UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA
Programa de Pós-Graduação em Ciências Biológicas
(Bioquímica)

Ariane Ferreira Nunes Alves

Simulações computacionais de desenovelamento de proteína e complexação de ligantes com amostragem aumentada

Versão original da tese defendida

São Paulo

Data do depósito na SPG:
05/10/2017

Ariane Ferreira Nunes Alves

Simulações computacionais de desenovelamento de proteína e complexação de ligantes com amostragem aumentada

Tese apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioquímica)

Orientador: Prof. Dr. Guilherme Menegon Arantes

São Paulo
2017

Ficha Catalográfica
Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP

A474s

Alves, Ariane Ferreira Nunes
Simulações computacionais de desenovelamento de proteína e complexação de ligantes com amostragem aumentada / Ariane Ferreira Nunes Alves. - São Paulo, 2017.
145 p.

Tese (doutorado) - Instituto de Química da Universidade de São Paulo. Departamento de Bioquímica.
Orientador: Arantes, Guilherme Menegon

1. bioquímica. 2. proteínas. 3. molécula. I. T. II. Arantes, Guilherme Menegon, orientador.

Dedico este trabalho aos meus pais,
Maria Elisa e Heli, e ao meu
marido, Javier.
Obrigada por todo amor, apoio e
incentivo.

Agradecimentos

Agradeço ao meu orientador, prof. Dr. Guilherme Menegon Arantes, por me propor projetos desafiadores e interessantes, por acompanhar o meu trabalho e por contribuir com sugestões, críticas construtivas e recomendações de leitura. Agradeço por todas as críticas e contribuições às minhas apresentações, relatórios e manuscritos. Além disso, a orientação do Guilherme foi muito importante para o meu crescimento intelectual. Durante nossos anos de convívio aprendi a ser paciente e perseverante no meu trabalho.

Agradeço ao prof. Dr. Daniel M. Zuckerman, da Oregon Health & Science University, que foi meu orientador durante o doutorado sanduíche. Fui muito bem recebida no laboratório dele, que na época da minha visita se situava na University of Pittsburgh. Sou muito grata pela sua paciência, pelas sugestões e críticas construtivas ao meu trabalho e pelos seus ensinamentos sobre o método *weighted ensemble* (WE).

Agradeço ao meu marido, Javier, que foi um dos primeiros revisores de muitos relatórios e apresentações que fiz durante o doutorado. Obrigada pelo carinho, paciência, incentivo e críticas construtivas.

Agradeço aos meus colegas e ex-colegas de laboratório, Vanesa, Raphael, Murilo, Felipe, André, Sofia e Rodrigo, pela boa convivência e pelas discussões e conversas científicas. Agradecimentos especiais ao Murilo, por ter revisado um dos meus manuscritos e alguns projetos que escrevi durante o doutorado e por dividir comigo alguns de seus códigos em bash.

Agradeço também aos meus colegas de laboratório durante o meu doutorado sanduíche, Ernesto, Rory, Ramu, Justin e Ian, pela boa convivência e pelas ótimas conversas sobre WE. Agradecimentos especiais ao Ernesto, que deu sugestões para o meu trabalho e com quem tive muitas conversas sobre as vantagens e defeitos de WE. Agradecimentos especiais também ao Rory, por dar sugestões para o meu trabalho e por dividir comigo alguns de seus códigos em python.

Agradeço a profa. Dra. Lillian Chong, da University of Pittsburgh, pelas sugestões para melhorar o meu trabalho. Agradeço também a um de seus alunos de doutorado, Adam Pratt, por dar sugestões para o meu trabalho e por me ajudar a resolver questões técnicas do WESTPA, programa usado para implementar o método WE.

Agradeço a minha família, em especial os meus pais, Maria Elisa e Heli, e meu irmão, Léo, por todo carinho e incentivo. Agradeço também a família que eu ganhei ao casar com o Javier (Jorge, Veronica, Christian, Ingrid, Pamela, Susana, Pablo, Maik e Kevin).

Agradeço também a todos os meus amigos (André, Liv, Estela, Bia, Claudinha, Lígia, Mônica, Lucyanne, Renato, Rodolfo, Ju, Thais) pela convivência e pelas risadas. Agradeço também aos meus amigos de Pittsburgh (Tales, Pedro, Anne, Eduardo, Kate, Jean, Vanessa, Cristiane), que ajudaram a tornar a minha estadia lá mais divertida.

Agradeço aos meus colegas e ex-colegas do Departamento de Bioquímica e do Departamento de Química do Instituto de Química, em especial Bruno Chausse, Bisson e meus colegas de representação discente, pelas conversas e pela motivação.

Agradeço ao Instituto de Química da Universidade de São Paulo por prover um bom ambiente para a realização do meu doutorado.

Agradeço ao Department of Computational and Systems Biology da University of Pittsburgh por ceder parte dos recursos computacionais usados para realizar o trabalho com o método WE e por prover um bom ambiente durante a realização do meu doutorado sanduíche. Agradeço também ao University of Pittsburgh Center for Research Computing por prover parte dos recursos computacionais usados para realizar o trabalho com o método WE.

Agradeço aos criadores do abnTeX2, uma classe LaTeXpara a criação e formatação de documentos conforme as normas ABNT.

Por fim, agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), que financiou meu doutorado sanduíche e grande parte do meu doutorado, e me proporcionou recursos para ir em congressos de alto nível científico, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que financiou o início do meu doutorado.

"What I cannot create, I do not understand."

