrete rotamer library of Tuffery *et al.* (enhanced by allowing several orientations for SH and OH groups).[59,65] To alleviate the rotamer approximation, each side chain pair is initially positioned using two library rotamers; then, 15 steps of conjugate-gradient energy minimization are performed, where only the pair of interest can move and the energy is limited to the interaction within the pair, plus the pair's interaction with the backbone and solvent.[41,64] The final interaction energy is stored in an "energy matrix" for later use.[52] Thanks to the NEA, the GB interaction energy only depends on the types and rotamers of the pair of interest.

F. Replica exchange MC protocol

The MC moves included side chain mutations and rotamer changes at one or two positions. For two-position moves, the first position was chosen randomly and the second within its neighborhood, defined by an interaction energy threshold of 3 kcal/mol. Simulations were done with Replica Exchange MC,[45] with 4 replicas at thermal energies of kT = 0.6, 0.9,1.3, and 1.8 kcal/mol. A conformation swap between a random pair of replicas (at neighboring temperatures) was attempted every 2000 steps. In the adaptive MC calculations, the bias update period was T = 1000 steps. The bias increment amplitude was e_0 = 0.2 kcal/mol. The scaling energy E_0 was 29.4 kcal/mol. The simulations used to determine the single-position biases, E_i^b, lasted 10^8 steps (per replica). Once the biases were optimized, a production simulation was run with the total bias E^B for 10^9 steps (per replica), both for the unbound and bound peptide. The positions allowed to mutate were positions −4 to −1 of the peptide (position 0 being the C-terminal position). Each position could adopt any side chain type except Gly or Pro, for a total of 18 possible types per position and 18^4 = 104 976 possible sequences overall. During the production simulation of the bound state, 101 159 sequences were sampled, and 74 963 (71%) were visited at least 1000 times during the PDZ:peptide simulation. For these, the relative binding free energies were estimated. The production simulations lasted about 1/2 h each on a desktop computer, thanks to the precomputed energy matrix and a shared memory, OpenMP parallelization.

III. RESULTS

The Sdc1 peptide was simulated in the unbound state, with its backbone fixed in the extended, β sheet geometry it has in the PDZ:peptide complex (Fig. 1). An adaptive bias potential E_i^b was constructed separately for each position i = −4, . . ., −1, such that when position i was the only position allowed to mutate, the distribution of side chain types was approximately flat. An overall bias potential was then defined as $E^B(s_,s_3, s_2, s_1) = \sum_i E_i^b(s_i)$. The peptide was simulated for one billion steps with positions −4, . . ., −1 allowed to mutate, in the presence of the overall bias E^B, using REMC with four replicas (see Sec. II). The C-terminal position 0 was not allowed to mutate, but kept its wildtype Ala side chain. All of the 18^4 = 104 976 possible sequences were sampled during the simulation. Table I reports the mean energy differences between each side chain type and the Ala type, taken as a reference, with the bias energy contribution removed. The values are averaged over the simulation. Values for positions −4, . . ., −1 are similar. These energies can be thought of as mutation energies for an extended, unfolded peptide chain. They contain contributions from intra-side chain interactions, side chain:backbone and side chain-side chain interactions, and side chain-solvent interactions (estimated with the GBSA implicit solvent model). Some of the energy differences are large and the corresponding types would not be sampled without using a bias potential E^B to flatten the energy landscape.

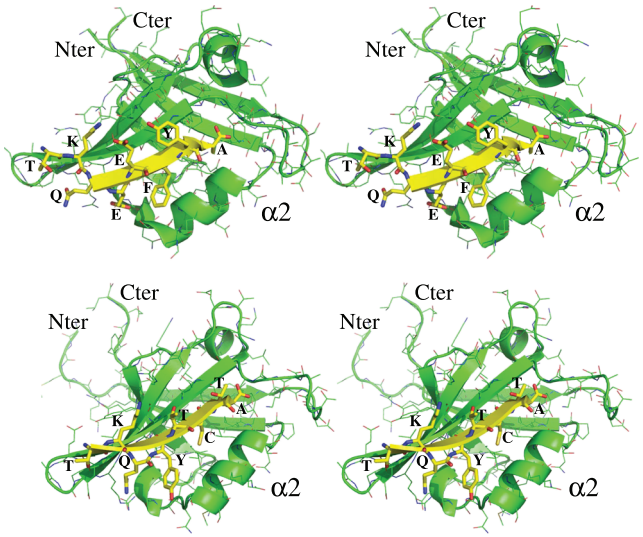


FIG. 1. The Tiam1-PDZ:Sdc1-peptide complex (cross-eyed stereo). Top: Wildtype complex, X-ray structure. Bottom: Complex with the TKQYTCTA peptide, which has the strongest computed binding affinity, representative MC structure. The peptide is yellow; its residues are labeled with their type.

A bias potential for the bound state was estimated in the same way, one position at a time (i = −4, . . ., −1). The complex was then simulated for one billion REMC steps with positions −4, . . ., −1 allowed to mutate, in the presence of the bias. 101 159 out of 104 976 possible sequences were sampled during the simulation; 74 963 (71%) were visited at least 1000 times. For these 74 963, we estimated the binding free energy difference $\Delta\Delta G$, relative to Sdc1, by subtracting the mutation free energies in the bound and unbound states. The mutation free energies were estimated from the probability ratios between each sequence and the wildtype Sdc1 (see Sec. II). Populations were taken from the room temperature

TABLE I. Mutation energies for the unbound peptide (kcal/mol). Values for each amino acid type and peptide position, with Ala as the reference, averaged over the unbound peptide simulation. "Mean" indicates an average over positions −4, . . ., −1.

	⟨ΔE⟩				
Type	Mean	P_{-4}	P_{-3}	P_{-2}	P_{-1}
ALA	0.0	0.0	0.0	0.0	0.0
ARG	−46.9	−47.3	−48.1	−48.7	−48.6
ASN	−15.8	−16.1	−16.9	−16.2	−16.3
ASP	−18.2	−18.5	−19.1	−17.5	−17.8
CYS	−0.4	0.1	−0.7	−0.4	−0.5
GLN	−12.6	−12.5	−14.1	−12.7	−12.8
GLU	−15.7	−15.1	−16.3	−14.3	−14.7
HIS	14.8	15.2	15.4	14.5	14.1
ILE	2.9	1.8	1.3	2.2	2.2
LEU	−3.9	−4.2	−4.8	−4.7	−4.7
LYS	−5.5	−6.0	−6.7	−7.3	−7.1
MET	−1.9	−2.2	−2.9	−2.4	−2.4
PHE	−1.1	−1.3	−2.2	−0.9	−1.2
SER	−4.4	−4.4	−4.7	−4.3	−4.5
THR	−7.1	−6.8	−7.9	−7.1	−7.4
TRP	−0.7	−2.2	−1.9	−1.2	−1.6
TYR	−6.8	−7.1	−8.4	−6.7	−7.2
VAL	−3.1	−3.8	−4.5	−3.5	−3.9

replica. Some of the sequences led to very unfavorable binding free energies, $\Delta\Delta G > 4$ kcal/mol, because of steric clashes between a mutated peptide side chain and the PDZ protein. Indeed, the protein and peptide backbones were held fixed in their X-ray conformation during the MC simulation and were unable to shift and make room for some of the larger side chain types. These large values can therefore be considered artefacts of the fixed backbone simulation. Figure 2 reports the $\Delta\Delta G$ values for the 50 000 sequences with the strongest binding, which span the range from −1.5 to 4 kcal/mol. For six sequences, reference values are available from experiment or, in two cases, from high-level, molecular dynamics (MD) free energy simulations.[66] The MD free energy simulations used explicit solvent and were found earlier to give good accuracy, within 1 kcal/mol of experiment. The reference values are also

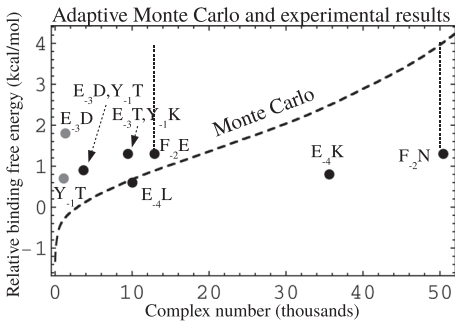


FIG. 2. Relative binding free energies $\Delta\Delta G$ from adaptive Monte Carlo (dashed line), experiment (black dots), and MD free energy simulations (gray dots). The MC values correspond to 50 000 sampled sequences, numbered by increasing $\Delta\Delta G$ values. For two sequences, only an experimental lower bound is known (interval indicated by vertical dashed lines). Labels indicate the mutations that define the experimental sequences (relative to the wildtype Sdc1).

TABLE II. Experimental and computed binding free energies for PDZ:peptide complexes (kcal/mol), with wildtype Sdc1 (WT) as the reference.

Peptide	Mutations	Expt.	MC	|Error|
TKQEEFYA	WT	0.0	0.0	. . .
TKQEEFTA	$Y_{-1}T$	0.7[a]	−0.3	1.0
TKQEDFYA	$E_{-3}D$	1.8[a]	−0.3	2.1
TKQEDFTA	$E_{-3}DY_{-1}T$	0.9	0.1	0.8
TKQETFKA	$E_{-3}TY_{-1}K$	1.3	0.5	0.8
TKQLEFYA	$E_{-4}L$	0.6	0.6	0.0
TKQKEFYA	$E_{-4}K$	0.8	1.8	1.0
TKQEENYA	$F_{-2}N$	>1.3[b]	2.6	0.0
TKQEEEYA	$F_{-2}E$	>1.3[b]	0.8	0.5
TKQEEIYA	$F_{-2}I$	0.8	7.7	6.9

[a]From MD free energy simulations.[66]
[b]Only an experimental lower bound is available.[54]

listed in Table II. Comparing the MC to the reference values, there are just two large errors. For the $E_{-3}D$ peptide single mutant, the MC value is −0.3 kcal/mol, compared to an MD value of +1.8 kcal/mol. For the $F_{-2}I$ single mutant, the MC value is very large, $\Delta\Delta G = 7.7$ kcal/mol, and is clearly an artefact of the fixed backbone approximation. Excluding this last, $F_{-2}I$ sequence, the mean unsigned deviation between the MC and reference values is 0.8 kcal/mol. A Null model where all the sequences have the same affinity gives a mean deviation of 0.5 kcal/mol.

2884 of the sampled peptide sequences have negative $\Delta\Delta G$ values, indicating a stronger affinity than Sdc1. 557 have

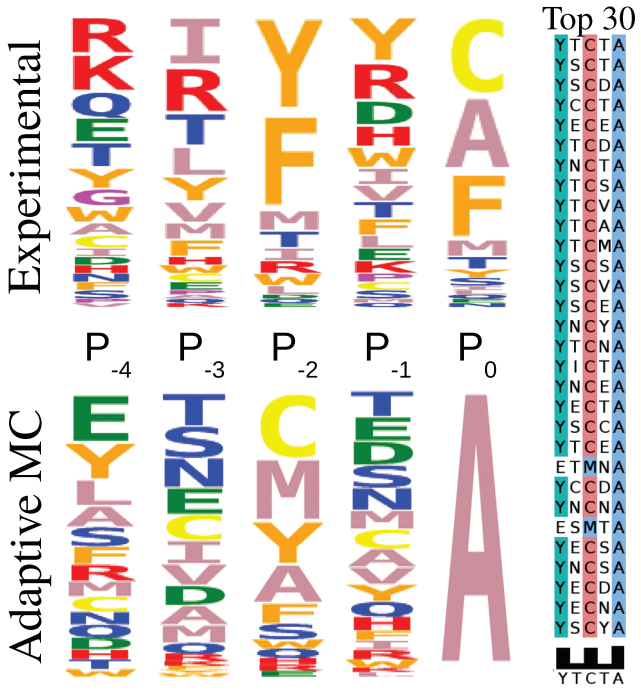


FIG. 3. Left: Sequence logos from Monte Carlo and from an experimental library of peptides that bind Tiam1.[53] Each column corresponds to a peptide position; position P0 is the C-terminus; the last five positions are shown. Amino acid types are shown with heights proportional to their abundancy and colored by physico-chemical categories. Right: The 30 strongest-binding MC sequences (ClustalW colors). The consensus sequence is shown at the bottom.

values of −0.5 or less. The top 30 sequences ($\Delta\Delta G \leq -1.2$ kcal/mol) are shown in Fig. 3. The structure of the top complex ($\Delta\Delta G = -1.5$ kcal/mol) is shown in Fig. 1. Its peptide sequence is TKQYTCTA, which differs from Sdc1 (TKQEEFYA) by four mutations. Figure 3 shows the MC sequences in the form of a sequence logo, where the sequences are Boltzmann-weighted according to their MC binding affinities. An experimental logo is shown for comparison, corresponding to a combinatorial library of peptides, selected experimentally for Tiam1 binding.[53] Positions P_0 and P_{-2} are the most important for PDZ binding specificity. Position P_0 occupies three main types experimentally, C, A, and F, but was held fixed during the simulations. Of the four positions allowed to mutate, P_{-1}, P_{-3}, and P_{-4} are highly variable in both the MC and the experimental logos. Of the top ten MC types at these positions, 7 or 8 are present in the experimental logo, and vice versa, with somewhat different occupancies. Position P_{-2} is more conserved, both experimentally and in the simulations. Of the top four experimental types, Y, F, M, and T, all but T are in the top five MC types. While T has less than 1% occupancy in the MC sequences, the chemically similar types A, C, and S are highly populated. Overall, the MC logo is in reasonable agreement with the experimental one. The top scoring sequence TKQYTCTA is a candidate for enhanced Tiam1 binding.

IV. CONCLUDING DISCUSSION

To design ligand binding, the quantity to optimize is the binding free energy, which is a Boltzmann average over many states. Nearly all CPD applications to-date have simply tried to identify one or a few low energy states. In previous studies, we computed protein:ligand relative binding free energies within the CPD framework that were formally rigorous using constant-activity MC.[41,42] Here, we went further, designing sequences based on binding affinity in a high throughput manner. We used a Wang-Landau adaptive MC approach, proposed earlier for protein:peptide binding[43] and for backbone conformational free energies.[40] The method is analogous to the recent metadynamics and lambda dynamics methods. We modified the earlier protein:peptide binding method in three ways. First, we used a more powerful, Replica Exchange Monte Carlo sampling method. Second, we introduced two adaptive bias potentials instead of one: one for the unbound peptide and a new one for the protein:peptide complex. This leads to enhanced sampling of more weakly binding sequences. Third, we implemented the method in our Proteus software, which is a flexible, user-friendly, and freely available package for CPD that allows very fast Monte Carlo simulations. This will make the method easier to apply and to combine with various solvent models or force fields.

To perfectly flatten the energy landscape in amino acid sequence space, a complex bias energy would be needed. Here, we used a simple bias that is a sum of one-position terms, where each term depends only on the side chain type of a single amino acid. As a result, we achieve only a partial landscape flattening. This is not a difficulty since we can easily compute the relative binding free energy taking into account unequal type distributions; this is done by Eq. (6). Despite the imperfect landscape flattening, almost 75 000 peptide sequences were characterized in a single REMC simulation of the bound peptide and another of the unbound peptide, which required just 1/2 h each of CPU time, thanks to the precomputed energy matrix. For more complex applications, it would be straightforward to include selected side chain:side chain interactions in the bias potential to increase flattening.

Sequences were ranked by their binding free energies and those with large, positive $\Delta\Delta G$ values (about 1/3 of the sequences) were deemed artefacts of the fixed backbone approximation and discarded. For eight sequences, comparing to experimental or high quality MD free energy simulations gave a mean unsigned deviation of 0.8 kcal/mol, 1/3 of the experimental range and close to a simple Null model. The MC sequence logo was also in reasonable agreement with experiment. Among the predicted strong binders, the peptide variant $E_{-3}D$ was characterized recently by extensive molecular dynamics simulations,[66] which predicted a weaker binding. This variant may thus represent a false positive of the design method (one out of eight variants tested). Additional predicted strong binders should be characterized experimentally to test for other false positives. Ideally, we would like to have experimental binding data for a full combinatorial library of peptide positions −5, . . ., −1, or an equivalent library for another PDZ domain. With decreasing costs for high throughput experiments, such data may become available in the future.

Among the discarded sequences, we expect there are some that would actually give strong binding, which can be considered false negatives of the fixed backbone design method. Similar problems were seen earlier[45] when designing four specificity positions within the Tiam1 PDZ domain. Depending on the backbone conformation of the peptide that was bound, large side chains were not always sampled at the designed protein positions. We expect that this artefact can be alleviated or eliminated by considering a few different peptide backbone conformations. The design calculations can either be done separately for each peptide backbone conformation or the conformations can be used together in a single, multi-backbone MC simulation using a hybrid MC method proposed recently.[40] For a system of the same size as Tiam1, with the Proteus multi-backbone approach (which samples a Boltzmann distribution), we estimate it would take a few days to run 10^9 MC steps using a few 16-core machines. While much slower, the method remains feasible on a laboratory computer cluster.

Here, as a proof of principle, we used one CPU hour on a desktop machine to explore four mutating positions. In future applications, more resources could be used, allowing sampling to be increased by 2–3 orders of magnitude. This would allow us to explore larger numbers of mutating positions, with a multi-backbone model and a more rigorous Generalized Born model that eliminates the Native Environment Approximation.[49] The method can be applied to any protein:peptide or protein:ligand binding problem, allowing the binding free energy to be used directly as the design criterion. For the present, Tiam1:Sdc1 system, the top-scoring sequence TKQYTCTA (which is also the consensus sequence derived from the top 30 sequences) is a

promising candidate for enhanced Tiam1 binding and experimental testing.

ACKNOWLEDGMENTS

Discussions with David Mignon (Ecole Polytechnique) and Ernesto J. Fuentes (U. of Iowa) are gratefully acknowledged.

[1]G. L. Butterfoss and B. Kuhlman, "Computer-based design of novel protein structures," Annu. Rev. Biophys. Biomol. Struct. **35**, 49 (2006).
[2]K. Feldmeier and B. Hoecker, "Computational protein design of ligand binding and catalysis," Curr. Opin. Chem. Biol. **17**, 929 (2013).
[3]C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard, and D. Baker, "Computational design of ligand-binding proteins with high affinity and selectivity," Nature **501**, 212 (2013).
[4]*Methods in Molecular Biology: Design and Creation of Ligand Binding Proteins*, edited by B. Stoddard (Springer Verlag, New York, 2016).
[5]N. Pokala and T. M. Handel, "Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation," Protein Sci. **13**, 925 (2004).
[6]I. Samish, C. M. MacDermaid, J. M. Perez-Aguilar, and J. G. Saven, "Theoretical and computational protein design," Annu. Rev. Phys. Chem. **62**, 129 (2011).
[7]Z. Li, Y. Yang, J. Zhan, L. Dai, and Y. Zhou, "Energy functions in de novo protein design: Current challenges and future prospects," Annu. Rev. Biochem. **42**, 315 (2013).
[8]X. Hu, H. Hu, D. N. Beratan, and W. Yang, "A gradient-directed Monte Carlo approach for protein design," J. Comput. Chem. **31**, 2164 (2010).
[9]A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. W. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y.-E. Andrew Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovic, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley, "Rosetta3: An object-oriented software suite for the simulation and design of macromolecules," Methods Enzymol. **487**, 545 (2011).
[10]R. F. Goldstein, "Evolutionary perspectives on protein thermodynamics," Lect. Notes Comput. Sci. **3039**, 718 (2004).
[11]E. J. Hong, S. M. Lippow, B. Tidor, and T. Lozano-Perez, "Rotamer optimization for protein design through MAP estimation and problem size reduction," J. Comput. Chem. **30**, 1923 (2009).
[12]I. Georgiev, R. H. Lilien, and B. R. Donald, "The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles," J. Comput. Chem. **29**, 1527 (2008).
[13]S. Traore, D. Allouche, I. André, S. de Givry, G. Katsirelos, T. Schiex, and S. Barbe, "A new framework for computational protein design through cost function network optimization," Bioinformatics **29**, 2129 (2013).
[14]D. Simoncini, D. Allouche, S. de Givry, C. Delmas, S. Barbe, and T. Schiex, "Guaranteed discrete energy optimization on large protein design problems," J. Chem. Theory Comput. **11**, 5980 (2015).
[15]C. Viricel, D. Simoncini, D. Allouche, S. de Givry, S. Barbe, and T. Schiex, "Approximate counting with deterministic guarantees for affinity computation," in*Advances in Intelligent Systems and Computing*, edited by H. A. LeThi, T. P. Dinh, and N. T. Nguyen (Springer Verlag, 2015), Vol. 360, p. 165.
[16]K. E. Roberts, P. R. Cushing, P. Boisguerin, D. R. Madden, and B. R. Donald, "Design of protein-protein interactions with a novel ensemble-based scoring algorithm," Lect. Notes Bioinf. **6577**, 361 (2011).
[17]Q. Shen, H. Tian, D. Tang, W. Yao, and X. Gao, "Ligand-K* sequence elimination: A novel algorithm for ensemble-based redesign of receptor-ligand binding," IEEE/ACM Trans. Comput. Biol. Bioinf. **11**, 573 (2014).
[18]J. Gorodkin, I. L. Hofacker, and W. L. Ruzzo, "Concepts and introduction to RNA bioinformatics," in *Methods in Molecular Biology*, edited by J. Gorodkin and W. L. Ruzzo (Springer Verlag, 2014), Vol. 1097, pp. 1–31.
[19]S. Findless, M. Etzel, S. Will, M. Moerl, and P. F. Stadler, "Design of artificial riboswitches as biosensors," Sensors **17**, 1990 (2017).
[20]J. Waldispuehl, C. W. O'Donnell, S. Devadas, P. Clote, and B. Berger, "Modelling ensembles of transmembrane beta-barrel proteins," Proteins **71**, 1097 (2008).
[21]J. Valleau and G. Torrie, "A guide to Monte Carlo for statistical mechanics," in *Modern Theoretical Chemistry*, edited by B. Berne (Plenum Press, New York, 1977), Vol. 5, pp. 137–194.
[22]V. Helms and R. C. Wade, "Computational alchemy to calculate absolute protein-ligand binding free energy," Biophys. J. **120**, 2710 (1998).
[23]Y. Deng and B. Roux, "Computations of standard binding free energies with molecular dynamics simulations," J. Phys. Chem. B **113**, 2234 (2009).
[24]T. Simonson, "Free energy calculations," in *Computational Biochemistry and Biophysics*, edited by O. Becker, A. Mackerell, Jr., B. Roux, and M. Watanabe (Marcel Dekker, NY, 2001), Chap. 9.
[25]C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology* (Springer Verlag, NY, 2007).
[26]T. Lelievre, G. Stoltz, and M. Rousset, *Free Energy Computations: A Mathematical Perspective* (World Scientific, Singapore, 2010).
[27]B. Roux and T. Simonson, "Implicit solvent models," Biophys. Chem. **78**, 1 (1999).
[28]X. Kong and C. L. Brooks, "Lambda-dynamics: A new approach to free energy calculations," J. Chem. Phys. **105**, 2414 (1996).
[29]Z. Y. Guo, J. Durkin, T. Fischmann, R. Ingram, A. Prongay, R. M. Zhang, and V. Madison, "Application of the lambda-dynamics method to evaluate the relative binding free energies of inhibitors to HCV protease," J. Med. Chem. **46**, 5360 (2003).
[30]R. L. Hayes, K. A. Armacost, J. Z. Vilseck, and C. L. Brooks, "Adaptive landscape flattening accelerates sampling of alchemical space in multisite lambda dynamics," J. Phys. Chem. B **121**, 3626 (2017).
[31]A. Laio and M. Parrinello, "Escaping free-energy minima," Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[32]A. Laio and F. Gervasio, "Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," Rep. Prog. Phys. **71**, 126601 (2008).
[33]A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: A smoothly converging and tunable free-energy method," Phys. Rev. Lett. **100**, 020603 (2008).
[34]J. F. Dama, M. Parrinello, and G. A. Voth, "Well-tempered metadynamics converges asymptotically," Phys. Rev. Lett. **112**, 240602 (2014).
[35]F. G. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," Phys. Rev. Lett. **86**, 2050 (2001).
[36]B. A. Berg and T. Neuhaus, "Multicanonical ensemble: A new approach to simulate 1st-order phase transitions," Phys. Rev. Lett. **68**, 9 (1992).
[37]N. Nakajima, H. Nakamura, and A. Kidera, "Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides," J. Phys. Chem. B **101**, 817 (1997).
[38]C. Bartels, M. Schaefer, and M. Karplus, "Determination of equilibrium properties of biomolecular systems using multidimensional adaptive umbrella sampling," J. Chem. Phys. **111**, 8048 (1999).
[39]F. Lou and P. Clote, "Thermodynamics of RNA structures by Wang-Landau sampling," Bioinformatics **26**, i278 (2010).
[40]K. Druart, J. Bigot, E. Audit, and T. Simonson, "A hybrid Monte Carlo method for multibackbone protein design," J. Chem. Theory Comput. **12**, 6035 (2017).
[41]T. Simonson, T. Gaillard, D. Mignon, M. Schmidt am Busch, A. Lopes, N. Amara, S. Polydorides, A. Sedano, K. Druart, and G. Archontis, "Computational protein design: The Proteus software and selected applications," J. Comput. Chem. **34**, 2472 (2013).
[42]K. Druart, Z. Palmai, E. Omarjee, and T. Simonson, "Protein:ligand binding free energies: A stringent test for computational protein design," J. Comput. Chem. **37**, 404 (2016).
[43]A. Bhattacherjee and S. Wallin, "Exploring protein-peptide binding specificity through computational peptide screening," PLoS Comput. Biol. **9**, e1003277 (2013).
[44]X. Yang and J. G. Saven, "Computational methods for protein design and protein sequence variability: Biased Monte Carlo and replica exchange," Chem. Phys. Lett. **401**, 205 (2005).
[45]D. Mignon and T. Simonson, "Comparing three stochastic search algorithms for computational protein design: Monte Carlo, replica exchange Monte Carlo, and a multistart, steepest-descent heuristic," J. Comput. Chem. **37**, 1781 (2016).
[46]S. Polydorides, E. Michael, D. Mignon, K. Druart, G. Archontis, and T. Simonson, "Proteus and the design of ligand binding sites," in *Methods in Molecular Biology: Design and Creation of Ligand Binding Proteins*, edited by B. Stoddard (Springer Verlag, New York, 2016), Vol. 1414, pp. 77–97.

[47]S. Polydorides and T. Simonson, "Monte Carlo simulations of proteins at constant pH with generalized Born solvent, flexible sidechains, and an effective dielectric boundary," J. Comput. Chem. **34**, 2742 (2013).
[48]E. Michael, S. Polydorides, T. Simonson, and G. Archontis, "Simple models for nonpolar solvation: Parametrization and testing," J. Comput. Chem. **38**, 2509 (2017).
[49]F. Villa, D. Mignon, S. Polydorides, and T. Simonson, "Comparing pairwise-additive and many-body generalized born models for acid/base calculations and protein design," J. Comput. Chem. **38**, 2396 (2017).
[50]A. E. Mertens, R. C. Roovers, and J. G. Collard, "Regulation of Tiam1-Rac signalling," FEBS Lett. **546**, 11 (2003).
[51]T. R. Shepherd, S. M. Klaus, X. Liu, S. Ramaswamy, K. A. DeMali, and E. J. Fuentes, "The Tiam1 PDZ domain couples to syndecan1 and promotes cell-matrix adhesion," J. Mol. Biol. **398**, 730 (2010).
[52]B. I. Dahiyat and S. L. Mayo, "De novo protein design: Fully automated sequence selection," Science **278**, 82 (1997).
[53]T. R. Shepherd, R. L. Hard, A. M. Murray, D. Pei, and E. J. Fuentes, "Distinct ligand specificity of the Tiam1 and Tiam2 PDZ domains," Biochemistry **50**, 1296 (2011).
[54]N. Panel, Y. J. Sun, E. J. Fuentes, and T. Simonson, "A simple PB/LIE free energy function accurately predicts the peptide binding specificity of the Tiam1 PDZ domain," Front. Mol. Biosci. **4**, 65 (2017).
[55]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys. **21**, 1087 (1953).
[56]D. Frenkel and B. Smit, in *Understanding Molecular Simulation* (Academic Press, New York, 1996), Chap. 3.
[57]G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes* (Oxford University Press, Oxford, United Kingdom, 2001).
[58]X. Liu, T. R. Shepherd, A. M. Murray, Z. Xu, and E. J. Fuentes, "The structure of the Tiam1 PDZ domain/phospho-syndecan1 complex reveals a ligand conformation that modulates protein dynamics," Structure **21**, 342 (2013).
[59]P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery, "A new approach to the rapid determination of protein side chain conformations," J. Biomol. Struct. Dyn. **8**, 1267 (1991).
[60]A. T. Brünger, *X-Plor Version 3.1: A System for X-Ray Crystallography and NMR* (Yale University Press, New Haven, 1992).
[61]W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," J. Am. Chem. Soc. **117**, 5179 (1995).
[62]G. D. Hawkins, C. Cramer, and D. Truhlar, "Pairwise descreening of solute charges from a dielectric medium," Chem. Phys. Lett. **246**, 122 (1995).
[63]A. Lopes, A. Aleksandrov, C. Bathelt, G. Archontis, and T. Simonson, "Computational sidechain placement and protein mutagenesis with implicit solvent models," Proteins **67**, 853 (2007).
[64]M. Schmidt am Busch, A. Lopes, D. Mignon, and T. Simonson, "Computational protein design: Software implementation, parameter optimization, and performance of a simple model," J. Comput. Chem. **29**, 1092 (2008).
[65]T. Gaillard and T. Simonson, "Pairwise decomposition of an MMGBSA energy function for computational protein design," J. Comput. Chem. **35**, 1371 (2014).
[66]N. Panel, F. Villa, E. J. Fuentes, and T. Simonson, "Accurate PDZ:peptide binding specificity with additive and polarizable free energy simulations," Biophys. J. **114**, 1091 (2018).

1 PB/LIE analysis

We computed binding free energies for 14 of the high scoring designed peptide sequences using PB/LIE. Sequences were extracted from the set of "top 30" designed binders (listed in Villa *et al.* [2018b], figure 3). We give first information about the PB/LIE method, from the published work of Panel *et al.* [2017].

1.1 Semi-empirical Free Energy Function

The binding free energy between Tiam1 and each designed peptide was estimated using the scoring function

$$\Delta G = \alpha \Delta E_{vdW} + \beta \Delta G_{PB} + \gamma \Delta A + \delta \tag{2.1}$$

where α, β, γ are adjustable constants. The other terms ΔE_{vdW}, ΔG_{elec} and ΔA are differences between bound and unbound states, averaged over snapshots of MD simulations of the protein-peptide complex. More precisely, ΔE_{vdW} is the average van der Walls interaction energy between protein and peptide. ΔG_{elec} is the electrostatic free energy difference between bound and unbound states obtained with Poisson Boltzmann (PB) continuum electrostatics, and ΔA is the change in molecular surface upon binding. The weights α, β and γ were optimized (Panel *et al.* [2017]) to reproduce affinities for a target data set composed of 35 variants of the Tiam1:Sdc1 complex (including both peptide and protein mutations). Optimized parameters were α=0.02 (vdW), β=0.25 (elec) and γ=-4 cal/mol/Å^2. After optimization the authours predicted experimental binding affinities for the target dataset, obtaining a mean unsigned error of 0.43 kcal/mol and Pearson correlation of 0.64 between experimental and calculate binding free energies.

1.2 Structural models and simulations setup

3D structures of the Tiam1:peptide complexes were obtained after reconstruction of the designed sequences with Proteus. For each peptide sequence under analysis, we extracted its rotamer configurations visited during the MC sampling in the bound state (hundreds for each sequence). Configurations were then sorted according to their Proteus energy (after subtracting the bias potential). The best rotamer configuration of each sequence was finally applied on the fixed backbone of the complex, and minimized with XPLOR. Histidines were set to be neutral, after inspection of hydrogen bonds in the 3D structures and pK_a calculations performed with PropKa (Bas *et al.* [2008]).

Each reconstructed complex was then solvated in an octahedral box, filled with explicit solvent (TIP3P water). The system was then neutralized with sodium or chloride ions. After equilibration, production MD was run using NAMD 2.12. Simulations were done at room temperature and pressure with the Amber ff14SB force field (Maier *et al.* [2015]), using Langevin dynamics with a Langevin Piston Nosé-Hoover barostat (Feller *et al.* [1995]). Long range electrostatic interactions were computed using PME (Darden, [2001]). Large energy fluctuations during dynamics were avoided using a flat-bottomed harmonic restraint of force constant 3.0 kcal/mol/Å^2 applied on the N-terminal peptide residue. Using such a restraint, the harmonic potential started to increase beyond a threshold distance of 3 Å between peptide and protein. For each complex, production MD was run for 100 ns.

1.3 PB calculations

MD snapshots were extracted from trajectories (1000 frames for each complex). For each snapshot, water molecules were deleted and the system was described using continuum electrostatics. We used a protein dielectric constant ϵ_p=8 and a solvent dielectric constant ϵ_w=80. The electrostatic free energy difference was obtained taking the difference between the electrostatic free energy of the complex and of the peptide, with structures built from the same snapshot. The ΔG_{elec} term was finally obtained averaging the electrostatic free energy differences over all the MD frames. PB calculations were run using CHARMM (Im *et al.* [1998]), solving the PB equation on a grid of 181 planes in each direction, with a 0.4 Å spacing between planes.

1.4 Results

Results of the PB/LIE analysis for the 14 selected peptides are resumed in table 2.1. We include, in the same table, the relative binding free energies obtained with Monte Carlo (from Villa *et al.* [2018b]). PB/LIE affinities were within 0.7 kcal/mol, while the ΔΔG obtained with MC were within 1.5 kcal/mol. We also note that with all MC ΔΔGs were negative (they belong to the top 30 designed binders), while PB/LIE gived two variants (sequences ETMNA and EEMNA) that have essentially the same affinity of the WT, while all other PB/LIE variants have positive ΔΔG. It was recognized that the PB/LIE model has a small systematic error, overestimating the experimental binding affinities by 0.30 kcal/mol on average (Panel *et al.* [2017]). However, this cannot explain the observed discrepancies. Moreover, a real comparision is difficult because for the selected variants experimental data is unknown. It would be interestig to test some of these variants with more sophisticated, alchemical free energy

simulations (Panel *et al.* [2018]) which allow to extract free energies without any adjustable parameter.

Peptide sequence	PB	vdW	SA	ΔG	PB/LIE ΔΔG	MC ΔΔG
ETMNA	-3.25	-48.10	-1310.1	3.46	-0.05	-1.2
EEMNA	-2.99	-49.48	-1307.6	3.49	-0.02	-1.1
EEFYA	-2.84	-58.67	-1349.3	3.51	0.00	0.0
YECEA	-3.01	-55.62	-1371.0	3.61	0.11	-1.3
YTCDA	-2.26	-50.67	-1313.2	3.77	0.16	-1.3
ESMTA	-2.24	-48.65	-1305.9	3.70	0.18	-1.2
YSCDA	-1.88	-49.70	-1295.0	3.71	0.20	-1.3
ETMTA	-2.50	-49.47	-1331.0	3.71	0.20	-1.1
ETMEA	-1.06	-46.81	-1247.8	3.79	0.28	-1.1
EEMSA	-1.64	-49.27	-1329.5	3.92	0.41	-1.2
YNCTA	+1.20	-38.10	-1111.0	3.98	0.47	-1.3
YTCVA	-1.87	-57.31	-1398.6	3.98	0.47	-1.3
YTCTA	-0.93	-50.07	-1321.8	4.05	0.54	-1.5
YNCYA	-1.56	-57.28	-1401.7	4.07	0.56	-1.3
YSCTA	-0.46	-49.27	-1311.0	4.14	0.63	-1.4

Table 2.1 – **PB/LIE binding free energies** for 14 selected peptide sequences. Each sequence specifies the amino acids at positions $P_{-4,-3,-2,-1,0}$. Specific contributions to the empirical ΔG are given on different columns. All terms are in kcal/mol, except values in the SA column (the surface area) which are in Å^2. PB, vdW, and SA contributions were weighted by optimized parameters to compute ΔG [eqn. 2.1]. Optimized weights were α=0.02 (vdw), β=0.25 (elec) and γ=-4 cal/mol/Å^2 (Panel *et al.* [2017]). Relative binding free energies of each sequence were obtained subtracting the ΔG of the WT sequence EEFYA.

2 Bias convergence

The relative binding free energies in Villa *et al.* [2018b] were obtained by difference, simulating the ligand pool twice. Two separate bias potentials were constructed, running adaptive Monte Carlo in bound and unbound state. In a second round, the bias potentials were used in two separate production simulations. The final free energy differences ΔG_{bound} and $\Delta G_{unbound}$ were obtained removing the respective bias potentials from the flattened distributions.

The method is formally exact, and does not require perfect landscape flattening [fig. 2.1]. However, let us suppose that the bias converges perfectly, flattening the sequence space of the simulated system. In the long time limit, the free energy difference between any couple of

sequences goes to zero, and the bias potential converges to

$$E^b(s,t \to \infty) = -\frac{E^0}{E^0 + kT} F(s) + C(t \to \infty) \tag{2.2}$$

where kT=0.6 kcal/mol is the thermal energy, E^0 is the energy factor used to define the bias increment rule (equation 4 in the article) and C is a constant that does not depend on the sequence but grows as $t \to \infty$. In equation 2.2, the quantity F(s) is the *unbiased* free energy of a sequence s at temperature T, that is completely compansated by the bias in the long time limit, up to a factor $E^0/(E^0 + kT)$ and an arbitrary constant, as expected. A detailed derivation of equation 2.2 can be found in Barducci *et al.* [2008].

With an exponentially decreasing bias update rule (equation 4 in the article), the bias potential converges to the free energy $F(s)$ of an ensemble at shifted temperature respect to the temperature T used in the adaptive MC. To compute free energy differences at temperature T, the bias potential must be multiplied by a factor that approaches 1 as kT is negligible respect to E^0. For very large E^0, one retrieves the result of standard metadynamics where $E^b(s) \to -F(s) + C$.

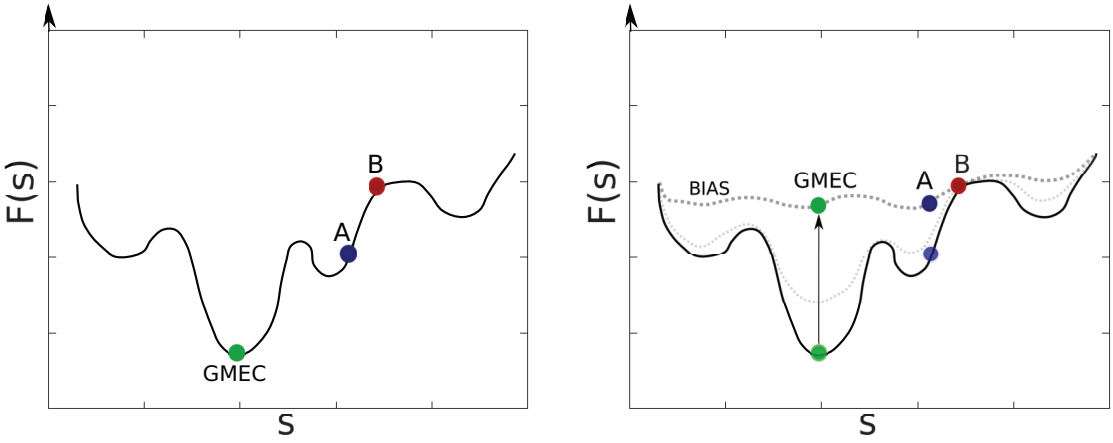


Figure 2.1 – **Free energy surface** (pictorial representation) as function of sequence s. **Left**: unbiased landscape, where the GMEC sits on the bottom and sequences A and B are separated by a free energy difference of several kcal/mol. **Right**: Biased simulation, with an almost flatten landscape. All sequences have similar probabilities and small free energy difference. Supposing perfect bias convergence, all sequences have the same probability (in the flattened landscape) and all their free energy differences are equal to zero.