Richard Feynman

Resumo

Alves, A.F.N. **Simulações computacionais de desenovelamento de proteína e complexação de ligantes com amostragem aumentada**. 2017. 145p. Tese - Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

Simulações moleculares podem fornecer informações e detalhes mecanísticos que são difíceis de obter de experimentos. No entanto, fenômenos bioquímicos como formação de complexos proteína-ligante e desenovelamento de proteína são lentos e difíceis de amostrar na escala de tempo geralmente atingida por simulações de dinâmica molecular (MD) convencionais. Esses fenômenos moleculares foram estudados aqui pela combinação de simulações de MD com diversos métodos e aproximações para aumentar a amostragem configuracional: método de energia de interação linear (LIE), a aproximação de *ensemble* ponderado (WE) e dinâmica molecular dirigida (SMD). Uma equação foi parametrizada para prever afinidades entre pequenas moléculas e proteínas baseada na aproximação LIE, que foca a amostragem computacional nos estados complexado e não-complexado do ligante. A flexibilidade proteica foi introduzida usando *ensembles* de configurações obtidos de simulações de MD. Diferentes esquemas de média foram testados para obter afinidades totais de complexos proteína-ligante, revelando que muitas configurações de complexo contribuem para as afinidades de proteínas flexíveis, enquanto as afinidades de proteínas rígidas são dominadas por uma configuração de complexo. O mutante L99A da lisozima T4 (T4L) é provavelmente a proteína mais frequentemente usada para estudar complexação de ligantes. Estruturas cristalográficas mostram que a cavidade de ligação artificial criada pela mutação é pouco acessível, portanto movimentos proteicos ou uma "respiração" conformacional são necessários para permitir a entrada e saída de ligantes. Simulações de MD foram combinadas aqui com a aproximação de WE para aumentar a amostragem de eventos infrequentes de saída do benzeno de T4L. Quatro possíveis caminhos foram encontrados e movimentações de alfa-hélices e cadeias laterais envolvidas na saída do ligante foram caracterizadas. Os quatro caminhos correspondem a túneis da proteína previamente observados em simulações de MD longas de T4L *apo*, sugerindo que a heterogeneidade de caminhos ao longo de túneis intrínsecos é explorada por pequenas moléculas para sair de cavidades de ligação enterradas em proteínas. Experimentos de microscopia de força atômica revelaram informações detalhadas do desenovelamento forçado e da estabilidade mecânica da rubredoxina, uma proteína ferro-enxofre simples. O desenovelamento completo da rubredoxina envolve a ruptura de ligações covalentes. Portanto, o processo de desenovelamento foi simulado aqui por simulações de SMD acopladas a uma descrição clássica da dissociação de ligações. A amostragem de eventos de desenovelamento forçado foi aumentada pelo uso de velocidades rápidas de esticamento. Os resultados foram analisados usando um modelo teórico válido para regimes de desenovelamento forçado lentos e rápidos. As simulações revelaram que mudanças no ponto de aplicação de força ao longo da sequência da rubredoxina levam a diferentes mecanismos de desenovelamento, caracterizados por variáveis graus de rompimento de ligações de hidrogênio e estrutura secundária da proteína.

Palavras–chave: formação de complexos proteína-ligante, cinética de ligação, desenovelamento de proteína, dinâmica molecular, amostragem aumentada

Abstract

Alves, A.F.N. **Computer simulations of protein unfolding and ligand binding with enhanced sampling.** 2017. 145p. PhD Thesis - Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Molecular simulations may provide information and mechanistic insights that are difficult to obtain from experiments. However, biochemical phenomena such as ligand-protein binding and protein unfolding are slow and hard to sample on the timescales usually reached by conventional molecular dynamics (MD) simulations. These molecular phenomena were studied here by combining MD simulations with several methods or approximations to enhance configurational sampling: linear interaction energy (LIE) method, weighted ensemble (WE) approach and steered molecular dynamics (SMD). An equation was parametrized to predict affinities between small molecules and proteins based on the LIE approximation, which focus computational sampling in ligand bound and unbound states. Protein flexibility was introduced by using ensembles of configurations obtained from MD simulations. Different averaging schemes were tested to obtain overall affinities for ligand-protein complexes, revealing that many bound configurations contribute to affinities for flexible proteins, while affinities for rigid proteins are dominated by one bound configuration. T4 lysozyme (T4L) L99A mutant is probably the protein most often used to study ligand binding. Crystal structures show the artificial binding cavity created by the mutation has low accessibility, so protein movements or conformational "breathing" are necessary to allow the entry and egress of ligands. MD simulations were combined here with the WE approach to enhance sampling of infrequent benzene unbinding events from T4L. Four possible pathways were found and motions on alpha-helices and side chains involved in ligand egress were characterized. The four pathways correspond to protein tunnels previously observed in long MD simulations of *apo* T4L, suggesting that pathway heterogeneity along intrinsic tunnels is explored by small molecules to egress from binding cavities buried in proteins. Previous atomic force microscopy experiments revealed detailed information on the forced unfolding and mechanical stability of rubredoxin, a simple iron-sulfur protein. Complete unfolding of rubredoxin involves rupture of covalent bonds. Thus, the unfolding process was simulated here by SMD simulations coupled to a classical description of bond dissociation. Sampling of forced unfolding events was increased by using fast pulling velocities. Results were analyzed using a theoretical model valid for both slow and fast forced unfolding regimes. Simulations revealed that changing the points of force application along the rubredoxin sequence leads to different unfolding mechanisms, characterized by variable degrees of disruption of hydrogen bonds and secondary protein structure.

Keywords: ligand-protein binding, binding kinetics, protein unfolding, molecular dynamics, enhanced sampling

List of abbreviations and symbols

AFM	atomic force microscopy
α_{LIE}	coefficient to scale the contribution from van der Waals interactions to ΔG_b^{LIE}
β_{LIE}	coefficient to scale the contribution from electrostatic interactions to ΔG_b^{LIE}
ΔG_b	binding free energy for ligand-protein complex
ΔG_b^{LIE}	binding free energy for ligand-protein complex predicted by the LIE approach
ΔH_b	change in enthalpy
ΔLc^{AFM}	contour length increment from AFM experiments
ΔLc^{PDB}	contour length increment calculated from crystal structures
ΔS_b	change in entropy
ΔU_{pot}	change in potential energy
$\Delta x^{\ddagger}$	distance between the folded state and transition configurations
F_{AFM}	force generated by the resistance offered by the molecule to extension in AFM experiments
FeS	iron-sulfur
FKBP12	FK506 binding protein 12
$\bar{F}_{unf}$	average unfolding force
HIV	human immunodeficiency virus
k_B	Boltzmann constant
k_c	force constant of cantilever
K_d	equilibrium dissociation constant for ligand-protein complex
k_{off}	dissociation rate constant for ligand-protein complex
k_{on}	association rate constant for ligand-protein complex
k_p	force constant of additional term in SMD
k_{unf}	spontaneous unfolding rate
$L_0(t)$	equilibrium distance between the cantilever and the surface
$L(t)$	current distance between the cantilever and the surface

LIE	linear interaction energy
MD	molecular dynamics
NMR	nuclear magnetic resonance
R	universal gas constant
SMD	steered molecular dynamics
T	temperature
τ_{dt}	dwell time
τ_{ed}	transition event duration
U_{add}	term added to the potential energy of the system in SMD
U_{elec}	potential energy of electrostatic interactions
U^L	interaction energy between the ligand and its environment when the ligand is in the unbound state
U^{LP}	interaction energy between the ligand and its environment when the ligand is in the bound state
U_{pot}	potential energy of the system
U_{vdW}	potential energy of van der Waals interactions
v_c	pulling velocity of stage in AFM
v_p	pulling velocity of additional term in SMD
WE	weighted ensemble
$\xi_0(t)$	reference value of the progress coordinate
$\xi(t)$	current value of the progress coordinate

Contents

1 Introduction

Computer simulations are helpful to provide information and mechanistic insights that cannot be obtained from experiments. The relevance of simulations was recognized by the Nobel Prize in Chemistry in 2013, which was attributed to the main developers of computational methods to model and simulate chemical and biochemical systems [1]. For instance, simulations were applied in the development of vaccines with increased stability [2] and in drug design [3].

The general aim of this thesis was to model biochemical phenomena slow in the timescales usually reached by computer simulations. The next sections present these biochemical phenomena (section 1.1), the proteins used as model systems to study these phenomena (section 1.2), the computational methods and approximations used to model these phenomena (section 1.3) and the specific aims of this thesis (section 1.4).

Besides the introduction, this thesis contains three chapters equivalent to manuscripts. Chapter 2 describes a method to estimate binding affinities based on the linear interaction energy (LIE) approach and including protein flexibility. This manuscript was published in the Journal of Chemical Information and Modeling in 2014. Chapter 3 characterizes unbinding pathways for benzene from the binding site of T4 lysozyme L99A mutant and the associated protein conformational changes, obtained by combining molecular dynamics (MD) simulations with the weighted ensemble (WE) approach. Finally, chapter 4 describes a method to couple covalent bond cleavage with molecular mechanics and steered molecular dynamics (SMD) simulations and the application of this method to study the forced unfolding of rubredoxin. This manuscript is currently under review in the Journal of Chemical Theory and Computation. The thesis finishes with a general conclusion (chapter 5).

1.1 Biochemical phenomena

The next sections describe the biochemical phenomena studied, binding of small molecules to proteins (section 1.1.1) and forced protein unfolding (section 1.1.2).

1.1.1 Protein-small molecule binding

In a system composed by protein (P), a small molecule or ligand (L) and surrounding solvent, binding can be modeled as a two-state process:

$$P + L \rightleftharpoons PL \tag{1.1}$$

where the unbound state corresponds to ligand and protein free in solvent, and the bound state corresponds to the ligand-protein complex in solvent. A state is a group of microstates (geometries or configurations) belonging to the same energy basin and separated by low energetic barriers compared to the thermal energy available to the system. On the other hand, the states or conformations of a system are separated by high energetic barriers.

The thermodynamics of the binding process is characterized by the equilibrium dissociation constant (K_d), which measures the affinity of the ligand for the protein. K_d is given by:

$$K_d = \frac{[P][L]}{[PL]} \tag{1.2}$$

where $[X]$ stands for the concentration of X in equilibrium. The affinity of the ligand for the protein can also be expressed by the binding free energy (ΔG_b), which is related to K_d by:

$$\Delta G_b = RTlnK_d \tag{1.3}$$

$$\Delta G_b = \Delta H_b - T\Delta S_b \tag{1.4}$$

where R is the universal gas constant, T is the temperature in Kelvin and ΔH_b and ΔS_b are the changes in enthalpy and entropy of the system due to ligand-protein binding, respectively. ΔG_b is a state function, since it depends on the end states of the binding

process only. The change in enthalpy is given by:

$$\Delta H_b = \Delta U_{pot} + \mathcal{P}\Delta V \tag{1.5}$$

where ΔU_{pot} is the change in potential energy, $\mathcal{P}$ is pressure and V is volume. In biological systems, ΔV is usually small and can be neglected. So, changes in the enthalpy are given by changes in the potential energy, which is given by the sum of covalent and noncovalent interactions in the system (details in section 1.3.2.1). Changes in enthalpy upon binding usually result from loss of noncovalent interactions, such as hydrogen bonds and electrostatic and van der Waals interactions, between water and protein or water and ligand and gain of noncovalent interactions between protein and ligand. Moreover, changes in enthalpy can also come from gain or loss of intramolecular interactions. Water molecules are usually released from stable interactions with protein or ligand upon binding, increasing their translational and rotational degrees of freedom, while protein and ligand may have increased restrictions in their configurational, translational or rotational degrees of freedom. Such changes lead to increase and decrease in the entropy of the system, respectively.

The kinetics of the binding process is characterized by the association (k_{on}) and dissociation rate constants (k_{off}), which indicate the timescales for binding and unbinding to happen. Under steady-state conditions:

$$K_d = \frac{k_{off}}{k_{on}} \tag{1.6}$$

Rate constants are proportional to the free energy barrier for unbinding ($\Delta G^{\ddagger}_{off}$) or binding ($\Delta G^{\ddagger}_{on}$), according to Eyring's equation [4, 5]:

$$k_{on} \propto exp\left(\frac{-\Delta G^{\ddagger}_{on}}{RT}\right) \tag{1.7}$$

$$k_{off} \propto exp\left(\frac{-\Delta G^{\ddagger}_{off}}{RT}\right) \tag{1.8}$$

Figure 1 shows an energy landscape and the associated ΔG_b, $\Delta G^{\ddagger}_{off}$ and $\Delta G^{\ddagger}_{on}$ values. $\Delta G^{\ddagger}_{off}$ and $\Delta G^{\ddagger}_{on}$ are not state functions, since they depend not only on the end states of the process, but also on the pathway used by the system to move from one state to the

other. The higher the value of $\Delta G^{\ddagger}_{on}$ or $\Delta G^{\ddagger}_{off}$, the lower will be the value of k_{on} or k_{off} and the lower will be the number of transition events for a fixed amount of time.