3 Perspectives: advanced sampling

The bias potential defined in the article (equation 5) is made of "diagonal terms" or single-position biases. As a consequence, it cannot fully compensate the different interactions between couples of residues of fluctuating amino acid types. To better flat the $F(s)$ surface, one may consider a more complex (but still simple) bias potential, with a funcional form that includes double-position biases:

$$\begin{aligned} E^b_{ij}(s_i(t),s_j(t);t) &= \sum_{n;nT<t} e^b_{ij}(s_i(nT),s_j(nT);nT)\delta_{s_i(t),s_i(nT)}\delta_{s_j(t),s_j(nT)} \\ e^b_{ij}(s_i(t),s_j(t);t) &= e_0 exp\left[-E^b_{ij}(s_i(t),s_j(t);t)/E^0\right] \end{aligned} \tag{2.3}$$

Here e_0 and E^0 are adjustable parameters, that control the growth of the bias potential during the adaptive MC simulation. As for the single-position bias, the increment $e^b_{ij}(s_i,s_j;t)$ is added at a particular MC step t only to one couple of side chains of type s_i and s_j. Such a double-position bias can be combined with the previously defined single-position bias (equation 3 and 4), to give:

$$E^b(s_1,s_2,...,s_p;t) = \sum_1^p E^b_i(s_i;t) + \sum_{i<j} E^b_{ij}(s_i(t),s_j(t);t) \tag{2.4}$$

The Proteus implementation is under development, and it has been verified that this "combined" bias better flattens the sequence space in various situations. For example, it can be used to flat the free energy landscape of the unbound state of a protein:peptide complex. After convergence and supposing almost perfect flattening, the difference in bias potential for two couple of sequences can be used to directly obtain their free energy difference in the unbound state. Using equation 2.2, we write

$$\Delta G_{unbound}(s \to s_r) = \left[E^b(s_r,t \leftarrow \infty) - E^b(s,t \to \infty)\right] \frac{E^0 + kT}{E^0} \stackrel{def}{=} \tilde{E}^b_{unbound}(s_r) - \tilde{E}^b_{unbound}(s) \tag{2.5}$$

The resulting, "scaled" bias potential $\tilde{E}^b_{unbound}(s)$ effectively flattens the free energy landscape of the unbound state at temperature T. It can be used to sample sequences of the bound state. The interesting result is that sequences sampled in this *biased* bound state will be populated according to their relative binding affinities:

$$\Delta\Delta G_{bind}(s \to s_r) = \Delta G_{bound}(s \to s_r) - \Delta G_{unbound}(s \to s_r) = \Delta G_{bound}(s \to s_r) - \Delta \tilde{E}^b_{unbound}(s \to s_r) \tag{2.6}$$

This represents a clear improvement respect to the former protocol described in Villa *et al.* [2018b], since using the last equation one avoids to run the production MC simulations in

the unbound state. Moreover, the design by affinity opens interesting potential applications for the engineering of PPI networks. Once a ligand pool has been flattened in the unbound state, the resulting bias can be applied to different complexes in separate simulations. This allows to screen binding affinities for many peptide variants binding to different receptors. In this case, one may also consider to compute an average $\tilde{E}^b_{unbound}(s)$ for many unbound state conformations, to better catch the peptide conformational changes upon binding.

Another interesting application is the design by *specificity*. It is based on the idea that flattening the sequence space for a given protein:peptide complex, the resulting potential $\tilde{E}^b_{bound}(s)$ can be used to sample sequences in the bound state of a different receptor. Indeed, as for the unbound state we can write

$$\Delta G_{bound}(s \to s_r) = \left[E^b(s_r,t \leftarrow \infty) - E^b(s,t \to \infty)\right] \frac{E^0 + kT}{E^0} \stackrel{def}{=} \tilde{E}^b_{bound}(s_r) - \tilde{E}^b_{bound}(s) \quad (2.7)$$

Considering the relative binding free energies for a generic couple of receptors A and B

$$\begin{aligned} \Delta\Delta G^A_{bind}(s \to s_r) &= \Delta G^A_{bound}(s \to s_r) - \Delta G_{unbound}(s \to s_r) \\ \Delta\Delta G^B_{bind}(s \to s_r) &= \Delta G^B_{bound}(s \to s_r) - \Delta G_{unbound}(s \to s_r) \end{aligned} \quad (2.8)$$

and indicating with $\tilde{E}^b_{bound,A}(s)$ the bias potential that flattens the bound state for receptor A, one has

$$\Delta\Delta G^{A,B}_{bind}(s \to s_r) = \Delta\Delta G^B_{bind}(s \to s_r) - \Delta\Delta G^A_{bind}(s \to s_r) = \Delta G^B_{bound}(s \to s_r) - \Delta\tilde{E}^b_{bound,A}(s \to s_r) \quad (2.9)$$

where the unbound state contributions cancel out. The interesting result is that sequences sampled in the bound state of B using $\tilde{E}^b_{bound,A}(s)$ will be populated according to their *binding specificity* for receptors B and A.

Chapter 3

Polarizable free energy simulations

1 Introduction

Molecular simulations need accurate energy functions to describe complex electrostatic interactions in crowded environments. Molecular dynamics (MD) needs also a simple, empirical representation of molecule's degrees of freedom. Most popular force fields like Amber (Maier *et al.* [2015]), Charmm (MacKerell *et al.* [1998]), Gromos (Oostenbrink *et al.* [2004]) and OPLS (Jorgensen *et al.* [1996]) are carefully parametrized to describe both bonded and non-bonded interactions via a similar energy function. Their functional forms differ for few terms, usually related to internal degrees of freedom. They are parametrized following slightly different strategies, even if they use similar approaches to fit QM and experimental data. All these well-established force fields are able to reproduce both equilibrium states and thermal fluctuations of a large number of molecules, from small model compounds to big polymers. They are also able, in some situations, to fold small proteins in a reasonable amount of time (Lindorff-Larsen *et al.* [2011]). However, all these force fields share a common simplification in their description of electrostatic interactions.

$$E_{nonbonded}(\{\bar{r}\}) = \sum_{\text{nonbonded pairs}} \left\{ \epsilon_{ij}^{min} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \right\} \tag{3.1}$$

where $\{\bar{r}\}$ is the collection of distances r_{ij} between nonbonded pairs of the system. The electrostatic interactions are computed as a sum over non-bonded pairs of the classical, Coulomb energy. Because of that, all these force fields are called *additive*. They have a limited represenation of the real molecular system because the charge of each atom i is condensed at one point in space r_i and has a constant magnitude, neglecting induced polarization effects. This common feature makes all these force fields problematic when describing interactions between ionic groups in a heterogeneous environment, for example when charges are buried inside the

protein core or are transferred from the bulk solvent to binding pockets. These force fields have also a limited accuracy in describing complex networks of salt bridges, when many ionic interactions fluctuate at thermodynamic equilibrium (Debiec *et al.* [2014]). In many applications this level of accuracy is acceptable, but there are cases where consequences can be dramatic.

Free energy calculations are among the most challenging applications in computational biochemistry. Despite the development of advanced sampling techniques and increasing computational power, there is still a great need of improved accuracy. Limitations are especially obvious when computing free energy differences for charging processes,

$$\Delta F(q_0 \rightarrow q_1) \qquad\qquad q_0 \neq q_1 \tag{3.2}$$

The precise calculation of electrostatic interactions for this kind of transormation is required when studying solvation, mutation free energies or ligand binding. For example, relative binding free energies for several point and double mutations in PDZ:peptide complexes were recently obtained from MD simulations, using a free energy perturbation protocol (Panel *et al.* [2018]). Many peptide variants were studied using the Amber 99SB force field, computing their relative affinities and comparing them to experimental data. Good agreement was obtained for non-ionic mutations, giving a root mean square deviation of 0.37 kcal/mol, a mean unsigned error of 0.32 kcal/mol, a maximum error of 0.6 kcal/mol and a correlation coefficient of 0.84. The situation was quite different considering a set of *ionic* mutations, where the mutated side chains have a different charge. In this case errors were larger than 1 kcal/mol. This result is comparable to other studies in literature (Deng & Roux [2006] ; Boyce *et al.* [2009]).

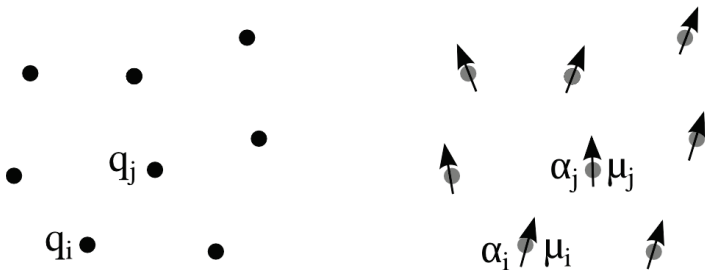
A new generation of molecular-mechanics force fields aims to overcome difficulties related to an additive, point-charge representation of electrostatic interactions. Because they include an empirical representation of electronic polarization, these *polarizable* force fields are more complex both in their functional form and in their parametrization strategies. Since they model the reponse of a molecule's electron charge distribution in different environments, their tuning requires extensive testing, carried out over several years of development. In the field of biomolecular simulation there are still many applications where polarizable MD has not yet been used. There are few examples of polarizable force fields that can be used routinelly to perform MD simulations at the $100 - 1000$ nanosecond timescale for free energy calculations, where one has to sample many microstates of systems composed of tons of thousands of atoms. The community recognises (Lemkul *et al.* [2016]) that there are actually three main models, implemented in different force fields, that have been shown to be more suitable and reliable for biomolecules: induced point-dipoles, fluctuating charges and the Drude oscillator model. The first is the one used in AMOEBA (Shi *et al.* [2013]), the second in the Charmm Fluctuating

Charge Force Field (Patel & Brooks [2003] ; Patel *et al.* [2004]) and the latter in the Drude force field (Lamoureux & Roux [2003]). The present thesis focuses on Drude, even if a short description of AMOEBA and Charmm Fluctuating Charge force fields will be given.

This chapter starts with a theoretical introduction about classical models used to study interaction between induced dipoles in molecular systems. The point-dipole model and key quantities will be presented. After a brief discussion of existing polarizable force fields suitable for molecular simulation, we will describe the Drude Force field. Several important concepts for the calculation of binding free energies will be discussed, with particular attention to charging processes. Since free energy differences are extracted from molecular simulations performed in infinite, periodic simulation boxes, several artefacts need to be taken into account.

2 Induced dipole model

2.1 Fields and dipoles



Modeling induced electronic polarization with an empirical, classical model dates back to 1972. The original work of Applequist *et al.* [1972] was based on the older study from Silberstein [1917] that introduced the point-dipole model for polyatomic molecules. *Atomic* polarizabilities are assigned to each atom. Considering a system composed of N atomic dipoles $\bar{\mu_j}$ located at positions $\bar{r_j}$, the electric field at position $\bar{r}_i$ is

$$\bar{E}_i = -\sum_{j=1}^{N} T_{ij}\bar{\mu_j} \tag{3.3}$$

where T_{ij} is the dipole field tensor for atomic centers i and j, which is a 3x3 matrix. If r is the distance between atoms i and j and where $x = x_i - x_j$, $y = y_i - y_j$ and $z = z_i - z_j$, then

$$T_{ij} = -\frac{3}{r^5}\begin{bmatrix} x^2 - 1/3r^2 & xy & xz \\ yx & y^2 - 1/3r^2 & yz \\ zx & zy & z^2 - 1/3r^2 \end{bmatrix} \tag{3.4}$$

The induced dipole for atom i at position $\bar{r}_i$ is computed using

$$\bar{\mu}_i = \alpha_i \bar{E}_i = \alpha_i \left[\bar{E}_i^{ext} - \sum_{j=1;j\neq i}^{N} T_{ij}\bar{\mu}_j \right] \tag{3.5}$$

where $\bar{E}_i^{ext}$ is an external electric field at position r_i and α_i is the *polarizability* tensor

$$\alpha_i = \begin{bmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{bmatrix} \tag{3.6}$$

For a molecule composed of N atoms in the presence of an applied electric field $\bar{E}$, the induced molecular dipole moment $\bar{\mu}_{mol}$ can be expressed as function of the molecular polarizability α_{mol}

$$\bar{\mu}_{mol} = \sum_{i=1}^{N} \bar{\mu}_i = \left[\sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij} \right] \bar{E} = \alpha_{mol}\bar{E} \qquad \alpha_{mol} = \sum_{i=1}^{N}\sum_{j=1}^{N} B_{ij} \tag{3.7}$$

The matrix B_{ij} contains atomic polarizabilities and the dipole field tensor. Its mathematical expression depends on the functional form of T_{ij} and the interaction is parametrized by molecular polarizabilities. Indeed, equation 3.5 can be seen as a system of equations for each $\bar{\mu}_i$ where $i = 1, ..., N$. In matrix form, it corresponds to

$$\begin{bmatrix} \alpha_1^{-1} & T_{12} & ... & T_{1N} \\ T_{21} & \alpha_2^{-1} & ... & T_{2N} \\ ... & & & \\ ... & & & \\ T_{N1} & ... & ... & \alpha_N^{-1} \end{bmatrix} \begin{bmatrix} \bar{\mu}_1 \\ ... \\ ... \\ ... \\ \bar{\mu}_N \end{bmatrix} = A\tilde{\mu} = \begin{bmatrix} \bar{E}_1 \\ ... \\ ... \\ ... \\ \bar{E}_N \end{bmatrix} = \tilde{E} \tag{3.8}$$

if $\tilde{B} = A^{-1}$ then $\tilde{\mu} = \tilde{B}\tilde{E}$, so that for each atom

$$\bar{\mu}_i = \sum_{j=1}^{N} B_{ij}\bar{E}_j \tag{3.9}$$

This simple model proposed by Applequist makes good predictions of mean, molecular polarizabilities. However, this point-like description suffers from several problems. First, it produces a large anisotropy in molecular systems. Second, in some situations it leads to infinite polarizabilities. This problem arises when the polarizable atoms are too close. It can be avoided by modification of the dipole field tensor. Thole [1981] proposed an alternative interaction model in order to damp interactions between dipoles when they get closer than a threshold value. More precisely, he proposed to compute dipole-dipole interactions using a slightly different dipole field tensor, which is built by modification of charge distributions according to a *shape function*. This shape function, which can be linear or exponential, is parametrized by atomic polarizabilities and a general screening parameter τ:

$$\rho = \frac{\tau^3}{8\pi}\exp\left[-\frac{\tau r_{ij}}{(\alpha_i\alpha_j)^{1/6}}\right] \tag{3.10}$$

Using the threshold distance $r^*_{ij} = \tau(\alpha_i\alpha_j)^{1/6}$ the modified dipole filed tensor is then computed using

$$T(u = r_{ij}/r^*_{ij}) = -\frac{3}{r_{ij}(r^*)^4}\begin{bmatrix} x^2(1-\frac{4}{3u}) & xy & xz \\ yx & y^2(1-\frac{4}{3u}) & yz \\ zx & zy & z^2(1-\frac{4}{3u}) \end{bmatrix} \tag{3.11}$$

when $r_{ij} < r^*_{ij}$

2.2 Electrostatic energy

The energy of a system of point dipoles is composed of several contributions. Its general prototype is

$$E_{elec}(\{\bar{r}\}) = E_{q-q} + E_{q-\mu} + E_{\mu-\mu} + E_{pol} \tag{3.12}$$

The first terms are charge-charge (q-q), charge-dipole (q-μ) and dipole-dipole ($\mu-\mu$) interactions. E_{pol} is the work of polarization, also called "self polarization energy". Electronic degrees of freedom, parametrized by atomic polarizabilities α_i, are represented using the smeared charge distribution. The dipole field-tensor T_{ij} is the mathematical object used to calculate

non-additive interactions (q-μ) and ($\mu - \mu$). It is function of the charge distributions of atoms i and j which depend, following equation 3.10, on the distance between centers i and j. A more explicit expression of equation 3.12 is

$$E_{elec}(\{\bar{r}\}) = \sum_{i<j} \left(\frac{q_i q_j}{r_{ij}} - q_i \sum_{\alpha=x,y,z} T_{ij}^{\alpha} \mu_{j\alpha} + \sum_{\alpha=x,y,z} \mu_{i\alpha} T_{ij}^{\alpha} q_j - \sum_{\alpha,\beta=x,y,z} \mu_{i\alpha} T_{ij}^{\alpha\beta} \mu_{j\beta} \right) + U_{pol} \quad (3.13)$$

where T_{ij} is the interaction tensor, such that

$$\begin{aligned} T_{ij}^{\alpha} &\stackrel{def}{=} \nabla_{\alpha} T_{ij} = -\frac{r_{ij,\alpha}}{r_{ij}^3} \\ T_{ij}^{\alpha,\beta} &\stackrel{def}{=} \nabla_{\alpha}\nabla_{\beta} T_{ij} = \left(3 r_{ij,\alpha} r_{ij,\beta} - r_{ij}^2 \delta_{\alpha,\beta}\right) r_{ij}^{-5} \end{aligned} \quad (3.14)$$

We immediately see that, respect to the simple charge-charge Coulomb interaction, additional terms appear in the sum. These terms are parametrized by molecular polarizability and Thole factors, which control strength of induced dipoles and the shape of the charge distribution. The last term is the work of polarization, and can be derived considering the induced dipole $\bar{\mu}_i$ of atom i in presence of electric field $\bar{E}_i$ at point $\bar{r}_i$ (Böttcher [1973]). The energy of the induced dipole is

$$U = -\bar{\mu}_i \cdot \bar{E}_i + U_{pol} \quad (3.15)$$

At equilibrium, for small changes $d\bar{\mu}_i$ of the induced dipole moment, one has dU=0. So we can write

$$dU_{pol} = -d\left[-\bar{\mu}_i \cdot \bar{E}_i\right] = \bar{E}_i \cdot d\bar{\mu}_i = \frac{\bar{\mu}_i}{\alpha_i} \cdot d\bar{\mu}_i = \frac{1}{2\alpha_i} d\left[\bar{\mu}_i^2\right] \quad (3.16)$$

The work of polarization is then obtained easily:

$$U_{pol} = \frac{1}{2\alpha_i} \int_0^{\bar{\mu}_i} d\left[\bar{\mu}_i^2\right] = \frac{1}{2\alpha_i} \bar{\mu}_i^2 \quad (3.17)$$

For the system of dipoles $\bar{\mu}_i$, the total work of polarization or total "self-polarization energy" is finally

$$U_{pol} = \frac{1}{2} \sum_i \frac{\bar{\mu}_i \cdot \bar{\mu}_i}{\alpha_i} \quad (3.18)$$

2.3 Induced dipole force fields

AMOEBA (Ren & Ponder [2003]) belongs to the class of force fields that make use of equation 3.13, as well as Amber ff02 (Cieplak *et al.* [2001]). AMOEBA includes also permanent, traceless quadrupoles, which are 3x3 tensors. So for each atom i it associates properties of the point-dipole model (charge q and dipole moment $\bar{\mu}_i$) but also additional, quadrupole $Q_{\alpha\beta}$ components. All atoms are polarizable, including hydrogens. The direct use of the induced-dipole representation is considered useful especially for its ability to reproduce anisotropy effects (Baker [2015]). However, increased computational cost is demanded to compute interactions, as well as more-expensive force evaluations in MD. To give an idea, one can have a look at recent timings of all-atom, explicit solvent molecular dynamics simulations performed using AMOEBA, implemented on an optimized version of Tinker (Lagardère *et al.* [2018]). Calculations were run on the *Occigen* supercomputer, hosted at the CINES centre of the French supercomputing agency GENCI. For plain MD of one solvated protein (dihydrofolate reductase, DHFR) composed of 23558 atoms the performances are of 7.2 ns/day on 960 cores. Recent calculations on a similar system performed in our laboratory using 224 cores using NAMD (Phillips *et al.* [2005]) and an additive Amberff99SB force field show performances of $\sim$ 74 ns/day, over 40 times faster. Despite the strong effort in software development and the accuracy in the calculation of electrostatic interactions, long simulations at timescales of the order of 1 μs with AMOEBA have an enormous computational cost.

3 Point charge models

3.1 Fluctuating charge model

Given the high computational cost demanded by induced-dipole representations, a lot of effort has been made to build more empirical but less expensive energy functions suitable for MD. In the fluctuating charge model, charges are considered as dynamical variables in an extended lagrangian framework. It is based on the *principle of electronegativity equalization*. Given a molecular system, charges fluctuate according to conformational changes that alter the local value of the electrostatic potential. Each atom has an associated *electronegativity* and within a molecule the system is constrained to keep the same electronegativity for each atom. In pratice, each charge has an associated mass M_Q that is small compared to the real, atomic mass. In this way, electronic degrees of freedom (which evolve in the extended lagrangian) are able to relax quickly in reponse to conformational changes. The main advantage of the method relies is that the energy function used to compute interactions is unchanged compared to the standard, additive force fields. However, it suffer from a strong limitation: intramolecular

electronic charge distributions are exchanged between atoms, so that charge transfer occurs mostly along bond lines. This is a strong limitation when one wants to model polarization in the direction perpendicular to the molecule plane.

3.2 Drude pseudo-particle model

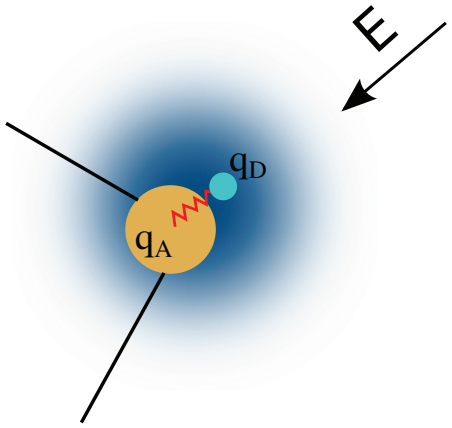


Figure 3.1 – **Drude system** composed of a polarizable atom q_A and its Drude particle q_D, displaced according to an external field

In the Drude pseudo-particle model, the electronic cloud surrounding a polarizable, heavy atom is represented by a negative charge q_D attached to the nucleus with a harmonic oscillator of force constant k_D [fig. 3.1]. All atoms except hydrogens are by default polarizable. The harmonic spring has the same force constant for all atoms. The total charge Q of the atom is sum of the nucleus and Drude charges. Q is split between the nucleus A and its Drude particle D depending on the atomic polarizability α:

$$Q = q_A + q_D \qquad \alpha = \frac{q_D^2}{k_D} \qquad M = m_A + m_D \qquad m_D = 0.4 \quad a.m.u. \tag{3.19}$$

where the mass is expressed in atomic mass units. In general, each atom has an isotropic polarizability $\alpha_{xx} = \alpha_{yy} = \alpha_{zz}$. In the absence of an external field, each Drude particle oscillates around its reference nucleus. Otherwise it oscillates around a displaced position, defined by the vector

$$\bar{d} = \frac{q_D \bar{E}}{k_D} \tag{3.20}$$

which depends on the Drude particle charge, the force constant and the applied field. From the formula we see that for a given $\bar{E}$, a bigger atomic polarizability α (a more negative charge q_D) produces a bigger displacement. The atom and Drude particle behave like a point charge

Q of polarizability α, in the spirit of the point-dipole model.

Interactions between atoms and Drude oscillators that are separated by three or more bonds are computed using the standard, Coulomb energy function, while dipole-dipole interactions between 1-2 and 1-3 atom pairs are screened with the previously discussed method proposed by Thole. In the Drude force field, instead of using a general screening parameter τ for all atoms, interactions between dipoles i and j are screened by multiplying their interactions by

$$S(r_{ij}) = 1 - \left(1 + \frac{\tau_{ij} r_{ij}}{2(\alpha_i \alpha_j)^{1/6}}\right) \exp\left(-\frac{\tau_{ij} r_{ij}}{(\alpha_i \alpha_j)^{1/6}}\right) \qquad \tau_{ij} = \tau_i + \tau_j \tag{3.21}$$

where τ_i and τ_j are screening (or Thole) parameters for atoms i and j. This formula corresponds to a smeared charge distribution

$$\rho = \frac{\tau_{ij}^3}{8\pi} \exp\left[-\frac{\tau_{ij} r_{ij}}{(\alpha_i \alpha_j)^{1/6}}\right] \tag{3.22}$$

The potential function includes also a self-polarization energy

$$U_{self}(\bar{d}) = \frac{1}{2} k_D \bar{d}^2 \tag{3.23}$$

The 2013 version of the Drude force field includes also anisotropic polarizabilities, where $\alpha_{xx} \neq \alpha_{yy} \neq \alpha_{zz}$. In this case, the self-polarization energy is calculated using the more general formula

$$U_{self}(\bar{d}) = \frac{1}{2}\left(k_{xx} d_x^2 + k_{yy} d_y^2 + k_{zz} d_z^2\right) \tag{3.24}$$

Drude has also other auxiliary particles, called *lone pairs* [fig. 3.2], not present in the first release of the force field. Lone pairs represent electron pairs in valence shells of atoms that do not participate in bonding. Their role is to include a charge density that is delocalized with respect to their reference atoms, using *virtual* rigid interaction sites. In fact, lone particles are massless and their forces are transferred to their reference atoms during MD simulations. The introduction of lone pairs in Drude improved agreement with QM calculations, both for energetics and for optimized geometries (Harder *et al.* [2006]).

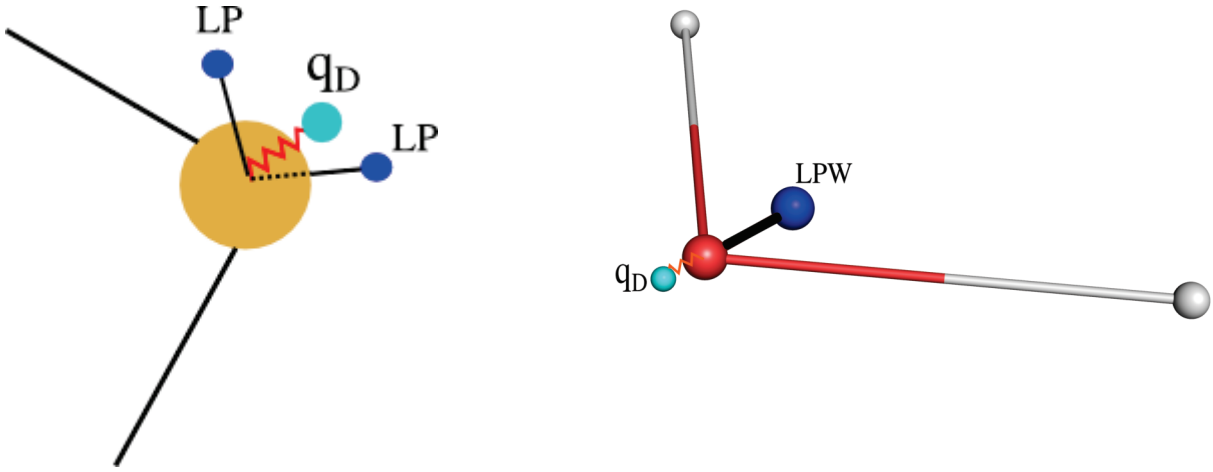


Figure 3.2 – Left: atom-drude + lone pair system. Right: SWM4-NDP Drude polarizable water model

3.3 Drude polarizable water models

The first Drude water model was published in 2003 (Lamoureux *et al.* [2003]) and named *SWM4*: *Simple Water Model with 4 interaction sites*. In a former parametrization, it was called *SWM4-DP* because the oxygen carried a positive Drude charge. Along with the usual oxygen and hydrogen atoms, a rigid particle called "LPW" was attached to the oxygen in order to better reproduce the gas-phase dipole moment and quadrupole moments. In analogy with the TIP4P water model, this lone charge is negative and its magnitude is twice the charge assigned to hydrogen atoms.

To be more consistent with the physical model, where negative Drude charges represent electronic degrees of freddom, a new model called SWM4-NDP (Lamoureux *et al.* [2006], *Negative Drude Polarization*) was introduced. This is the "default" Drude water in the current version of the force field. When compared to the non-polarizable and widely used TIP3P water, SWM4-NDP better reproduces several thermodynamic properties, especially in situations where a reponse to different environments is needed. In particular, the molecular dipole moment increases from the gas phase to a hydrogen-bonding environment. Another model, called SWM6-NDP (Yu *et al.* [2013]) adds two lone pairs out of the molecular plane. It reproduces water proporties even better than SWM4, but its simulation is more expensive. An important feature of polarizable water is not just related to average properties, but also to dynamics. Polarizable models reproduce water self-diffusion and viscosity better than fixed-charge models and this should give more realistic timescales in molecular dynamics simulations. However, more interaction sites need more parametrization, especially in charged enviroments. One

example is the case of divalent ions, in particular the interaction between SWM4-NDP and Mg^{2+} (Lemkul & MacKerell [2016a]).

3.4 Simulating the Drude force field

To carry out consistent polarizable MD simulations, electronic degrees of freedom $\{\bar{d}\}$ need to correctly relax depending on the nuclei positions $\{\bar{r}\}$ in order to reproduce the response to variations of conformation-dependent electric fields. In the original Drude model electrons were represented by oscillating *massless particles*. For each oscillator there is a unique displacement $\bar{d}_i^\star$ such that the energy is minimized:

$$\frac{\partial U}{\partial \bar{d}_i} = \frac{\partial U_{self}(\{\bar{d}\})}{\partial \bar{d}_i} + \frac{\partial U_{int}(\{\bar{d},\bar{r}\})}{\partial \bar{d}_i} = 0 \tag{3.25}$$

This is referred to as the Self Consistent Field condition, or SCF condition. The SCF condition can be achieved by minimizing the Drude positions at each MD step. This is the most straightforward solution to produce conistent trajectories, and is compatible with massless Drude particles. However, the procedure needs accurate implementation, is computationally expensive and can be affected by numerical errors.

The energy minimization can be avoided by following an alternative procedure. In the current Drude force field, Drude particles are not massless, so their motion can be integrated as the rest of the system. One "physical" temperature T is used to thermalize the real atoms while Drude oscillators are kept at a low temperature $T_D \ll T$ using a dual Nosé-Hoover thermostat. They evolve almost adiabatically in response to nuclear conformational changes which modify the local field. The SCF condition is approximated, following Lamoureux and Roux, using the expansion

$$U(\{\bar{r},\bar{d}\}) = U^{SCF}(\{\bar{r}\}) + \sum_i \delta\bar{d}_i \frac{\partial U}{\partial \bar{d}_i}(d_i^\star) + \frac{1}{2}\sum_{ij} \delta\bar{d}_i \frac{\partial U}{\partial \bar{d}_i}(d_i^\star)\frac{\partial U}{\partial \bar{d}_j}(d_j^\star)\delta\bar{d}_j + ... \tag{3.26}$$

where $d_i^\star$ are the relaxed displacements for which the SCF condition is satisfied, and $\delta\bar{d}_i = \bar{d}_i - d_i^\star$.

3.5 MD implementation

Drude systems are preparated using CHARMM, which implements the routines necessary to transform a system described with an additive force field to a Drude system. Drude topology files have additional keywords: we give, as an example, the SWM4-NDP water topology

```
RESI SWM4           0.000    ! generate using noangle nodihedral
GROUP
ATOM  OH2  ODW      0.00000  TYPE DOH2  ALPHA -0.97825258 THOLE 1.3
ATOM  OM   LPDW    -1.11466
ATOM  H1   HDW      0.55733
ATOM  H2   HDW      0.55733
BOND  OH2  H1
BOND  OH2  H2
BOND  OH2  OM
BOND  H1   H2 ! for SHAKE
ANGLE H1 OH2 H2
LONEPAIR bisector OM OH2 H1 H2 distance 0.24034492 angle 0.0 dihe 0.0
```

As in the additive force fields, the residue is declared using the RESI keyword followed by its name and its charge. The optional keyword GROUP is used to indicate that the following atoms have a total charge equal to zero. Then the residue's atoms are listed, each one using the ATOM keyword followed by its name, its atom type and its charge. Polarizable atoms (in this case the oxygen OH2) have additional keywords, used to indicate the value of their polarizability ALPHA and their THOLE screening parameter. If THOLE is missing, a default value of 1.3 is used. Drude particles attached to polarizable atoms are not explicitly included in topology files, as they are automatically created by CHARMM. Their name is assigned using the letter D followed by the name of its reference atom. For example, OH2 carries a Drude particle of atom name "DOH2". CHARMM automatically assigns the default atom type "DRUD" to all oscillators, except in special cases where a different type is indicated using the optional keyword "TYPE". In SWM4, the oxygen Drude particle has such a special atom type, called "DOH2". This particular choice allow to specify special nonbonded interaction parameters (using NBFIX), explained later in the text. Virtual interaction sites of default atom type LP are indicated in the residue topology. Three atoms are used to define the particle position through the LONEPAIR line. In the example, coordinates of the OM lone pair particle are computed from positions of the OH2-H1-H2 atoms. Like the DOH2 oscillator, OM has its own (non-default) atom type called "LPDW". As in the additive water model TIP3P, SWM4 has an H1-H2 bond used to fix bond lengths and angles during dynamics.

Another Drude keyword called ANISOTROPY is used to specify magnitudes of diagonal elements α_{xx}, α_{yy}, α_{zz} in the polarizability matrix. Since anisotropy depends on the molecule internal geometry, (x,y,z) are defined in a reference frame constructed using positions of three

other atoms in the residue. In the following example, the polarizability of the first oxygen depends on the positions of atoms C, CL and N.

```
ANISOTROPY O C CL N A11 0.697 A22 1.219
```

Using the equation A11+A22+A33=3, matrix elements α_{xx}, α_{yy} and α_{zz} are set scaling ALPHA according to the ratios between A11, A12 and A13.

After reading the Drude topology and parameter files with CHARMM, then the system is built using the CHARMM *generate* command

```
generate @segname first @nterpatch last @cterpatch setup warn drude dmass 0.4
```

The sintax is similar to the one used to construct systems with additive force fields, the only difference is a *drude* keyword that creates vectors to manage Drude particles and LPs. The *dmass* $m_D = 0.4$ a.m.u. is subtracted from polarizable atomic centers and assigned to their Drude particle. Charges are split between atomic centers q_A and oscillators q_D, using the ratio between polarizability and force constant, according to equation 3.19.

Once the coordinates for atomic centers are read in, Drude and Lone particle positions are assigned using the CHARMM commands *coor sdrude* and *coor shake*. Displacements of all oscillators are initially set to zero, requiring minimization before dynamics. This minimization is usually performed first on Drude particles with restrained atomic centers, then relaxing the whole system without restraints. Parameter files are very similar to the ones of the additive force field C36. Internal parameters are speficied in the BOND, ANGLE, DIHEDRAL sections and LJ parameters are specified after the NONBONDED keyword. As in C36, NBFIX is used to modify LJ interactions between two particular atom types i and j, using a special value of r_{min} and e_{min}. A new keyword called NBTHOLE (in the spirit of NBFIX) is used to specify τ_{ij} for atom types i and j. It is usually introduced to balance particular interactions of charged groups, for example between nucleic acids and divalent ions.

The Drude force field is currently implemented in CHARMM, NAMD, OpenMM and GROMACS, even if it is only partially supported by the latter. NVT and NPT simulations can be run with CHARMM and NAMD exploiting the dual thermostat, while GROMACS completely supports just the NVT ensemble. Its implementation for the NPT ensemble is less efficient because it uses the older "SCF minimization" method to relax electronic degrees of freedom. Moreover, GROMACS does not support NBTHOLE parameters. In NAMD, pressure is controled using Langevin dynamics to control fluctuations in the barostat (Martyna *et al.* [1994] ; Feller *et al.* [1995]). Details about extended lagrangian implementation can be

found in Jiang *et al.* [2010]. It is clear that to simulate a system using Drude, extra computing time is demanded. In terms of CPU time, the Drude force field is almost four times more expensive than a non-polarizable force field like Charmm C36. A factor of two is introduced by the many auxiliary particles. SWM4 has four interaction sites, all heavy atoms have an additional Drude particle and several atoms carry lone pairs. In addition, harmonic forces (with the relative self-polarization energies) need to be calculated at each timestep for all electronic degrees of freedom. Another factor of two comes from the dual thermostat: equations of motions need to be integrated with a small timestep, which should never be greater than 1 fs.

As an example, in table 3.1 we compare MD performances of one additive force field (C36) and the Drude force field. We considered a protein:peptide complex, the Tiam1 PDZ domain (89 amino acids) with its natural peptide binder Syndecan1 (8 amino acids). The system was solvated in an octahedral box containing 10139 water molecules. We used TIP3P water for C36 simulations, SWM4-NDP water for Drude simulations. As expected, the number of atoms of each system was proportional to the number of water molecules. TIP3P is composed of three atoms, where SWM4-NDP has four interaction sites plus a Drude particle on the oxygen atom. As a consequence, the C36 system was composed of roughly 30k atoms and the Drude system was composed of roughly 50k atoms (including Drude and lone particles). MD was done using NAMD 2.12. Long-range electrostatic interactions were computed with PME, LJ interactions were shifted to zero for separations between 10 and 12 Å. We used a timestep of 2 fs for C36 and of 1 fs for Drude. Simulations were run on the Occigen supercomputer of the CINES center of the French computer agency GENCI. For both simulations we used 224 cores, distribuited over 8 nodes equipped with 28 processors. As expected, C36 was four time faster than Drude.

Description	C36	Drude
Number of atoms	31 994	53 393
Timestep	2 fs	1 fs
Number of cores	224 (8x28)	224 (8x28)
Performance	75 ns/day	18 ns/day

Table 3.1 – **Drude-C36 comparision** for MD run using NAMD 2.12. System details are given in the text.

Using the extended langrangian which exploit the dual-thermostat protocol described above, polarizable simulations at timescales up to microseconds can be run using massive parallel architectures. However, some considerations about the stability of dynamics are necessary. Since the SCF condition is satisfied *approximately*, there are situations where atoms are in very polar environments and the Drude particle may oscillate too far from the corresponding nucleus.

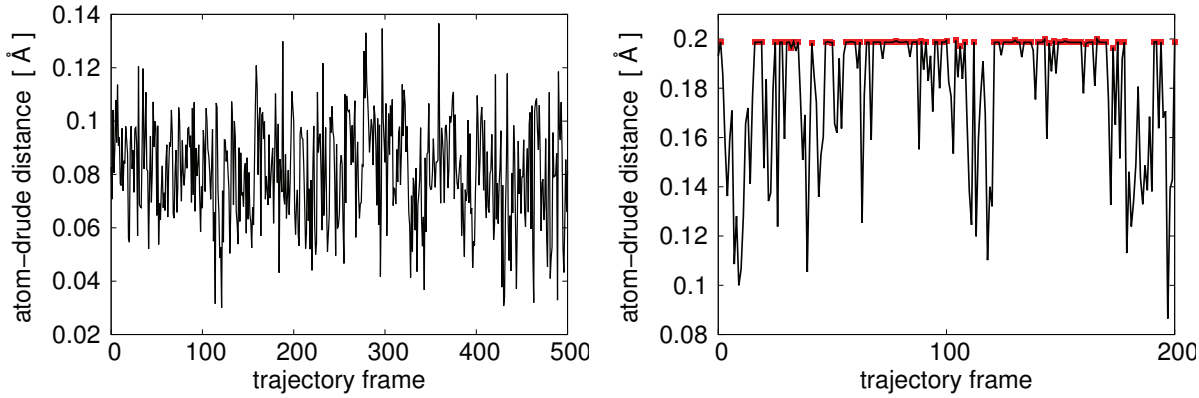


Figure 3.3 – **Atom-Drude distance during MD simulation** is shown in the two plots. **Left**: simulation where the Drude oscillator follows a normal regime, where the distance between q_A and q_D is always lower than 0.2 Å. **Right**: simulation where the Drude oscillator tends to cross the threshold value of 0.2 Å. A *DrudeHardWall* option is applied at frames shown in red.

In a stable system, Drude particles usually oscillate around their reference nucleus at average distance of 0.1 Å [fig. 3.3, left]. However, large forces, arising from strong electrostatic interactions, sometimes lead to the so-called *polarization catastrophe*. In the worst case, the system will be so unstable that the dynamics will be broken and the MD engine will stop the simulation. The polarization catastrophe can be avoided using a smaller timestep to integrate equation of motions, decreasing it to 0.75 fs or less. However, because this represents a further increase of CPU time, an alternative solution called *DrudeHardWall* has been introduced. When a Drude particle exceeds a threshold distance respect to its center, by default equal to 0.2 Å, an hard-wall potential is applied [fig. 3.3, right] and the Drude particle is reprojected along its bond direction. The *negative* displacement is generated using the projection of its velocity along the bond line, flipped and rescaled according to the oscillator temperature T_D. This should be sufficient to correct the trajectory and restore the correct regime. Since DrudeHardwall can be considered as a workaround, it should not be applied for large portions of trajectories, as in the example shown in figure 3.3. Because of its implementation, it leads to unrealistic frames and wrong energetics. The force field has been extensively tested and this problem should not occour too often. However, this is not always true for unofficial, in-house parametrized molecules. To avoid this situation, one may increase screening of electrostatic interactions by adjustment of Thole factors τ_i of particular atoms.

4 Standard and relative binding free energies

Considering a receptor R and a ligand L at equilibrium, their binding is governed by the equation

$$RL \leftrightarrows R + L \tag{3.27}$$

The binding constant (dimensions of a volume) is

$$K_{eq} = \frac{\rho_{RL}}{\rho_R \rho_L} = [V] \tag{3.28}$$

It depends on the concentrations of the complex RL and of the two unbound species R and L. We note that the use of equation 3.28 needs a definition of *bound* and *unbound* states.