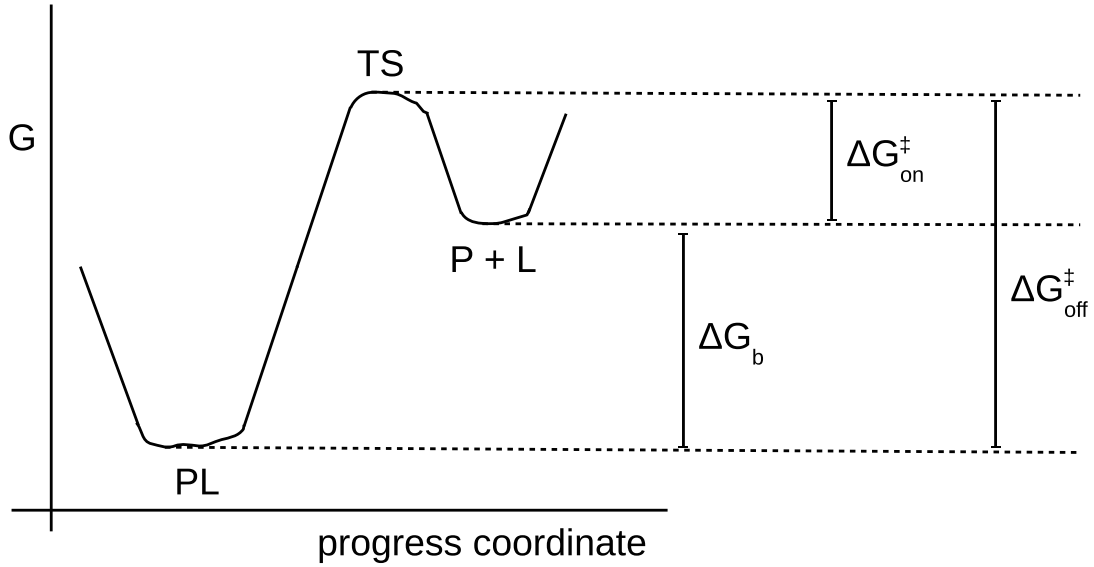


Figure 1 – Energy landscape for a two-state binding process (equation 1.1). G: free energy, L: ligand, P: protein, TS: group of transition structures, ΔG_b: binding free energy, $\Delta G^{\ddagger}_{off}$: free energy barrier for unbinding, $\Delta G^{\ddagger}_{on}$: free energy barrier for binding.

The rate constants k_{on} and k_{off} can also be described as mean first passage times (MFPT) [5]:

$$MFPT_{on} = \frac{1}{k_{on}[L]} \tag{1.9}$$

$$MFPT_{off} = \frac{1}{k_{off}} \tag{1.10}$$

$MFPT_{off}$ is also known as the residence time and describes the time a ligand spends bound to a protein [6–9]. A single first passage time (FPT) corresponds to the time it takes to happen one transition between states and can be expressed as [10, 11]:

$$FPT = \tau_{dt} + \tau_{ed} \tag{1.11}$$

where τ_{dt} is the dwell time, which is the waiting time for the start of the transition, and τ_{ed} is the transition event duration, the time it takes to complete a transition from one state to the other once it starts. During τ_{dt} the system is occupying the free energy basin corresponding to the bound or unbound state and may accumulate energy to change

states. As states are usually separated by high energetic barriers compared to the thermal energy available to the system, the τ_{dt} value is usually large. Moreover, τ_{dt} is usually much larger than τ_{ed} and represents the largest portion of the FPT. Once the system accumulates energy to change states, the duration of the transition event corresponding to such change is usually fast, leading to a small τ_{ed} value [10].

It should be noted that representing ligand-protein binding as a two-state process is a simplified picture. Intermediate metastable states may be involved in binding, what would lead to additional steps in equation 1.1 [5, 9]. Moreover, conformational changes after the formation of the ligand-protein complex can happen, leading to another stable state with increased affinity. This effect is known as induced fit and would also lead to an additional step in equation 1.1.

1.1.2 Forced protein unfolding

Proteins have flexible structures and can assume multiple native conformational states in solution. Unfolding is the process by which a protein moves from one of these native states to a non-native one. Protein unfolding experiments can reveal information about the molecular interactions underlying the stability of native states. Unfolding can be probed by thermal or chemical denaturation, which retrieve an average behavior for a group of molecules. On the other hand, unfolding can also be achieved by single-molecule techniques, such as fluorescence resonance energy transfer and force spectroscopy [12].

Force spectroscopy experiments using atomic force microscopy (AFM) [13] lead to protein unfolding by application of a mechanical force. Such experiments were used, for instance, to reveal the pathways and intermediate states of unfolding of membrane proteins [14–16] and to understand the extensible properties of the protein titin, which is responsible for the elasticity of muscle tissue cells [17–22].

In single-molecule AFM experiments one end of a molecule is adsorbed to a surface and the other end is attached to a cantilever (figure 2a). Motion of the stage containing the surface in the perpendicular direction leads to unfold of the molecule, generating a force-extension curve with a regular saw-tooth pattern (figure 2b) [23, 24]. The force (F_{AFM}) is

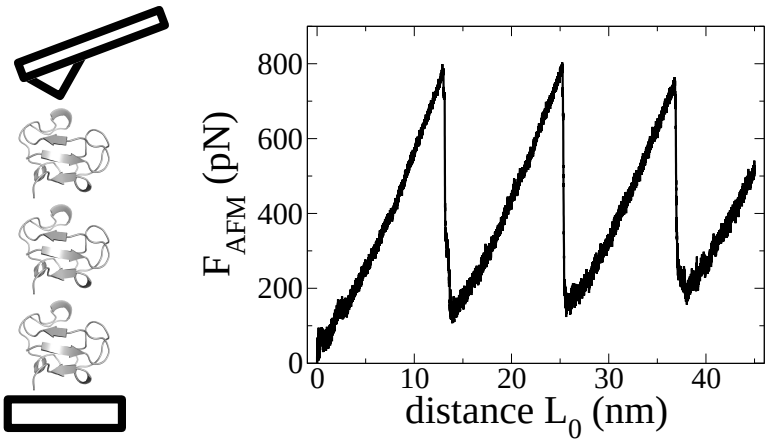
generated by the resistance offered by the molecule to extension, causing deflection of the cantilever from its equilibrium position, and is determined according to Hooke's law [24]:

$$F_{AFM}[L(t)] \;=\; -k_c[L(t) - L_0(t)] \tag{1.12}$$

where $L(t)$ and $L_0(t)$ are the current and equilibrium distances between the cantilever and the surface, and k_c is the force constant of the cantilever. $L_0(t)$ changes in time (t) according to the pulling velocity (v_c):

$$L_0(t) \;=\; L(0) + v_c t \tag{1.13}$$

Alternatively, forced protein unfolding can be obtained by manipulating the stage to obtain constant pulling force. The present section will focus on the results and interpretation of experiments obtained by motion of the stage at constant pulling velocity only.



(a) (b)

Figure 2 – Atomic force microscopy (AFM) experiments. (a) Scheme of a single molecule (polyprotein, in gray). One end of the polyprotein is adsorbed to a surface and the other end is attached to a cantilever. (b) Force-extension curve with a regular saw-tooth pattern. Each force peak corresponds to unfolding of a protein unit in the polyprotein.

Single proteins are small and hard to manipulate in AFM experiments [23]. Thus, polyproteins are built to generate a single molecule. Polyproteins are composed of multiple protein units in tandem (figure 2a), which are assembled by genetic engineering [25] or chemical cross-linking [26].

AFM experiments reveal force peaks and contour length or maximum extension increments (ΔLc^{AFM}). Each peak of the force-extension curve corresponds to the unfold of a protein unit in the polyprotein. The ΔLc^{AFM} value corresponds to the increase in the maximum extension of the polyprotein after one unfolding event. This value is obtained by fitting the unfolding peaks from force-extension curves to the worm like chain model [27] to estimate the contour length (Lc) and calculating the difference between fitted Lc values from successive peaks. The ΔLc^{AFM} value allows the prediction of the unfolded region by comparison with the contour length increments calculated from crystal structures ($\Delta \mathrm{Lc}^{PDB}$).

The average unfolding forces obtained from the peaks of several force-extension curves depend on the pulling velocity. AFM experiments run at different pulling rates depict the dependency of unfolding forces on pulling velocities, also known as the force spectrum [28, 29]. The force spectrum can be fitted to mathematical models [30–35], allowing the estimation of the spontaneous unfolding rate (k_{unf}), which is proportional to the free energy barrier for unfolding ($\Delta G^{\ddagger}_{unf}$), and the distance between the folded state and transition configurations ($\Delta x^{\ddagger}$) in an energy landscape where the progress coordinate corresponds to the pulling coordinate $L(t)$ (figure 3).

One of these models is the phenomenological model [30, 31], which is based on the observation of a linear relationship between average unfolding forces and the logarithm of v_c. According to this model, the average unfolding force ($\bar{F}_{unf}$) is given by:

$$\bar{F}_{unf}\beta \approx \frac{1}{\Delta x^{\ddagger}} ln\left(k_c \beta v_c \Delta x^{\ddagger} e^{-\gamma} \frac{1}{k_{unf}}\right) \tag{1.14}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant and γ is the Euler-Mascheroni constant.

However, the linear relationship between $\bar{F}_{unf}$ and the logarithm of v_c does not hold for high pulling velocities. Hummer and Szabo [34] proposed a microscopic model to address this issue, where $\bar{F}_{unf}$ is given by [34]:

$$\bar{F}_{unf} = -k_c\left(\Delta x^{\ddagger} - v_c \int_0^{\tau_x} S(t)dt\right) \tag{1.15}$$

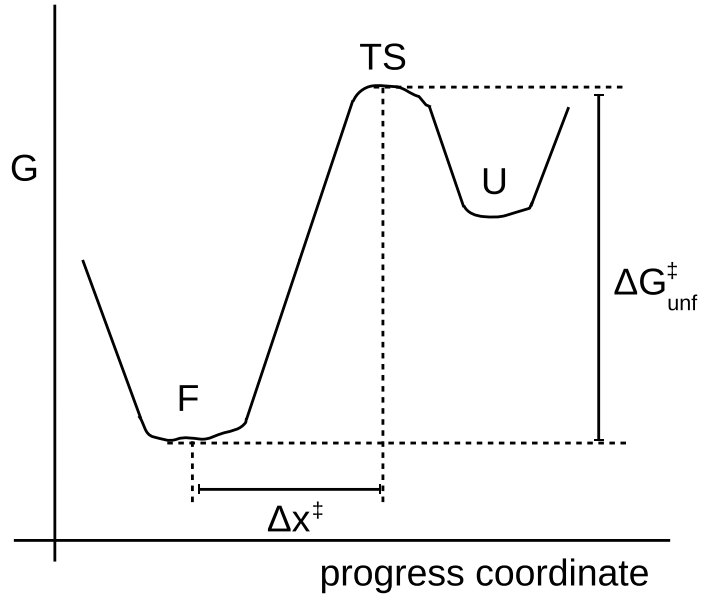


Figure 3 – Energy landscape for protein unfolding. G: free energy, F: folded state, U: unfolded state, TS: group of transition structures, $\Delta G^{\ddagger}_{unf}$: free energy barrier for unfolding, $\Delta x^{\ddagger}$: distance between the folded state and transition configurations.

where $S(t)$ is the survival probability or fraction of folded proteins at time t, given by [34]:

$$S(t) = exp\left[-\frac{k_{unf}e^{-k_c\beta(\Delta x^{\ddagger})^2/2}}{k_c\beta v_c\Delta x^{\ddagger}[k_c/(k_m+k_c)]^{3/2}}(e^{k_c\beta v_c\Delta x^{\ddagger}t-(k_c\beta v_c t)^2/[2\beta(k_m+k_c)]}-1)\right] \quad (1.16)$$

where k_m is the molecular force constant and τ_x is the time at which $\Delta x^{\ddagger}$ is equal to the average protein extension ($\bar{x}$), given by [34]:

$$\bar{x}(t) = \frac{v_c k_c\beta}{D[\beta(k_m+k_c)]^2}[Dt\beta(k_m+k_c)+e^{-Dt\beta(k_m+k_c)}-1] \quad (1.17)$$

where D is the diffusion coefficient. At intermediate pulling velocities, which are typical of most AFM experiments, this model predicts a nonlinear relationship between $\bar{F}_{unf}$ and the logarithm of v_c, differing from the phenomenological model. At high pulling velocities the model predicts a linear relationship between $\bar{F}_{unf}$ and $v_c^{1/2}$ [34]. Such prediction was recently supported by AFM experiments performed at high pulling velocities [28].