We want to connect the binding constant (which can be obtained from experiments) to the corresponding *binding free energy*, more precisely to the free energy that can be obtained from a computer simulation. We use X to indicate the different species in the solution: the ligand L, the receptor R or the complex RL. From the Gibbs free energy change

$$dG = VdP - SdT + \sum_X \mu_X dN_X \tag{3.29}$$

at the conditions of constant temperature and pressure ($dT = 0$ and $dP = 0$) we obtain the change in free energy upon binding of one molecule of R and one molecule of L, to form one molecule of the complex RL

$$\Delta G = \mu_{RL} - \mu_R - \mu_L \tag{3.30}$$

μ_X is the chemical potential of species X,

$$\mu_X = \left(\frac{\partial G}{\partial N_X}\right)_{N_{Y \neq X}, P, T} = -kT \left(\frac{\partial ln Q}{\partial N_X}\right)_{N_{Y \neq X}, P, T} \tag{3.31}$$

For a non-ionic, dilute solution it has a logarithmic dependence on concentration

$$\mu_X(P,T) = \mu_X^0(P,T) + kT \ln \left[\frac{\rho_X}{\rho_X^0}\right] \tag{3.32}$$

$\mu_X^0(P,T)$ depends on P and T and is the *standard chemical potential*. It is defined by an arbitrary *standard state* concentration $\rho_X^0 = \rho^0$ for all chemical species X, so that

$$\Delta G = \mu_{RL}^0 - \mu_R^0 - \mu_L^0 + kT\left(\frac{\rho_{RL}}{\rho^0}\right) - kT\left(\frac{\rho_R}{\rho^0}\right) - kT\left(\frac{\rho_L}{\rho^0}\right) = \Delta G^0 + kT\ln\left[\frac{\rho^0\rho^{RL}}{\rho^R\rho^L}\right] \quad (3.33)$$

We note that the right term in the formula depends on K_{eq}. The formula holds for a hypothetical equilibrium state where concentrations are constrained to be constant; we call it *reaction free energy*. Without constraints, the system goes to a state where $\Delta G = 0$ and so

$$\Delta G^0 = -kT\ln\left(\frac{\rho^0\rho_{eq}^{RL}}{\rho_{eq}^L\rho_{eq}^R}\right) = -kT\ln\left[\rho^0 K_{eq}(p,T)\right] \quad (3.34)$$

The standard chemical potential of equation 3.32 can be computed from the ratio of two partition functions. Indeed, we can write

$$\mu_X^0 = -kTln\frac{Z_{N,X}(V_{N,X})}{Z_{N,0}(V_{N,0})V_{N,X}\rho^0} + P^0\bar{V}_X \quad (3.35)$$

where $Z_{N,X}$ is the canonical partition function of a system containing one molecule of solute X solvated in N solvent molecules at volume $V_{N,X}$. $Z_{N,0}$ is the partition function of the same system without X at volume $V_{N,0}$. The two volumes are the *equilibrium volumes* of the two systems (with or without solute) when they are mantained at pressure P^0. $\bar{V} = V_{N,X} - V_{N,0}$ is the change in equilibrium volume when a molecule of solute X is added to the system with N solvent molecules, or partial molar volume.

In equation 3.34, the free energy has a logarithmic dependence on the binding constant K_{eq}. We also note that, in the equation 3.28, K_{eq} has dimensions of a volume and the logarithm argument in equation 3.34 is dimensionless. This trivial argument based on dimensional analysis (General [2010]) is foundamental to note that computer simulations and experiments may divide k_{eq} by two different concentration factors. The choice of standard state concentration is arbitrary, it is usually set to be 1M. It is important, when comparing K_{eq} obtained from an experiment and the one obtained from a simulation, that both equilibrium constants (and so free energies) refer to the same standard state. The relationship between the free energy difference ΔG^0 and the one measured in a simulation at $\rho = \rho^{comp}$ (which is usually run at a different concentration) is

$$\Delta G^0 = \Delta G^{comp} + kT\ln\left[\frac{\rho^{comp}}{\rho^0}\right] \quad (3.36)$$

For $\rho^0 = 1M$, a cubic simulation box with one solute molecule has a volume $V \simeq 1661$ Å^3, for $L \simeq 11.84$ Å. This is smaller than box size usually used to solvate proteins, and lead to correction terms that are significative: 1.9 kcal/mol for a box of $L = 30$ Å, to 3.9 kcal/mol for a box of $L = 90$ Å.

The above formulation is exact with classical mechanics and dilute solutions. Another requirement of [eqn. 3.34] is that the binding involves *nonionic* solutes, otherwise another correction must be applied. Indeed, the chemical potential of ions can be expressed analytically in the limit of infinite dilution (Hill, [1992])

$$\mu_X = \mu_X^0(p,T) + kT \ln \left[\frac{\rho_X}{\rho_X^0}\right] + kT \ln \left[\gamma_X(\rho_X,p,T)\right] \tag{3.37}$$

At infinite dilution and small ionic strength κ the activity coefficient γ can be approximated:

$$kT \ln \left[\gamma_X(\rho_X,p,T)\right] \simeq \kappa \frac{q_X^2}{2\epsilon_w} \qquad \kappa^2 = \frac{4\pi}{kT\epsilon_w}\left(q_R^2\rho_L + q_L^2\rho_L\right) \tag{3.38}$$

Even if we know an analytical expression for standard binding free energies [eqn. 3.34] which has a single standard-state dependence, more often we are interested in computing the *relative binding affinity* between a receptor R and two different ligands L and L'.

$$\Delta\Delta G^0(L \to L') = \Delta G^0(RL') - \Delta G^0(RL) = -kT\left\{\ln\left[\frac{Q_{RL'}}{Q_{RL}}\right] - \ln\left[\frac{Q_{L'}}{Q_L}\right]\right\} \tag{3.39}$$

We note that the standard state dependence cancels out and the calculation corresponds to measuring ratios of configuration integrals in bound and unbound states.

5 Free energy perturbation

From the last equations we note that free energy differences are obtained from the *ratio of configuration integrals*. In general, one is interested in the free energy change between two states A and B

$$\Delta G(A \to B) = -kTln\frac{Q_B}{Q_A} \tag{3.40}$$

For a system of hamiltonian

$$H = T + U \tag{3.41}$$

the canonical partition function is defined as

$$Z(N,V,T) = \int \exp\left[-\frac{H(x,p)}{kT}\right] dx^N dp^N \tag{3.42}$$

It is proportional to the configuration integral

$$Q(N,V,T) = \int \exp\left[-\frac{U(x)}{kT}\right] dx^N \tag{3.43}$$

In biological applications we are more often interested in the isothermal-isobaric ensemble where the configuration integral is

$$Q(N,P,T) = \int \exp\left[-\frac{U(x)+PV}{kT}\right] dx^N dV \tag{3.44}$$

The exact calculation of two partition functions Q_B and Q_A from computer simulations is normally too expensive. However, one can consider states A and B which are *close*, connected by a small perturbation in the potential function

$$U^{'} = U + \epsilon \tag{3.45}$$

The corresponding free energy difference

$$F^{'} - F = -kT \ln\left[\frac{Z^{'}}{Z}\right] = -kT \ln \frac{\int \exp\left[-\frac{U(x)}{kT}\right] \exp\left[-\frac{\epsilon(x)}{kT}\right] dx^N}{\int \exp\left[-\frac{U(x)}{kT}\right] dx^N} \tag{3.46}$$

can be obtained from

$$F^{'} - F = -kT \ln\left\langle \exp\left[-\frac{\epsilon(x)}{kT}\right]\right\rangle_U \tag{3.47}$$

For a perturbation ϵ the free energy difference $F^{'} - F$ can be measured as an average over an ensemble of configurations, collected from a canonical simulation carried out with potential U(x).

The formula [eqn. 3.47] holds for small perturbations, where ensembles generated with unperturbed and perturbed potentials overlap. In general, for perturbations larger than one kcal/mol the formula cannot be used directly. To solve this problem, a good strategy is to split the path between U and $U + \epsilon$ into several smaller steps $\epsilon = \sum_i \epsilon_i$ (called *windows*) and

then obtain the total free energy change as a sum of smaller differences. A generalization of this step-wise procedure consists in the definition of the energy function

$$V(\lambda) = (1-\lambda)V_A + \lambda V_B \tag{3.48}$$

where A and B are two distinct thermodynamic states of the same conformational space. The parameter λ can vary continuously between 0 and 1, connecting the two states. This situation is easily implemented in MD: when $\lambda = 0$ the interactions of B are turned off, while for $\lambda = 1$ the interactions of A are turned off. For intermediate values, the system is in a hybrid thermodynamic state between A and B. This kind of transformation with unphysical states is called *alchemical*, as it cannot be reproduced in the real world. Using the energy function [eqn. 3.48] one can split the path between any two states A and B into many, small windows characterized by a different value of λ and then obtain the free energy difference as sum of perturbations

$$\Delta F(A \to B) = F_B - F_A = \sum_i \Delta F(\lambda_i \to \lambda_{i+1}) \tag{3.49}$$

5.1 Thermodynamic integration

The derivative of F respect to λ can be computed:

$$\frac{\partial F(\lambda)}{\partial \lambda} = \frac{\int \frac{\partial U(x,\lambda)}{\partial \lambda} \exp\left[-\frac{U(x,\lambda)}{kT}\right] dx^N}{\int \exp\left[-\frac{U(x,\lambda)}{kT}\right] dx^N} = \left\langle \frac{\partial U(x,\lambda)}{\partial \lambda} \right\rangle_{U(x,\lambda)} \tag{3.50}$$

The conformations used to calculate the average value of $\partial U/\partial\lambda$ are collected from a simulation performed at fixed λ. Calculating the derivative for several values of the coupling parameter between 0 and 1, one can integrate the results and obtain the desired free energy difference

$$\Delta F(A \to B) = F_B - F_A = \int_0^1 \frac{\partial F}{\partial \lambda} d\lambda = \sum_{i=0}^{N-1} \int_{\lambda_i}^{\lambda_{i+1}} \frac{\partial F}{\partial \lambda} d\lambda \tag{3.51}$$

In general, several numerical integration schemes can be used to estimate the integral. For a simple, trapezoid approach

$$\Delta F(\lambda_i \to \lambda_{i+1}) = \frac{1}{2}(\lambda_{i+1} - \lambda_i)\left[\frac{\partial F}{\partial \lambda}(\lambda_{i+1}) + \frac{\partial F}{\partial \lambda}(\lambda_i)\right] \tag{3.52}$$

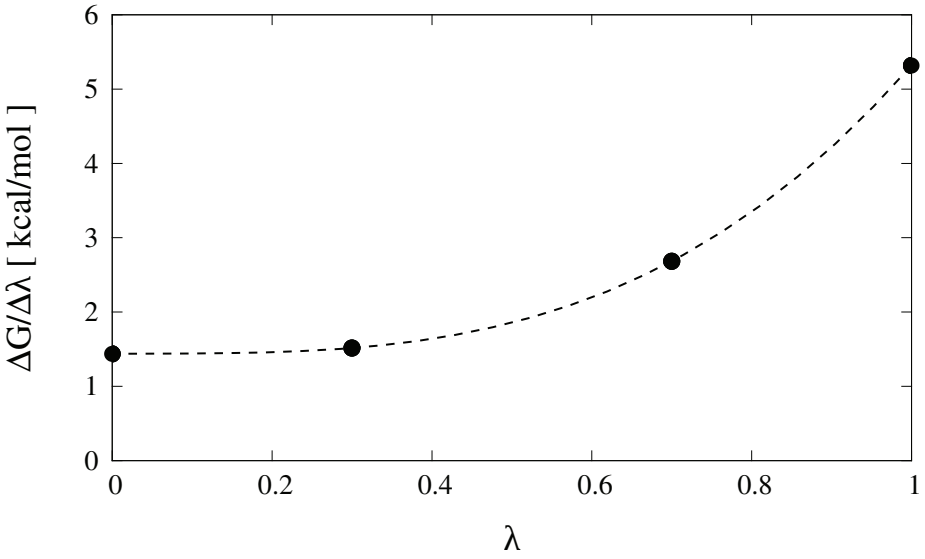


Figure 3.4 – **Interpolation of derivatives** in thermodynamic integration. Calculated derivatives are represented with circles, the interpolated curve using a dashed line.

This estimation can be inaccurate. A better interpolation scheme is the *cubic spline*, where the function $\partial G/\partial\lambda$ is interpolated using a polynomial [fig. 3.4], then integrated.

5.2 Bennet acceptance ratio

The BAR method was introduced in 1976 (Bennett [1976]). It estimates the ratio of configuration integrals Q_A/Q_B between two states defined by potential U_A and U_B, following a rule inspired by Monte Carlo simulations. A detailed formulation can be found in the original paper, here we summarize the main concepts. The idea is to consider a hypothetical MC simulation where the trial move does not change the system configuration but it switches the potential function from U_A to U_B. For this move, equilibrium requires that

$$M(U_B - U_A)\exp(-U_A) = M(U_A - U_B)\exp(-U_B) \tag{3.53}$$

where M is the Metropolis acceptance probability $M(x) = min(1,exp(-x))$ that gives Boltzmann probability. Integrating both sides of the equation and multiplying each term by 1

$$\frac{Q_A}{Q_A}\int_\Omega M(U_B - U_A)\exp(-U_A)dq = \frac{Q_B}{Q_B}\int_\Omega M(U_A - U_B)\exp(-U_B)dq \tag{3.54}$$

one obtains the ratio

$$\frac{Q_A}{Q_B} = \frac{< M(U_A - U_B) >_{U_B}}{< M(U_B - U_A) >_{U_A}} \tag{3.55}$$

As in the simple free energy perturbation method, the ratio between configuration integrals is computed using ensemble-averages. The formula can be generalized for a function f that satisfies $f(x) = f(-x)$; the free energy difference between states A and B is then estimated using

$$\Delta F_{AB} = kT \ln \frac{< f(U_A - U_B + C) >_{U_B}}{< f(U_B - U_A + C) >_{U_A}} + C \tag{3.56}$$

where $f(x) = 1/(1 + exp(x/kT))$ is the Fermi function. C is equal to

$$C = \ln \frac{Q_A N_B}{Q_B N_A} \tag{3.57}$$

where N_A and N_B are the number of configurations for the two states identified with U_A and U_B. The constant C is determined with an iterative procedure, until the convergence criterion $f(U_A - U_B + C) = f(U_B - U_A) + C$ is satisfied. It is easy to show that if $N_A = N_B$, after convergence one has

$$\Delta F_{AB} = C \tag{3.58}$$

5.3 BAR and TI with Drude

To obtain the free energy difference between two simulation windows with BAR, one has to compute ensemble-averages. System configurations are extracted from the trajectories of separate MD simulations, that were performed using two different potentials $U(\lambda)$ and $U(\lambda + \delta\lambda)$. Then two ensemble-averages are computed using equation 3.56

$$\langle f(U(\lambda + \delta\lambda) - U(\lambda) + C) \rangle_{U(\lambda)} \qquad \langle f(U(\lambda) - U(\lambda + \delta\lambda) + C) \rangle_{U(\lambda+\delta\lambda)} \tag{3.59}$$

In practice, for each MD snapshot extracted from the trajectory obtained using $U(\lambda)$, one has to compute the potential energy with $U(\lambda)$ and also with $U(\lambda + \delta\lambda)$. However, the Drude particle displacements of these snapshots $\{d_i^\star(\lambda)\}$ satisfy the SCF condition $\partial U/\partial d$=0 for the potential $U(\lambda)$ but not for the potential $U(\lambda + \delta\lambda)$. For this reason, the application of BAR with Drude is not formally correct. Before to compute $U(\lambda + \delta\lambda)$ for a shapshot of the $U(\lambda)$ trajectory, one should minimize the Drude particles using $U(\lambda + \delta\lambda)$. However, this solution is not practicable since it is too expensive.

The problem does not persist if free energy differences are obtained with TI, where one has to compute derivatives:

$$\frac{\partial F(\lambda)}{\partial \lambda} = \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{U(\lambda)} \tag{3.60}$$

In practice, the derivative $\partial U/\partial\lambda$ can be obtained by a finite-difference method, by taking the difference between energies computed at two slightly different λ values:

$$\frac{\partial U}{\partial \lambda} = \frac{U(\lambda + \delta\lambda) - U(\lambda)}{\delta\lambda} \tag{3.61}$$

Thus, with TI, we compute energies at, or very close to the λ value used to produce the trajectory. In this case, the relaxed displacements are supposed to be similar $d_i^\star(\lambda+\delta\lambda) \simeq d_i^\star(\lambda)$ and the SCF condition is still well approximated. Indeed, evaluating the Taylor expansion of equation 3.26 for small δd_i the first derivative is equal to zero, leaving a quadratic term that is negligible when

$$\delta d_i = d_i^\star(\lambda + \delta\lambda) - d_i^\star(\lambda) \simeq O(\delta\lambda^2) \tag{3.62}$$

Few comments can be made. First, we note that while BAR is not formally correct with Drude, it can be applied to the AMOEBA polarizable model, where the induced dipoles are automatically adjusted whenever the parameter λ is changed. Second, Drude simulations are run in the "*Cold oscillators regime*" where the oscillators are thermalized at a low temperature T_D in order to approximate the SCF condition $\partial U/\partial\bar{d} = 0$. When combined with a barostat, MD samples a (N,P,T,T_D) ensemble. Several studies describe methods to compute free energies for similar dual-temperature systems, where some degrees of freedom evolve adiabatically respect to others equilibrated at room temperature (Rosso *et al.* [2002] ; Maragliano & Vanden-Eijnden [2006] ; Abrams & Tuckerman [2008] ; Cuendet & Tuckerman [2014]). Even if their application for Drude is not straightforward, the problem could be studied mode in detail in the future.

5.4 Alchemical transformation

The calculation of binding free energies for the reaction [eqn. 3.27] has been extensively discussed in literature. The article from Boresch *et al.* [2003] includes an overview about techniques used to obtain absolute binding free energies, with a discussion about the standard state dependence. In figure 3.5 we show the thermodynamic cycle corresponding to the *Double Decoupling method*. It relies on the decomposition of the binding reaction

$$R_s + L_s \xrightarrow[\text{bind}]{\Delta G} (RL)_s \tag{3.63}$$

into two transformations

$$L_s \xrightarrow[1]{\Delta G} L_g$$
$$(RL)_s \xrightarrow[2]{\Delta G} R_s : L_g \xrightarrow[\text{restr}]{\Delta G} R_s + L_g \tag{3.64}$$

the subscript s is used to indicate the solvated system. In the first reaction, the ligand is decoupled from the solvent and transferred to the gas phase (indicated using the subscript g). The free energy cost for this transformation is called ΔG_1. The second reaction is composed of two parts. First, the ligand is transferred to the gas phase under specitic *restraints* introduced in order to keep L bound to R, even when their interactions are turned off. The free energy cost for this reaction is called ΔG_2. The endpoint is called R_s:L_g to indicate that R and L are bound just because of the introduced restraints. Then, the restraint potential is removed and the ligand is fully transferred to the gas phase. The free energy cost for this reaction is called ΔG_{restr}. The the energy to apply the restraint in the bound state $(RL)_s$ is usually neglected. Finally,

$$\Delta G_{bind} = \Delta G_2 + \Delta G_{restr} - \Delta G_1 \tag{3.65}$$

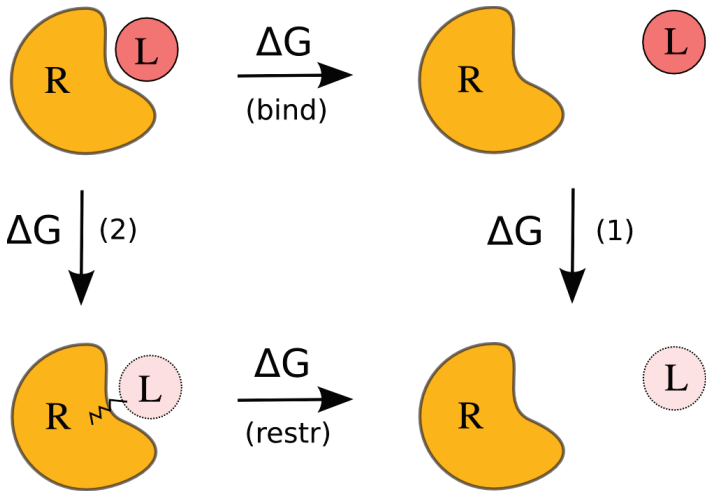


Figure 3.5 – **Absolute binding free energy** (ΔG_{bind}) thermodynamic cycle. **Top, left**: bound state in solution $(RL)_s$. **Top, right**: unbound state in solution, $R_s + L_s$. **Bottom, left**: restrained-bound state $R_s : L_g$ with ligand in the gas phase. **Bottom, right**: unrestrained-unbound state $R_s + L_g$ with the ligand in the gas phase.

To compute the free energy necessary to remove the restraint between receptor and ligand, one can use a simulation that starts from the endpoint of reaction (2), where the restraint potential is gradually turned off. In practice, this is always complicated (Boresch *et al.* [2003]). The most straightforward method to obtain this free energy difference is to compute it ana-

lytically: for a cubic box, $\Delta G_{restr} = -kTln(Q_{restr}/L^3)$ where $V = L^3$ is the volume of the simulation box. For example, for the simple case of an ion linked to the receptor with an harmonic spring one can estimate the configuration integral exploiting the spherical symmetry

$$Q_{restr} = \int_0^{+\infty} 4\pi r^2 dr \exp\left[-\beta U_R(r)\right] \tag{3.66}$$

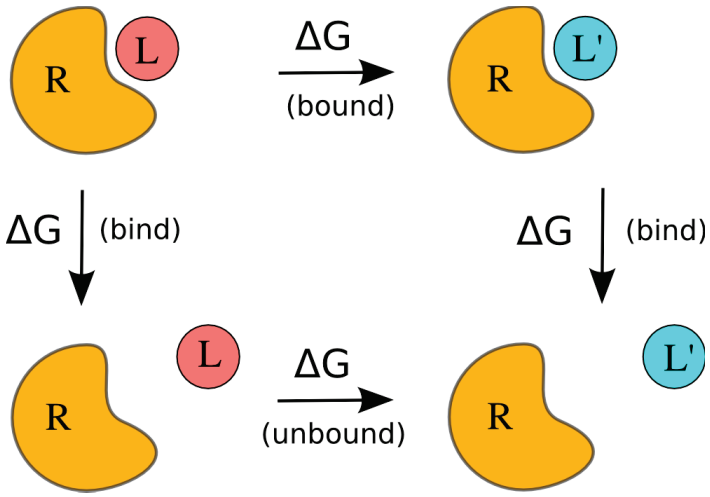


Figure 3.6 – **Relative binding free energy** $\Delta\Delta G(L \rightarrow L^{'})$ thermodynamic cycle. The transformations endpoints represent bound and unbound states with two different ligands L and $L^{'}$.

6 Artefacts in charging free energy calculations

When comparing simulation results with theory and experiments, several artefacts must be taken into account. In most cases, electrostatic interactions are computed on a periodic, infinite lattice using Particle Mesh Ewald (PME). As a result, alchemical free energies differ respect to the real quantities of interest depending on box size, charge distributions, the presence of counterions, and the nature of the solvent (Rocklin *et al.* [2013]). In general, one has to compute corrections such that the *real* free energy change is obtained from the *alchemical* free energy difference plus a term which can be computed analitically or numerically, depending on the situations. In few situations simulations need to be postprocessed, for example using continuum electrostatics to estimate several artificial contributions. Otherwise, a well-established theoretical background allows to estimate some of these corrections analytically.

6.1 PME with tinfoil boundary conditions: neutralizing gellium

We consider a simulation box and its infinite periodic images, where the total net charge of each box $Q \neq 0$. To avoid the (unphysical) situation where the infinite, periodic system has a divergent potential, the simulation box must be neutral. A possible strategy consists in introducing one (or more) counterions to balance the total net charge Q. For large boxes, the counterions will be sufficiently isolated, without affecting the simulation interacting with the system or perturbing each other's solvation. However, a more straightforward and practical solution is to introduce a uniform compensating charge density. For each charge q inserted in the system, a neutralizing charge $-q$ is spread uniformly over the entire box volume. This charge density is called the *gellium*. Its charge density depends on the volume of the box: $\rho = -q/V = -q/L^3$. It generates a uniform potential that now we call ϕ^{gel}. Since $\nabla\phi^{gel}(x) = 0$ for all points inside the box, the gellium does not introduce an electric field in the MD simulation or contribute to the forces. However, the role of the gellium is to shift the electrostatic potential ϕ in such a way that its average over the entire simulation box is zero:

$$\langle\phi(x)\rangle_{box} = 0 \tag{3.67}$$

To see this, we consider a charge q inserted in the system. It will give rise to infinite periodic images and an associated gellium, so that at a point r inside the box, the associated charge density is

$$\rho(r) = q\sum_{i=0}^{\infty}(\delta(r - r_i) - L^{-3}) \tag{3.68}$$

where r_i is the position of the perturbing charge or one of its images. The energy change per box is

$$\Delta U = \int_{box} dr\rho(r)\phi(r) = \int_{box} drq\left[\delta(r - r_i) - L^{-3}\right]\phi(r) = q(\phi(r_i) - \langle\phi\rangle_{box}) \tag{3.69}$$

The perturbing charge q thus experiences a shifted potential $\phi' = \phi(r_i) - \langle\phi\rangle_{box}$. In effect, the gellium has shifted the mean box potential ϕ to zero; in other words, the system is electrically grounded.

The potential shift $\langle\phi\rangle_{box}$ depends on the nature of the solute and the volume fraction of the solvent in the box (Lin *et al.* [2014]). When free energy calculations are performed exploiting thermodynamic cycles, different simulations can be run in different simulation boxes. Since different simulation boxes can be shifted by a different potential $\langle\phi\rangle_{box}$, this effect must be taken into account. One example is the calculation of relative binding affinities illustrated

in figure 3.6, where different shifts can artificially overstabilize bound or unbound states. A practical example to illustrate how we can correct for this effect will be given in the next chapter.

6.2 Solvent polarization artefacts in a periodic system

Other artefacts are introduced by the the fact that electrostatic interactions are calculated in a periodic system. We consider the simple case of a spherical ion placed at the center of a periodic, cubic box of size L. Because of periodicity, the ion will interact with its periodic images and the neutralizing gellium. These interactions arising from the lattice summation are also called ion *self-energy* and are related to the box volume through a $1/L$ dependency. The self energy must be considered when, computing alchemical charging free energies $\Delta G(L)$, one wants to extrapolate the so called *large box limit* $\Delta G(L \to \infty)$ where artefacts are removed. As pointed out by Hummer *et al.* [1996], the self contribution to the charging free energy of a small spherical ion of charge q is

$$\Delta G_{self} = q^2 \frac{\xi}{2L} \tag{3.70}$$

where $\xi = -2.837297$ is the Wigner constant for a cubic lattice. The term includes interactions between the ion, its images and the gellium. However, ΔG_{self} is not sufficient to remove all the artefacts. As observed by Figueirido *et al.* [1997], the total charging free energy has a weaker dependence on the simulation box volume because of the solvent screening. For high dielectric solvents, the real deviation between $\Delta G(L)$ and $\Delta G(L \to \infty)$ is smaller. It depends on the solvent dielectric constant and is proportional to $q^2/(\epsilon_W L)$. Another finite-size artefact can be removed considering another correction term which scales as L^{-3} (Hummer *et al.* [1997]), valid for spherical ions of radius R. The final correction formula that removes all these artefacts is

$$\Delta G(L \to \infty) = \Delta G(L) + \frac{q^2 \left|\xi\right|}{2\epsilon_W L} + \frac{2\pi q^2 R^2}{3L^3} \frac{\epsilon_w - 1}{\epsilon_w} + O(R^2 L^{-3} \epsilon_w^{-3}) \tag{3.71}$$

To obtain charging free energies in kcal/mol, the Wigner constant must be multiplied by a factor to convert $e^2/\text{Å}$ (in CHARMM this is the CCELEC constant = 332.0716). More information can be found in Simonson & Roux [2016]. Here we report some important observations. First, the charging free energy extracted from a simulation $\Delta G(L)$ is smaller than the one obtained in the large box limit. This is intuitive considering that the positive ion will interact with the neutralizing gellium, which is a uniform negative charge density spread over the simulation box volume. Second, this analytical formula is valid for spherical ions in cubic boxes. For more complex solutes and different simulation boxes, the correction must be computed using continuum theory. Numerical schemes can be found in Rocklin *et al.* [2013]. The

correction is obtained by computing a charging free energy for a periodic and a non-periodic system and then taking the difference:

$$\Delta G(L \to \infty) = \Delta G(L) + \Delta G_{PE}(L \to \infty) - \Delta G_{PE}(L) \tag{3.72}$$

Chapter 4

PDZ-peptide binding specificity with polarizable free energy simulations

1 Introduction

Despite the abundant experimental data, our understanding and ability to rationally engineer PDZ/peptide interactions is limited. For the best-studied domains, experiments have revealed the binding thermodynamics of 20-40 protein or peptide variants, but only a few x-ray structures are available. Even with x-ray structures, many details are unclear. Protein:peptide binding affinity and specificity depend on many factors, difficult to quantify using experiments. These are short-range interactions in the binding pocket, longer-range electrostatic interactions, dielectric shielding by protein and solvent, ordered waters in the binding site, the structure and flexibility of the unbound peptide, and the conformational dynamics of the receptor.

Computer simulations are a complementary tool and several approaches have been proposed to predict the binding energetics for protein/peptide complexes. Some of them are semi-empirical and use a Linear Interaction Energy or "PB/LIE" free energy function based on molecular mechanics plus an implicit solvent. This function has several contributions (usually van der Waals, continuum electrostatics and a solvent accessible area term) weighted by different factors, which are optimized to reproduce binding affinities for a training dataset.

Of particular interest are more expensive, nonempirical methods that predict binding free energy differences from molecular dynamics (MD) simulations without any adjustable parameters. These methods are formally exact and are not optimized for a particular test system. Nonempirical methods are based on thermodynamic perturbation theory, and extract free

energies from series of simulations where a mutating side chain of interest is alchemically mutated from one type to another. This class of methods is often referred to as free energy perturbation or FEP (Zwanzig [1954] ; Straatsma & McCammon [1992] ; Kollman [1993]).

There are mainly two sources of errors that can influence the prediction of binding free energies with the FEP. The first is related to sampling. FEP is expensive and usually needs long simulations which may last several hours or days, even when performed on a supercomputer. Sampling requirements may be especially high for ionic mutations, since they involve long range interactions and long timescales that characterize dielectric relaxation in proteins (Simonson [2003]). A second source of errors is the molecular mechanics force field used to run MD simulations and to extract the free energy differences. As anticipated in the last chapter, a strong limitation is the simple treatment of electronic polarization modeled implicitly through a fixed set of atomic partial charge, as in most of the widely-used, "additive" force fields. The use of a polarizable force field could be especially important when comparing ligands that differ by one or more ionic side chain mutations. However, when using more sophisticated models the computational time needed to accomplish sufficient conformational sampling is increased.

Experimental data is known for 50 Tiam1:peptide complexes, 25 of them are peptide variants (Shepherd *et al.* [2010, 2011] ; Liu *et al.* [2013]). Binding free energies are within a 2.2 kcal/mol range, with mean experimental uncertainty of 0.2 kcal/mol. Some peptides have high or non measurable dissociation constant. Previous FEP calculations were run using Amber ff99SB and produced good results in the prediction of relative binding affinities for six nonionic mutations. Results are listed in our recently published paper (Panel *et al.* [2018]). Errors were low, with mean unsigned error (MUE) of 0.32 kcal/mol, root mean square deviation (RMSD) of 0.37 kcal/mol and correlation coefficient of 0.84. For other six ionic mutations, mean errors with the additive force field were larger, and there was one very large error of 3 kcal/mol [tab. 4.1]. Among these six mutants, four of them displayed errors of more than 1 kcal/mol: three are mutations of the peptide position P_4 *E4K*, *E4L*, *K4L* and one mutation was in the protein *K912E*. We use italics to indicate mutations, for example *E4K* will refer to the alchemical Sdc1→ Sdc1-E4K mutation.

In the present chapter, we report alchemical free energy simulations of the Tiam1 PDZ domain using two additive force fields (Amber ff99SB and Charmm C36) and the Drude polarizable force field. Our main goal was to determine the accuracy of free energy perturbation (FEP) in characterizing the binding energetics of PDZ/peptide complexes. We were also interested in determining if the explicit threatment of electronic polarization can improve results for ionic mutations, for which additive force fields are known to be imprecise. In this work, we studied four ionic mutations using the Charmm additive force field C36 and the Drude pola-

rizable force field, in order to investigate if the explicit threatment of electronic polarizability can improve predictions. All calculations (with polarizable or additive force field) relied on the *Dual Topology Approach*. To our knowledge, FEP has never been used to study protein:ligand binding free energy changes with the Drude polarizable force field.

Simulations were done using periodic boundary conditions and particle mesh Ewald (PME) summation for long-range electrostatic interactions (Darden [2001]). Special care was taken to extrapolate the computed free energies correctly to the thermodynamic limit of an infinitely large simulation box. Applying the Drude polarizable force field to the four ionic mutations with the largest errors, two of the errors were significantly reduced and the largest one decreased from 3.0 to 1.3 kcal/mol. Thus, the polarizable force field led to a good accuracy for the ionic mutations, although not as good as that previously obtained for the nonionic mutations with ff99SB. Overall, we have established that FEP can be used successfully to understand and potentially engineer PDZ/peptide binding specificity, and that a polarizable force field is necessary to obtain good accuracy for some of the ionic mutations. The FEP methodology can thus be used in a predictive way, in synergy with experiment and other computational approaches, to design PDZ/peptide specificity.

2 Methods

2.1 Structural models and simulation setup

Complexes involving mutants of the Tiam1 PDZ domain or the Sdc1 peptide were built from the wild-type/Sdc1 x-ray structure (pdb access code 4GVD) using the SCWRL4 program (Krivov *et al.* [2009]). For each complex, the mutated side chain and nearby side chains were positioned by SCWRL4, with the rest of the structure held fixed. There were six missing residues in the wild-type/Sdc1 x-ray structure, in the exposed β1-β2 and β2-β3 loops, which were rebuilt using the program MODELLER (Fiser *et al.* [2000]). The unbound peptides were all modeled by extracting the bound conformation from the corresponding complex.

Simulation setup was carried out using the CHARMM program. The system was first built using the C36 force field. Each complex or unbound peptide was immersed in a cubic box of size L =80 Å of pre-equilibrated TIP3P water, deleting solvent molecules within the volume occupied by the solute, using a distance threshold of 3.0 Å from the complex heavy atoms. Then the box was trimmed, from cubic to truncated octahedron, deleting solvent molecules at the cube edges. The wild type/Sdc1 system was neutralized using sodium ions. Solvated systems were equilibrated using CHARMM and the additive force field C36, in a step-wise procedure.

500 ps of dynamics were run increasing timestep and system temperature, decreasing initial restraints applied to the solute heavy atoms. Then 40 ns of MD were run. After this first, "additive" equilibration, Drude particles and Lone pairs were added using a CHARMM script based on the *drude-prepper* tool of the CHARMM-GUI (Jo *et al.* [2008, 2016]). The script reads coordinates of the additive system, generates the Drude structure file (PSF) and centers each Drude particle at the position of its reference atom. At this point, it was necessary to relax Drude particles, since the generated Drude positions were unphysical. A first energy minimization was run (200 steps) to relax the system with restrained centers and unrestrained Drude particles. A second minimization (100 steps) was run relaxing the whole system without restraints. Starting from the minimized conformation, 20 ns of NPT equilibration followed by 100 ns of production dynamics were run. Minimization and MD were run using NAMD 2.12. Molecular dynamics simulations with the C36 and Drude force fields were done at room temperature and pressure, using Langevin dynamics with a Langevin Piston Nosé-Hoover barostat (Feller *et al.* [1995]). Long-range electrostatic interactions were treated with a PME approach with tin-foil boundary conditions (Darden [2001]). With Drude, the MD timestep was 1 fs, instead of 2 fs used for the simulations with C36. The Drude particles had the default mass of 0.4 atomic units, and their temperature was controlled with a Langevin thermostat with $T_D = 1$ K.

2.2 Alchemical MD simulations

The FEP approach (Simonson *et al.* [2002] ; Tuckerman [2007] ; Jorgensen & Thomas [2008]) was used to calculate the binding free energy differences between variants A and B using equation 3.39. Two separate alchemical simulations were performed. In the first simulation, the mutating side chain was alchemically transformed in the bound state, represented by the protein:peptide complex in water solution. In the second simulation, the transformation was performed in the unbound state, represented by the unbound peptide (or unbound protein) in solution. From these two simulations, we extracted free energy differences $\Delta G_{bound}(A \to B)$ and $\Delta G_{unbound}(A \to B)$. The relative binding free energy for a generic mutation $A \to B$ was then obtained by difference

$$\Delta\Delta G_{bind}(A \to B) = \Delta G_{bound}(A \to B) - \Delta G_{unbound}(A \to B)$$

Let us assume we are comparing the wild type Sdc1 peptide to one of its point mutants. For this pair, we consider a hybrid peptide, which has two side chains at the mutated position, one of each type. The energy function U depends on a coupling coordinate λ that scales selected electrostatic and van der Waals energy terms. When $\lambda = 0$ (respectively, 1), the

mutant (respectively, wild type) side chain is decoupled from its surroundings, retaining only its covalent interactions with its own backbone. For intermediate λ values, both side chains are present, with intermediate weights. U has the form

$$U(\lambda) = U_{LJ}(\lambda) + U_{elec}(\lambda) \tag{4.1}$$

Considering two atoms i and j at distance r_{ij}, their Lennard Jones interactions are modified using a coupling parameters λ_{LJ}

$$U_{LJ}(r_{ij};\lambda_{LJ}) = \lambda_{LJ}\epsilon_{ij}\left[\left(\frac{R_{ij}^{min,2}}{r_{ij}^2+\delta(1-\lambda_{LJ})}\right)^6 - \left(\frac{R_{ij}^{min,2}}{r_{ij}^2+\delta(1-\lambda_{LJ})}\right)^3\right] \tag{4.2}$$

The electrostatic potential was modified using another coupling paramer λ_{elec}. For simulations with C36:

$$U_{elec}(\lambda) = \lambda_{elec}\frac{q_i q_j}{r_{ij}} \tag{4.3}$$

while for the Drude polarizable force field were scaled charge-charge, charge-dipole and dipole-dipole interactions:

$$U_{elec}(\lambda) = \lambda_{elec}\left[\frac{q_{A,i}\cdot q_{A,j}}{|\bar{r}_{A,i}-\bar{r}_{A,j}|} + \frac{q_{D,i}\cdot q_{D,j}}{|\bar{r}_{D,i}-\bar{r}_{D,j}|} + \frac{q_{D,i}\cdot q_{A,j}}{|\bar{r}_{D,i}-\bar{r}_{A,j}|} + \frac{q_{A,i}\cdot q_{D,j}}{|\bar{r}_{A,i}-\bar{r}_{D,j}|}\right] \tag{4.4}$$

The van der Waals interactions are thus modified using a soft-core functional form, ensuring a gradual transformation. The shifting parameter δ was set to its default value in NAMD $\delta = 5$. Proceeding in this way, clashes at endpoints are avoided when λ is close to 0 or 1 (Beutler *et al.* [1994] ; Zacharias *et al.* [1994]).

Calculations were done using the *alch* facility of NAMD. When performing simulations with the Drude force field, particular attention is also needed to balance appearing and disappearing electrostatic interactions. This is related to the stability of MD simulations, since one has to avoid the *polarization catasfrophe* occurring when a Drude oscillator deviates too much from its reference atom. In NAMD, simulations often become unstable when one atom-Drude distance exceeds the default threshold of 0.2 Å. We observed that when the LJ repulsion between a group of atoms is annihilated *too early* compared to electrostatics, positive charges belonging to nearby groups can get too close to the negative Drude charges, attracting them away. As a consequence, these oscillators exceed the threshold distance and NAMD prints an error message interrupting the MD simulation. This situation shows up often when water

molecules interact with hydrogen bond acceptors [fig. 4.2].

We set the NAMD parameter *alchElecLambdaStart*=0.4 or 0.5, depending whether the simulations were run using the additive or the Drude force field. This means the van der Waals terms of the new side chain are introduced starting at $\lambda = 0$, but the electrostatic terms are introduced starting at $\lambda = 0.4$ or $\lambda = 0.5$. With a value of 0.4, electrostatic terms of the disappearing side chain are completely decoupled at $\lambda = 0.6$, while its van der Waals terms disappear at $\lambda = 1$. In figure 4.1, we show the protocol adopted for Drude simulations.

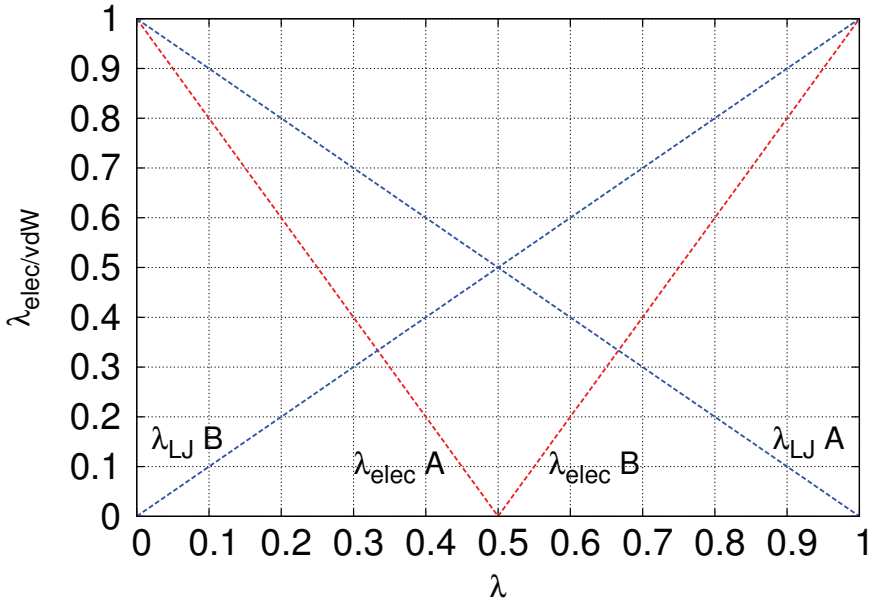


Figure 4.1 – **Alchemical transformation** of two side chains A and B with Drude. In the figure, se use the **red** dashed line to indicate electrostatic interactions which are decoupled/coupled at λ_{elec}=0.5. Scaling of LJ interactions is represented with **blue** dashed lines: side chain A is canceled between $\lambda_{LJ} = 0 \rightarrow 1$, while side chain B is inserted in the same interval.

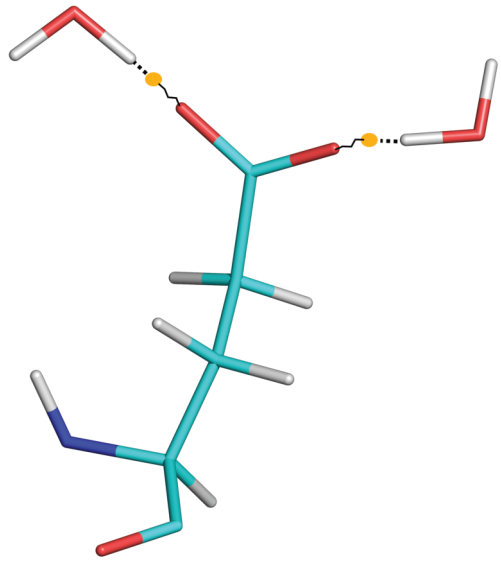


Figure 4.2 – **Water-acceptor** interaction with partially-annihilated LJ interactions. Water hydrogens approach the hydrogen bond acceptors, increasing displacements of their Drude particle.

2.3 Drude dual topology

The structure files (PSF) for alchemical MD were built using special topology files called *dual topologies* applied on the mutating residue. To simulate a mutation between two different side chains A and B, A and B were both attached to the same C_α atom. Bonded and nonbonded interactions between the two side chains were excluded. In figure 4.3 we give an example of dual residue called HEK, composed of glutamate and lysine side chains attached to the same backbone. In the first part of the file, we list atoms of the HEK backbone, shared by the two side chains. Backbone atoms form a neutral, zero-charge group. In the second part of the file we list the side chains, attached to the same C_α through two bonds CBE-CA and CBK-CA. At each ATOM line we indicate the atom name, type, charge, polarizability and Thole factor, as usual in the Drude force field. Then a list of atom names belonging to the other side chain is given. These define nonbonded interactions between GLU and LYS that will be excluded. These exclusions are applied for all values of λ. In the given example, we used "[...]" to abbreviate the lines present in the real topology file, where all the atoms of the other side chain are listed.

In HEK, the C_β atoms (CBE and CBK) are *depolarized* (they do not carry a Drude particle) even if according to the foorce field they should be polarizable. This modification was necessary to exclude spurious dipole-dipole interactions between these two atoms. The

```
RESI HEK  0.00

! Backbone

GROUP
ATOM N    ND2A2   -0.382  ALPHA -1.942  THOLE 0.250
ATOM HN   HDP1A    0.272
ATOM CA   CD31C    0.169  ALPHA -0.960  THOLE 1.078
ATOM HA   HDA1A   -0.017
ATOM C    CD2O1A   0.497  ALPHA -0.675  THOLE 0.295
ATOM O    OD2C1A   0.000  ALPHA -0.651  THOLE 0.310
ATOM LPOA LPDO1   -0.312
ATOM LPOB LPDO1   -0.227

! Glutamate sidechain
                                                  !excluded interactions:
ATOM CBE  CD32B   -0.066                          cbk hb1k hb2k cgk [...]
ATOM HB1E HDA2A    0.033                          cbk hb1k hb2k cgk [...]
ATOM HB2E HDA2A    0.033                          cbk hb1k hb2k cgk [...]
ATOM CGE  CD32A   -0.190  ALPHA -2.528  THOLE 1.414   cbk hb1k hb2k cgk [...]
ATOM HG1E HDA2A    0.004                          cbk hb1k hb2k cgk [...]
ATOM HG2E HDA2A    0.004                          cbk hb1k hb2k cgk [...]
ATOM CDE  CD2O2A   0.708  ALPHA -1.016  THOLE 0.899   cbk hb1k hb2k cgk [...]
ATOM OE1  OD2C2A   0.003  ALPHA -0.699  THOLE 2.399   cbk hb1k hb2k cgk [...]
ATOM OE2  OD2C2A   0.003  ALPHA -0.699  THOLE 2.399   cbk hb1k hb2k cgk [...]
ATOM LP1A LPD     -0.383                          cbk hb1k hb2k cgk [...]
ATOM LP1B LPD     -0.383                          cbk hb1k hb2k cgk [...]
ATOM LP2A LPD     -0.383                          cbk hb1k hb2k cgk [...]
ATOM LP2B LPD     -0.383                          cbk hb1k hb2k cgk [...]

! Lysine sidechain
                                                  !excluded interactions:
ATOM CBK  CD32A   -0.066                          cbe hb1e hb2e cge [...]
ATOM HB1K HDA2A    0.033                          cbe hb1e hb2e cge [...]
ATOM HB2K HDA2A    0.033                          cbe hb1e hb2e cge [...]
ATOM CGK  CD32A   -0.156  ALPHA -1.660  THOLE 1.300   cbe hb1e hb2e cge [...]
ATOM HG1K HDA2A    0.078                          cbe hb1e hb2e cge [...]
ATOM HG2K HDA2A    0.078                          cbe hb1e hb2e cge [...]
ATOM CDK  CD32A   -0.156  ALPHA -1.660  THOLE 1.300   cbe hb1e hb2e cge [...]
ATOM HD1K HDA2A    0.078                          cbe hb1e hb2e cge [...]
ATOM HD2K HDA2A    0.078                          cbe hb1e hb2e cge [...]
ATOM CE   CD32A    0.043  ALPHA -1.656  THOLE 0.895   cbe hb1e hb2e cge [...]
ATOM HE1  HDA2C    0.143                          cbe hb1e hb2e cge [...]
ATOM HE2  HDA2C    0.143                          cbe hb1e hb2e cge [...]
ATOM NZ   ND3P3A  -0.349  ALPHA -1.298  THOLE 0.895   cbe hb1e hb2e cge [...]
ATOM HZ1  HDP1B    0.340                          cbe hb1e hb2e cge [...]
ATOM HZ2  HDP1B    0.340                          cbe hb1e hb2e cge [...]
ATOM HZ3  HDP1B    0.340                          cbe hb1e hb2e cge [...]
```

Figure 4.3 – **Glutamate/Lysine** Dual Topology (residue HEK)

free energy to depolarize them is computed separetly, as explained below. Other undesired, spurious interactions between the two side chains must be removed when atom types of CA, CBE and CBK form an angle for which the force field expects an harmonic term. The same holds when bridging atoms are involved in a dihedral term of the energy function. This does not occour too often in C36 or Drude and is more frequent in Amber, since Amber has fewer atom types. In this case, undesired terms must be manually deleted within CHARMM before producing the system PSF.

2.4 Spurious interactions with Drude and dual topology

In the Drude polarizable force field, non-bonded lists are constructed in such a way that Drude particles have the same bond hierarchy of their parent atoms. We consider as an example a triatomic molecule composed of atoms A1-A2-A3 and their Drude particles DA1, DA2 and DA3, shown in figure 4.4, right. The Drude particle DA1 attached to the core atom A1 form a 1-3 pair with the Drude particle DA3 attached to the core atom A3, since their parent atoms are a 1-3 couple. The same property holds for diatomic molecules, where Drude particles DA1 and DA2 form a 1-2 couple, as shown in figure 4.4, left.

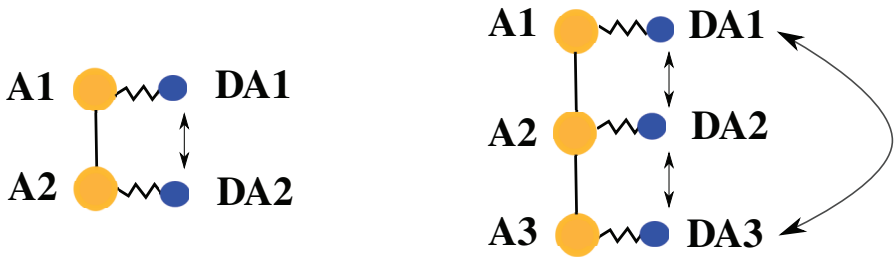


Figure 4.4 – **Bond hierarchy** in the Drude force field for a diatomic o triatomic molecule. Atoms are represented in yellow, while their Drude particles are represented in blue. Interactions are indicated with arrows.

The so defined non-bonded lists are exploited to compute non-bonded interactions as follows (Harder *et al.* [2006]). Interactions between core charges q_A are excluded between 1-2 and 1-3 pairs. Interactions between Drude particles q_D and core charges q_A are excluded between 1-2 and 1-3 pairs. Interactions between all core charges q_A and all Drude particles q_D are computed for 1-4 pairs and beyond. Interactions between Drude charges q_D of 1-2 and 1-3 pairs are computed and screened according to their Thole parameters, using equation 3.21. With NAMD, interactions between Drude particles are also screened for nonbonded 1-4 pairs and beyond, if their distance is lower than a threshold value set by default to 5.0 Å. However, the screening between 1-4 pairs is disabled when running the FEP using the *alch* keyword.

Summing up, as mentioned in the official CHARMM documentation "If two atoms do not see each other in the non-polarizable force field, their dipoles (and only their dipoles, not their net partial charges) will see each other in the polarizable force field". The particular bond hierarchy is used to compute the screened dipole-dipole interactions between 1-2 and 1-3 pairs, as proposed by Thole [1981]. However, this introduces a spurious interaction in the dual topologies. In the HEK example, C_β atoms CBE and CBK are in a 1-3 relationship as they are both bonded to the backbone atom C_α. As a consequence, their Drude particles DCBE and DCBK have a 1-3 relationship: their interaction is computed and screened according to their Thole parameters, using equation 3.21. This is a spurious interaction, since in the FEP the two side chains of the mutating residue should not interact. It occours because the Drude force field is only partially supported by the *alch* facility of Namd, which does not correctly modify the nonbonded-lists in order to exclude the dipole-dipole interaction between the two C_β atoms.

As a consequence, the current Namd implementation of *alch* does not allow to run FEP simulations with Drude using dual topologies where the bridging atom is linked to two atoms carrying Drude particles. However, the problem can be solved following an ad-hoc protocol. Our dual topologies were built *removing* the Drude particle attached to both CB atoms of the two side chains. These atoms were *depolarized*. In practice, we condensed the total charge of the CB+DCB couple to the CB atom $Q = q_A$ and we have set the CB polarizability and Thole factor to zero [fig. 4.5]. Alchemical simulations were then run using these modified topology files, where the spurious interaction were removed.

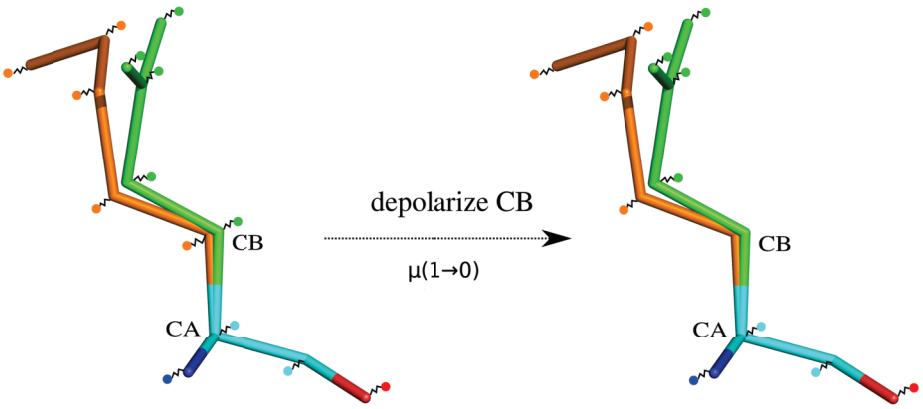


Figure 4.5 – **Model residue** used for the Drude dual topology definition. C_β atoms are *depolarized*, in the sense that its Drude particle is completely removed from the system. Polarization and Thole factors are set to zero, and the oscillator charge is condensed in the reference atom.

Removing C_β polarizability has a free energy cost that must be considered. We introduced additional transformations at the beginning and the end of the mutation process, where the

C_β Drude charges were gradually removed (for A) or introduced (for B). These simulations correct the alchemical free energy difference for the transformation with dual topology $\lambda = 0 \rightarrow 1$, as shown in figure 4.6. We defined coupling parameters μ(A) and μ(B) that scaled the polarizabilities on the two C_β atoms. When μ(A)=1, the C_β of A had its full polarizability; when μ(A)=0, its polarizability had been removed. To change the polarizability α of an atom, we adjusted the atom+Drude charges q_A and q_D using the relationship

$$\alpha = \frac{q_D^2}{k_D} \qquad q_A + q_D = Q \tag{4.5}$$

We varied μ(A) from 1 to 0 in four MD windows. During these windows, side chain A was fully coupled (the coupling coordinate λ was zero), whereas side-chain B was decoupled and the polarizability of its C_β was turned off. We then did the alchemical A→B transformation with both C_β polarizabilities turned off. Finally, we turned on the C_β polarizability of B,μ=0→1. For the polarizability transformations, we used MD windows where μ(A) (respectively, μ(B)) was set to 0, 0.3, 0.7 and 1 (20 ns each). The value of the polarizability was controlled by editing charges in the NAMD protein structure file. Free energies were computed with TI and added to the λ=0→1 results.

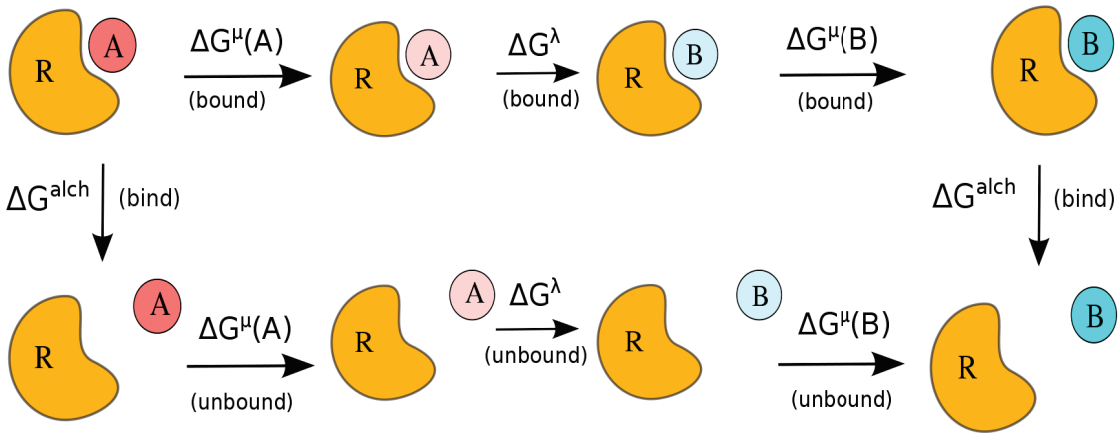


Figure 4.6 – **Thermodynamic cycle** for the alchemical relative binding free energy of two side chains A and B. The transformations ΔG^λ are carried out with the dual topology, where CB atoms are depolarized. The other legs in the thermodynamic cycle ΔG^μ are the transformations used to compute the additional contributions $\mu(0 \rightarrow 1)$, as described in the text.

2.5 PME correction

Using PME under periodic boundary conditions, the electrostatic potential of the infinite, periodic system is infinite unless the simulation boxes are grounded. In alchemical simulations

of ionic mutations, the transformation of the mutating side chain along the path $\lambda = 0 \rightarrow 1$ is a charging process and the endpoints cannot both be neutral. There are two possible solutions, previously discussed in the last chapter. The first is to introduce one or more counterions in the simulation box, which are alchemically transformed during the FEP in such a way to neutralize the system for each value of λ. The free energy can be easily corrected keeping the counterion far from the solute, since it will contribute independently to the charging free energy. A more common strategy is to introduce the *gellium*, that has an important effect: it leads to an electrostatic potential, averaged over the box volume, that is zero at all times. In other words, it grounds all simulation boxes to zero. However, the periodic system with the gellium is not a faithful representation of the true physical endpoints. These correspond to a more realistic situation, where the solute (the protein:peptide complex or the unbound peptide) is solvated in a macroscopic but finite solvent volume, such the large spherical cluster of water molecules, like shown in figure 4.7.

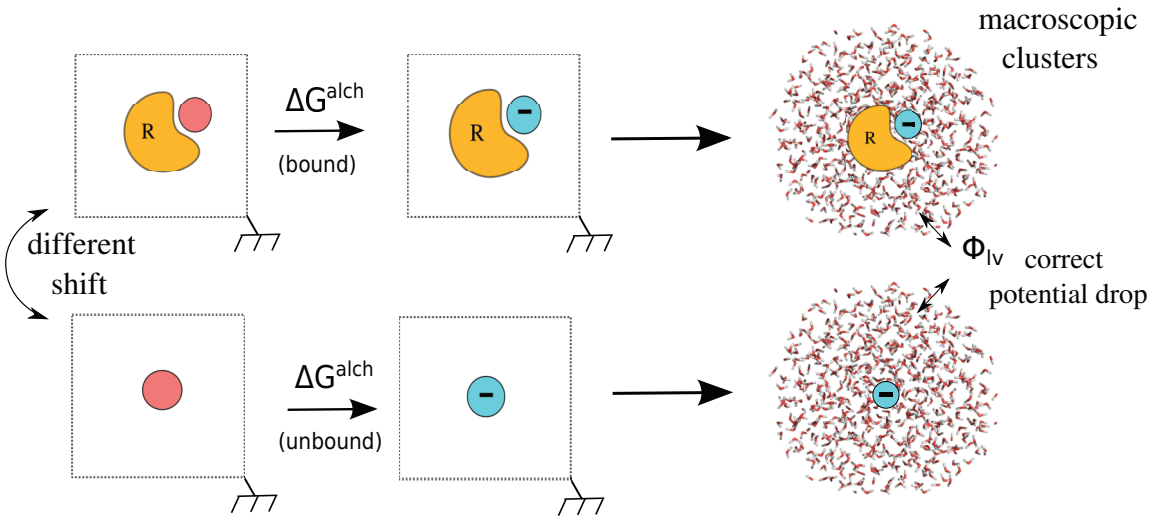


Figure 4.7 – **Grounded simulation boxes** and the more realistic macroscopic clusters, which are not grounded as described in the text. The two boxes used to simulate bound and unbound states are neutrilized by a different grounding potential. For the unbound state, the potential shift is close to Φ_{lv}.

In the macroscopic clusters, the average electrostatic potential per box $\langle\phi\rangle_{cluster}$ is not zero, it is close to the vapor/liquid boundary potential Φ_{lv}. For TIP3P water, $\Phi_{lv} = -520$ mV (Lin *et al.* [2014] ; Harder & Roux [2008]) while for SWM4-NDP its value is -545 mV (Lamoureux *et al.* [2006]). The grounded PME potential of a simulation box $\langle\phi\rangle_{box}$ is so shifted upwards, by a quantity close to this boundary potential. The exact value of this shift depends on the fraction of box occupied by the water and on the nature of the solute. Calculations (Lin *et al.* [2014]) showed that for a small spherical ion solvated in a cubic box, the shift scales as L^{-3} where L is the size of the box. For more heterogeneous and different solutes, the shift will be

different even if the same box size is used.

The potential shift for the protein/peptide complex and of the unbound peptide simulations will be different, since these simulations are performed in different boxes, and the nature of the solute is different. As a consequence, inserting a charge on the peptide in the simulation box of the complex (bound state), this charge will be subjected to a shifted potential which is different respect to the one in the simulation box of the unbound peptide. A mutation that inserts a net charge onto the peptide in both bound and unbound states requires a free energy correction, which is computed following the protocol described in Lin *et al.* [2014] ; Simonson & Roux [2016].

For a given solute and mutation endpoint we compute, from the MD trajectory, the mean potential in the solvent region of the box, defined to be > 12 Å away from the solute or its periodic images. For a small solute like a peptide, it is close to zero, the overall box average with PME. For a protein solute, the overall box average is an average over a solvent region (far from the solute) and a protein region (the rest of the box):

$$V_p \langle \Phi \rangle_p + V_w \langle \Phi \rangle_w = 0 \tag{4.6}$$

Here, V_w and V_p are the volumes of the solvent and protein regions and the brackets indicate a spatial average of the potential ϕ over either region. Thus,

$$\langle \Phi \rangle_w = \frac{V_p}{V_w} \langle \Phi \rangle_p \tag{4.7}$$

The potential in the solvent region of a protein solute is offset, instead of matching that in the peptide simulation (zero). With both TIP3P and Drude SWM4 water, the offset is positive, making it artificially easier to insert a positive charge into the protein than the peptide.

For a given solute, we denote the solvent potential $\langle \phi_w \rangle$(solute). If the mutation introduces a net charge q on the peptide, this charge will then experience a potential that is shifted by ϕ_w(complex) in the bound state and ϕ_w(peptide) in the unbound state. Multiplying q by the difference gives the free energy correction. In practice, we average the free energy correction over the simulations of the initial and final peptides (before and after the mutation): peptide, complex, peptide$'$ and complex$'$, respectively. The potential shift correction is thus:

$$\Delta\Delta G^{\Phi}_{PME} = \frac{q}{2} \left[\left(\Phi_w(\text{complex}) + \Phi_\text{w}(\text{complex}') \right) - \left(\Phi_w(\text{peptide}) + \Phi_\text{w}(\text{peptide}') \right) \right] \tag{4.8}$$

In effect, we average over a calculation where q is inserted into the initial peptide and one where q is removed from the final peptide. In practice, the two values are very similar. We verified that the potential is almost uniform throughout the solvent region, and very similar for complex and complex$'$, supporting the idea that ϕ_w(solute) accurately represents the PME potential shift.

The potential throughout the simulation box was obtained for each frame of each endpoint trajectory by the VMD plugin PMEpot (Humphrey *et al.* [1996]), which prints out the potential values on a cubic grid spanning the box. Time and spatial averaging were done with a shell script. MD simulations with a cubic box (of the same volume) were specifically done for each system to allow use of the plugin (which does not support a truncated icosahedral box). With Drude, for a +1 charge insertion an artificial contribution of $\Delta\Delta G^{\phi} = -0.7$ kcal/mol was obtained and must be corrected for. For the same process C36 has a smaller shift of -0.39 kcal/mol, close to the one obtained earlier for ff99SB which equals -0.45 kcal/mol.

3 Results

We computed binding free energies differences $\Delta\Delta G$ for four mutations using the additive force field C36 and the Drude polarizable force field. The same variants were previously studied using the Amber ff99SB additive force field (Panel *et al.* [2018]). Results are shown in table 4.1, including the $\Delta\Delta G$s of two other ionic mutations previously obtained using ff99SB. Peptide positions are numbered backward from the C-terminus (position 0 or P0). In the table we list all contributions to the binding free energy differences. Care was taken here to extrapolate the MD results to the thermodynamic limit of a very large simulation box. The Drude PME correction Φ_{PME} was 0.70 kcal/mol for *E4L*, -0.70 for *K4L* and 1.40 kcal/mol for *E4K*. For *K912E*/Sdc1, the correction was zero, since the V_p/V_w ratio in bound and unbound states was essentially the same. PME corrections for additive force fields were smaller, because the mean electrostatic potential in TIP3P water is smaller than that in Drude SWM4-NDP water. With Drude, there were additional transformations μ_A, μ_B at the beginning and end of the mutation that removed/restored the C_β polarizabilities, computed with TI. These contributions were between -0.3 and 0.1 kcal/mol. The alchemical steps computed using BAR $\lambda = 0 \rightarrow 1$ contributed between -1.0 and 2.1 kcal/mol with Drude, while for C36 and ff99SB they were always positive and between 0.8 and 3.2 kcal/mol. Some of the Drude contributions were analyzed also with TI, obtaining the same result. Indeed, even if BAR is not formally correct when used with Drude (as described before in the text), the moderate errors for the bound and unbound states are mostly canceled when computing the binding free energy differences $\Delta\Delta G = \Delta G_{bound} -$

$\Delta G_{unfound}$. We note that all three contributions to the Drude binding free energy change $\lambda = 0 \rightarrow 1$, Φ_{PME} and μ_A, μ_B were of roughly the same order of magnitude.

Mutation	Contribution	Exp.	Drude	ff99SB	C36
E4K	μ_A,μ_B		0.1	-	-
	$\lambda = 0 \rightarrow 1$		-1.0	0.8	1.0
	Φ_{PME}		1.4	0.9	0.8
	$\Delta\Delta G_{bind}$	0.8	0.5	1.9	1.8
	err.	-	-0.3	1.1	1.0
E4L	μ_A,μ_B		-0.2	-	-
	$\lambda = 0 \rightarrow 1$		1.4(1.3)	3.1	3.2
	Φ_{PME}		0.7	0.5	0.4
	$\Delta\Delta G_{bind}$	0.6	1.9(1.8)	3.6	3.5
	err.	-	1.3	3.0	2.9
K4L	μ_A,μ_B		-0.3	-	-
	$\lambda = 0 \rightarrow 1$		2.1(2.1)	1.4	2.2
	Φ_{PME}		-0.7	-0.5	-0.4
	$\Delta\Delta G_{bind}$	-0.2	1.1(1.1)	0.9	1.8
	err.	-	1.3	1.1	2.0
K912E	μ_A,μ_B		0.7	-	-
	$\lambda = 0 \rightarrow 1$		2.1	2.0	1.6
	Φ_{PME}		0.0	0.0	0.0
	$\Delta\Delta G_{bind}$	1.0	2.6	2.0	1.6
	err.	-	1.6	1.0	0.6
	rmsd	-	1.23	1.76	1.86
	mue	-	1.12	1.55	1.62
E3T,Y1K	$\lambda = 0 \rightarrow 1$		-	1.5	-
	Φ_{PME}		-	0.9	-
	$\Delta\Delta G_{bind}$	1.3	-	2.3	-
	err.	-	-	1.0	-
K912E/Caspr4	$\lambda = 0 \rightarrow 1$		-	0.5	-
	Φ_{PME}		-	0.0	-
	$\Delta\Delta G_{bind}$	0.7	-	0.5	-
	err.	-	-	-0.2	-
	rmsd	-	-	1.50	-
	mue	-	-	1.23	-

Table 4.1 – **Relative binding free energies for ionic mutations** obtained with polarizable and additive force fields. Specific contributions are shown on separate lines. Alchemical free energy differences $\lambda = 0 \rightarrow 1$ were obtained using BAR. Two values were also computed with TI, values are shown in parenthesis. The table is separated into two parts: Rmsd and Mue are first computed considering the first 4 mutations for which Drude and C36 results are known (*E4K*, *E4L*, *K4L* and *K912E*). Then, rmsd and mue are calculated for the six ionic mutations previously studied using ff99SB (Panel *et al.* [2018]).

The overall Drude binding free energy changes had errors of −0.3, 1.3, 1.3, and 1.6 kcal/mol, respectively, for an RMS error of 1.2 kcal/mol. The same variants with ff99SB and C36 gave RMS error of 1.8 and 1.9 kcal/mol. Thus, the RMS error was reduced with Drude. The improvement was mainly due to *E4L*, which had the largest error with both the additive force fields. The C36 force field gave $\Delta\Delta G$ values for *E4K* and *E4L* within 0.1 kcal/mol of AMBER ff99SB and a similar RMS error. The closure errors for the thermodynamic cycle (*Sdc1/E4K/E4L/Sdc1*) with Drude and C36 were very small, 0.3 and 0.0 kcal/mol, respectively, suggesting convergence was good. The ff99SB closure error was slightly larger, 0.8 kcal/mol. For the *E4L* and *E4K* ionic mutations, previous very long simulations were run with ff99SB (495 and 770 ns, respectively) to establish that the errors were due to force-field limitations, not sampling. The agreement between C36 and ff99SB supports the idea that the large E4K errors were due to the lack of explicit polarizability.

For the other five ionic mutations, agreement with experiment was still rather good using AMBER ff99SB, with an RMSD of 0.90 kcal/mol. Errors of -1 *kcal/mol* were also seen recently with this force field for ionic mutations in the tyrosyl-tRNA synthetase enzyme binding its amino acid substrate (Simonson *et al.* [2016]).

The main interactions that stabilize the different complexes were identified by structure analysis. With the AMBER additive force field, the Sdc1-E4 side chain makes a salt bridge with either Arg871 or Lys912 half of the time [fig. 4.9], whereas with the Drude polarizable force field, a Glu4-Lys921 salt bridge is present only 1/4 of the time as shown in figure 4.10. No salt bridge is present in the x-ray complex. Thus, the additive force field overpopulates the salt-bridge conformations, leading to an overstabilization of wild-type/Sdc1 relative to the *E4K* and *E4L* mutants. In the mutant *E4K* and *E4L* complexes, there are no salt bridges involving position P4, either with AMBER ff99SB or with Drude. K4 mostly makes polar interactions with solvent and L4 is hydrophobic. Although the Drude model is less affected by salt-bridge overstabilization, it still predicts an E4-Lys912 interaction 1/4 of the time, and this appears to account for the positive $\Delta\Delta G$ errors for the *E4L* and *K912E* mutations obtained with Drude.

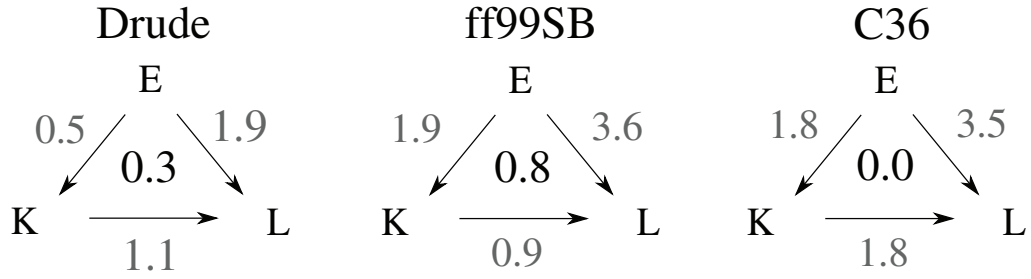


Figure 4.8 – **Closure errors** for relative binding free energies calculated with Polarizable or additive force fields

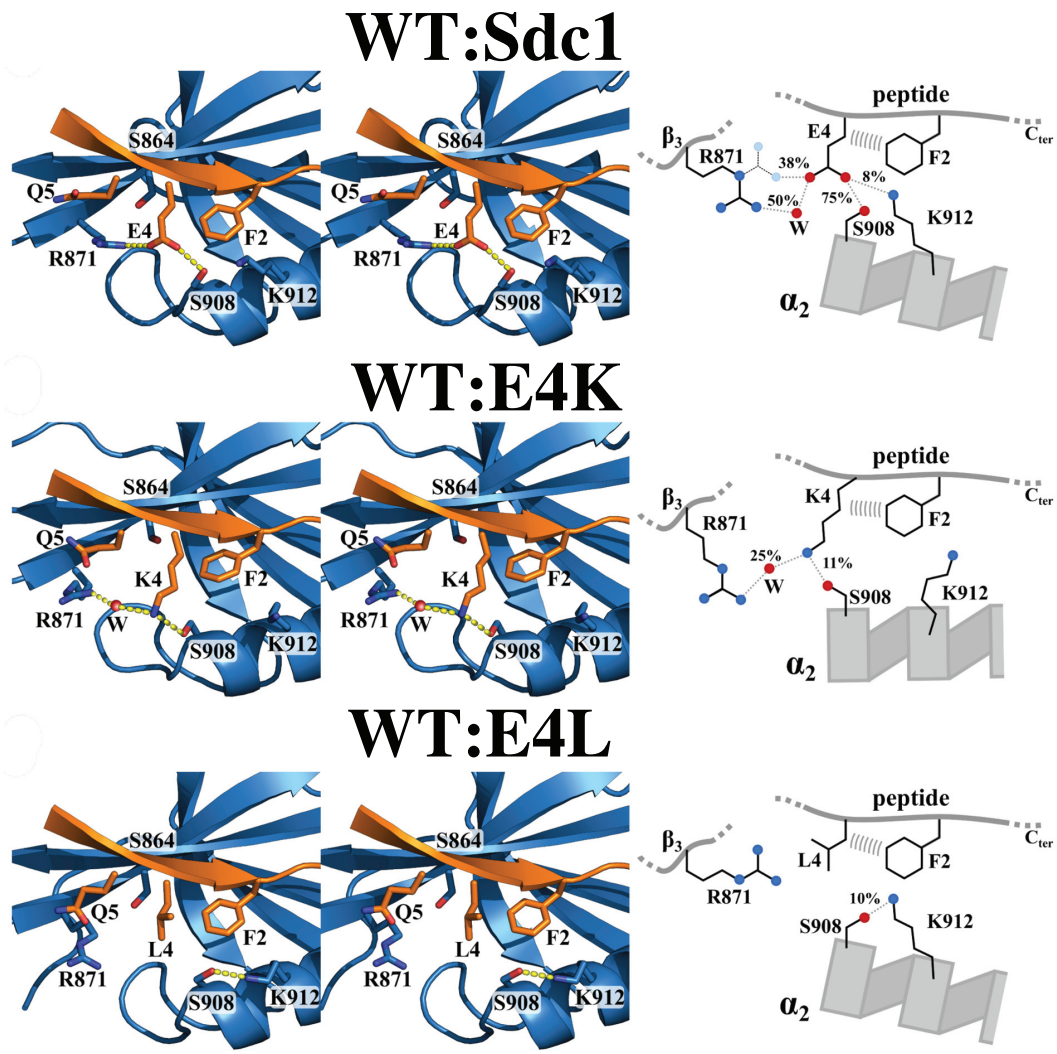


Figure 4.9 – **Structural views based on the Amber ff99SB MD simulations** for the wild-type/Sdc1 complex (top) and its E4K (middle) and E4L (bottom) mutants. The 3D structures are the mean structures from the MD.

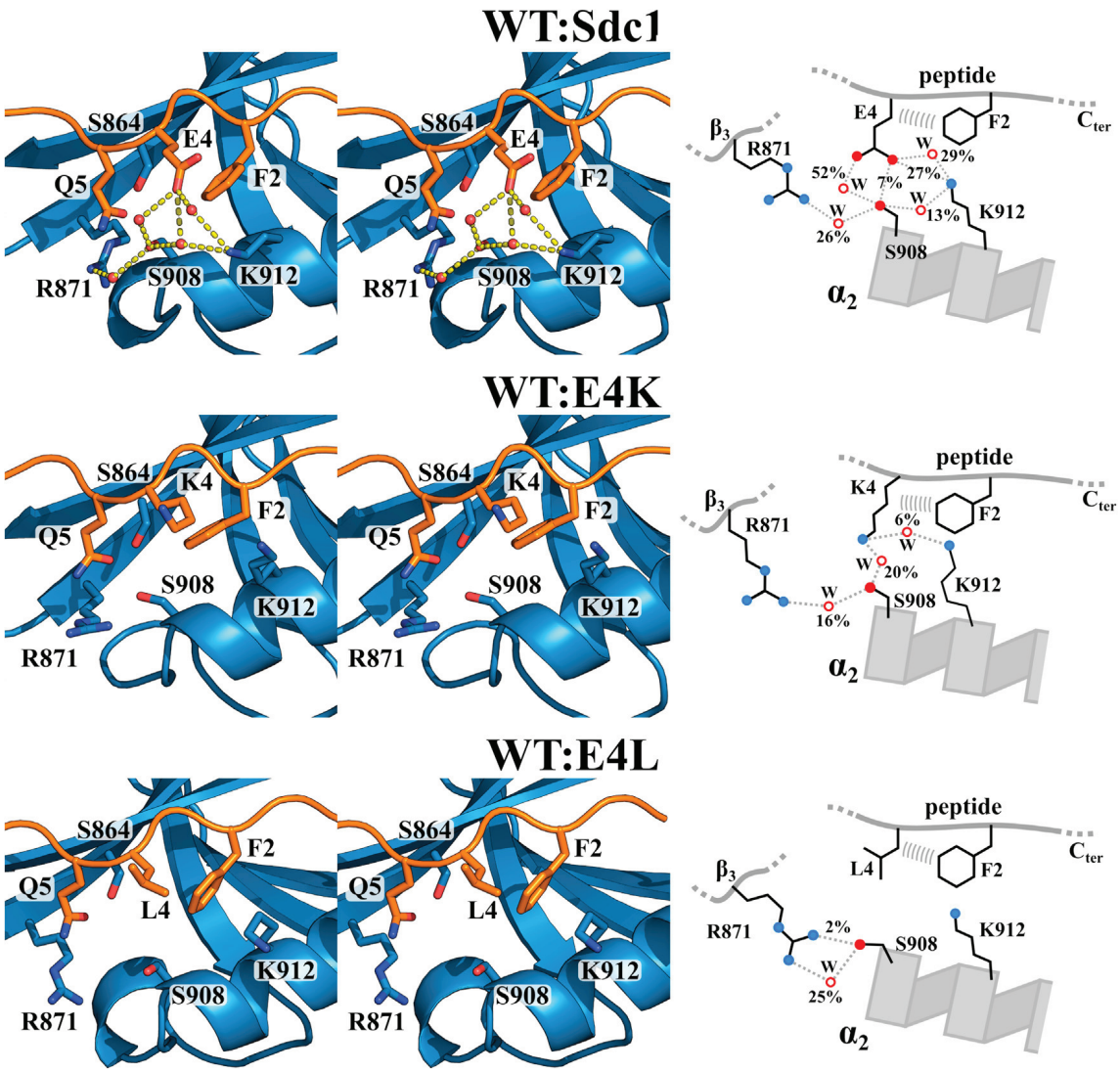


Figure 4.10 – **Structural views based on the Drude polarizable MD simulations** for the wild-type/Sdc1 complex (top) and its E4K (middle) and E4L (bottom) mutants. The 3D structures are the mean structures from the MD.

4 Conclusions

To our knowledge, this study was the first application of the Drude force field to compute relative binding affinities for a set of protein:ligand complexes. The described protocol can be applied to any alchemical FEP when a dual topology is preferable. Technical issues were highlighted, mainly related to the actual NAMD implementation of FEP. Simulations were needed to compute the additional contributions μ_A, μ_B to restore the deleted CB polarizabilities. This increases the computational cost of the Drude FEP, which is already bigger (a factor 4) compared to Amber ff99SB or C36. However, future versions of the *alch* facility of Namd could solve this problem, deleting the spurious interactions. Overall Drude results are

encouraging: the *E4L* error was reduced from 3.0 to 1.3 kcal/mol, the *E4K* error was reduced from 1.1 to 0.3 kcal/mol; the *K4L* error was reduced by 0.7 kcal/mol compared to C36, and was similar to the ff99SB error. Only the K912E/Sdc1 error increased somewhat with Drude, from 1.0 kcal/mol with ff99SB to 1.6 kcal/mol. K912 is highly exposed to solvent and does not form a salt bridge. Established additive force fields are well parameterized for this situation, whereas we can speculate that the more recent Drude parameters may still require some tuning. Further studies could focus on this aspect. We also recognize that the TIP3P water model employed in the additive simulations can play an important role, and may contribute to the observed errors. SWM4-NDP is a four-point model, which can also adapt its dipole moment in reponse to perturbations of nearby charges. Moreover, the possibility that a Drude FEP using SWM6-NDP could produce better results is an interesting open question, since SWM6 is recognized to reproduce water properties even better than SWM4 (Yu *et al.* [2013]).