1.2 Protein systems studied

Computational methods are usually validated by comparing the results obtained from simulations with those obtained from experiments. If the simulation is able to reproduce experimental results, this indicates that the simulation captures the microscopic

details necessary to model the biochemical phenomena studied. Therefore, proteins used as model systems in computer simulations are usually those with many experimental data available. Such protein systems may or may not have applications in biology. Once computational methods are validated using such proteins, these methods may be employed to study proteins with pharmaceutical or biotechnological interest. The next sections describe the protein systems used in this thesis to study or test computational methods.

1.2.1 T4 lysozyme mutants

Bacteriophage T4 lysozyme is a monomeric protein containing 164 amino acid residues. Its structure is globular and has two domains connected by an alpha helix (figure 4) [36, 37]. This protein contributes to the lytic cycle of the virus by catalyzing the hydrolysis of $\beta(1 \rightarrow 4)$ linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine, causing rupture of bacteria cell wall [37, 38].

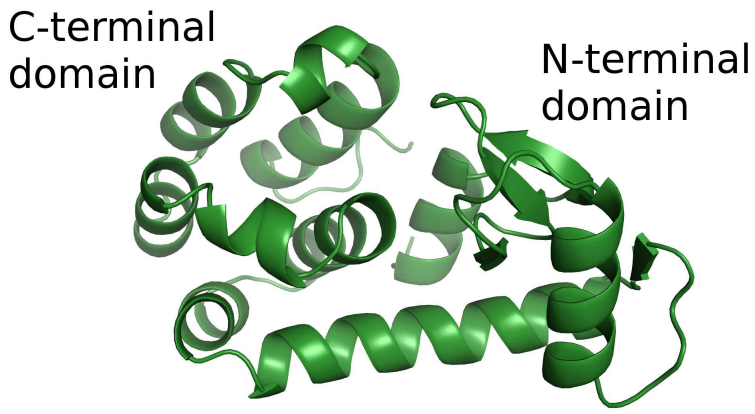
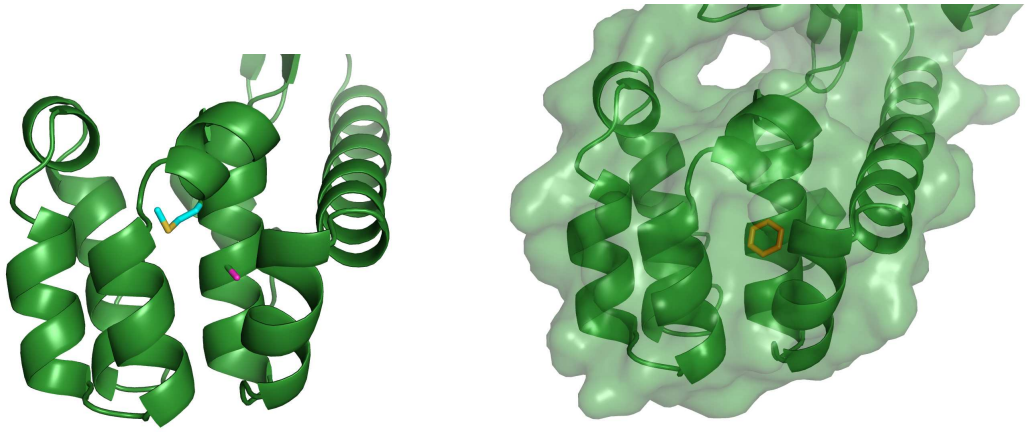


Figure 4 – Crystal structure of T4 lysozyme.

Several mutants of T4 lysozyme were created [39–41] after the determination of its structure by X-ray crystallography [42] to study the factors that determine the structure and stability of proteins. One of these mutants, L99A (figure 5) [43], contains a hydrophobic cavity of 150 $Å^3$ in the C-terminal domain. This cavity is absent in the wild type protein and was shown to bind to noble gases [44] and small nonpolar molecules such as benzene (figure 5b) [43]. Moreover, another mutant, L99A/M102Q (figure 5a) [45], was

designed to introduce a polar group in the engineered cavity, allowing binding of small polar molecules such as phenol and aniline.



(a) (b)

Figure 5 – Crystal structure of T4 lysozyme L99A mutant. (a) The amino acid residues of positions 99 and 102 are highlighted by pink and cyan carbons, respectively. (b) T4 lysozyme L99A mutant bound to benzene (orange). The protein is represented with its molecular surface (green transparency), showing the ligand is fully buried. Only the C–terminal domain is shown.

T4 lysozyme L99A and L99A/M102Q mutants are often used as model systems in computational and experimental studies of binding thermodynamics [37,45–56] due to the simplicity of the engineered binding site. Crystal structures of T4 lysozyme mutants with (*holo*) or without ligands (*apo*) [43,45,47,48,51,52,57] revealed that the engineered cavity is hidden from solvent (figure 5b) and is empty in the absence of ligands, indicating that a desolvation step for ligand binding is not necessary. Moreover, small rotameric changes or shifts in alpha helix F are enough to accommodate ligands. Such situation differs from binding events for most proteins, which may involve displacement of water molecules in the binding site by the ligand and large protein conformational changes before binding, imposing difficulties to the prediction of binding affinities. T4 lysozyme mutants were used in my master's thesis as a model system to develop a computational method to predict binding affinities including protein flexibility [37].