Sophisticated methods have been proposed to accelerate conformational sampling. They include adaptive biasing of the energy surface with metadynamics (Laio & Parrinello [2002]), λ-dynamics (Hayes *et al.* [2017]), or the orthogonal space random walk (Zheng *et al.* [2008]). Other methods like replica exchange (Kokubo *et al.* [2013]) increase sampling by thermally activating some or all degrees of freedom. Here, we used standard MD with long simulations, and we verified convergence by doing redundant calculations. As anticipated before in the text, the closure errors suggest that the C36 and Drude FEP were well converged, even if this result does not give information about the convergence of the μ_A, μ_B simulations. Indeed, these last contributions cancel out in the thermodynamic cycle since they were added and then substracted. Overall, the accuracy obtained indicates FEP can be used in a design strategy, where many variants obtained with empirical energy functions are filtered further with more rigorous FEP simulations before being tested experimentally. For ionic mutations in buried regions, electronic polarization is likely to contribute and the use of the Drude force field could be preferred.

Chapter 5

Classical Drude Model for methyl phosphate and phosphotyrosine

1 Introduction

1.1 Phosphotyrosine in Tiam1 binding

Phosphorylation is among the most diffused post-translational modifications that regulate PPIs in cell signalling networks (Khoury *et al.* [2011] ; Pawson & Scott [2005]). Syndecan1 is a member of the *syndecans* family, high conserved transmembrane and cytoplasmic domains containing four conserved tyrosine residues that may be phosphorylated (Reiland *et al.* [1996]). In the case of Syndecan1, the phosphorylation of a tyrosine at C-terminal position P_{-1} has an important role in cell adhesion (Sulka *et al.* [2009]).

The Tiam1 PDZ domain can bind both the wildtype Syndecan1 peptide (Sdc1, sequence TKQEEFYA) and its phosphorylated form (pSdc1, sequence TKQEEF(pTyr)A). Their affinities to Tiam1 were studied experimentally, revealing that the phosphorylation of Tyr_{-1} in Sdc1 does not change considerably its binding affinity to Tiam1 $\Delta\Delta G_{bind}(Sdc1 \rightarrow pSdc1) = -0.21$ kcal/mol (Liu *et al.* [2013]). Crystal structures of the wildtype Tiam1:Sdc1 complex (pdb access code 4GVD) and of the Tiam1:pSdc1 complex (pdb access code 4GVC) show that there is no Tiam1 rearrangement after binding, with an rmsd between the two backbone conformations of just 0.26 Å. However, the interactions between amino acid side chains in the binding pocket are different [fig. 5.1]. The tyrosine at position P_{-1} rotates of almost 90° after phosphorylation, where the phosphate group interacs strongly with the Tiam1 residue K879 and also with T857 [fig. 5.1]. On the other hand, in the Tiam1:Sdc1 complex these interactions are different with Tyr_{-1} interacting with N876, E_{-4} and K_{-6} thorugh an hydrogen bonding network.

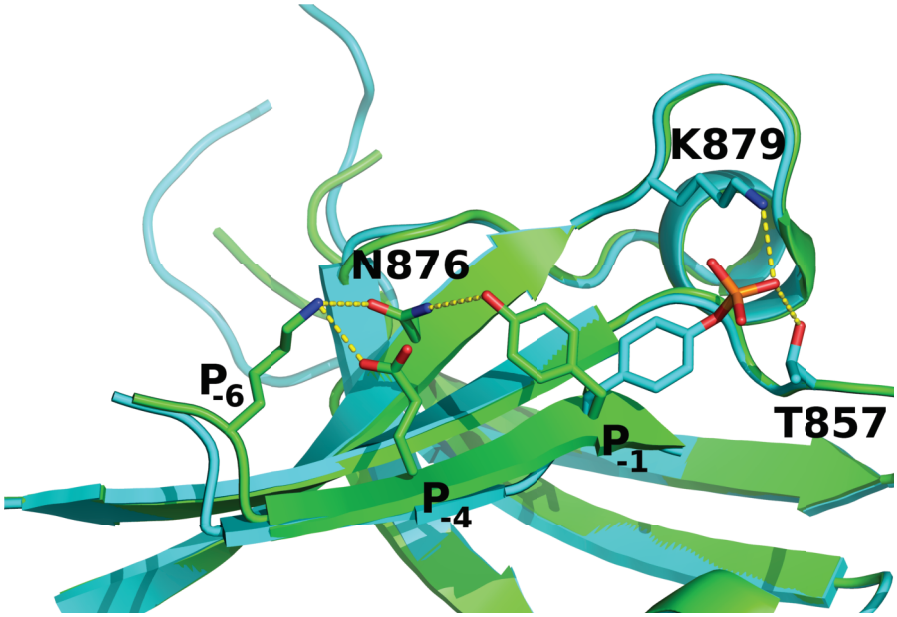


Figure 5.1 – **Tiam1-Sdc1 and Tiam1-pSdc1 structures superposition.** The wildtype complex is shown in green, the phosphorylated complex is shown in cyan. From crystal structures 4GVD and 4GVD.

Previous molecular dynamics simulations of the Tiam1:Sdc1 and Tiam1:pSdc1 complex (Nicolas Panel, results are not yet published) suggested that in the bound state the dianionic phosphotyrosine $pTyr^{2-}$ is dominant respect to its monoanionic form $pTyr^{-}$. During MD, monoanionic phosphotyrosine did not conserved the crystal structure configuration, moving towards the tyrosine orientation in the WT complex. On the other hand, dianionic phosphotyrosine conserved the crystal structure orientation, interacting strongly with K879.

1.2 Phosphates in biology

Phosphate groups are essential in biochemistry. They are components of nucleic acids and also present in certain protein side chains, whose interactions with solvent, metal ions and ionic side chains help control protein folding and binding (Westheimer [1987]). Phosphates are in ATP, they are added/removed enzymatically as a function of cellular condition, and they facilitate important PPIs in a number of signaling pathways (Pawson [1995]). Often, the presence or absence of the phosphate group makes the protein competent or incompetent to bind another molecule (Shepherd *et al.* [2011]). Indeed, the phosphorylated species act as chemical entities different from natural amino acids, making intramolecular and intermolecular hydrogen bonds with PPI partners. During evolution, nature chose phosphates because they were abundant on Earth, because they are solubile in water, and because they have versatile chemical properties (Hunter [2012]). Quoting the author, “Life as we know it could not have

evolved without phosphate".

Computer simulations can give detailed insights into the structure and interactions of phosphate groups. To exploit this capability it is essential to have accurate force field models of these moieties. However, developing accurate simulation models is challenging for ionic interactions such as those involving phosphate groups.

1.3 Earlier additive results, need for polarizability

Phosphate-containing molecules will usually be more polarized when exposed to solvent or close to an ion, and less polarized when removed from solvent or ions. For this reason, additive force fields can give large errors, since they are usually parameterized to describe ionic interactions between groups that are highly-exposed to aqueous solvent. Previous calculations of the standard binding free energy between Mg^{2+} and hydrogen phosphate $P_i^{2-} = HPO_4^{-}$ were done using the Charmm additive force field C27, overstimating of an order of magnitude the experimental value of -3.7 kcal.mol (Satpati *et al.* [2011]). Other previous calculations of the relative binding free energy for the phosphorylation of Tyr_{-1} in the Tiam1:Sdc1 complex (not published) revealed that the additive force field Amber ff99SB strongly overstimates the stability of pTyr in the bound state, giving a $\Delta\Delta G(Sdc1 \rightarrow pSdc1) = -12.1$ kcal.mol, where the experimental value is -0.2 kcal/mol.

If additive force fields give large deviations from experiments, remarkable results were obtained using polarizable models. Mg^{2+}:phosphate binding free energies were studied with AMOEBA, giving excellent agreement with experiment (Kumar *et al.* [2014, 2018]). The AMOEBA force field was also used to reproduce the phosphate binding mode of a phosphate-binding protein, giving binding affinities in excellent agreement with experimental measurements (Qi *et al.* [2018]). Dimethyl phosphate (DMP) and its interactions with Mg^{2+} were studied with the Drude polarizable force field, giving good agreement with experiment for structure and dynamics (Lemkul & MacKerell [2016a]). Another recent study revealed that when compared to additive force fields, the Drude polarizable force field better estimates relative binding free energies for a series of ionic mutations of the Tiam1:Sdc1 complex (Panel *et al.* [2018]). All these studies confirm that to describe phosphate interactions, the explicit treatment of electronic polarization becomes necessary.

1.4 Methyl phosphate as a model

Here, we extended the Drude polarizable force field to three other phosphate compounds and their interactions with Mg^{2+}, then we developed a model for the dianionic phosphotyrosine

residue ($pTyr^{2-}$) [fig. 5.2]. We fist developed Drude parameters for methyl phosphate (MP), monoanionic and dianionc $MP^{-} \equiv CH_4PO_4^{-}$, $MP^{2-} \equiv CH_3PO_4^{2-}$. If DMP is analogous to the phosphate groups within DNA and RNA polymers, methyl phosphate (MP) is analogous to the phosphate groups at DNA and RNA termini. We validated our MP model computing MP:Mg^{2+} standard binding free energies, for which the additive force field gave large errors. We also developed a model for mono hydrogen phosphate $P_i^{2-} = HPO_4{}^{-}$, for which experimental Mg^{2+} binding data is available. After validation, MP was used to derive parameters for dianionic phosphotyrosine, for which the additive force field overstimated its interactions in the binding pocket of the Tiam1 PDZ domain. Our parametrization strategy was inspired by the method used to develop the $pTyr^{2-}$ model for the existing Charmm C36 force field (Feng *et al.* [1996]), where dianionic methyl phosphate was used to optimize a residue patch that transforms the standard tyrosine to $pTyr^{2-}$. The same approach could be used in the future to develop a Drude polarizable model for monoanionic phosphotyrosine ($pTyr^{-}$) but also for other phosphate-containing residues, like phosphoserine and phosphothreonine.

Figure 5.2 – **Methyl phosphate, hydrogen phosphate and phosphotyrosine** structures.

2 Methods: overview

Parametrization was done following the established Drude parametrization strategy described in Anisimov *et al.* [2005] ; Lemkul *et al.* [2016]. The empirical, molecular-mechanics (MM) Drude energy function was used to fit target quantum-mechanical (QM) quantities [fig. 5.3]. Several programs and tools were used to obtain and to fit the QM data. We finally validated our models computing Mg^{2+}:phosphate standard binding free energies in aqueous solution.

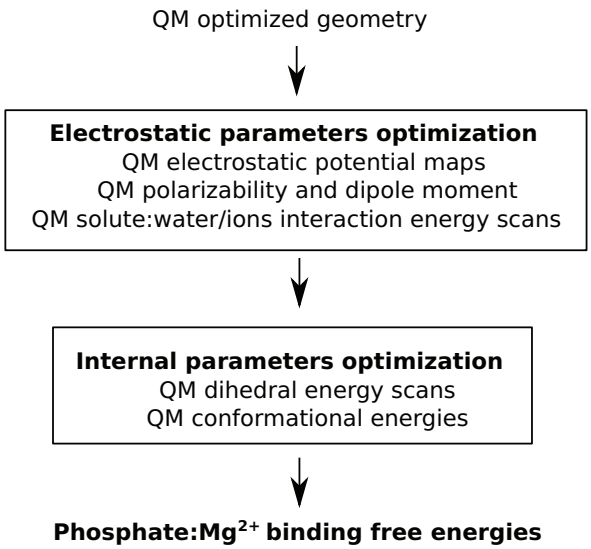


Figure 5.3 – **Parametrization procedure overview.**

2.1 QM quantities and software

QM optimized geometries were used to compute the QM quantities listed in figure 5.3: electrostatic potential (ESP) maps (unperturbed and perturbed, as discussed below), molecular polarizability and dipole moment, solute-water and solute-ion interaction energies, QM conformational energies. Details are given in the next section (see Methods: QM calculations). All QM quantities were computed using the GAUSSIAN version g09 (Frisch *et al.* [2009]) or PSI4 (Turney *et al.* [2011]). GAUSSIAN is the program traditionally used to compute target QM quantities for Charmm C36 and Drude force field parametrization. Several scripts to produce the QM data were already available, suggesting us to use the GAUSSIAN to compute most of the quantities. On the other hand, PSI4 is more recent and is becoming more and more used. It is free and has a flexible python interface. Some of the GAUSSIAN scripts were so rewritten for Psi4, and will be part of our parametrization toolbox.

2.2 The GAAMP and DGENFF tools

Parameter optimization was done using several tools. Some of them are programs borrowed from the GAAMP distribution (Huang & Roux [2013]). GAAMP "*General Automated Atomic Model Parameterization*" is a webserver for automatic force-field parametrization of small molecules. It has a simple interface, allowing the user to load an input structure and then automatically obtain the desired parameters for various additive force fields (for example Charmm C36 or Amber) and for the Drude polarizable force field. However, we preferred to run a more careful manual parametrization, where several utilities written in C++ were extracted from the GAAMP server and then run locally on our computers. Other steps of the

parametrization procedure were carried out computing molecular-mechanics quantities with the CHARMM program (version 42b2), together with bash and python scripts. CHARMM scripts were extracted from an ongoing first release of the Drude "DGENFF", analogous to the well-known toolbox for the Charmm force field parametrization CGENFF (Vanommeslaeghe *et al.* [2009]).

2.3 Parametrization strategy

Starting from a PDB file containing structure information we generated the initial Drude topology and parameter files using GAAMP. Input geometries were optimized with GAUSSIAN, and then used to compute the target QM data (see Methods: QM calculations).

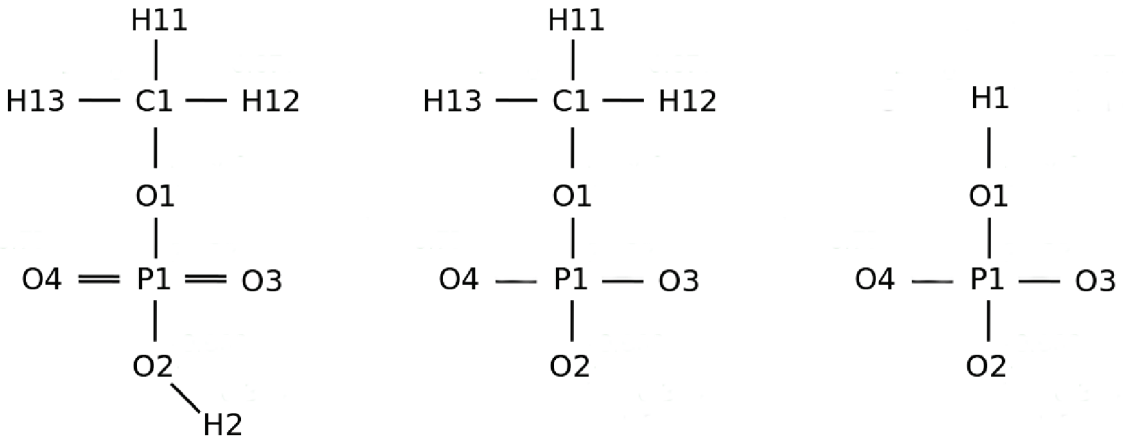


Figure 5.4 – **Methyl phosphate and hydrogen phosphate** 2D structures. Left: MP^{-}. Center: MP^{2-}. Right: P_i^{2-}

Electrostatic parameters were first optimized for MP^{-} (2D structure shown in figure 5.4, left), proceeding in two main phases. In the first phase, electrostatic potential maps and polarizability from QM were fitted with the MM energy function by adjusting the atomic charges, polarizabilities and Thole factors. In a second phase, these optimized parameters were further adjusted in order to reproduce the QM solute:water interaction energies, together with the QM molecular dipole moment. For details about the parameter fitting procedure, see the section "Methods: fitting the QM quantities". Once electrostatic parameters of MP^{-} were obtained, we optimized those of MP^{2-} (2D structure shown in figure 5.4, center). We fitted the MP^{2-} atomic charges to all the QM data at once (electrostatic potential maps, water scans, molecular dipole moment and polarizability) by means of a grid search in MM parameter space. Once electrostatic parameters of MP^{-} and MP^{2-} were obtained, we transferred them to $P_i^{2-} = HPO_4^{2-}$ (2D structure shown in figure 5.4, right) and then reoptimized with a grid

search.

In this study we did not optimized LJ parameters. Indeed, most atom types were taken from the existing dimethyl phosphate (DMP) in the nucleic acid Drude force field; the main exception was an oxygen (O1) typed according to the lipid Drude force field. We also studied DMP with QM and with the existing Drude force field, to better evaluate the quality of our fit for methyl phosphate. For both MP^{-} and MP^{2-}, the overall QM/MM agreement was similar to that obtained for DMP.

For both MP^{-} and MP^{2-}, we then went on and combined the optimized electrostatic parameters with existing Drude bonded parameters. All steps stayed as close as possible to existing Drude types/parameters, with input also from C36 where relevant (eg, a few bonded parameters). Few dihedrals were scanned with QM. The total energy for several dihedral configurations was computed with QM and then compared to MM. In this phase, few parameters were optimized to improve agreement. At the end of the parametrization, we tested the overall MM model further by comparing MM conformational geometries and energies to QM.

Final MP and P_i^{2-} models were validated computing their standard binding free energies with Mg^{2+}. The magnesium binding model is described below. Affinities were computed with alchemical free energy simulations (see Methods: phosphate:Mg^{2+} binding free energies).

Dianionic phosphotyrosine was parametrized starting from existing Drude force field information and the MP^{2-} final set of parameters. The $pTyr^{2-}$ side chain was modeled using a model compound, the phosphorylated form of the p-Cresol (see Results: Drude model for dianionic phosphotyrosine). QM ESP maps (unperturbed+perturbed) and solute-water interaction energies were fitted. Phosphotyrosine parametrization is still underway, but the model was partially validated running polarizable molecular dynamics of the Tiam1:pSdc1 complex.

2.4 Phosphate:Mg^{2+} binding model

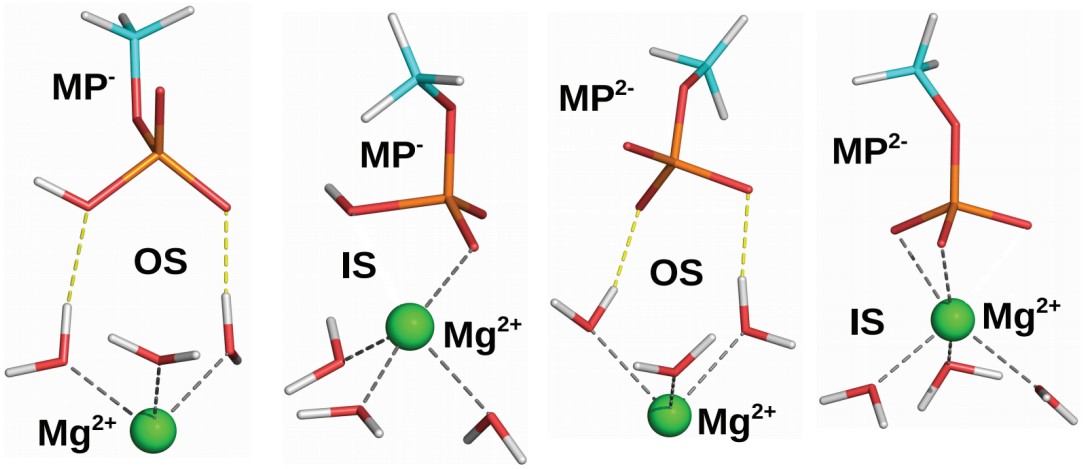


Figure 5.5 – **Internal and Outer shell binding** for methyl phosphate.

The $MP^{-/2-} : Mg^{2+}$ binding is governed by the equation:

$$(Mg^{2+})_w + (MP^{-/2-})_w \xrightarrow[\text{bind}]{\Delta G} (Mg^{2+}MP^{-/2-})_w \tag{5.1}$$

where the subscript w is used to indicate that the binding reaction takes place in pure water. For P_i^{2-}, the equation is similar. There are two possible bound states: Outer Shell (from now called OS) and Internal Shell (from now called IS) shown in figure 5.5. In IS, the phosphate interacts directly with magnesium ion, while in OS this interaction is mediated by a shell of water molecules. Experimental affinities for anionic and dianionic phosphates are known for $H_2PO_4^-$ (-1.7 kcal/mol) and $P_i^{2-} = HPO_4^{2-}$ (-3.7 kcal/mol) (Verbeeck *et al.* [1984], Alberty & Goldberg [1992]). To obtain the binding free energies, we have run a series of alchemical MD simulations. The detailed protocol is explained below in the text (see Methods: phosphate:Mg^{2+} binding free energies).

This study is motivated by two facts. First, methyl phosphate parameters were built using the established parametrization procedure, where most of the calculations involved constrained geometries and interactions in vacuum. To better validate our model it is important to check its behaviour during molecular dynamics in SWM4-NDP water. Binding affinities can give important insights about reiability of the developed parameters, and how this simple polarizable model can reproduce biochemical properties. Second, despite several phosphate-containing moieties were already parametrized in the official Drude force field, the correct threatment of their interactions with charged ions are still challenging. In previous studies, a strong effort has been made to develop consistent Drude force field parameters to balance

interactions between Mg^{2+} and nucleic acids (Lemkul & MacKerell [2016a]). Several *ad-hoc* parameters were used to optimize interaction energies and geometries of complexes involving nucleic acid moieties, water and magnesium. However, they were not extensively tested. Magnesium binding free energy simulations are thus a stringent test not just for our new models, but also for the existing Drude force field parameters.

3 Methods: QM calculations

Quantum calculations were performed using as reference the QM optimized geometry. Optimization was done with GAUSSIAN, using MP2/6-31G*. We extracted the optimized geometry from the GAUSSIAN output and then prepared several input files, used to calculate the following QM data.

3.1 QM electrostatic potential maps

Ten concentric grids based on Connolly surfaces were positioned around each solute using the program CGRID from GAAMP, totalling around 2000 grid points and ranging up to about 5 Å from the solute atoms. An example of such a grid is given in figure 5.6, left. GAUSSIAN allows to automatically generate an uniformly distribute cubic grid with the molecule centered in it, however the present strategy allow to reduce the number of points and to better adapt to the solvent accessible surface of different molecules. Points which are far from the center are inserted with a lower density factor respect to points distributed over closer Connelly surfaces. 67 perturbing charges were distributed around MP^- and 69 around MP^{2-} [fig. 5.6, right]. Grid layers were alterned with perturbing charges layers. The electrostatic potential was then evaluated for each grid point i and for each perturbing charge j, totalling 67(69) perturbed ESP maps $\sum_i \phi_{ji}^{QM}$ for $i = 1, ..., n_{grid}$ plus one unperturbed map $\sum_i \phi_i^{QM}$. During the QM calculation, the model compound was kept fixed at its optimized geometry, which is not relaxed for different perturbations. As grid points, perturbing charges were distributed uniformly on connelly surfaces, except some of them which were placed at specific positions following the protocol described by Anisimov *et al.* [2005]. A first group of charges sits along molecule's bond lines, at the intersection with the surfaces. Then a second group of charges was positioned around oxygen or nitrogen atoms when they form covalent bonds with two other atoms. Finally, the charge layers were covered respecting a minimum distance threshold between couples of charges of 1.5 Å. The electrostatic potential was evaluated with GAUSSIAN, using MP2/6-31G*.

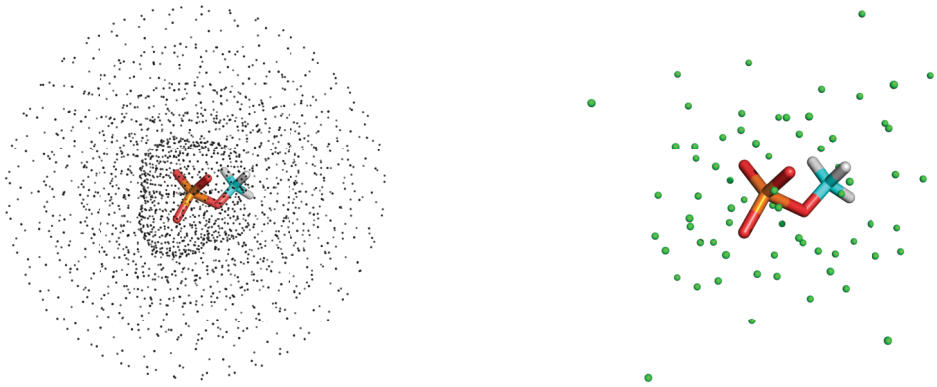


Figure 5.6 – **QM electrostatic potential maps** for MP^{2-}. **Left**: Grid for evaluation of electrostatic potential. **Right**: perturbing charges, in green.

3.2 QM polarizability and dipole moment

We computed molecular polarizability of MP^{-}, MP^{2-} and P_i^{2-} using MP2/6-31G*. The "exact polarizability" (au^3) was extracted from the GAUSSIAN output and converted to $Å^3$ (1 au = 0.529177249 Å). The dipole moment was computed with PSI4 using RIMP2/6-31G*. Only diagonal elements of the $3x3$ polarizability matrix were considered because they are considerably greater than off-diagonal elements, which were considered negligible.

3.3 QM solute–water and solute–ion interaction energy scans

We start describing the protocol for water scans. Water poses were generated with GAUSSIAN (*scan* keyword) or with a CHARMM script, varying *only* the solute–water distance in the [1.4:3.0] Å range, in 0.1 Å steps. During each scan, the solute and water internal geometries were held fixed. Internal coordinates were specified in a Z-matrix and the solute–water distance was controlled by an r_{OH} variable. For each pose, the interaction energy was computed with PSI4 using RIMP2/cc-pVQZ with BSSE corrections. To define water poses relative to a specific acceptor/donor, a DUMMY particle was added to the solute-water system [fig. 5.7, left]. This auxiliary atom was then used to specify angles and dihedrals that define the internal geometry of the compound:water complex. In the figure, the water hydrogen H1, the phosphate oxygen O2 and the dummy atom DUM form together an angle of 90°.

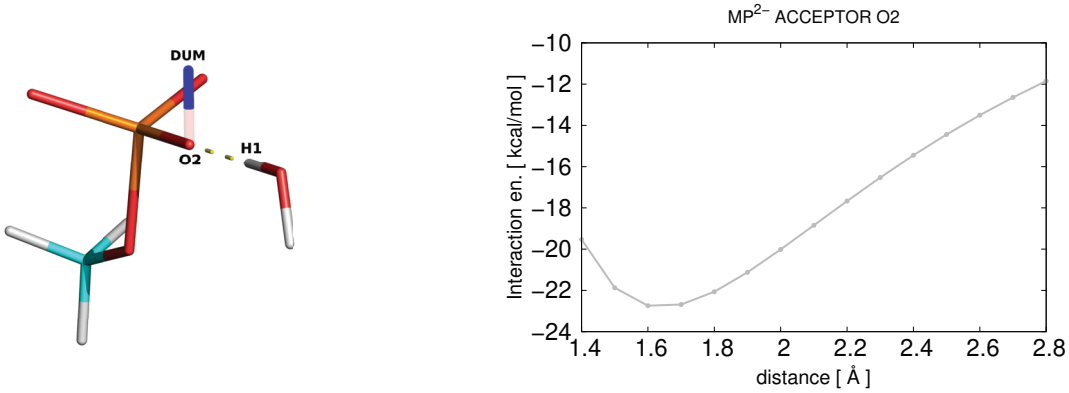


Figure 5.7 – **Water-solute scans. Left**: water pose built using a dummy particle (shown in blue) on the MP^{2-} acceptor O2. **Right**: QM interaction energy scan for the MP^{2-} acceptor O2.

Solute–ion scans for magnesium (Mg^{2+}) and sodium (Na^{+}) were performed following an equivalent protocol, except that QM calculations with PSI4 were done using MP2/aug-cc-pvtz with BSSE corrections.

MP^{-} has four hydrogen bond acceptors and one donor: O1, 02, O3, O4, and H2. Although O3 and O4 have equivalent types, they are not symmetrical in the optimized structure so we did a separate scan for each atom. For MP^{2-}, we scanned the three acceptors: O1, O2, O3 (equivalent to O4), while for P_i^{2-} we scanned acceptors O1, O2 and O3. Interactions were also scanned for the bridging oxygen of dimethyl phosphate (DMP).

3.4 QM dihedral scans

Starting from the QM optimized geometry, selected dihedrals were scanned using angles $\psi \in [-\pi : \pi]$ with a step of 15°. We scanned O1-P1-O2-H2 and C1-O1-P1-O2 dihedrals for MP^{-} and the C1-O1-P1-O2 dihedral for MP^{2-}. Scans were performed with GAUSSIAN (*scan* keyword) or with PSI4, using MP2/6-31+G(d). In the GAUSSIAN input file, geometries were defined by a z-matrix where the target dihedral angle a variable. For each configuration visited during the scan, the system energy was computed after geometry relaxation, performed keeping the scanned dihedral fixed at its value ϕ. The QM energy and geometries were saved for eachi value of ψ, to be compared afterwards with MM.

3.5 QM vibration frequencies

We performed vibrational analysis using Gaussian with the *freq* keyword. We used MP2/aug-cc-pVTZ for geometry optimization and frequencies calculation. From the QM output we

extracted frequences and vectors for all the normal modes and also force constants in cartesian coordinates, represented in matrix form (the Hessian). This $3Nx3N$ matrix has entries H_{ijkl} where i,$j = 1$, ..., N is are atom IDs and k,$l = x$, y ,z.

$$H_{ijkl} = \frac{\partial U^{QM}}{\partial k \partial l} \tag{5.2}$$

The QM force constants matrices of MP^{-} and MP^{2-} were used to compare QM and MM energies for random displacements generated by combination of the extracted QM modes. Indeed, for small coordinate displacements $\sum_i d_i$ (where $i = 1$,...,Natoms) around the molecule's optimized geometry, the QM energy can be estimated from the force constants matrix using the harmonic approximation

$$\tilde{U}^{QM} = \frac{1}{2} \sum_{ijkl} d_{ik} H_{ijkl} d_{jl} \tag{5.3}$$

where i,j are atoms IDs and k,$l = x$, y ,z.

4 Methods: fitting the QM quantities

4.1 Fitting QM potential maps

The QM electrostatic potential maps were fitted using the GAAMP "*drude-fitcharge*" module. In addition to the above constraints on the molecule geometry, restraints were applied to charges, polarizabilites and Thole parameters to avoid large deviations from physically reasonable reference values. The target function χ^2 is composed of several terms:

$$\chi^2 = \chi^2_{ESP} + \chi^2_{CG} + \chi^2_{ALPHA} + \chi^2_{THOLE} \tag{5.4}$$

The first term tries to match the QM and MM potential maps. In the presence of n_{grid} points and n_{pert} perturbation charges,

$$\chi^2_{ESP} = \frac{10^4}{2 n_{pert} n_{grid}} \left[n_{pert} \sum_{i=1}^{ngrid} (\phi_i^{MM} - \phi_i^{QM})^2 + \sum_{j=1}^{n_{pert}} \sum_{i=1}^{n_{grid}} (\phi_{ji}^{MM} - \phi_{ji}^{QM})^2 \right] \tag{5.5}$$

The other terms are the restraints, weighted by the constant factors w_{CG}, w_α and w_τ (set to default values). Reference values for charges are those specified in the initial topology file $\{q_i^0\}$:

$$\chi^2_{CG} = \frac{w_{CG}}{n} \sum_{i=1}^{n} f(q_i,q_i^0) \tag{5.6}$$

where n is the number of atoms and

$$f(q_i,q_i^0) = \begin{cases} 0 & |q_i - q_i^0| \leq 0.03 \\ \left(|q_i - q_i^0| - 0.03\right)^2 & \text{otherwise} \end{cases} \tag{5.7}$$

Polarizabilities α_i and Thole parameters τ_i were restrained using

$$\chi^2_{ALPHA} = \frac{w_\alpha}{n} \sum_{i}^{n_p} \left(\alpha_i - \alpha_i^0\right)^2 \qquad \chi^2_{THOLE} = \frac{w_\tau}{n} \sum_{i}^{n_p} \left(\tau_i - \tau_i^0\right)^2 \tag{5.8}$$

where n is the number of polarizable atoms. Reference polarizabilities α_i^0 were taken from (Miller [1990]) then scaled by 0.7. Reference Thole factors τ_i^0 were set to 1.3 for all atoms; this corresponds to average screening parameter $\tau_{ij} = \tau_i + \tau_j = 2.6$, as originally proposed by Thole.

4.2 Fitting the molecular polarizability

Using GAUSSIAN, one obtains $\hat{\alpha}$ in the $\{xyz\}$ frame. Its diagonal components are denoted α_{xx}, α_{yy}, α_{zz}. Atomic polarizabilities and Thole factors were manually adjusted to reproduce the QM values of α_{xx}, α_{yy}, α_{zz} (and thus their sum, the polarizability trace). The atomic charges were held fixed (since $\hat{\alpha}$ is independent of the charges). During adjustment, we cross-checked that the polarizability and Thole modifications did not affect too much the quality of the potential maps (the value of χ^2_{ESP}). The Drude molecular polarizability was obtained using the CHARMM script *polar_efield* (from DGENFF). Diagonal elements of $\hat{\alpha}_{mol}$ in the xyz frame were computed by measuring the variation of the molecular dipole moment under the influence of an external electric field of intensity $E = 10^{10}$ V/m. When measuring the electronic polarizability, the atoms were kept fixed in the molecule's optimized geometry. Drude particles were relaxed before computing the dipole moment in the presence or absence of the applied field. The field was applied using the CHARMM command

```
pull efield 1E10 xdir <float> ydir <float> zdir <float>
```

The molecular mechanics calculation of α_{xx}, α_{yy}, α_{zz} was done in 4 steps. First, the molecular dipole moment $\bar{\mu}^0$ was computed in the absence of an external field, with the CHARMM command "coor dipole oxyz". Next, values of the molecular dipole moment $\bar{\mu}^x$, $\bar{\mu}^y$, $\bar{\mu}^z$ were computed in the presence of three different external fields, oriented along x, y, and z, respectively. The nine components of the molecular polarizability tensor were then obtained using

$$\alpha_{ij} = \frac{\mu_j^i - \mu_j^0}{E} \tag{5.9}$$

4.3 Fitting solute–water interaction energies and the molecular dipole

Two strategies were used. For MP^-, we adjusted the atomic charges manually, without changing the polarizabilites or Thole values. For one atom (the bridging oxygen O1) we departed from the Drude DMPN vdW types (see Results). After adjusting the charges, we checked the quality of the potential maps, the dipole and the polarizability. For MP^{2-}, we optimized the atomic charges to fit all the QM data at once, through a simple grid search in parameter space. Starting from MP^- parameters, we transferred atomic polarizabilities and thole factors to MP^{2-}. Atomic charges were readjusted to match the total charge. Then we fitted molecular polarizability (not dependent on atomic charges), adjusting polarizability and Thole factors manually. As for MP^-, targeted molecular polarizabilities were scaled by a factor 0.85. Once α and τ factors were obtained, we went on performing a grid search for many values of atomic charges to fit water interaction energy scans for acceptors O1, O2, O3. An equivalent protocol was used to derive electrostatic parameters of P_i^{2-}.

5 Methods: phosphate:Mg^{2+} binding free energies

5.1 Free energy perturbation protocol

To obtain the $MP^{-/2-}$:Mg^{2+} and P_i^{2-}:Mg^{2+} binding free energies, we simulated the alchemical annihilation or "decoupling" of magnesium alone in solution and bound to the phosphate solute, over a series of MD simulations. For MP^- and MP^{2-}, we considered both IS and OS binding. For P_i^{2-}, we considered OS binding. From each simulation, the binding free energy was obtained using the Double Decoupling Method (DDM). Here we describe the procedure for methyl phosphate, given that for P_i^{2-} we used exactly the same protocol. The thermodynamic

cycle is displayed in figure [5.8]. The starting point was a standard-state ideal-dilute solution of the unbound $MP + Mg^{2+}$ at concentration $\rho^0 = 1M$, the endpoint was a standard-state ideal-dilute solution of the complex ($MP^{-/2-}Mg^{2+}$) (bound state).

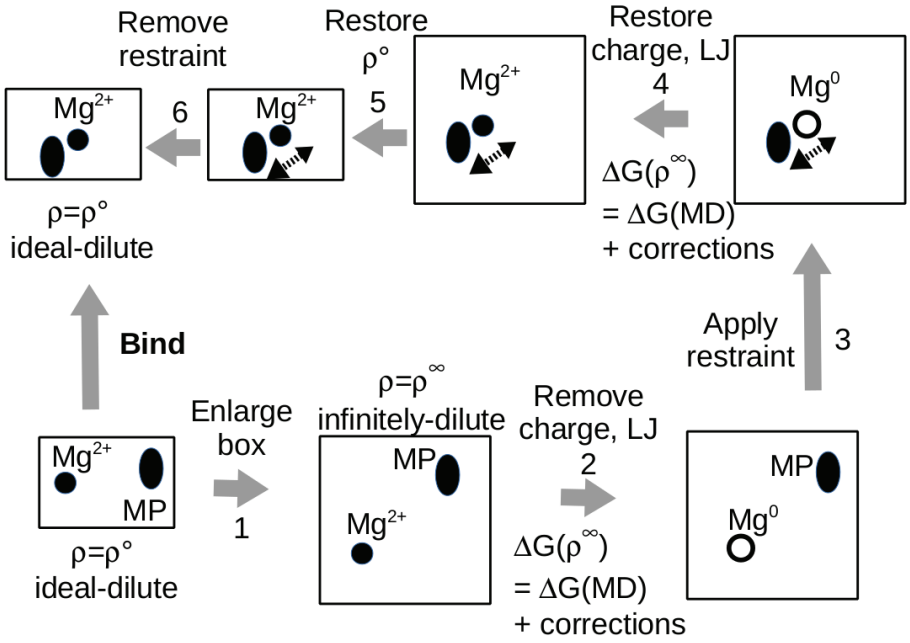


Figure 5.8 – **Thermodynamic cycle** used to compute absolute binding free energies for Mg^{2+} and methyl phosphate complexes (vertical arrow on the left, indicated using the label *Bind*). The reaction is decomposed in several steps labelled with integers from **1** to **6** and described in the text.

The binding reaction was decomposed in six steps, connecting starting and ending states:

[**1**] We changed the concentration to the very low value ρ^∞. The corresponding free energy change was:

$$\Delta G(\rho^0 \rightarrow \rho^\infty) = -2KTln\rho^\infty/\rho^0 \tag{5.10}$$

[**2**] We removed the Mg^{2+} charge and LJ interactions alchemically, in two steps. First, magnesium charge and polarizability were turned off. The corresponding free energy difference was ΔG_{elec}. Then, Lennard-Jones interactions of the uncharged Mg were removed giving a second contribution ΔG_{LJ}.

[**3**] A flat-bottomed harmonic restraint was introduced to keep Mg^{2+} close to MP. The re-

straint potential was, for both OS and IS states:

$$U_R(r) = \begin{cases} \frac{1}{2}c(r-a)^2 & r < a \\ \frac{1}{2}c(r-b)^2 & r > b \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

where r is the distance between Mg^{2+} and the MP phosphorus atom P. For IS, $[a;b]$ was set to [0.0 : 3.5Å] and c=4 $kcal/mol/Å^2$. For OS we used [4.1 : 5.9Å] and c=8 $kcal/mol/Å^2$. The restraint free energy was obtained analitically, using

$$\Delta G_{restr} = KTln(\rho^{\infty}Q(c)) \tag{5.12}$$

where

$$Q(c) \simeq \int_0^{\infty} 4\pi r^2 \exp[-\beta U_R(r)]\, dr = \tag{5.13}$$

$$= \int_0^{a} 4\pi r^2 \exp[-\frac{\beta c}{2}(r-a)^2]\, dr + \frac{4}{3}\pi(b^3-a^3) + \left(\frac{2\pi kT}{c}\right)^{3/2} + \frac{8\pi kTb}{c} + (2\pi)^{3/2}b^2\sqrt{\frac{kT}{c}}$$

[**4**] We alchemically reintroduced interactions between Mg and MP

[**5**] We restored the standard-state concentration, with the corresponding free-energy contribution $KTln(\rho^{\infty}/\rho^0)$.

[**6**] We removed the restraint and the corresponding free energy cost was negligible.

At the end, all terms involving ρ^{∞} (steps 1,3,4) cancel out, leaving a logarithmic dependency on the standard-state concentration $kTln\rho^0$. Just free energies for steps 2 and 4 were computed with alchemical free energy differences. Details about these simulations are given below.

5.2 Simulations setup

System preparation was done with the CHARMM program. For unbound state simulations (point 2) the solute was one Mg^{2+} ion placed at the center of the box. For the bound state

simulations (point 4) the solute was represented by the MP-Mg^{2+} complex. OS and IS states were prepared separetly, putting the Mg^{2+} ion at average distance d=(a+b)/2 from the P atom, where a and b vary between OS and IS. Each solute was then solvated in a cubic box of side L=31 Å, containing around 1000 pre-equilibrated SWMD4-NDP water molecules. Solvent within the solute volume was deleted using a distance threshold of 3.0 Å from the solute heavy atoms. The system was then minimized with NAMD, first with restrained heavy atoms and unrestrained Drude particles, then without restraints. The minimized system was finally equilibrated, running 200 ps of MD.

All MD similations (equilibration and alchemical) were carried out using NAMD (version 2.12) at room temperature and pressure, using Langevin dynamics with a Langevin Piston Nosé-Hoover barostat (Feller *et al.* [1995]). The simulations used periodic boundary conditions and a Particle Mesh Ewald electrostatic treatment (Darden [2001]). Bonds to hydrogen atoms were constrained with the SHAKE method (Ryckaert *et al.* [1977]). LJ interactions were shifted to zero for separations between 10 and 12 Å. The MD timestep was 1 fs and the oscillators temperature was T_D=1 K. Throughout the simulations, the center of mass of MP (or P_i^{2-}) was restrained to the box center using an harmonic potential.

5.3 Alchemical free energy simulations

Alchemical free energy changes for steps 2 and 4 were extracted with TI [eqn. 3.51] from series of MD simulations. Decoupling of electrostatic interactions of Mg^{2+} in the bound or unbound state was performed by scaling its charges (the atom charge q_A and its Drude charge q_D), its polarizability α and Thole factor τ. MD simulations were run using the potential function of equation 4.2, for different values of the coupling parameter λ_{elec}=1, 0.9, 0.8, ..., 0.1, 0. Decoupling of LJ interactions was done using the *alch* facility of NAMD, with a series of simulations using λ_{LJ}=1, 0.9, 0.8, 0.6, 0.4, 0.2, 0. The electrostatic uncharging(respectively, charging) free energies at concentration ρ^∞ were extrapolated analitically for $L \to \infty$, using equation 3.71

$$\Delta G(L \to \infty) = \Delta G(L) + \frac{q^2|\zeta|}{2\epsilon_W L} + \frac{2\pi q^2 R^2}{3L^3}\frac{\epsilon_W - 1}{\epsilon_W} + O(R^2 L^{-3}\epsilon_W^{-3}) \tag{5.14}$$

where ϵ_W is the solvent dielectric constant, L is the box side length in the NPT simulation (on average 31.2Å) and $\zeta = -2.837297$. The correction was computed considering two distinct values of Mg^{2+} radius for bound and unbound states, respectively R=1.2 Å and R=1.5 Å.

6 Results

6.1 Initial MP model and atom types

Starting topology and parameter files for MP^- and MP^{2-} borrowed as much information as possible from the Drude dimethyl phosphate (DMPN) of the official Drude *nucleic acids* topology files. We give existing C36 and Drude information first.

```
----------------c36 methyl phosphate information-----------------------------------------------------------------------
RESI MP_1          -1.000       H11          RESI MP_2          -2.000
ATOM C1   CG331    -0.170        |           ATOM P1   PG2       1.100
ATOM O1   OG303    -0.620  H13--C1--H12      ATOM O1   OG303    -0.400        H11
ATOM P1   PG1       1.500        |           ATOM O2   OG2P1    -0.900         |
ATOM O2   OG311    -0.670       O1           ATOM O3   OG2P1    -0.900   H13--C1--H12
ATOM O3   OG2P1    -0.820        |           ATOM O4   OG2P1    -0.900         |
ATOM O4   OG2P1    -0.820 (-)O4==P1==O3      GROUP                           O1
ATOM H11  HGA3      0.090        |           ATOM C1   CG331    -0.270         |
ATOM H12  HGA3      0.090       O2           ATOM H11  HGA3      0.090 (-)O4==P1==O3(-)
ATOM H13  HGA3      0.090         \          ATOM H12  HGA3      0.090        ||
ATOM H2   HGP1      0.330          H2        ATOM H13  HGA3      0.090        O2



---------Drude dimethyl phosphate, from nucleic acids----------------------------------
RESI DMPN -1.000
ATOM P    PD1AN     1.191  ALPHA -0.974  THOLE 2.098         H11
ATOM O13  OD2C2C   -0.776  ALPHA -0.931  THOLE 1.083          |
ATOM O14  OD2C2C   -0.776  ALPHA -0.931  THOLE 1.083    H13--C1--H12
ATOM O11  OD30BN   -0.470  ALPHA -0.901  THOLE 0.181          |
ATOM O12  OD30BN   -0.470  ALPHA -0.901  THOLE 0.181         O11
ATOM C1   CD33C     0.0725 ALPHA -1.642  THOLE 0.862          |
ATOM H11  HDA3A     0.026                                (-)O14==P==O13
ATOM H12  HDA3A     0.026                                       |
ATOM H13  HDA3A     0.026                                      O12
ATOM C2   CD33C     0.0725 ALPHA -1.642  THOLE 0.862           \
ATOM H21  HDA3A     0.026                                      C2--H21
ATOM H22  HDA3A     0.026                                     /  \
ATOM H23  HDA3A     0.026                                    H23  H22
```

The initial MP^- is given below, where atom O2 (the hydroxyl oxygen) has a Lone Pair. Atom types of the initial Drude model were transferred from "nucleic acid" DMPN.

```
---------Initial Drude methyl phosphate MP----------------------------------------------
RESI MP_1 -1.000
ATOM P1     PD1AN   1.190  ALPHA -0.974  THOLE 2.098            H11
ATOM O1     OD30BN -0.438  ALPHA -0.901  THOLE 0.831             |
ATOM O2     OD31A  -0.150  ALPHA -0.927  THOLE 1.300       H13--C1--H12
ATOM LPAO2  LP     -0.180                                       |
ATOM LPBO2  LP     -0.180                                      O1
ATOM H2     HDP1A   0.160                                       |
ATOM O3     OD2C2C -0.776  ALPHA -0.931  THOLE 1.083      (-)O4==P1==O3
ATOM O4     OD2C2C -0.776  ALPHA -0.931  THOLE 1.083             |
ATOM C1     CD33C   0.072  ALPHA -1.642  THOLE 0.862            O2
ATOM H11    HDA3A   0.026                                       \
ATOM H12    HDA3A   0.026                                        H2
ATOM H13    HDA3A   0.026
ATOM H23    HDA3A   0.026
```

We start giving results for the MP^- parametrization, describing in detail all the intermediate sets of parameters. Results for MP^{2-} and P_i^{2-} will be given later, as they were obtained all at once with a grid search. MP^- parametrization was carried out following three steps executed in sequence, resumed in table 5.1.

Set	Construction	Target data
1	adjusted q_i, α_i, τ_i to minimize χ^2 with GAAMP	QM electrostatic potential maps
2	charges from set 1, manually adjusted α_i and τ_i	QM molecular polarizability
3	α_i and τ_i from set 2, manually adjusted charges	QM dipole moment and water scans

Table 5.1 – **Electrostatic parameters optimization workflow** for MP^-, with intermediate and final parameter sets. Each set was optimized fitting different QM target data, as described in the text.

6.2 Fitting the MP^- potential maps

Atom	Parameters	Atom	Parameters
O1	polarizability, Thole	O3	charge, polarizability, Thole
P1	charge, polarizability, Thole	O4	- redundant -
O2	charge, polarizability, Thole	C1	- fixed -
LPAO2	charge	H11	- fixed -
LPBO2	charge	H11	- redundant -
H2	charge	H11	- redundant -

Table 5.2 – **Electrostatics Parameters** for MP^-. Some of the parameters were fixed, other are redundant for atoms belonging to equivalent groups, as described in the text.

In table 5.2 we indicate the list of MP^- electrostatic parameters to optimize. We imposed two groups of equivalent atoms: (H11, H12, H13) and (O3, O4). Each group shares the same charge, polarizability and Thole parameters. For the restrained electrostatic potential (RESP) fit, we further reduced the number of parameters by fixing some of the charges at their DMPN values: q(H11)=q(H12)=q(H13)=0.026 and q(C1)=0.0725. The charge of atom O1 is a function of all the other charges. Initial Thole factors were set to 1.3 and weights for restraints were $w_{CG} = 3.0$, $w_\alpha = 0.4$, $w_\tau = 0.2$ (the defaults). The performance and parameters after fitting (set 1) are given below in table 5.3 and table 5.4.

χ^2	ESP	Charge	α	τ	Std. Err. (ESP)
0.9046	0.8042	0.0705	0.0192	0.0107	0.009

Table 5.3 – **Restrained fit quality** for MP^- including all contributions to the target function from equations 5.5, 5.6 and 5.8. The target function is close to the ESP fit quality, meaning that optimized parameters don't deviate too much respect to applied restraints.

Atom	charge	α	τ	Atom	Charge	α	τ
O1	-0.479	-0.360	1.305	O3	-0.738	-0.497	1.391
P1	1.122	-1.254	1.395	O4	-0.738	-0.497	1.391
O2	0.097	-0.457	1.194	C1	0.072	-1.642	0.767
LPAO2	-0.290			H11	0.026		
LPBO2	-0.290			H12	0.026		
H2	0.166			H13	0.026		

Table 5.4 – **Intermediate parameters set 1** for MP^-, "post-ESP" fit.

6.3 Fitting the MP^- polarizability

We first computed the diagonal elements of the polarizability tensor using parameter set 1 (post-ESP fit). All the components were understimated compared to the target value, which is the QM result scaled by 0.85. We manually adjusted polarizabilities and Thole parameters. We found that $\hat{\alpha}$ was well reproduced by transferring parameters from DMPN and modifying only the O1 Thole value. Fit results are in table 5.5 and optmized parameters (set2) in table 5.6.

	α_{xx}	α_{yy}	α_{zz}	α
QM	8.2846	7.2605	6.9212	7.4888
QM×0.85	7.0419	6.1714	5.8830	6.3655
MM, set 1	5.3153	4.4205	4.004	4.7121
MM, set 2	7.0912	6.0173	5.9377	6.3487

Table 5.5 – **Molecular polarizabilities** in ($Å^3$) fitted using initial values (intermediate parameters set 1, "post-ESP"fit) and then optimized with the intermediate parameters set 2.

6.4 Fitting the MP^-–water radial interaction energy scans

Starting from the set 2 parameters, we manually modified charges to better reproduce energy scans and also the molecular dipole moment. During adjustment, we periodically checked the quality of the RESP fit. Below, we report QM (target) and MM interaction energies e_{min} at

Atom	Charge	α	τ	Atom	Charge	α	τ
O1	-0.479	-0.901	0.831	O3	-0.738	-0.931	1.083
P1	1.122	-0.974	2.098	O4	-0.738	-0.931	1.083
O2	0.097	-0.927	1.300	C1	0.072	-1.642	0.862
LPAO2	-0.290			H11	0.026		
LPBO2	-0.290			H12	0.026		
H2	0.166			H13	0.026		

Table 5.6 – **Intermediate parameters set 2** for MP^-, after fitting the molecular polarizability.

the minimum distance r_{min} for the MP^- acceptor and donor atoms [tab. 5.7]. We also report the optimized charges, "set 3" [tab. 5.9]. The solute–water interaction for acceptor O1 was poorly fitted initially. The problem was related to the O1 type/vdW parameters we had transferred from the existing Drude atom type OD30BN. Indeed, the results suggest that the vdW parameters of atom type OD30BN (Drude FF, nucleic acids) are not ideal for the MP^-–water interaction when scanning the acceptor O1. Therefore, we studied the same interaction in DMP^-, which carries the same atom type on its bridging oxygen.

We optimized the DMP geometry with Gaussian using HF/6-31G* and then we scanned the water interaction energy, focussing on the the same acceptor O11 (equivalent to O1 in MP^-). Interaction energies were obtained with PSI4 using RIMP2/cc-pVQZ with BSSE corrections. Computing interaction energies with the MM energy function and the existing Drude DMPN parameters, we obtained the same error: the interaction e_{min} was too low and r_{min} was strongly underestimated. Minimum energy at minimum distance were more like those of MP^- $r_{min} =$ 1.5 Å and $e_{min} = -11.544$ kcal/mol, where QM values for DMPN were $r_{min} = 1.9$ Å and $e_{min} =$ -10.996 kcal/mol. Results are shown in figure 5.9. At this point, we repeated the MM scan using a different set of parameters: we assigned the vdW parameters of a different atom type OD30B (Drude FF, lipids) and we transferred to DMP the electrostatic parameters obtained at the end of the MP^- optimization (set 3). This transfer required a small charge adjustment to respect the total charge of DMP^-. With this substitution, the error was completely removed and the MM radial scan fit well the QM target data. Values were $r_{min} = 1.8$ Å and $e_{min} =$ -10.784 kcal/mol, more like those obtained with QM.

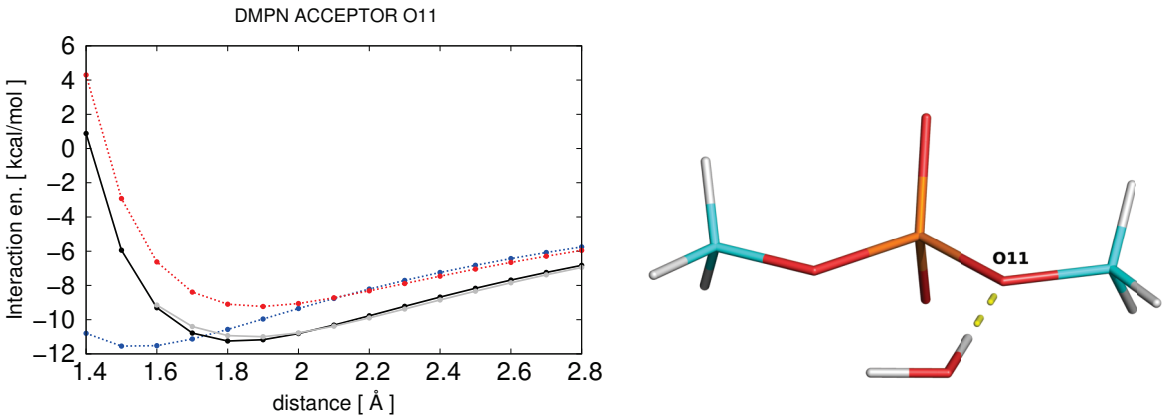


Figure 5.9 – **Water scan** for DMP. **Left**: interaction energy as function of the water-acceptor O11 distance. Several lines represent the scan performed using different parameters set. In blue, we show the scan performed using the official Drude parameters for DMPN (nucleic acids topology). In red, we show the scan performed after substitution of the atom type OD30BN with the lipid atom type OD30B. Charges were reoptimized, giving perfect agreement showed in the black line. QM is shown in grey. **Right**: geometry of the complex used for the scan.

	QM		MM, set 2		MM, set 3	
atom	r_{min}	e_{min}	r_{min}	e_{min}	r_{min}	e_{min}
O1	**1.9**	-11.663	**1.5**	-12.675	**1.5**	-12.663
O2	1.9	-10.583	1.9	-9.456	1.8	-10.892
O3	1.8	-13.433	1.7	-13.030	1.7	-13.759
O4	1.8	-12.934	1.7	-13.206	1.7	-14.076
H	2.0	0.073	2.3	3.036	2.0	0.5518

Table 5.7 – **Solute-water interaction energies** for MP^- , fitted using the intermediate set (after fitting the molecular polarizability) and then optimized with the intermediate parameters set 3. We highlight the strong deviation of r_{min} for acceptor O1, the methyl phosphate bridging oxygen.

	μ_x	μ_y	μ_z	μ
QM	-4.2395	2.5428	-1.0626	5.0565
MM, set 2	-2.7443	1.6167	-0.5710	3.2358
MM, set 3	-3.1737	2.5779	-0.5337	4.1234

Table 5.8 – **Molecular dipole moment** for MP^- (Debyes), obtained using the intermediate set 2 parameters and then optimized with the parameters set 3.

At this point, we replaced the atom type OD30BN with the atom type OD30B also for MP^-. Then repeated the charge adjustment. With this new set of charges, the radial interaction energy scans were better fitted for O1, O3, O4 and the donor H, while for O2, the result

	Set 2			Set 3		
Atom	Charge	α	τ	Charge	α	τ
O1	-0.479	-0.901	0.831	-0.470	-0.901	0.831
P1	1.122	-0.974	2.098	1.162	-0.974	2.098
O2	0.097	-0.927	1.300	0.000	-0.927	1.300
LPAO2	-0.290			-0.310		
LPBO2	-0.290			-0.310		
H2	0.166			0.360		
O3	-0.738	-0.931	1.083	-0.796	-0.931	1.083
O4	-0.738	-0.931	1.083	-0.796	-0.931	1.083
C1	0.072	-1.642	0.862	-0.008	-1.642	0.862
H11	0.026			0.056		
H12	0.026			0.056		
H13	0.026			0.056		

Table 5.9 – **Parameters set 3** for MP^-, after fitting the water interaction energy scans. The intermediate set 2 is shown for comparison.

was slightly worse but still acceptable. To better evaluate the quality of the final set of parameters, the resulting fit for MP^- will be given below, together with results for MP^{2-} and P_i^{2-}.

6.5 Electrostatic parameters optimization for MP^{2-} and P_i^{2-}

Parameters for dianionic methyl phosphate (MP^{2-}) and hydrogen phosphate P_i^{2-} were obtained performing a grid search in parameter space. Initial models were created transfering MP^- atom types and electrostatic parameters, rapidly adjusted to match the total charge. Polarizabilities and Thole factors were manually adjusted to reproduce the QM polarizability. Then the charges were optimized performing a grid search, that fits all QM data at once. The full QM/Drude comparisions are shown below MP^{2-} and P_i^{2-}, including MP^- data for comparision. In table 5.10, we give the fit of molecular polarizabilities and dipole moments. In table 5.11 we give the result of the fitted water scans. The configurations used for the scans are given in figure 5.10 for MP^- (MP^{2-} and P_i^{2-} water poses were similar). In figures 5.11 and 5.12 we plot all the scans for MP^- and MP^{2-}, QM vs MM.

molecule	method	α_{xx}	α_{yy}	α_{zz}	α	method	μ_x	μ_y	μ_z	μ
MP^-	QM	8.28	7.26	6.92	7.49	QM	-4.24	2.54	-1.06	5.06
	QM×0.85	7.04	6.17	5.88	6.36	MM	-2.74	1.62	-0.57	3.23
	MM	7.09	5.96	5.88	6.31					
MP^{2-}	QM	9.83	7.80	7.92	7.49	QM	-4.57	-0.45	0.01	4.59
	QM×0.85	8.36	6.63	6.73	7.24	MM	-4.23	0.47	0.00	4.26
	MM	7.59	6.50	6.49	6.86					
P_i^{2-}	QM	7.05	6.72	6.69	6.82	QM	0.12	-1.10	-2.36	2.61
	QM×0.85	5.99	5.71	5.68	5.80	MM	0.08	-0.67	-1.10	1.29
	MM	4.75	4.80	4.47	4.67					

Table 5.10 – MP^-, MP^{2-} and P_i^{2-} molecular polarizabilities (Å^3) and dipole moments (D)

		QM		MM	
molecule	atom	r_{min}	e_{min}	r_{min}	e_{min}
MP^-	O1	1.9	-11.66	1.8	-11.40
	O2	1.9	-10.58	1.8	-10.93
	O3	1.8	-13.43	1.7	-13.66
	O4	1.8	-12.93	1.7	-13.93
	H	2.0	0.07	2.0	0.55
MP^{2-}	O1	1.7	-21.08	1.7	-21.18
	O2	1.6	-22.75	1.6	-22.70
	O3	1.6	-22.88	1.6	-22.80
P_i^{2-}	O1	1.7	-20.77	1.6	-21.00
	O2	1.6	-22.41	1.6	-23.15
	O3	1.6	-22.61	1.6	-22.62

Table 5.11 – MP^-, MP^{2-} and P_i^{2-} water interaction distances (Å) and energies (kcal/mol)

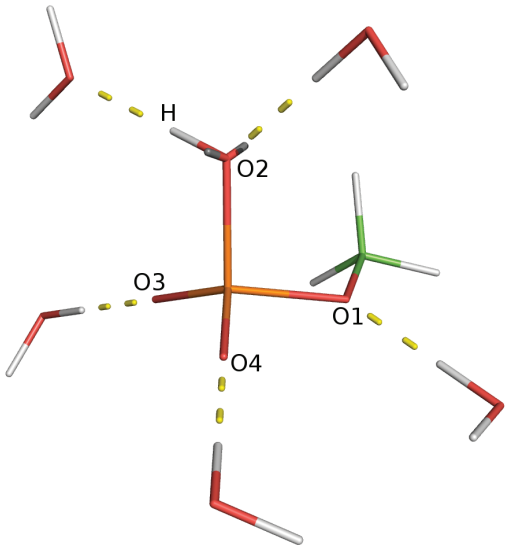


Figure 5.10 – Solute-water energy scan geometry for MP^-.

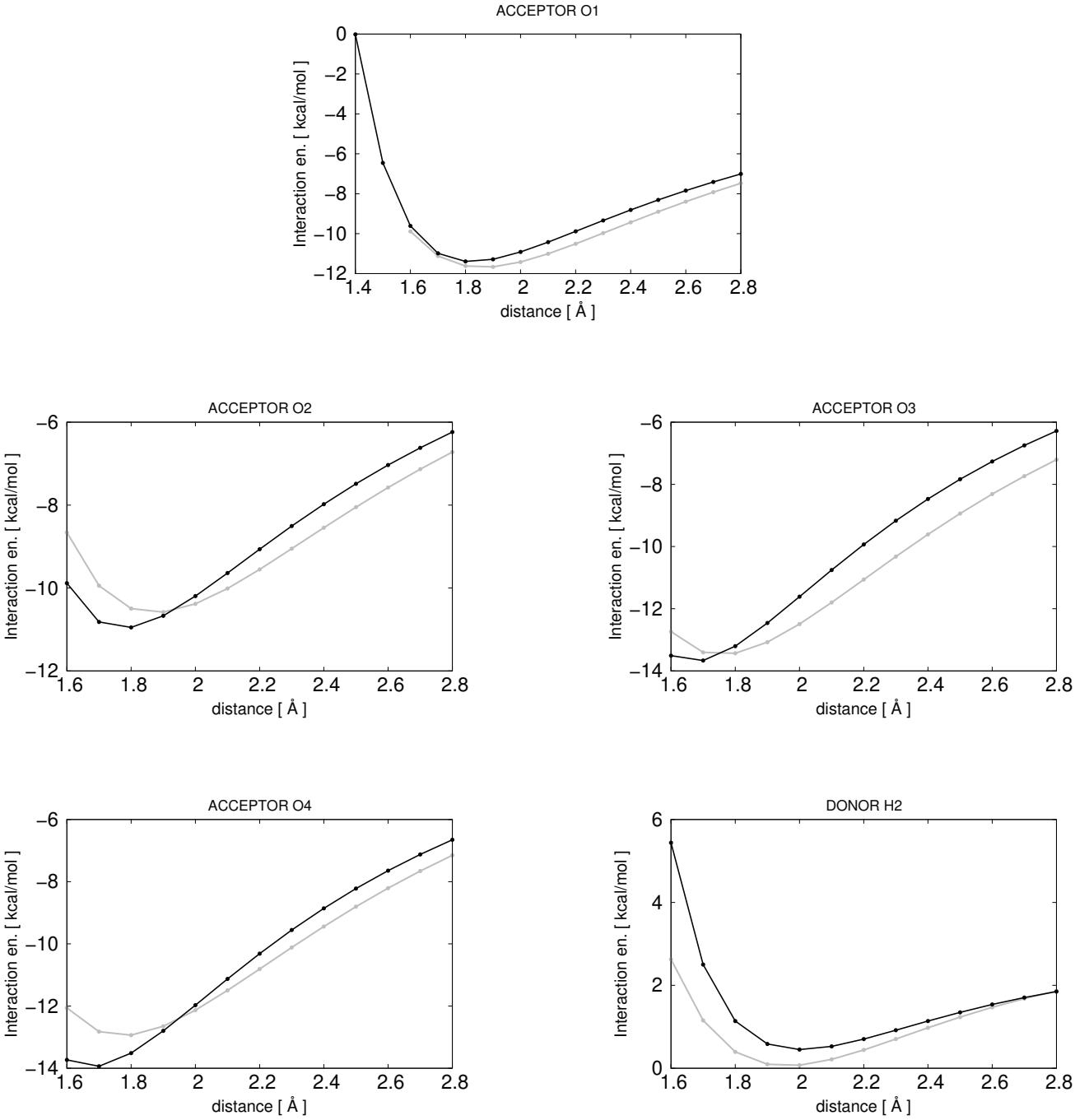


Figure 5.11 – **MP$^-$-water interaction** Grey: QM; Black: MM

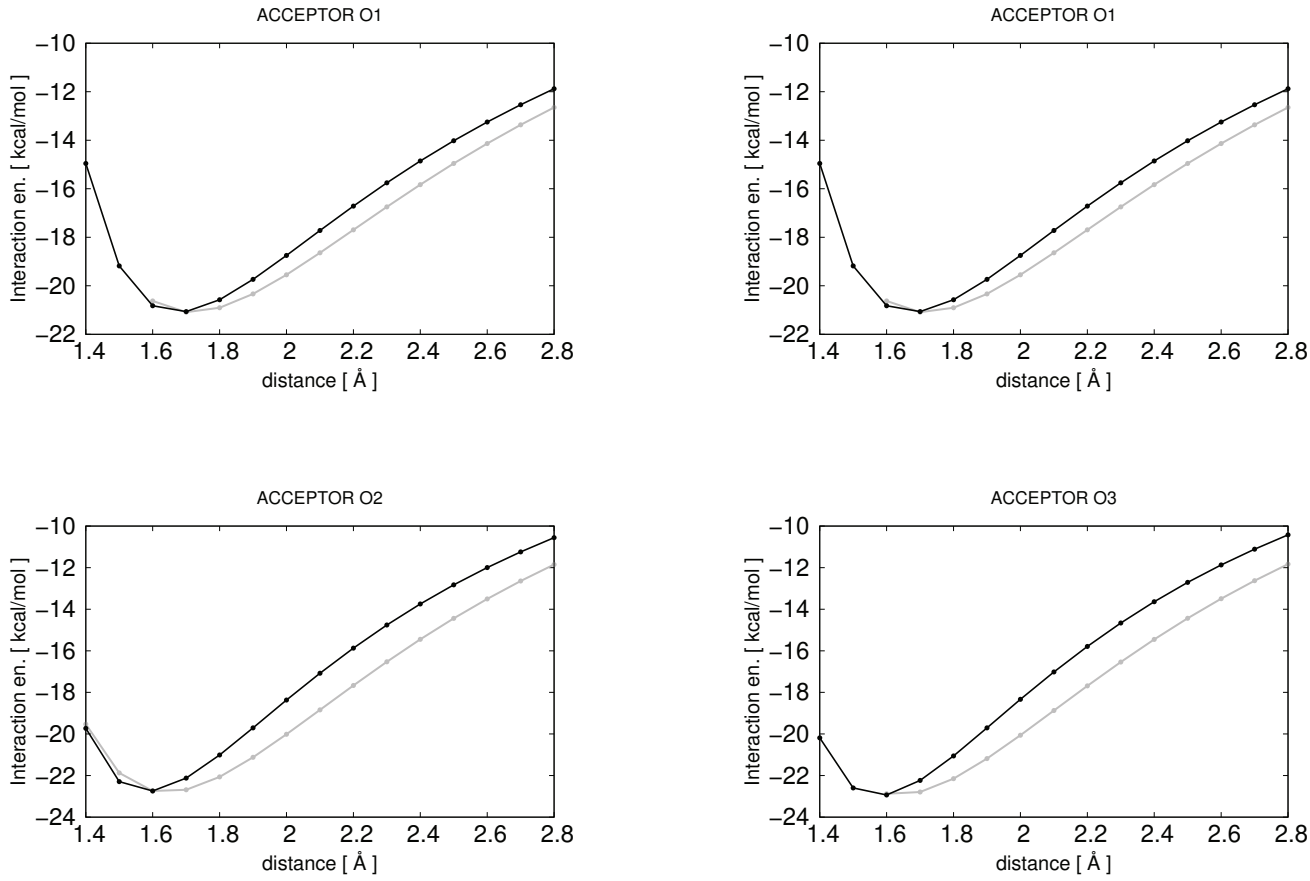


Figure 5.12 – **MP^2−-water interaction** Grey: QM; Black: MM

6.6 Fitting the dihedral energy scans

Electrostatic parameters to compute the MM energy are those optimized above. Internal parameters are those of DMPN except for the hydroxyl torsion (no present in the existing Drude force field) which has been borrowed from C36 and then reoptimized. In all scans, the molecule geometry energy was relaxed at each value of ψ in the selected interval, both in QM and MM calculations. More precisely, the scan has been performed first using QM. For each dihedral value, the QM relaxed conformation has been saved and then re-minimized using MM. We obtained perfect agreement for MP^{2-} and small deviations for MP^{-} dihedrals. Scans (QM vs MM) are given in figure 5.13. Local minima differs for $\sim$ 1 kcal/mol and energy barriers are well reproduced using the MM energy function.

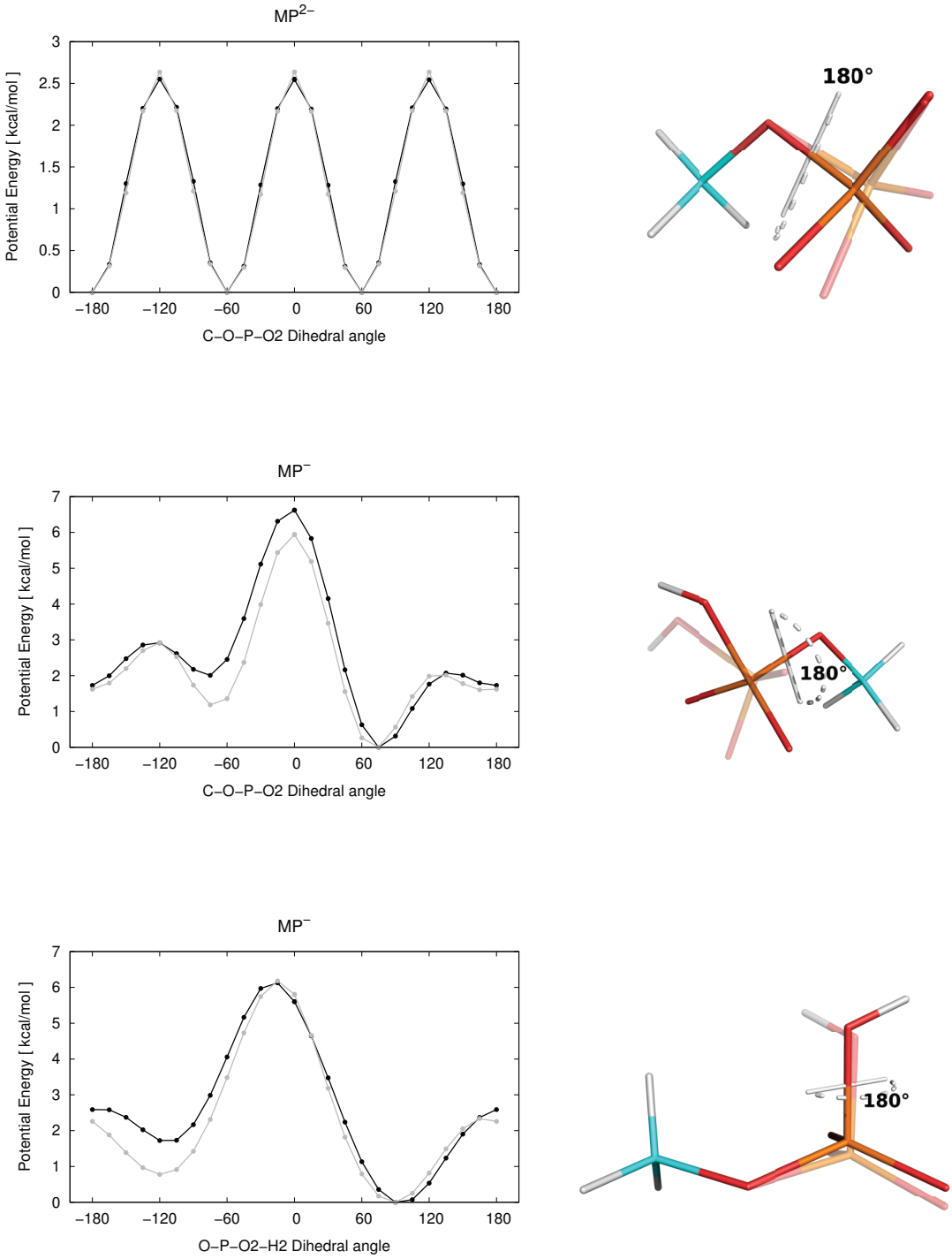


Figure 5.13 – **QM/MM torsion energy scans. Left**: result of the scans, Grey: QM; Black: MM. **Right**: geometries used for the scan (two structures of different ϕ, the first fixed at 180 degrees).

6.7 Conformational energies of MP^- and MP^{2-}

For both MP^- and MP^{2-} we used optimized QM geometries and the force constant matrix in cartesian coordinates [eqn. 5.2] to estimate the total QM energy $\tilde{U}^{QM}$ [eqn. 5.3] for 200 random conformations. The obtained energies were then compared to the ones calculated with the MM energy function applied on the same 200 random conformations. For each conformation, displacements d_i were obtained as random combination of the QM normal modes, extracted from the Gaussian output. We used all modes except the trivial ones (rotations, translations of the molecule) and the ones relative to hydrogen(s) stretching.

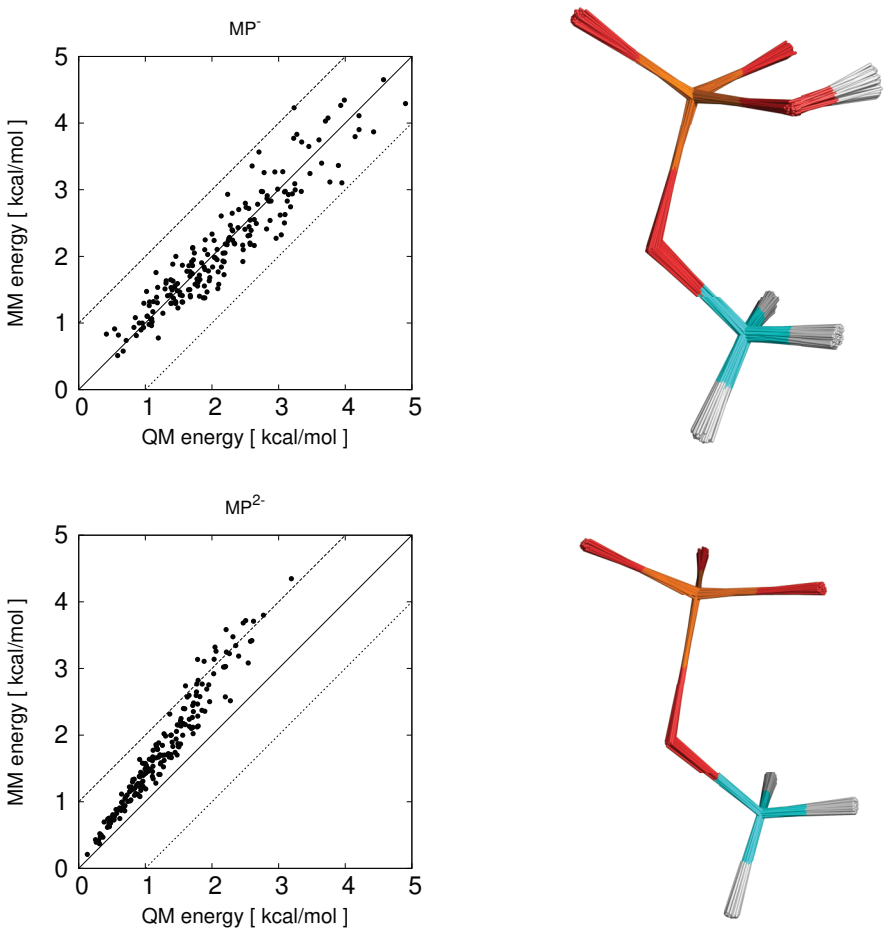


Figure 5.14 – **Conformational energies. Left**: QM vs MM comparision. **Right**: superposition of the 200 random configurations generated by random linear combination of the QM normal modes. For both MP^- and MP^{2-} QM/MM energies have correlation coefficient of 0.9. Root mean square deviations are 0.3 kcal/mol for MP^- and 0.6 kcal/mol for MP^{2-}.

6.8 Fitting QM interactions with Mg and Na ions

Pair-specific NBFIX and NBTHOLE terms were optimized by the Drude force field developers to tune Lennard-Jones and electrostatic interactions between magnesium and nucleic acids compounds in water (Lemkul & MacKerell [2016a]). Others specific terms for sodium were developed earlier by Savelyev & MacKerell [2014]. Details about parameter optimization and QM target data can be found in the publications. We recall that NBFIX parameters define pair-specific LJ interactions for a couple of *atom types* i and j. Two variables indicating minimum energy ϵ^{ij}_{min} and minimum radius r^{ij}_{min} are used to compute the LJ interaction between atoms of type i and j at distance r. The use of NBFIX does not alter LJ interactions between atoms of type i or j and all other types $k \neq i,j$ in the force field. NBTHOLE is analogous to NBFIX. A pair-specific screening parameter τ_{ij} is used to define the smeared charge distribution [eqn. 3.22] for the calculation of electrostatic interaction between a couple of atoms of type i and j.

We give the list of these optimized parameters in table 5.12. A first NBFIX was applied between magnesium (atom type MAGD) and the Drude charge attacched to the oxygen atom of SWM4-NDP water (atom type DOH2). The authors modified this interaction after observing undesired behaviour of Mg^{2+}-water complexes during MD simulations. More precisely, they noticed that after few nanoseconds of dynamics the standard Drude force field parameters produced frequent exchange of water molecules in complex with Mg^{2+}, where one ion interacts with six SWM4-NDP molecules. This should not happen, because these exchanges should take place at the microsecond timescale (Neely & Connick [1970]). The introduced NBFIX was then used to damp the too-strong dipole-dipole repulsion between water molecules, which is induced by the too-strong dipole reponse of water molecules when interact with Mg^{2+}. The authors also observed that the new parameters improved the agreement between QM and MM Mg^{2+}-water interaction energies. We conclude that this NBFIX should always be applied in our simulations. Two couples of NBFIX and NBTHOLE parameters were introduced to better reproduce QM geometries for several nucleic acid moieties (Ade, Gua, Cyt, Ura, DMP) in complex with Mg^{2+}. The authors considered Inner-shell and Outer-shell coordinated complexes, parameters were optimized by a grid search. We note that the value of r_{min} for the phosphate bridging oxygen (O1) is considerably big when compared with the one for the nonbridging oxygen.

We performed QM/Drude scans with magnesium and sodium on methyl phosphate and for the bridging oxygen of the existing DMP from the official nucleic acids Drude force field (DMPN). For DMPN, we scanned the bridging oxygen in two directions. First, imposing "direct interaction" with Mg^{2+}, then in the opposite direction when magnesium is closer to the phosphate nonbridging oxygens [fig. 5.15]. Each MM scan was performed using two slightly

different set of parameters. First, we used all the special NBFIX and NBTHOLE parameters from the existing force field, listed in table 5.12. Then we have run a second round. Instead of using the official NBFIX/NBTHOLE for the bridging oxygen, we applied LJ parameters of the Drude lipid atom type (OD30B) as we did to correct the water scans. Results are given in table 5.13.

		type	**NBFIX**		**NBTHOLE**
			$\epsilon(kcal/mol)$	r(Å)	
Mg^{2+}	Drude charge of water	DOH2	-0.07500	2.65500	-
	bridging oxygen	OD30B(N)	-0.40592	4.40157	1.151
	nonbridging oxygen	OD2C2C	-0.31211	2.59990	1.449
Na^{+}	nonbridging oxygen	OD2C2C	-0.04696	3.49668	1.449
	hydroxyl oxygen	OD31A	-0.06000	3.07000	1.820

Table 5.12 – **Pair-specific parameters** for Mg^{2+} and Na^{+} in Drude, to balance their interactions with phosphates and water. Values are extracted from Lemkul & MacKerell [2016a] ; Savelyev & MacKerell [2014]

			QM		NBFIX		OD30B	
molecule	atom	type	r_{min}	e_{min}	r_{min}	e_{min}	r_{min}	e_{min}
DMPN	O1	OD30BN	1.8	-278.27	3.35	-152.20	1.9	-265.45
	O1-inv	OD30BN	3.0	-313.31	3.2	-254.16	2.45	-303.32
	O3	OD2C2C	1.8	-299.47	1.9	-261.07	1.9	-260.87
MP^{-}	O1	OD30B	1.9	-282.93	3.25	-168.71	1.9	-297.55
	O2	OD31A	1.9	-258.80	2.2	-258.40	1.95	-262.60
	O3	OD2C2C	1.8	-290.17	2.15	-272.85	1.95	-272.75
	O4	OD2C2C	1.8	-298.76	2.16	-295.15	1.95	-295.10
MP^{2-}	O1	OD30B	1.8	-480.05	3.1	-315.90	1.9	-519.62
	O2	OD2C2C	1.7	-483.41	1.9	-508.07	1.9	-469.15
	O3	OD2C2C	2.0	-485.08	1.9	-502.42	1.9	-463.35

Table 5.13 – **Ions scans with Mg^{2+}** Distances are in Å and interaction energies in kcal/mol.