Although the structural and microscopic details underlying ligand binding thermodynamics for T4 lysozyme mutants are well characterized, binding kinetics is not fully

understood yet. Crystal structures of the mutants complexed with ligands [43, 45, 47, 48, 51, 52, 57] show that the opening on the protein surface for ligand entry and escape from the engineered binding site is small (figure 5b). Nuclear magnetic resonance (NMR) spectroscopy experiments [58] were used to study the binding kinetics of small ligands, determining k_{off} values of 325 s^{-1} and 800 s^{-1} for indole and benzene respectively and a k_{on} value of 10^6 M^{-1} s^{-1} for both ligands. Recent computer simulations found five transient tunnels connecting the engineered binding site to the solvent in the *apo* L99A mutant [59]. Computer simulations also revealed that one of these tunnels is used for benzene entry in the binding site [60] and another tunnel is used for benzene exit [61]. Moreover, three tunnels were identified for O_2 to exit or access the binding site [62], among which two were previously described [59]. So, it remains to be tested if all the transient tunnels found in the *apo* L99A mutant are used as exit routes for ligands.

Since the engineered binding site of the mutants is hidden from solvent, protein conformational changes are expected to allow ligand excursion to the binding site [58]. Spin nuclear relaxation experiments [63] showed the existence of two conformational states for the L99A mutant: a highly populated state (97%) similar to the crystal structure and a less populated state (3%) that was suggested as the state that opens the cavity to allow ligand entry. A structure of this less populated state was proposed with the use of chemical shifts and computer simulations [64]. In this structure alpha helix F is aligned with alpha helix G and one amino acid residue is occupying the engineered binding site. Therefore, this structure does not make the cavity accessible to ligands. Motions in alpha helix F were suggested [36, 58] to contribute to the binding process, as previous data from crystal structures [43, 45, 47, 48, 51, 52, 57] and NMR [63, 64] showed this alpha helix is more disordered than the other structural elements in the C-terminal domain of T4 lysozyme. However, it remains to be tested if motions in alpha helix F are useful for ligand binding. Pathways for ligand unbinding from T4 lysozyme and the associated protein conformational changes will be addressed in chapter 3.

1.2.2 HIV reverse transcriptase

Reverse transcriptase of the human immunodeficiency virus (HIV) 1 is a heterodimeric protein containing a 560-residue subunit known as p66 and a 440-residue subunit known as p51 (figure 6). This protein contributes to the HIV cycle by synthesizing a double-stranded deoxyribonucleic acid (DNA) using the virus ribonucleic acid (RNA) as template, allowing integration of the viral genome in the host chromosome. The catalytic site is contained in the p66 subunit [65]. HIV reverse transcriptase is a major target in drug design due to its role in the replication of HIV, which causes the acquired immune deficiency syndrome (AIDS) [66].

HIV-1 reverse transcriptase is used as a model system in computational studies of ligand binding thermodynamics [67–70] due to the availability of half maximal inhibitory concentrations, which are proportional to binding affinities, for many inhibitors [71–74] and *holo* and *apo* crystal structures [75–78].

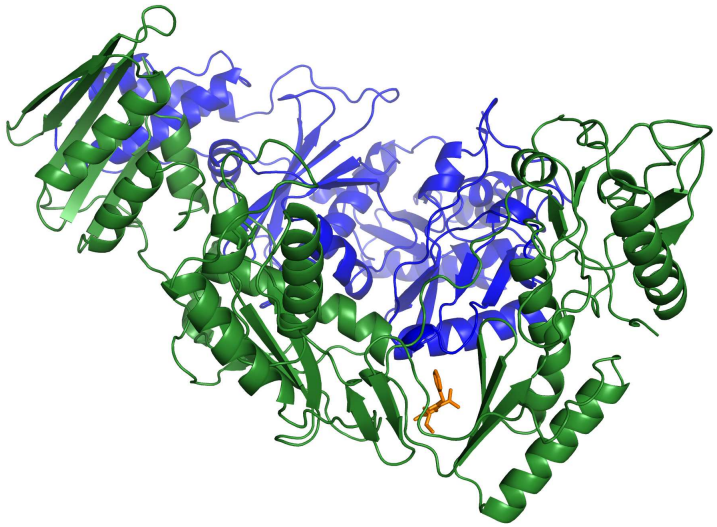


Figure 6 – Crystal structure of HIV-1 reverse transcriptase bound to an inhibitor (orange). The p66 and p51 subunits are depicted in green and blue, respectively.

1.2.3 Human FK506 binding protein

Human FK506 binding protein 12 (FKBP12) is a monomeric protein containing 108 amino acid residues (figure 7). This protein has peptidylprolyl cis/trans isomerase

activity and is a major target in drug design due to its participation in immunosuppressant effects when bound to drugs such as FK506 [79].

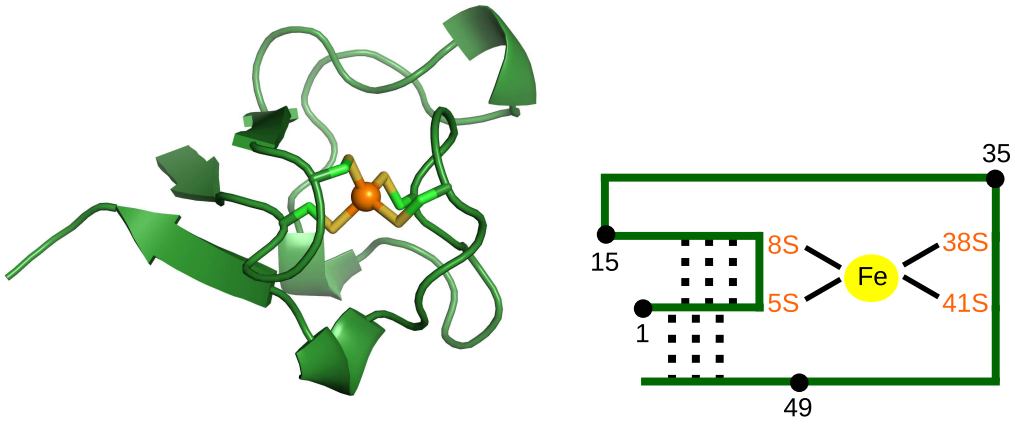
FKBP12 is used as a model system in computational studies of ligand binding thermodynamics [80–82] due to the availability of binding affinities for many ligands [83,84] and *holo* and *apo* crystal structures [83–86]. Although experimental rate constants are unknown for the binding of ligands to FKBP12, this protein is also used as a model system in computational studies of ligand binding kinetics [87, 88] because the binding site is shallow and exposed to solvent (figure 7), facilitating ligand dissociation.

Figure 7 – Crystal structure of FKBP12 bound to a ligand (orange).