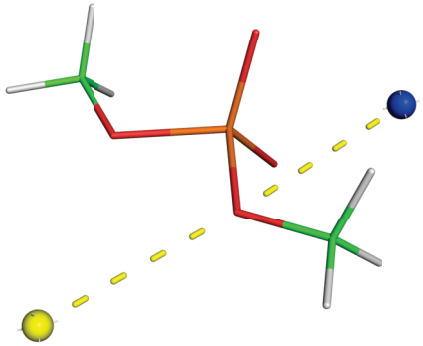


Figure 5.15 – **Ion scan** with magnesium performed on dimethyl phosphate (DMPN). Two positions were considered: direct interaction (Yellow) and inverse direction (Blue)

Results show that the NBFIX/NBTHOLE parameters of table 5.12 were optimized to reproduce the inverse scan for the bridging oxygen, while they produce poor agreement with QM for the "direct" scan of the same interaction. This was a reasonable choice, since the inverse interaction is more favorable compared to the direct one. Moreover, minimum energies for both directions were not well reproduced. On the other hand, using the lipid atom type the energy is better fitted, even if the minimum distance could be improved. O1 scans shows improvement also for MP. This suggested us to use the lipid atom type instead of the NBFIX.

To better check model transferability we also computed scans for sodium, using the parameters developed by Savelyev & MacKerell [2014]. Results are shown in table 5.14. We obtained good QM/Drude agreement, for both distances and interaction energies.

			QM		NBFIX	
molecule	atom	type	r_{min}	e_{min}	r_{min}	e_{min}
DMPN	O1	OD30BN	2.2	-102.21	2.1	-107.58
MP^{-}	O1	OD30B	2.2	-105.69	2.0	-121.85
	O2	OD31A	2.2	-101.12	2.2	-106.66
	O3	OD2C2C	2.1	-111.73	2.2	-110.37
	O4	OD2C2C	2.1	-115.81	2.2	-120.63
MP^{2-}	O1	OD30B	2.1	-193.84	1.9	-229.89
	O2	OD2C2C	2.0	-199.09	2.1	-204.10
	O3	OD2C2C	2.0	-200.10	2.1	-201.90

Table 5.14 – **Ions scans with Na^{+}** Distances are in Å and interaction energies in kcal/mol.

6.9 Mg^{2+}:phosphate binding free energies

Experimental Mg^{2+} binding free energies are available for HPO_2^- and H2PO$^-$ but not for methyl phosphate. Nevertheless, we expect that the methyl substitution has a modest effect on the free energy, since the magnesium ion binds on the opposite side of the molecule, as shown in figure 5.5 where IS and OS binding modes were extracted from snapshots of the MD simulations. This is further supported by the similar OS binding free energies computed below for MP^{2-} and P_i^{2-}. The MP$^-$ IS complex mostly coordinated Mg^{2+} through a single oxygen (mean Mg^{2+}-O distance of 2.07±0.05 Å). The MP^{2-} IS complex mostly had bifurcated magnesium coordination, with two of the phosphate oxygens interacting closely with Mg^{2+} (mean distances of 2.09±0.07 Å).

The IS complexes were stable during 1 ns of unrestrained MD, indicating that they are separated from OS by a free energy barrier. The MP^{2-} OS complex was also stable during 1 ns of MD, with no tendency to unbind or shift to IS. The MP$^-$ OS complex had a shorter lifetime, estimated from a 10 ns unrestrained MD simulation [fig. 5.16]. At the simulation concentration (55 mM), the MP$^-$ OS binding free energy was 0.4 kcal/mol (see below), giving an expected long-time limit of 33% for the OS population. The observed population during the 10 ns of MD was 27%, indicating that rough convergence was achieved for the binding/unbinding rates (which is sufficient for our purpose). The complex remained bound for intervals of about 275 ps on average, rebinding a few hundred picoseconds later. Evidently, the barrier separating the OS bound and unbound states is small, probably on the order of 2 kcal/mol. Transitions to IS were not observed. The OS lifetime is roughly comparable to the rate of OS/unbound transitions obtained for P_i^- with the AMOEBA force field (which was not measured precisely, Kumar *et al.* [2014, 2018]).

For P_i^{2-}, we only considered OS binding, for several reasons. OS binding was predominant for MP^{2-} in the present simulations, see below; we expect P_i^{2-} to behave similarly. It also proved difficult in this work to obtain a good force field model for short-range P_i^{2-}-Mg^{2+} interactions using the same van der Waals parameters as for MP. More extensive parameter optimization with one or two new atom types would be necessary. Since modeling inorganic phosphate was not our goal in this work, we left a full parameter optimization and IS binding study for future work. The computed free energies and their components are given in table 5.15. The unbound Mg^{2+} decoupling free energy, 412.3 kcal/mol, is 1.8 kcal/mol larger than the result (corrected for box size effects) of earlier simulations (Lemkul & MacKerell [2016b]) that used a smaller model and shorter runs. Repeating the entire decoupling calculation (51 ns) for the Mg^{2+}-MP$^-$ complex gave the same result within 0.1 kcal/mol, indicating good convergence.

Table 5.15 – Mg^{2+}–phosphate binding free energies from simulations and experiment

solute or complex	recharge Mg^{2+} ΔG_{elec} (ρ_{MD})	size correction	restore LJ ΔG_{LJ}	[a]apply restraint	total	[b]ΔG_{bind}	Expt.[c]
Mg^{2+}:MP^{-}(OS)	-415.6	0.0	+0.8	+0.4	-414.4	-2.1	-1.7
Mg^{2+}:MP^{-}(IS)	-414.4	0.0	+1.0	+1.1	-412.3	0.0	
Mg^{2+}:MP^{2-}(OS)	-417.5	+1.0	+1.0	+0.4	-415.1	-2.8	-3.7
Mg^{2+}:MP^{2-}(IS)	-416.5	+1.0	+1.6	+1.1	-412.8	-0.5	
Mg^{2+}:P_i^{2-}(OS)	-418.5	+1.0	+0.8	+0.4	-416.3	-4.0	-3.7
Mg^{2+}	-414.1	+0.9	+0.9	–	-412.3	–	–

In kcal/mol. [a]At the standard state concentration. [b]Obtained by subtracting the Mg^{2+} value from the value in "total" column. [c]Values are for $H_2PO_4^- \equiv P_i^-$ and $HPO_4^{2-} \equiv P_i^{2-}$ Verbeeck *et al.* [1984] ; Alberty & Goldberg [1992].

For MP^-, the computed binding free energies were -2.1 kcal/mol for the OS state, 0.0 kcal/mol for the IS state, and -2.1 kcal/mol overall. All three values are within 0.8 kcal/mol of the AMOEBA predictions for H2PO$^-$. The overall binding free energy is 0.4 kcal/mol away from the experimental H2PO$^-$ value of -1.7 kcal/mol (Verbeeck *et al.* [1984], Alberty & Goldberg [1992]). Binding is predicted to be dominated by the OS state, which has a 97% occupancy, close to the AMOEBA H_2PO^- prediction (95% OS). For MP^{2-}, computed binding free energies were -2.8 kcal/mol for the OS state, -0.5 kcal/mol for the IS state, and -2.8 kcal/mol overall. The overall value is 0.9 kcal/mol away from the experimental HPO^{2-} value of -3.7 kcal/mol. The bound state is predicted to be 87% OS and 13% IS. Earlier ab initio calculations with a dielectric continuum solvent also predicted that OS was favored over IS, but with a smaller difference. For P_i^{2-}, the OS binding free energy was -4.0 kcal/mol, close to the experimental value. The OS values for MP^{2-} and P_i^{2-} show that the methyl substitution has a modest effect, as expected, weakening binding by 1.2 kcal/mol. We note that for MP^{2-}, P_i^{2-}, and unbound Mg^{2+}, the correction for the MD box size was small but distinctly non-negligible (over 1 kcal/mol). It was zero for MP^-, since introducing Mg^{2+} changes the total charge from -1 to +1 (same absolute charge magnitude). Overall, for all three complexes considered, the computed binding free energies are in good agreement with experiment. This contrasts dramatically with the additive, Charmm C36 force field, which overestimated the Mg^{2+} binding free energy by an order of magnitude (Satpati *et al.* [2011]). The results are consistent with mostly OS binding for all the complexes.

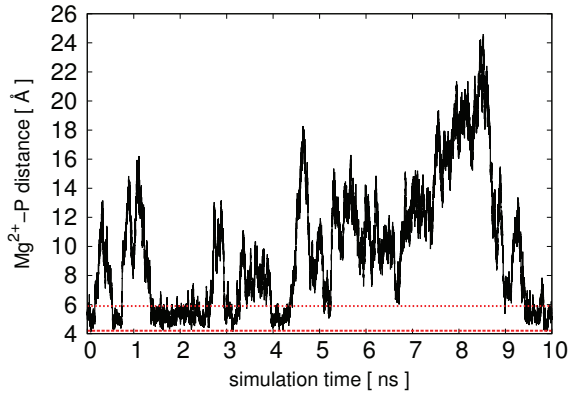


Figure 5.16 – **Magnesium-phosphate binding-unbinding** Black line : distance between Mg^{2+} and P atoms, extracted from a 10ns simulation of the MP^{-}:Mg^{2+} complex without restrain potential. **Red ,dashed line**: limit for OS bound state where the magnesium-phosphorus distance is between 4.1 and 5.9Å

6.10 Final MP and P_i^{2-} topology and parameters

We give first the final model for hydrogen phosphate P_i^{2-} =HPO_4^{2-}, for which vdW optimization should be improved

```
RESI HP_2 -2.000 ! HPO4 Hydrogenphosphate, dianionic  !        H
GROUP                                                 !        |
ATOM P1     PD1AN   1.020  ALPHA -1.244   THOLE 1.478 !        O1
ATOM O2     OD2C2C -0.860  ALPHA -0.921   THOLE 0.803 !        |
ATOM O3     OD2C2C -0.860  ALPHA -0.921   THOLE 0.803 ! (-)O4--P1--O3(-)
ATOM O4     OD2C2C -0.860  ALPHA -0.921   THOLE 0.803 !        |
ATOM O1     OD31D   0.000  ALPHA -0.927   THOLE 0.900 !        O2
ATOM LPA    LPDNA1 -0.320
ATOM LPB    LPDNA1 -0.320
ATOM H      HDP1A   0.200
LONEPAIR relative LPA O1 P1 H distance 0.35 angle 110.00 dihe 90.00
LONEPAIR relative LPB O1 P1 H distance 0.35 angle 110.00 dihe 270.00
BOND O1 P1
BOND O2 P1
BOND O3 P1
BOND O4 P1
BOND O1 H
BOND O1 LPA
BOND O1 LPB
PATCH FIRST NONE LAST NONE
```

The final models for methyl phosphate were included in the official Drude polarizable force field. We give now topologies and parameters, as they are in the new force field release

```
RESI MP_1       -1.00    ! CH4O4P Methylphosphate, anionic
GROUP                    ! Villa et al., 2018
ATOM O1      OD30D    -0.546  ALPHA -0.901  THOLE 0.811   !       H11
ATOM P1      PD1AN     1.122  ALPHA -0.974  THOLE 2.098   !        |
ATOM O2      OD31D     0.000  ALPHA -0.927  THOLE 1.100   !  H13--C1--H12
ATOM LPA     LPDNA1   -0.300                               !        |
ATOM LPB     LPDNA1   -0.300                               !        O1
ATOM H2      HDP1A     0.360                               !        |
ATOM O3      OD2C2C   -0.778  ALPHA -0.921  THOLE 1.083   !    O4==P1==O3 (-)
ATOM O4      OD2C2C   -0.778  ALPHA -0.921  THOLE 1.083   !        |
ATOM C1      CD33C    -0.008  ALPHA -1.642  THOLE 0.862   !        O2
ATOM H11     HDA3A     0.076                               !         \
ATOM H12     HDA3A     0.076                               !         H2
ATOM H13     HDA3A     0.076
BOND O1     P1
BOND O1     C1
BOND O2     P1
BOND O3     P1
BOND O4     P1
BOND O2     H2
BOND C1     H11
BOND C1     H12
BOND C1     H13
BOND O2     LPA
BOND O2     LPB
LONEPAIR relative LPA O2 P1 H2 distance 0.35 angle 110.00 dihe 90.00
LONEPAIR relative LPB O2 P1 H2 distance 0.35 angle 110.00 dihe 270.00
ANISOTROPY O2 P1 LPA LPB A11 0.76473 A22 1.16239
PATCH FIRST NONE LAST NONE

RESI MP_2         -2.00 ! CH3O4P Methylphosphate, dianionic
GROUP                   ! Villa et al., 2018
                                                        !        H11
ATOM O1     OD30D    -0.685  ALPHA -1.291  THOLE 1.091   !         |
ATOM P1     PD1AN     1.050  ALPHA -1.244  THOLE 1.678   !   H13--C1--H12
ATOM O2     OD2C2C   -0.860  ALPHA -0.951  THOLE 1.083   !         |
ATOM O3     OD2C2C   -0.860  ALPHA -0.951  THOLE 1.083   !        O1
ATOM O4     OD2C2C   -0.860  ALPHA -0.951  THOLE 1.083   !         |
ATOM C1     CD33C     0.020  ALPHA -1.642  THOLE 0.862   ! (-)O4==P1==O3(-)
ATOM H11    HDA3A     0.065                               !        ||
ATOM H12    HDA3A     0.065                               !        O2
ATOM H13    HDA3A     0.065
BOND O1 P1
BOND O1 C1
BOND O2 P1
BOND O3 P1
BOND O4 P1
BOND C1 H11
BOND C1 H12
BOND C1 H13
PATCH FIRST NONE LAST NONE
```

```
BONDS
OD30D  PD1AN  240.00  1.600
CD33C  OD30D  335.00  1.440
OD31A  PD1AN  237.00  1.610
OD2C2C PD1AN  525.00  1.500
HDP1A  OD31A  536.50  0.970
CD33C  HDA3A  322.00  1.111

ANGLES
CD33C  OD30D  PD1AN   40.00  120.50
OD30D  PD1AN  OD31A   48.10  108.00
OD30D  PD1AN  OD2C2C  98.90  113.00
HDA3A  CD33C  OD30    60.00  109.50
HDP1A  OD31A  PD1AN   30.00  115.00   40.00  2.3500 ! Urey-Bradley term
OD31A  PD1AN  OD2C2C  98.90  111.00
OD2C2C PD1AN  OD2C2C 120.00  125.00
HDA3A  CD33C  HDA3A   35.50  108.40

DIHEDRALS
CD33C  OD30D  PD1AN   OD2C2C  0.100  3    0 .0  ! from nucleic acids
CD33C  OD30D  PD1AN   OD31A   0.500  3    0.0
CD33C  OD30D  PD1AN   OD31A   0.950  2    0.0
PD1AN  OD30D  CD33C   HDA3A   0.000  3    0.0
OD30D  PD1AN  OD31A   HDP1A   0.2000 1  180.0
OD30D  PD1AN  OD31A   HDP1A   1.0000 2    0.0
HDP1A  OD31A  PD1AN   OD2C2C  0.2000 3    0.0
CD33C  OD30D  PD1AN   OD2C2C  0.100  3    0.0

NONBONDED E14FAC 1.000000
OD30D  0.0     -0.1700   1.7700   ! from DMP @ lipids: new type for MP
PD1AN  0.0000  -0.3200   1.9000
CD33C  0.0000  -0.0780   1.9400
OD31A  0.0000  -0.1500   1.7650
OD2C2C 0.0000  -0.0700   1.8650
HDP1A  0.0000  -0.0100   0.4000
HDA3A  0.0000  -0.0240   1.3400
MAGD   0.0000  -0.0500   1.1264156

NBFIX
MAGD   DOH2   -0.07500  2.65500 ! Lemkul & MacKerell, J.Phys.Chem.B 2016
MAGD   OD2C2C -0.31211  2.59990 ! Lemkul & MacKerell, J.Phys.Chem.B 2016
MAGD   LPD    -0.16504  2.20163 ! Lemkul & MacKerell, J.Phys.Chem.B 2016
SODD   OD2C2C -0.04696  3.49668 ! nucleic acids, Savelyev-2014
SODD   OD30BN -0.02510  3.38168 ! nucleic acids, Savelyev-2014
SODD   LPD    -0.09000  2.92000 ! nucleic acid bases
SODD   OD31A  -0.06000  3.07000 ! ethanol, ser, thr, Hui-2015
MAGD   OD30D  -0.40592  2.25157 ! new, Villa et al, 2018
SODD   OD30D  -0.02510  3.38168 ! transferred from DMP OD30BN

NBTHOLE
MAGD   OD2C2C  1.44900 ! Lemkul & MacKerell, J.Phys.Chem.B 2016
MAGD   OD30D   1.15100 ! Lemkul & MacKerell, J.Phys.Chem.B 2016 (OD30BN) )
MAGD   ODW     1.51567 ! Lemkul & MacKerell
SODD   OD31A   1.82000 ! ethanol, ser, thr, Hui-2015
MAGD   OD31A   1.95000 ! new, Villa et al, 2018
```

We note that a couple of atom types differ respect to those described during optimization. Their names were changed to avoid overlapping parameters/types in the official force field, including those for interactions with Mg^{2+} and the existing force field parameters for the nucleic acids bridging (OD30BN) and hydroxyl (OD31A) oxygens. The first was replaced by the new type OD30D, the second by the new atom type OD31D. They have the same LJ parameters of the existing OD30B (lipids) and OD31A (nucleic acids) types.

7 Classical Drude model for dianionic phosphotyrosine

As anticipated in the Methods section, after validation of the methyl phosphate and P_i model we parametrized the dianionic phosphotyrosine $pTyr^{2-}$. The tyrosine residue in the existing Drude force field was derived from the model compound p-Cresol C_7H_8O (CRES, 3D structure shown in figure 5.18). The official Drude topology files for CRES and TYR are given in figure 5.17. Atom types and electrostatic parameters of CRES were completely transferred to the TYR side chain, then the charges of C_β hydrogens HB1 and HB2 were readjusted.

```
! Topologies from the Drude force field
!
!
RESI CRES          0.000 ! p-Cresol
!
!
!
!
!
!
!
!
GROUP
ATOM CB   CD33A   -0.267  ALPHA -1.802  THOLE 0.148
ATOM HB1  HDA3A    0.075
ATOM HB2  HDA3A    0.075
ATOM HB3  HDA3A    0.075
ATOM CG   CD2R6A   0.086  ALPHA -1.418  THOLE 1.270
ATOM CD1  CD2R6A  -0.202  ALPHA -1.427  THOLE 1.270
ATOM HD1  HDR6A    0.118
ATOM CD2  CD2R6A  -0.202  ALPHA -1.427  THOLE 1.270
ATOM HD2  HDR6A    0.118
ATOM CE1  CD2R6A  -0.136  ALPHA -1.444  THOLE 1.270
ATOM HE1  HDR6A    0.123
ATOM CE2  CD2R6A  -0.136  ALPHA -1.444  THOLE 1.270
ATOM HE2  HDR6A    0.123
ATOM CZ   CD2R6A   0.297  ALPHA -1.174  THOLE 1.270
ATOM OZ   OD31C    0.000  ALPHA -0.683  THOLE 0.601
ATOM HZ   HDP1A    0.319
ATOM LP1A LPD     -0.233
ATOM LP1B LPD     -0.233
```

```
RESI TYR           0.000
GROUP ! tyrosine backbone
ATOM N    ND2A2   -0.382  ALPHA -1.942  THOLE 0.250
ATOM HN   HDP1A    0.272
ATOM CA   CD31C    0.169  ALPHA -0.960  THOLE 1.078
ATOM HA   HDA1A   -0.017
ATOM C    CD2O1A   0.497  ALPHA -0.675  THOLE 0.295
ATOM O    OD2C1A   0.000  ALPHA -0.651  THOLE 0.310
ATOM LPOA LPDO1   -0.312
ATOM LPOB LPDO1   -0.227
GROUP ! tyrosine side chain
ATOM CB   CD32A   -0.192  ALPHA -1.802  THOLE 0.148
ATOM HB1  HDA2A    0.075
ATOM HB2  HDA2A    0.075
!
ATOM CG   CD2R6A   0.086  ALPHA -1.418  THOLE 1.270
ATOM CD1  CD2R6A  -0.202  ALPHA -1.427  THOLE 1.270
ATOM HD1  HDR6A    0.118
ATOM CD2  CD2R6A  -0.202  ALPHA -1.427  THOLE 1.270
ATOM HD2  HDR6A    0.118
ATOM CE1  CD2R6A  -0.136  ALPHA -1.444  THOLE 1.270
ATOM HE1  HDR6A    0.123
ATOM CE2  CD2R6A  -0.136  ALPHA -1.444  THOLE 1.270
ATOM HE2  HDR6A    0.123
ATOM CZ   CD2R6A   0.297  ALPHA -1.174  THOLE 1.270
ATOM OH   OD31C    0.000  ALPHA -0.683  THOLE 0.601
ATOM HH   HDP1A    0.319
ATOM LP1A LPD     -0.233
ATOM LP1B LPD     -0.233
```

Figure 5.17 – **Drude topology files** for p-Cresol (left) and tyrosine (right). Equivalent atoms are shown on the same line.

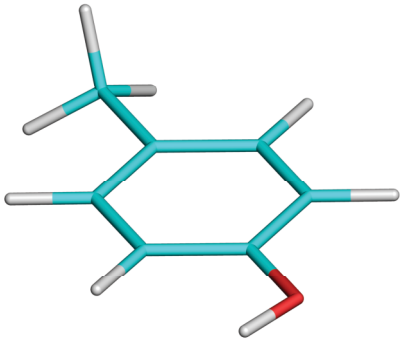


Figure 5.18 – **p-Cresol** 3D structure

To parametrize dianionic phosphotyrosine, we adopted a similar strategy. We considered the dianionic, phosphorylated form of pCRESOL (residue CREP), 3D structure shown in figure 5.19. We created the initial model combining atom types and electrostatic parameters of the existing p-Cresol and of our MP^{2-} model.

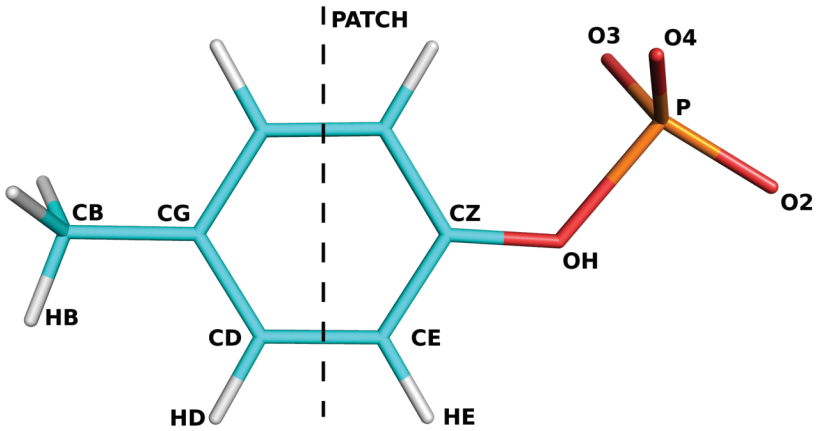


Figure 5.19 – **Phosphorylated p-CRESOL** (CREP) from non-phosphorylated p-CRESOL (CRES). The phosphorylated form can be seen as a patch applied to the non-phosphorylated compound. Atoms on the right of the dashed line are modified, while atoms on the left are kept as in CRES. Groups of equivalent atoms are (HB1-HB2-HB3), (CD1,CD2), (CE1-CE2), (HD1-HD2), (HE1-HE2) and (O2-O3-O4).

7.1 QM quantities

We computed target QM quantities, based on the QM optimized geometry. All calculations were run with the same QM softwares and methods used for MP. During optimization, the CZ-OH-P-O2 dihedral was constrained to 180°. Then we computed unperturbed and perturbed QM ESP maps using a grid of 2178 points and 64 perturbing charges. We scanned

the solute:water interaction energy for acceptors OH and O3, then for donors P and HE1. Geometries used for the scans are shown in figure 5.20. Using PSI4, we scanned the dihedral CE1-CZ-OH-P. During the scan we kept the CZ-OH-P-O2 dihedral constrained to 180°, as for geometry optimization.

7.2 Electrostatic parameters optimization

Parameter optimization was done following an existing procedure, used in a previous study to develop the phosphotyrosine model of the additive force field C36 (Feng *et al.* [1996]). We considered pTyr like a patch for the tyrosine residue, which shares with TYR the atoms belonging to the methyl group and the first part of the TYR ring, as shown in figure 5.19. Charges of atoms CE1(2), HE1(2), CZ, OH, P, O2(3,4) were readjusted performing a grid search in order to obtain the better agreement between Drude and the QM ESP maps and QM solute-water interaction energy scans. After optimization, we obtained a χ^2_{ESP}=0.25 in good agreement with the QM data. Results for the water scans are listed in table 5.16.

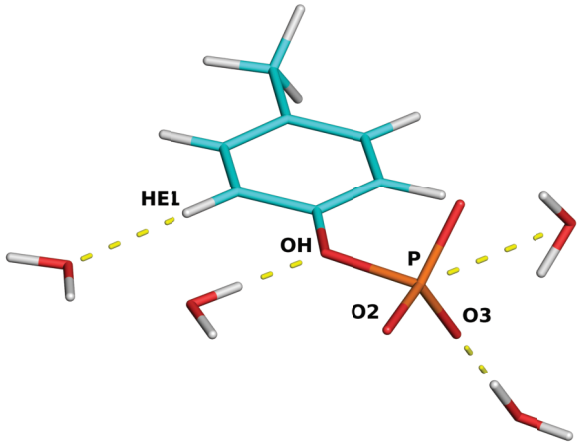


Figure 5.20 – **Phosphorylated p-Cresol** Water scans

	QM		MM	
atom	r_{min}	e_{min}	r_{min}	e_{min}
OH	1.8	-18.02	1.8	-17.93
O3	1.7	-19.50	1.8	-19.48
P	3.3	-24.18	3.2	-24.15
HE1	2.8	3.74	2.6	3.50

Table 5.16 – CREP–water interaction distances (Å) and energies (kcal/mol)

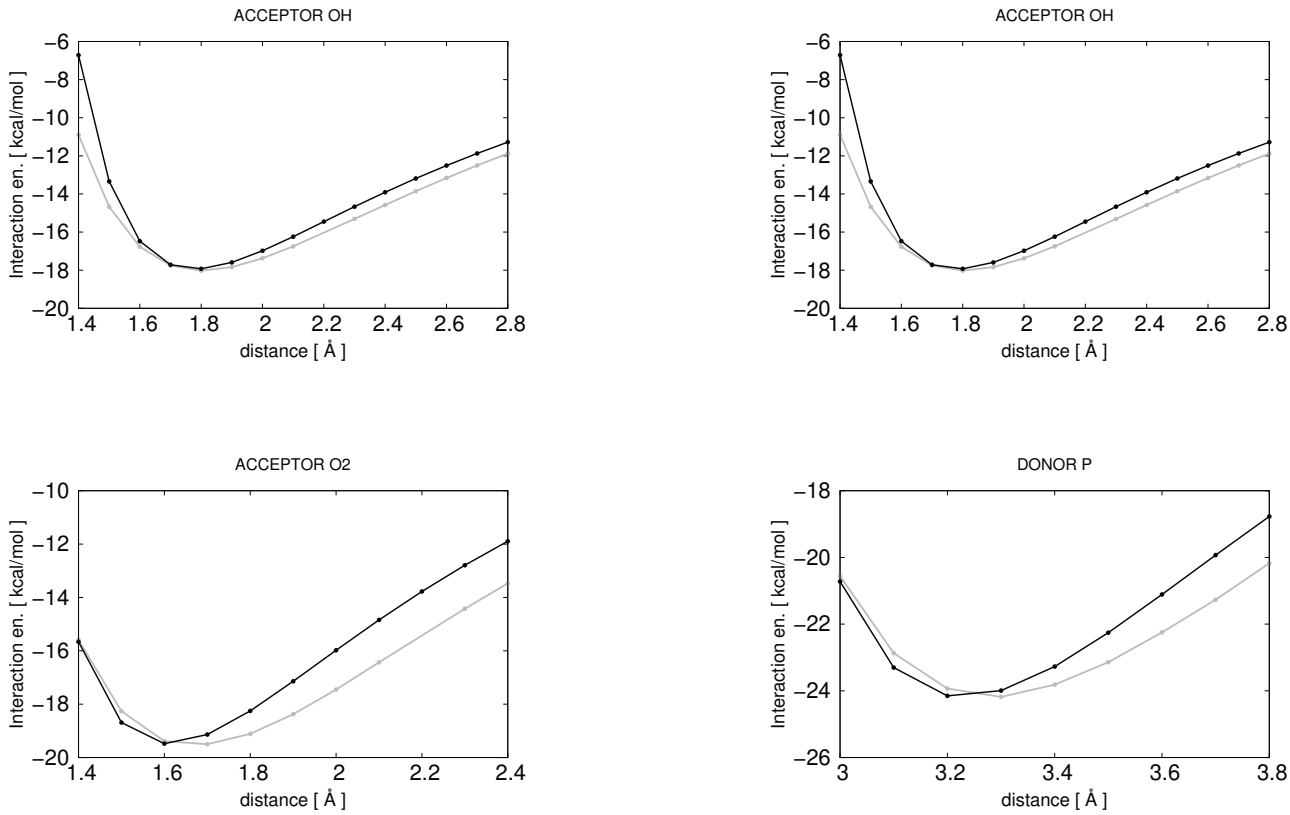


Figure 5.21 – **CREP^{2-}–water interaction** Grey: QM; Black: MM

7.3 Dihedral parameters and fit

Internal parameters were borrowed from the existing additive model and then readjusted to fit dihedral scans. For the bridging oxygen OH we imposed the atom type OD30D, as in MP. We have run a dihedral scan for the dihedral angle CE1-CZ-OH-P using PSI4. The relaxed scan was performed keeping CZ-OH-P-O2 fixed at the value of 180°, as in the optimized geometry. Results are shown in figure 5.22

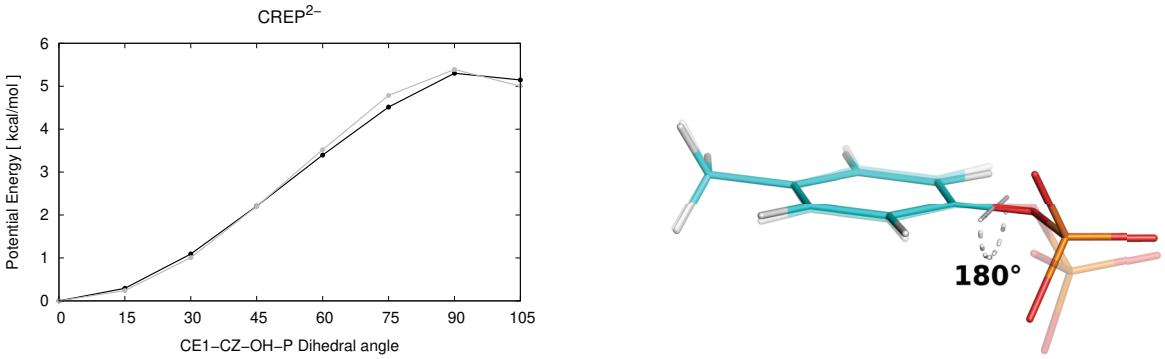


Figure 5.22 – **QM/MM torsion energy scans** Grey: QM; Black: MM

7.4 Final pCRES and pTyr^{-2} topology and parameters

We give topology information for the final, optimized CREP and pTyr^{2-} models in the extended format. Atom names are the ones used for the pTYr (TP2) residue of the C36 force field (Feng *et al.* [1996]).

```
RESI CREP         -2.000 ! Phosphorylated p-Cresol
!
!                       HD1  HE1
!                        |    |
!              HB1       CD1--CE1        O3(-)
!                 \     /      \       /
!              HB2-CB--CG       CZ--OH--P--O4
!                 /     \      /       \
!              HB3       CD2--CE2        O2(-)
!                        |    |
!                       HD2  HE2
!
GROUP
ATOM CB     CD33A   -0.267    ALPHA -1.802    THOLE 0.148
ATOM HB1    HDA3A    0.075
ATOM HB2    HDA3A    0.075
ATOM HB3    HDA3A    0.075
ATOM CG     CD2R6A   0.086    ALPHA -1.418    THOLE 1.270
ATOM CD1    CD2R6A  -0.202    ALPHA -1.427    THOLE 1.270
ATOM HD1    HDR6A    0.118
ATOM CD2    CD2R6A  -0.202    ALPHA -1.427    THOLE 1.270
ATOM HD2    HDR6A    0.118
ATOM CE1    CD2R6A  -0.220    ALPHA -1.444    THOLE 1.270
ATOM HE1    HDR6A    0.180
ATOM CE2    CD2R6A  -0.220    ALPHA -1.444    THOLE 1.270
ATOM HE2    HDR6A    0.180
ATOM CZ     CD2R6A   0.174    ALPHA -1.174    THOLE 1.270
ATOM OH     OD30D   -0.600    ALPHA -0.683    THOLE 0.901
ATOM P      PD1AN    1.090    ALPHA -1.344    THOLE 1.678
ATOM O2     OD2C2C  -0.820    ALPHA -0.931    THOLE 1.083
ATOM O3     OD2C2C  -0.820    ALPHA -0.931    THOLE 1.083
ATOM O4     OD2C2C  -0.820    ALPHA -0.931    THOLE 1.083
BOND CB  HB1  CB  HB2  CB  HB3
BOND CB  CG   CG  CD1  CG  CD2
BOND CD1 HD1  CD1 CE1  CD2 HD2
BOND CD2 CE2  CE1 HE1  CE1 CZ
BOND CE2 HE2  CE2 CZ   CZ  OH
BOND OH  P    P   O2   P   O3   P  O4
PATCH FIRST NONE LAST NONE
```

```
RESI TP2   -2.000 ! Dianionic Phosphotyrosine
!
!                  |        HD1  HE1
!               HN-N         |    |
!                  |   HB1  CD1--CE1         O3(-)
!                  |   |    //     \\       /
!               HA-CA--CB--CG      CZ--OH--P--O4
!                  |   |    \  __  /        \
!                  |   HB2  CD2--CE2         O2(-)
!                O=C          |    |
!                  |        HD2  HE2
!
GROUP
ATOM N     ND2A2   -0.382   ALPHA -1.942   THOLE 0.250
ATOM HN    HDP1A    0.272
ATOM CA    CD31C    0.169   ALPHA -0.960   THOLE 1.078
ATOM HA    HDA1A   -0.017
ATOM C     CD2O1A   0.497   ALPHA -0.675   THOLE 0.295
ATOM O     OD2C1A   0.000   ALPHA -0.651   THOLE 0.310
ATOM LPOA  LPDO1   -0.312
ATOM LPOB  LPDO1   -0.227
GROUP
ATOM CB    CD32A   -0.192   ALPHA -1.802   THOLE 0.148
ATOM HB1   HDA2A    0.075
ATOM HB2   HDA2A    0.075
ATOM CG    CD2R6A   0.086   ALPHA -1.418   THOLE 1.270
ATOM CD1   CD2R6A  -0.202   ALPHA -1.427   THOLE 1.270
ATOM HD1   HDR6A    0.118
ATOM CD2   CD2R6A  -0.202   ALPHA -1.427   THOLE 1.270
ATOM HD2   HDR6A    0.118
ATOM CE1   CD2R6A  -0.220   ALPHA -1.444   THOLE 1.270
ATOM HE1   HDR6A    0.180
ATOM CE2   CD2R6A  -0.220   ALPHA -1.444   THOLE 1.270
ATOM HE2   HDR6A    0.180
ATOM CZ    CD2R6A   0.174   ALPHA -1.174   THOLE 1.270
ATOM OH    OD30D   -0.600   ALPHA -0.683   THOLE 0.901
ATOM P     PD1AN    1.090   ALPHA -1.344   THOLE 1.678
ATOM O2    OD2C2C  -0.820   ALPHA -0.931   THOLE 1.083
ATOM O3    OD2C2C  -0.820   ALPHA -0.931   THOLE 1.083
ATOM O4    OD2C2C  -0.820   ALPHA -0.931   THOLE 1.083
BOND N  CA    CA C     C + N      CA HA
BOND N  HN    C  O     O   LPOA   O  LPOB
BOND CA CB    CB HB1   CB  HB2    CB CG
BOND CG  CD1  CG CD2   CD1 HD1    CD1 CE1
BOND CD2 HD2  CD2 CE2  CE1 HE1    CE1 CZ
BOND CE2 HE2  CE2 CZ   CZ  OH     OH  P
BOND P   O2   P   O3   P   O4
IMPR C   CA  +N  O   N  -C  CA HN
CMAP -C  N   CA  C   N  CA  C  +N
LONEPAIR relative LPOA O C CA distance 0.30 angle 91.0 dihe   0.0
LONEPAIR relative LPOB O C CA distance 0.30 angle 91.0 dihe 180.0
ANISOTROPY O C LPOA LPOB  A11  0.82322 A22  1.14332
DONOR HN N
ACCEPTOR O C
```

```
BONDS
CD2R6A OD30D  340.00  1.390

ANGLES
CD2R6A CD2R6A OD30D   75.00  119.40
CD2R6A OD30D  PD1AN   90.00  124.20

DIHEDRALS
HDR6A  CD2R6A  CD2R6A   OD30D    4.2000     2   180.00
CD2R6A CD2R6A  CD2R6A   OD30D    3.1000     2   180.00
CD2R6A CD2R6A  OD30D    PD1AN    1.5500     2   180.00
CD2R6A CD2R6A  OD30D    PD1AN    0.2000     3   180.00
CD2R6A OD30D   PD1AN    OD2C2C   0.100      3     0.00
```

7.5 Simulating Drude phosphotyrosine in the Tiam1:pSdc1 complex

Starting from the crystal stucture, the Tiam1:pSdc1 complex was solvated in a large box of size L=80 Å containing pre-equilibrated TIP3P water. Then we deleted solvent molecules within the volume occupied by the solute, using a distance threshold of 3.0 Å from the complex heavy atoms. The box was trimmed, from cubic to truncated octahedron, deleting solvent molecules at the cube edges. Drude and lone particles were added with CHARMM, then the Drude particles were minimized with restrained atomic centers. The system was finally relaxed without restraints.

Equilibration MD was done using the Drude force field, in a step-wise procedure: first restraints were applied on the solute backbone and side chain atoms, then just on the solute backbone. After equilibration, we have run 20 ns of producion MD. Throughout the simulations, the center of mass of Tiam1 was restrained to the box center using an harmonic potential.

MD was done using NAMD (version 2.12) at room temperature and pressure, using Langevin dynamics with a Langevin Piston Nosé-Hoover barostat (Feller et al. [1995]). We used periodic boundary conditions and a PME electrostatics (Darden [2001]). Bonds to hydrogen atoms were constrained with SHAKE (Ryckaert et al. [1977]). LJ interactions were shifted to zero for separations between 10 and 12 Å. MD was run using a timestep of 1 fs and Drude oscillators temperature T_D=1 K.

We extracted configurations from the production simulation, where $pTyr^{2-}$ strongly interacts with K879. Average and crystal structures are showed in figure 5.23. The average phosphorus-nitrogen distance during the 20 ns of MD was 3.42 Å, close to the 3.61 Å of the

crystal structure. The average value of the χ_1 dihedral (N-CA-CB-CG) was 168.4 degrees, slightly bigger than in the crystal structure where it is 158.5 degrees. In the crystal structure, phosphotyrosine interacts also with the Tiam1 residue T857. In the simulation, this interaction is weaker, with an average phosphate-threonine side chain distance which is more than 3 Å larger respect to the crystal structure.

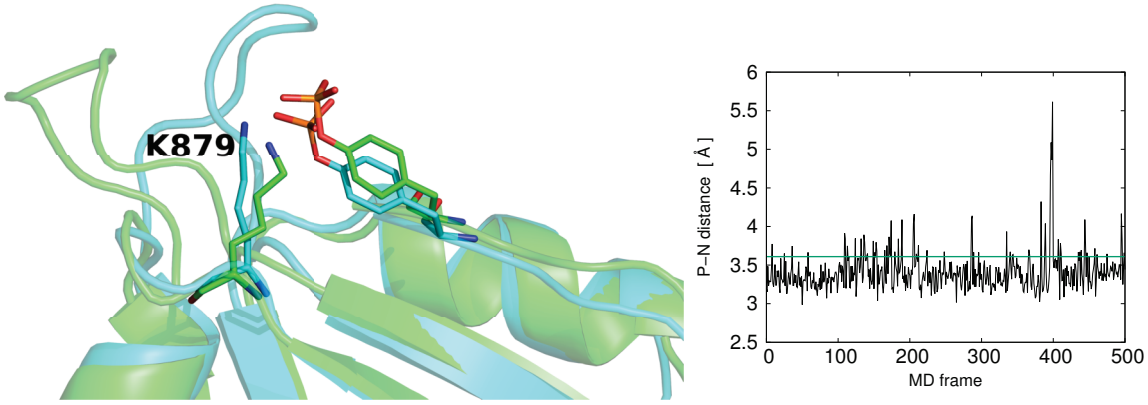


Figure 5.23 – **pTyr^{2-}-K879 interactions** during 20 ns of MD. **Left**: crystal structure (light blue) and average structure from the simulation (green). **Right**: phosphorus-nitrogen distance during MD. Frames were extracted each 40 ps. The straight line indicates the distance in the crystal structure (3.61 Å).

8 Conclusion

A classical Drude model for methyl phospate and its interactions with Mg^{2+} was developed. The parameters obtained were similar to those derived earlier for DMP, Lemkul & MacKerell [2016a] with some significant differences. Agreement between the QM and MM data was similar to that obtained earlier for DMP.

The MP model was used to compute Mg^{2+} standard binding free energies. Results were in good agreement with experiments, indicating that OS state was predominant for both anionic forms of MP. This is the first time the predominant bound state has been clearly identified for MP^{2-} and MP^{-}. For this and similar binding processes, atomistic simulations are a powerful tool, and the polarizable force fields perform definetively better then additive additive force fields. This has now been confirmed for phosphate-Mg^{2+} binding free energies, and was confirmed also for several other ionic interactions (Panel *et al.* [2018]).

The parametrization and simulation of MP with the Drude polarizable force field is a step towards modeling phosphorylated side chains in proteins. The Drude force field is applicable to

large biomolecules and provides the speed and accuracy for long and wellconverged simulations. The accuracy obtained here for phosphate- Mg2+ binding indicates that the Drude force field can be applied to elucidate molecular recognition processes involving phosphorylated proteins, which are essential for biological signalling.

Bibliography

Abrams J.B. & Tuckerman M.E. (2008). Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations. *The Journal of Physical Chemistry B* **112**, 15742–15757.
cited page 65

Aguilar B., Shadrach R. & Onufriev A.V. (2010). Reducing the secondary structure bias in the generalized born model via r6 effective radii. *Journal of Chemical Theory and Computation* **6**, 3613–3630.
cited page 13

Alberty R.A. & Goldberg R.N. (1992). Standard thermodynamic formation properties for the adenosine 5′-triphosphate series. *Biochemistry* **31**, 10610–10615.
cited pages 98 and 123

Alford R.F., Leaver-Fay A., Jeliazkov J.R., O'Meara M.J., DiMaio F.P., Park H., Shapovalov M.V., Renfrew P.D., Mulligan V.K., Kappel K., Labonte J.W., Pacella M.S., Bonneau R., Bradley P., Dunbrack R.L., Das R., Baker D., Kuhlman B., Kortemme T. & Gray J.J. (2017). The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation* **13**, 3031–3048.
cited page 16

Allouche D., Andre I., Barbe S., Davies J., de Givry S., Kastirelos G., O′Sullivan B., Prestwich S., Sciex T. & Traore S. (2014). Computational protein design as an optimization problem. *Artificial Intelligence* **212**, 59–79.
cited page 16

Anisimov V.M., Lamoureux G., Vorobyov I.V., Huang N., Roux B. & MacKerell A.D. (2005). Determination of electrostatic parameters for a polarizable force field based on the classical drude oscillator. *Journal of Chemical Theory and Computation* **1**, 153–168.
cited pages 94 and 99

Applequist J., Carl J. & Fung K. (1972). Atom dipole interaction model for molecular polarizability. application to polyatomic molecules and determination of atom polarizabilities. *Journal of American Chemical Society* **94**, 2952–2960.
cited page 45

Archontis G. & Simonson T. (2005). A residue-pairwise generalized born scheme suitable for protein design calculations. *The Journal of Physical Chemistry B* **109**, 22667–22673.
cited page 14

Baker C.M. (2015). Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **5**, 241–254.
cited page 49

Barducci A., Bussi G. & Parrinello M. (2008). Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters* **100**.
cited page 39

Bas D.C., Rogers D.M. & Jensen J.H. (2008). Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins: Structure, Function, and Bioinformatics* **73**, 765–783.
cited page 36

Basdevant N., Weinstein H. & Ceruso M. (2006). Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *Journal of the American Chemical Society* **128**, 12766–12777.
cited page 3

Bendas G. & Borsig L. (2012). Cancer cell adhesion and metastasis: Selectins, integrins, and the inhibitory potential of heparins. *International Journal of Cell Biology* **2012**, 1–10.
cited page 3

Bennett C.H. (1976). Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics* **22**, 245–268.
cited page 63

Beutler T.C., Mark A.E., van Schaik R.C., Gerber P.R. & van Gunsteren W.F. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters* **222**, 529–539.
cited page 75

Blöchliger N., Xu M. & Caflisch A. (2015). Peptide binding to a pdz domain by electrostatic steering via nonnative salt bridges. *Biophysical Journal* **108**, 2362–2370.
cited page 3

Böhm H. & Schneider G. (2003). Protein-ligand interactions: From molecular recognition to drug design. *Wiley, Methods and Principles in Medicinal Chemistry* .
cited page 1

Boresch S., Tettinger F., Leitgeb M. & Karplus M. (2003). Absolute binding free energies: a quantitative approach for their calculation. *The Journal of Physical Chemistry B* **107**, 9535–9551.
cited pages 65 and 66

Böttcher C. (1973). Theory of electric polarization. *Elsevier* .
cited page 48

Boyce S., Mobley D., Rocklin G., Graves A., Dill K. & BK. S. (2009). Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *Journal of Molecular Biology* **394**, 747–763.
cited page 44

Cieplak P., Caldwell J. & Kollman P. (2001). Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of Computational Chemistry* **22**, 1048–1057.
cited page 49

Cuendet M.A. & Tuckerman M.E. (2014). Free energy reconstruction from metadynamics or adiabatic free energy dynamics simulations. *Journal of Chemical Theory and Computation* **10**, 2975–2986.
cited page 65

Dahiyat B.I. & Mayo S.L. (1997). De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87.
cited pages 9 and 13

Darden T. (2001). Computational biochemistry and biophysics, treatment of long-range forces and potential. *Taylor & Francis Group* **1**.
cited pages 73, 74 and 107

Debiec K., Gronenborn A. & Chong L. (2014). Evaluating the strength of salt bridges: A comparison of current biomolecular force fields. *Journal of Physical Chemistry B* **118**, 6561–6569.
cited page 44

Deng Y. & Roux B. (2006). Calculation of standard binding free energies: Aromatic molecules in the t4 lysozyme l99a mutant. *Journal of Chemical Theory and Computation* **2**, 1255–1273.
cited page 44

Ding Y., Chen B., Wang S., Zhao L., Chen J., Ding Y., Chen L. & Luo R. (2009). Overexpression of tiam1 in hepatocellular carcinomas predicts poor prognosis of HCC patients. *International Journal of Cancer* **124**, 653–658.
cited page 3

Druart K., Palmai Z., Omarjee E. & Simonson T. (2015). Protein:ligand binding free energies: A stringent test for computational protein design. *Journal of Computational Chemistry* **37**, 404–415.
cited page 24

Dunbrack R.L. & Karplus M. (1993). Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of Molecular Biology* **230**, 543–574.
cited page 13

Ernst A., Gfeller D., Kan Z., Seshagiri S., Kim P.M., Bader G.D. & Sidhu S.S. (2010). Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing. *Molecular BioSystems* **6**, 1782.
cited page 2

Feller S.E., Zhang Y., Pastor R.W. & Brooks B.R. (1995). Constant pressure molecular dynamics simulation: The langevin piston method. *The Journal of Chemical Physics* **103**, 4613–4621.
cited pages 37, 55, 74 and 107

Feng M.H., Philippopoulos M., MacKerell A.D. & Lim C. (1996). Structural characterization of the phosphotyrosine binding region of a high-affinity SH2 domain-phosphopeptide complex by molecular dynamics simulation and chemical shift calculations. *Journal of the American Chemical Society* **118**, 11265–11277.
cited pages 94, 129 and 131

Figueirido F., Buono G.S.D. & Levy R.M. (1997). On finite-size corrections to the free energy of ionic hydration. *The Journal of Physical Chemistry B* **101**, 5622–5623.
cited page 69

Finkelstein A.V. & Ptitsyn O.B. (1977). Theory of protein molecule self-organization. i. thermodynamic parameters of local secondary structures in the unfolded protein chain. *Biopolymers* **16**, 469–495.
cited page 13

Fiser A., Do R. & Svali A. (2000). Modeling of loops in protein structures. *Protein Science* **9**, 1753–1773.
cited page 73

Frisch M.J., Trucks G.W., Schlegel H.B., Scuseria G.E., Robb M.A., Cheeseman J.R., Scalmani G., Barone V., Mennucci B., Petersson G.A., Nakatsuji H., Caricato M., Li X., Hratchian H.P., Izmaylov A.F., Bloino J., Zheng G., Sonnenberg J.L., Hada M., Ehara M., Toyota K., Fukuda R., Hasegawa J., Ishida M., Nakajima T., Honda Y., Kitao O., Nakai H., Vreven T., Montgomery Jr. J.A., Peralta J.E., Ogliaro F., Bearpark M., Heyd J.J., Brothers E., Kudin K.N., Staroverov V.N., Kobayashi R., Normand J., Raghavachari K., Rendell A., Burant J.C., Iyengar S.S., Tomasi J., Cossi M., Rega N., Millam J.M., Klene M., Knox J.E., Cross J.B., Bakken V., Adamo C., Jaramillo J., Gomperts R., Stratmann R.E., Yazyev O., Austin A.J., Cammi R., Pomelli C., Ochterski J.W., Martin R.L., Morokuma K., Zakrzewski V.G., Voth G.A., Salvador P., Dannenberg J.J., Dapprich S., Daniels A.D., Farkas O., Foresman J.B., Ortiz J.V., Cioslowski J. & Fox D.J. (2009). Gaussian09 revision e01. Gaussian Inc. Wallingford CT.
cited page 95

Gaillard T. & Simonson T. (2014). Pairwise decomposition of an MMGBSA energy function for computational protein design. *Journal of Computational Chemistry* **35**, 1371–1387.
cited page 13

General I.J. (2010). A note on the standard state's binding free energy. *Journal of Chemical Theory and Computation* **6**, 2520–2524.
cited page 59

Harder E. & Roux B. (2008). On the origin of the electrostatic potential difference at a liquid-vacuum interface. *The Journal of Chemical Physics* **129**, 234706.
cited page 82

Harder E., Anisimov V., Vorobyov I., Lopes P., Noskov S., MacKerell A. & Roux B. (2006). Atomic level anisotropy in the electrostatic modeling of lone pairs for a polarizable force field based on the classical drude oscillator. *Journal of Chemical Theory and Computation* **2**, 1587–1597.
cited pages 51 and 79

Hayes R.L., Armacost K.A., Vilseck J.Z. & Brooks C.L. (2017). Adaptive landscape flattening accelerates sampling of alchemical space in multisite λ dynamics. *The Journal of Physical Chemistry B* **121**, 3626–3635.
cited page 89

Huang L. & Roux B. (2013). Automated force field parameterization for nonpolarizable and polarizable atomic models based on ab initio target data. *Journal of Chemical Theory and Computation* **9**, 3543–3556.
cited page 95

Hummer G., Pratt L.R. & García A.E. (1996). Free energy of ionic hydration. *The Journal of Physical Chemistry* **100**, 1206–1215.
cited page 69

Hummer G., Pratt L.R. & García A.E. (1997). Ion sizes and finite-size corrections for ionic-solvation free energies. *The Journal of Chemical Physics* **107**, 9275–9277.
cited page 69

Humphrey W., Dalke A. & Schulten K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38.
cited page 84

Hunter T. (2012). Why nature chose phosphate to modify proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 2513–2516.
cited page 92

Im W., Beglov D. & Roux B. (1998). Continuum solvation model: Computation of electrostatic forces from numerical solutions to the poisson-boltzmann equation. *Computer Physics Communications* **111**, 59–75.
cited page 37

Janin J., Wodak S., Levitt M. & Maigret B. (1978). Conformation of amino acid side-chains in proteins. *Journal of Molecular Biology* **125**, 357–386.
cited page 13

Jiang W., Hardy D.J., Phillips J.C., MacKerell A.D., Schulten K. & Roux B. (2010). High-performance scalable molecular dynamics simulations of a polarizable force field based on classical drude oscillators in NAMD. *The Journal of Physical Chemistry Letters* **2**, 87–92.
cited page 56

Jo S., Kim T., Iyer V.G. & Im W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859–1865.
cited page 74

Jo S., Cheng X., Lee J., Kim S., Park S.J., Patel D.S., Beaven A.H., Lee K.I., Rui H., Park S., Lee H.S., Roux B., MacKerell A.D., Klauda J.B., Qi Y. & Im W. (2016). CHARMM-GUI 10 years for biomolecular modeling and simulation. *Journal of Computational Chemistry* **38**, 1114–1124.
cited page 74

Jorgensen W. & Thomas L. (2008). Perspective on free energy perturbation calculations for chemical equilibria. *Journal of Chemical Theory and Computation* **4**, 869–876.
cited page 74

Jorgensen W., Maxwell D. & Tirado-Rives J. (1996). Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of American Chemical Society* **118**, 11225.
cited page 43

Khoury G.A., Baliban R.C. & Floudas C.A. (2011). Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports* **1**.
cited page 91

Kokubo H., Tanaka T. & Okamoto Y. (2013). Two-dimensional replica-exchange method for predicting protein-ligand binding structures. *Journal of Computational Chemistry* **34**, 2601–2614.
cited page 89

Kollman P. (1993). Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **93**, 2395–2417.
cited page 72

Krivov G.G., Shapovalov M.V. & Dunbrack R.L. (2009). Improved prediction of protein side-chain conformations with scwrl4. *Proteins: Structure, Function, and Bioinformatics* **77**, 778–795.
cited page 73

Kumar M., Simonson T., Ohanessian G. & Clavaguéra C. (2014). Structure and thermodynamics of mg:phosphate interactions in water: A simulation study. *ChemPhysChem* **16**, 658–665.
cited pages 93 and 122

Kumar M., Simonson T., Ohanessian G. & Clavaguéra C. (2018). Corrigendum: Structure and thermodynamics of mg:phosphate interactions in water: A simulation study. *ChemPhysChem* **19**, 1117–1117.
cited pages 93 and 122

Lagardère L., Jolly L.H., Lipparini F., Aviat F., Stamm B., Jing Z.F., Harger M., Torabifard H., Cisneros G.A., Schnieders M.J., Gresh N., Maday Y., Ren P.Y., Ponder J.W. & Piquemal J.P. (2018). Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chemical Science* **9**, 956–972.
cited page 49

Laio A. & Parrinello M. (2002). Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99**, 12562–12566.
cited page 89

Lamoureux G. & Roux B. (2003). Modeling induced polarization with classical drude oscillators: Theory and molecular dynamics simulation algorithm. *Journal of Chemical Physics* **119**, 3025.
cited page 45

Lamoureux G., MacKerell A. & Roux B. (2003). A simple polarizable model of water based on classical drude oscillators. *Journal of Chemical Physics* **119**, 5185.
cited page 52

Lamoureux G., Harder E., Vorobyov I., Roux B. & MacKerell A. (2006). A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters* **418**, 245–249.
cited pages 52 and 82

Lazaridis T. & Karplus M. (1999). Effective energy function for proteins in solution. *Proteins: Structure, Function, and Genetics* **35**, 133–152.
cited page 12

Lemkul J. & MacKerell A. (2016a). Balancing the interactions of mg2+ in aqueous solution and with nucleic acid moieties for a polarizable force field based on the classical drude oscillator model. *Journal of Physical Chemistry B* **120**, 11436–11448.
cited pages 53, 93, 99, 119, 120 and 134

Lemkul J., Huang J., Roux B. & MacKerell Jr A. (2016). An empirical polarizable force field based on the classical drude oscillator model: Development history and recent applications. *Chemical Reviews* **116**, 4983–5013.
cited pages 44 and 94

Lemkul J.A. & MacKerell A.D. (2016b). Balancing the interactions of $mg2^{2+}$ in aqueous solution and with nucleic acid moieties for a polarizable force field based on the classical drude oscillator model. *The Journal of Physical Chemistry B* **120**, 11436–11448.
cited page 122

Letunic I. & Bork P. (2017). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46**, D493–D496.
cited page 1

Lin Y.L., Aleksandrov A., Simonson T. & Roux B. (2014). An overview of electrostatic free energy computations for solutions and proteins. *Journal of Chemical Theory and Computation* **10**, 2690–2709.
cited pages 68, 82 and 83

Lindorff-Larsen K., Piana S., Dror R. & DE. S. (2011). How fast-folding proteins fold. *Nature* **28**, 517–520.
cited page 43

Liu X., Shepherd T.R., Murray A.M., Xu Z. & Fuentes E.J. (2013). The structure of the tiam1 PDZ domain/ phospho-syndecan1 complex reveals a ligand conformation that modulates protein dynamics. *Structure* **21**, 342–354.
cited pages 3, 72 and 91

MacKerell A.D., Bashford D., Bellott M., Dunbrack R.L., Evanseck J.D., Field M.J., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., Kuchnir L., Kuczera K., Lau F.T.K., Mattos C., Michnick S., Ngo T., Nguyen D.T., Prodhom B., Reiher W.E., Roux B., Schlenkrich M., Smith J.C., Stote R., Straub J., Watanabe M., Wiórkiewicz-Kuczera J., Yin D. & Karplus M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The Journal of Physical Chemistry B* **102**, 3586–3616.
cited page 43

Maier J., Martinez C., Kasavajhala K., Wickstrom L., Hauser K. & Simmerling C. (2015). ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation* **11**, 3696–3713.
cited pages 37 and 43

Maragliano L. & Vanden-Eijnden E. (2006). A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical Physics Letters* **426**, 168–175.
cited page 65

Martyna G.J., Tobias D.J. & Klein M.L. (1994). Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **101**, 4177–4189.
cited page 55

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. & Teller E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.
cited page 21

Miller K.J. (1990). Additivity methods in molecular polarizability. *Journal of the American Chemical Society* **112**, 8533–8542.
cited page 103

Minard M.E., Ellis L.M. & Gallick G.E. (2006). Tiam1 regulates cell adhesion, migration and apoptosis in colon tumor cells. *Clinical & Experimental Metastasis* **23**, 301–313.
cited page 3

Neely J. & Connick R. (1970). Rate of water exchange from hydrated magnesium ion. *Journal of the American Chemical Society* **92**, 3476–3478.
cited page 119

Oostenbrink C., Villa A., Mark A. & van Gunsteren W. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry* **25**, 1656–1676.
cited page 43

Panel N., Sun Y.J., Fuentes E.J. & Simonson T. (2017). A simple PB/LIE free energy function accurately predicts the peptide binding specificity of the tiam1 PDZ domain. *Frontiers in Molecular Biosciences* **4**.
cited pages 27, 36, 37 and 38

Panel N., Villa F., Fuentes E.J. & Simonson T. (2018). Accurate PDZ/peptide binding specificity with additive and polarizable free energy simulations. *Biophysical Journal* **114**, 1091–1102.
cited pages 4, 38, 44, 72, 84, 85, 93 and 134

Patel S. & Brooks C.L. (2003). CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *Journal of Computational Chemistry* **25**, 1–16.
cited page 45

Patel S., Mackerell A.D. & Brooks C.L. (2004). CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *Journal of Computational Chemistry* **25**, 1504–1514.
cited page 45

Pawson T. & Scott J.D. (2005). Protein phosphorylation in signaling – 50 years and counting. *Trends in Biochemical Sciences* **30**, 286–290.
cited page 91

Phillips J.C., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R.D., Kalé L. & Schulten K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781–1802.
cited page 49

Polydorides S. & Simonson T. (2013). Monte carlo simulations of proteins at constant pH with generalized born solvent, flexible sidechains, and an effective dielectric boundary. *Journal of Computational Chemistry* **34**, 2742–2756.
cited page 25

Ponder J.W. & Richards F.M. (1987). Tertiary templates for proteins. *Journal of Molecular Biology* **193**, 775–791.
cited page 13

Qi R., Jing Z., Liu C., Piquemal J.P., Dalby K.N. & Ren P. (2018). Elucidating the phosphate binding mode of phosphate-binding protein: The critical effect of buffer solution. *The Journal of Physical Chemistry B* **122**, 6371–6376.
cited page 93

Qin X., Hayashi F. & and S.Y. (2006). Solution structure of the PDZ domain of t-cell lymphoma invasion and metastasis 1 varian.
cited page 3

Reiland J., Ott V.L., Lebakken C.S., Yeanman C., McCarty J. & Rapraeger A.C. (1996). Pervanadate activation of intracellular kinases leads to tyrosine phosphorylation and shedding of syndecan-1. *Biochemical Journal* **319**, 39–47.
cited page 91

Ren P. & Ponder J.W. (2003). Polarizable atomic multipole water model for molecular mechanics simulation. *Journal of Physical Chemistry B* **104**, 5933–5947.
cited page 49

Rocklin G.J., Mobley D.L., Dill K.A. & Hünenberger P.H. (2013). Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *The Journal of Chemical Physics* **139**, 184103.
cited pages 67 and 69

Rosso L., Mináry P., Zhu Z. & Tuckerman M.E. (2002). On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *The Journal of Chemical Physics* **116**, 4389–4402.
cited page 65

Ryckaert J.P., Ciccotti G. & Berendsen H.J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327–341.
cited page 107

Satpati P., Clavaguera C., Ohanessian G. & Simonson T. (2011). Free energy simulations of a gtpase: Gtp and gdp binding to archaeal initiation factor 2. *The Journal of Physical Chemistry B* **115**, 6749–6763.
cited pages 93 and 123

Savelyev A. & MacKerell A.D. (2014). All-atom polarizable force field for DNA based on the classical drude oscillator model. *Journal of Computational Chemistry* **35**, 1219–1239.
cited pages 119, 120 and 121

Schaefer M. & Karplus M. (1996). A comprehensive analytical treatment of continuum electrostatics. *The Journal of Physical Chemistry* **100**, 1578–1599.
cited pages 9 and 10

Schultz J. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Research* **28**, 231–234.
cited page 1

Sensoy O. & Weinstein H. (2015). A mechanistic role of helix 8 in GPCRs: Computational modeling of the dopamine d2 receptor interaction with the GIPC1–PDZ-domain. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1848**, 976–983.
cited page 3

Shepherd T.R., Klaus S.M., Liu X., Ramaswamy S., DeMali K.A. & Fuentes E.J. (2010). The tiam1 PDZ domain couples to syndecan1 and promotes cell–matrix adhesion. *Journal of Molecular Biology* **398**, 730–746.
cited pages 3 and 72

Shepherd T.R., Hard R.L., Murray A.M., Pei D. & Fuentes E.J. (2011). Distinct ligand specificity of the tiam1 and tiam2 PDZ domains. *Biochemistry* **50**, 1296–1308.
cited pages 3, 72 and 92

Shi Y., Xia Z., Zhang J., Best R., Wu C., Ponder J. & Ren P. (2013). Polarizable atomic multipole-based amoeba force field for proteins. *Journal of Chemical Theory and Computation* **9**, 4046–4063.
cited page 44

Silberstein L. (1917). Molecular refrectivity and atomic interaction. *Philosophical Magazine* **33**, 521.
cited page 45

Simonson T. (2003). Electrostatics and dynamics of proteins. *Reports on Progress in Physics* **66**, 737–787.
cited page 72

Simonson T. & Roux B. (2016). Concepts and protocols for electrostatic free energies. *Molecular Simulation* **42**, 1090–1101.
cited pages 69 and 83

Simonson T., Archontis G. & Karplus M. (2002). Free energy simulations come of age: Protein-ligand recognition. *Accounts of Chemical Research* **35**, 430–437.
cited page 74

Simonson T., Carlsson J. & Case D.A. (2004). Proton binding to proteins: pKaCalculations with explicit and implicit solvent models. *Journal of the American Chemical Society* **126**, 4167–4180.
cited page 25

Simonson T., Gaillard T., Mignon D., am Busch M.S., Lopes A., Amara N., Polydorides S., Sedano A., Druart K. & Archontis G. (2013). Computational protein design: The proteus software and selected applications. *Journal of Computational Chemistry* **34**, 2472–2484.
cited pages 4 and 16

Simonson T., Ye-Lehmann S., Palmai Z., Amara N., Wydau-Dematteis S., Bigan E., Druart K., Moch C. & Plateau P. (2016). Redesigning the stereospecificity of tyrosyl-tRNA synthetase. *Proteins: Structure, Function, and Bioinformatics* **84**, 240–253.
cited page 86

Songyang Z. (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* **275**, 73–77.
cited page 1

Spiegel I., Salomon D., Erne B., Schaeren-Wiemers N. & Peles E. (2002). Caspr3 and caspr4, two novel members of the caspr family are expressed in the nervous system and interact with pdz domains. *Mol. Cell. Neurosci.* **20**, 283–297.
cited page 3

Steiner S. & Caflisch A. (2012). Peptide binding to the pdz3 domain by conformational selection. *Proteins: Structure, Function, and Bioinformatics* **80**, 2562–2572.
cited page 3

Stiffler M.A., Chen J.R., Grantcharova V.P., Lei Y., Fuchs D., Allen J.E., Zaslavskaia L.A. & MacBeath G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **317**, 364–369.
cited page 1

Still W.C., Tempczyk A., Hawley R.C. & Hendrickson T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **112**, 6127–6129.
cited page 10

Straatsma T.P. & McCammon J.A. (1992). Computational alchemy. *Annual Review of Physical Chemistry* **43**, 407–435.
cited page 72

Street A.G. & Mayo S.L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Folding and Design* **3**, 253–258.
cited page 14

Sulka B., Lortat-Jacob H., Terreux R., Letourneur F. & Rousselle P. (2009). Tyrosine dephosphorylation of the syndecan-1 PDZ binding domain regulates syntenin-1 recruitment. *Journal of Biological Chemistry* **284**, 10659–10671.
cited page 91

Thole B. (1981). Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics* **59**, 341–350.
cited page 80

Tian F., Lv Y., Zhou P. & Yang L. (2011). Characterization of pdz domain-peptide interactions using an integrated protocol of qm/mm, pb/sa, and cfea analyses. *Journal of Computer-Aided Molecular Design* **25**, 947–958.
cited page 3

Tonikian R., Zhang Y., Sazinsky S.L., Currell B., Yeh J.H., Reva B., Held H.A., Appleton B.A., Evangelista M., Wu Y., Xin X., Chan A.C., Seshagiri S., Lasky L.A., Sander C., Boone C., Bader G.D. & Sidhu S.S. (2008). A specificity map for the PDZ domain family. *PLoS Biology* **6**, e239.
cited page 2

Traoré S., Allouche D., André I., de Givry S., Katsirelos G., Schiex T. & Barbe S. (2013). A new framework for computational protein design through cost function network optimization. *Bioinformatics* **29**, 2129–2136.
cited page 16

Tuckerman M.E. (2007). Free energy calculations: Theory and applications in chemistry and biology. springer series in chemical physics, 86 edited by christophe chipot and andrew pohorille. *Journal of the American Chemical Society* **129**, 10963–10964.
cited page 74

Tuffery P., Etchebest C., Hazout S. & Lavery R. (1991). A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure and Dynamics* **8**, 1267–1289.
cited page 13

Turney J.M., Simmonett A.C., Parrish R.M., Hohenstein E.G., Evangelista F.A., Fermann J.T., Mintz B.J., Burns L.A., Wilke J.J., Abrams M.L., Russ N.J., Leininger M.L., Janssen C.L., Seidl E.T., Allen W.D., Schaefer H.F., King R.A., Valeev E.F., Sherrill C.D. & Crawford T.D. (2011). Psi4: an open-source ab initio electronic structure program. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 556–565.
cited page 95

Vanommeslaeghe K., Hatcher E., Acharya C., Kundu S., Zhong S., Shim J., Darian E., Guvench O., Lopes P., Vorobyov I. & Mackerell A.D. (2009). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry* NA–NA.
cited page 96

Verbeeck R.M.H., Bruyne P.A.M.D., Driessens F.C.M. & Verbeek F. (1984). Solubility of magnesium hydrogen phosphate trihydrate and ion-pair formation in the system magnesium hydroxide-phosphoric acid-water at 25°c. *Inorganic Chemistry* **23**, 1922–1926.
cited pages 98 and 123

Villa F., Mignon D., Polydorides S. & Simonson T. (2017). Comparing pairwise-additive and many-body generalized born models for acid/base calculations and protein design. *Journal of Computational Chemistry* **38**, 2396–2410.
cited pages 4, 14 and 21

Villa F., MacKerell A., Roux B. & Simonson T. (2018a). Classical drude polarizable force field model for methyl phosphate and its interactions with mg2+. *The Journal of Physical Chemistry A* .
cited page 4

Villa F., Panel N., Chen X. & Simonson T. (2018b). Adaptive landscape flattening in amino acid sequence space for the computational design of protein:peptide binding. *The Journal of Chemical Physics* **149**, 072302.
cited pages 4, 27, 36, 37, 38 and 40

Wernisch L., Hery S. & Wodak S.J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization 1 1edited by j. thorton. *Journal of Molecular Biology* **301**, 713–736.
cited page 15

Wiedemann U., Boisguerin P., Leben R., Leitner D., Krause G., Moelling K., Volkmer-Engert R. & Oschkinat H. (2004). Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *Journal of Molecular Biology* **343**, 703–718.
cited page 1

Yu W., Lopes P., Roux B. & MacKerell A. (2013). Six-site polarizable model of water based on the classical drude oscillator. *Journal of Chemical Physics* **138**, 34508.
cited pages 52 and 89

Zacharias M., Straatsma T.P. & McCammon J.A. (1994). Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration. *The Journal of Chemical Physics* **100**, 9025–9031.
cited page 75

Zhao L., Liu Y., Sun X., He M. & Ding Y. (2010). Overexpression of t lymphoma invasion and metastasis 1 predict renal cell carcinoma metastasis and overall patient survival. *Journal of Cancer Research and Clinical Oncology* **137**, 393–398.
cited page 3

Zheng L., Chen M. & Yang W. (2008). Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *Proceedings of the National Academy of Sciences* **105**, 20227–20232.
cited page 89

Zwanzig R.W. (1954). High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics* **22**, 1420–1426.
cited page 72

Résumé

Les interactions protéine-protéine (IPPs) médient la signalisation cellulaire. Leur ingénierie peut fournir des informations et conduire au développement de molécules thérapeutiques. Les IPPs sont souvent médiés par domaines spécialisés, avec différents modes de reconnaissance (grandes surfaces de contact, peptides). Leur caractérisation avec des modèles physiques est complexe à cause des nombreux événements qui ont lieu après le binding: changements de conformation, réorganisation des molécules de solvant, redistribution des charges ou protonation-déprotonation des chaînes latérales.

Le travail de thèse est focalisé sur les domaines PDZ, importants médiateurs de IPPs. Elles lient les 4-10 résidus C-terminaux de protéines cibles. Elles lient aussi les peptides correspondants, qui peuvent servir de systèmes modèle ou d'inhibiteurs. Nous avons développé deux approches computationnelles basé sur des modèles physiques et les avons appliquées au domaine PDZ de la protéine Tiam1, un facteur d'échange pour la protéine Rac, impliqué dans la protrusion neuronale. Sa cible est la protéine Syndecan1. Des affinités expérimentales sont connues pour le peptide C-terminal, noté Sdc1, et plusieurs mutants; elles ont servi pour tester les calculs.

Dans une première phase, nous avons développé des modèles performants "haut débit" basé sur l'échantillonnage Monte Carlo adaptatif. Contrairement aux différentes méthodes computationnelles existantes, les nouveaux modèles permettent l'étude d'un grand nombre de variants de protéines (grands ensembles des séquences et des structures) et de réduire significativement les temps de calcul. Nous avons appliqué notre méthodes pour de dessin computationnel haut débit de Sdc1 dans le complexe Tiam1-Sdc1. Une simulation Monte Carlo est faite où les chaines latérales de la protéine et du peptide peuvent changer de conformères et certaines positions peuvent muter. Le solvant est implicite. Le paysage énergétique est applati par la méthode adaptative de Wang-Landau, de sorte qu'un vaste ensemble de variants est échantillonné. Effectuant des simulations distinctes du complexe et du peptide seul nous avons obtenu les énergies libres relatives d'association de 75,000 variants en heure CPU sur une machine de bureau. Les valeurs sont compatibles avec les quelques données expérimentales disponibles. Notre modèle haut débit est implémenté dans la nouvelle version de Proteus, le

programme pour de dessin computationnel de protéines développé à l'École polytechnique. Cependant, notre approche est générale et peut être implémentée dans d'autres logiciels largement utilisé.

Dans une seconde phase, nous avons développé une approche beaucoup plus détaillée et réaliste, avec un coût de calcul plus élevé. Cette approche est basée sur des simulations de dynamique moléculaire et implique l'utilisation de techniques de calcul d'énergie libre. Soluté et solvant sont décrits par un champ de force atomique, qui représente explicitement la polarisation électronique: le champ de force Drude de Charmm. La polarisabilité peut être importante car les résidus de l'interface PDZ:peptide passent, lors de l'association, d'un environement riche en solvant à un autre pauvre en solvant. Nous avons fait des simulations alchimiques d'énergie libre pour comparer quatre variants du peptide qui diffèrent par une ou deux chaines latérales ioniques. Pendant chaque transformation alchimique, les chaînes latérales d'intérêt sont mutées en utilisant un chemin d'états non physiques. Les transformations sont réalisés avec plusieurs simulations indépendantes, nécessitent l'utilisation de plusieurs processeurs en parallèle. Nos calculs sont principalement effectués sur le superordinateur OCCIGEN du CINES. Les résultats sont en bon accord avec l'expérience. Les champs de force "additifs" Charmm et Amber, qui représentent la polarisabilité implicitement, donnent un moins bon accord. Ces calculs sont le premier exemple de simulations alchimiques d'énergies libre d'association relatives protéine:ligand avec un champ de force polarisable. Enfin, pour une modélisation future de peptides phophorylés, nous avons étendu le champ de force Drude pour inclure le méthyl phosphate et la phospho tyrosine. Il en résulte un excellent accord vaec les affinités expérimentales phosphate:magnésium.

université PARIS-SACLAY
ÉCOLE DOCTORALE
INTERFACES
Approches Interdisciplinaires : Fondements, Applications et Innovation

Titre: Simulations numériques pour le dessin des interactions PDZ:peptide

Mots clés: Dessin de protéine, dynamique moléculaire, simulation Monte Carlo, champ de force polarisable, calcul d'énergie libre, PDZ

Les interactions protéine-protéine (IPPs) médient la signalisation cellulaire. Leur ingénierie peut fournir des informations et conduire au développement de molécules thérapeutiques. Les domaines PDZ sont des médiateurs importants de IPPs. Elles lient les 4--10 résidus C-terminaux de protéines cibles. Elles lient aussi les peptides correspondants, qui ppeuvent servir de systèmes modèle ou d'inhibiteurs. Nous avons développé deux approches computationnelles et les avons appliquées au domaine PDZ de la protéine Tiam1, un facteur d'échange pour la protéine Rac, impliqué dans la protrusion neuronale. Sa cible est la protéine Syndecan1. Des affinités expérimentales sont connues pour le peptide C-terminal, noté Sdc1, et plusieurs mutants; elles ont servi pour tester les calculs. Nous avons d'abord développé une méthode de dessin computationnel haut débit. Une simulation Monte Carlo est faite où les chaines latérales de la protéine et du peptide peuvent changer de conformères et certaines positions peuvent muter. Le solvant est implicite. Le paysage énergétique est applati par la méthode adaptative de Wang-Landau, de sorte qu'un vaste ensemble de variants est échantillonné. Effectuant des simulations distinctes du complexe et du peptide seul nous avons obtenu les énergies libres relatives d'association de 75,000 variants en heure CPU sur une machine de bureau. Les valeurs sont compatibles avec les quelques données expérimentales disponibles. Ensuite, nous avons développé une approche beaucoup plus détaillée et réaliste. Soluté et solvant sont décrits par un champ de force atomique, qui représente explicitement la polarisation électronique: le champ de force Drude de Charmm. La polarisabiité peut être importante car les résidus de l'interface PDZ:peptide passent, lors de l'association, d'un environement riche en solvant à un autre pauvre en solvant. Nous avons fait des simulations alchimiques d'énergie libre pour comparer quatre variants du peptide qui diffèrent par une ou deux chaines latérales ioniques. Les résultats sont en bon accord avec l'expérience. Les chamops de force "additifs" Charmm et Amber, qui représentent la polarisabilité implicitement, donnent un moins bon accord. Ces calculs sont le premier exemple de simulations alchimiques d'énergies libre d'association relatives protéine:ligand avec un champ de force polarisable. Enfin, pour une modélisation future de peptides phophorylés, nous avons étendu le champ de force Drude pour inclure le méthyl phosphate et la phospho tyrosine. Il en résulte un excellent accord vaec les affinités expérimentales phosphate:magnésium.

Title: Computer simulations to engineer PDZ-peptide recognition

Keywords: Computational Protein Design, Molecular dynamics, Monte Carlo simulation, Polarizable force field, Free energy calculation, PDZ

Protein-protein interactions (PPIs) mediate complex signaling networks in cells. Engineering them can provide understanding and is a recognized strategy for drug design. PDZ domains are among the most widespread domains mediating PPIs. They recognize the 4-10 C-terminal amino acids of their target proteins. They can also bind the corresponding peptides, which can thus serve as inhibitors or model systems. We have developed and tested two computational approaches to characterize and engineer PDZ:peptide recognition. We applied them to the PDZ domain of the Tiam1 protein, a Rac GTP exchange factor involved in neuronal protrusion and axon guidance. Its natural target protein is the Syndecan1 protein. Experimental affinities are available for the C-terminal Syndecan1 peptide, denoted Sdc1, and several mutants; this data was used for testing and benchmarking. We first developed a novel high throughput strategy for protein and peptide design. A Monte Carlo simulation was done where protein and peptide side chains could change conformations and selected positions could mutate. Solvent was modeled implicitly. The energy landscape was adaptively flattened, following the Wang-Landau method, so that a very diverse set of sequences was sampled. By performing separate simulations for the PDZ:peptide complex and the unbound peptide, we could recover the relative binding free energies of around 100,000 peptide variants using just one hour of CPU time on a desktop machine. The computed affinities were consistent with (sparse) available experimental data. Next, we developed a much more detailed and accurate simulation model. The solute and solvent were described in full atomic detail, using molecular dynamics simulations and a state-of-the-art force field that explicitly accounts for electronic polarization: the Charmm Drude force field. Polarizability is expected to be important for PDZ:peptide binding, since residues that form the binding interface are transferred from a solvent-rich to a solvent-poor environment upon binding. We performed alchemical free energy simulations to compare the binding free energies of four peptide variants, which all differ in one or two ionic side chains. The calculations gave good agreement with experiment. In contrast, the Charmm and Amber "additive" force fields, which treat electronic polarization implicitly, gave poorer agreement. These calculations represent the first example of relative protein:ligand binding free energies computed with alchemical free energy simulations and a polarizable force field. Finally, to allow future modeling of phosphorylated peptide variants, we extended the Drude force field to include both methyl phosphate and phosphotyrosine. The force field gave excellent agreement with experimental phosphate--magnesium binding free energies.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France