1.2.4 Rubredoxin

Rubredoxin from the hyperthermophilic archaeon *Pyrococcus furiosus* is a monomeric protein containing 53 amino acid residues. It is the smallest protein to show an iron-sulfur (FeS) center, which is composed of four cysteine side chains S bound to one Fe atom in a tetrahedral orientation (figure 8) [89]. This protein participates in electron transfer reactions to reduce superoxide to hydrogen peroxide [90].

Rubredoxin from *Pyrococcus furiosus* is considered a hyperthermostable protein, since it unfolds at temperatures beyond 100 °C [91,92]. Computational and experimental



(a) (b)

Figure 8 – Rubredoxin. (a) Crystal structure. Cysteines of the FeS center are shown as sticks, iron is shown in orange. (b) Scheme of protein structure, showing the positions of the FeS center, beta-sheets (hydrogen bonds depicted as dotted lines) and point mutations (black dots). The protein backbone is represented by green lines.

studies [91–97] of this protein alone or with its counterpart, the mesophilic rubredoxin from *Clostridium pasteurianum*, have been done to understand the microscopic reasons underlying thermal stability in proteins. Such studies showed that salt bridges and hydrophobic interactions help in the achievement of increased thermal stability.

The structural stability of rubredoxin has been extensively studied by AFM [98–103]. Initial work [98] used a polyprotein composed of rubredoxin units assembled by the N and C-terminal residues using genetic engineering [25]. Force-extension curves obtained for this polyprotein revealed an average ΔLc^{AFM} value of 12.6 nm. Such value indicates rupture of the FeS center and complete unfolding of rubredoxin, which requires rupture of at least two of the four ferric-thiolate (Fe-S) covalent bonds. Moreover, fitting of the force spectrum to the phenomenological model resulted in a k_{unf} value of 0.15 s^{-1} and a $\Delta x^{\ddagger}$ value of 0.11 nm. Later [100], polyproteins were constructed by chemical cross-linking [26] of cysteine residues introduced in the rubredoxin sequence by point mutations. Mutations were introduced in positions 1 and 49, 15 and 49, 15 and 35 or 1 and 35 (figure 8b), resulting in different points of force application along the rubredoxin sequence. ΔLc^{AFM} values obtained indicate rupture of the FeS center in all mutants. Rubredoxins mutated in

positions 1 and 49, 15 and 49, or 15 and 35 presented k_{unf} and $\Delta x^{\ddagger}$ values similar to the ones obtained in the initial work, while rubredoxins mutated in positions 1 and 35 had a slower k_{unf} value (3 10^{-6} s^{-1}) and a larger $\Delta x^{\ddagger}$ value (0.30 nm). The molecular reasons for the dependence of rubredoxin unfolding kinetics on the point of force application are unknown.

Electronic structure calculations conducted in our research group [103–105] revealed details of the Fe-S bond rupture in AFM, showing that Fe-S bond cleavage is homolytic and that water substitution leads to faster Fe-S bond rupture. Further microscopic details of the unfolding mechanism of rubredoxin in AFM remain to be elucidated. This issue will be addressed in chapter 4.

1.3 Computational methods

The next sections present the two computational methods used to model the biochemical phenomena considered previously, molecular docking (section 1.3.1) and molecular dynamics (MD) simulations (section 1.3.2), and the methods used to enhance configurational sampling (section 1.3.3).

1.3.1 Molecular docking

Molecular docking [106] generates complexes between proteins and small molecules or ligands and estimates a score for these complexes using the structures of a target protein and of a ligand, and a grid determining the region in the protein where potential binding sites will be searched. A search algorithm is used to explore different orientations and configurations of the ligand in the protein. This search algorithm retrieves the best poses of the ligand guided by a scoring function, which aims at mimicking experimental affinities [107].

Due to its low computational cost, molecular docking is the most common computational method used in rational drug design efforts. One of its uses is in predicting ligand poses for target proteins with a crystal structure available [108–110]. Knowledge of

the ligand-protein complex structure shows which intermolecular interactions contribute for binding, providing information for the design of ligands with improved affinities.

Docking can also be employed in virtual screening [107,108,111–115]. In this case, libraries containing thousands of molecules or candidate ligands are tested. These molecules are docked to a target protein and ranked according to the score attributed to the complex. Then, the top molecules of this ranking are chosen to be tested experimentally.

Although very popular, docking presents two major approximations that can be sources of error in the search for ligand poses and in the scoring function. One of them is keeping the protein rigid (section 1.3.1.1) and the other is using an approximate scoring function (section 1.3.1.2), which neglects important contributions for binding [37,108,115]. These approximations will be discussed in the next sections.

1.3.1.1 Rigid protein approximation

In docking the protein structure is usually represented as rigid. This helps to keep the computational cost low. However, it is known from experimental results that proteins are flexible. Such flexibility is pointed out, for instance, by increased B-factors or alternative side chain conformations in crystal structures, and by the use of an ensemble to represent structures determined by NMR. So, protein structures are better represented not by one configuration, but by an ensemble or group of configurations. Moreover, induced fit effects are also neglected in docking due to lack of protein flexibility.

Some errors can be generated by representing the protein as rigid, such as not recognizing that a ligand fits in the binding site or generating a poor ligand-protein complex, that do not resemble the crystallographic one.

Previous works addressed the challenge of including protein flexibility in docking. Soft docking [116] allows some superposition between ligand and protein structure during docking. So, protein flexibility is addressed in a limited way. Side chain flexibility can be incorporated using a rotamers library [117] or allowing rotation of selected side chains during docking [118]. However, unfeasible configurations, which are not accessible in solution, can be generated and protein backbone moves are not included.

On the other hand, there are methods which allow the inclusion of flexibility of the protein backbone and side chains. In such cases, docking is performed using not one protein configuration, but a group of configurations obtained from MD simulations [37, 107, 119–121], different crystal structures [122] or NMR studies [123]. For instance, a group of configurations from MD simulations was used in our group to represent a phosphatase [121] and in my master's thesis to represent T4 lysozyme mutants [37]. When MD simulations are used to obtain groups of configurations the simulations should be long enough to guarantee that all the configurations important for ligand binding were visited (section 1.3.2.2).

1.3.1.2 Scoring function

The scores attributed to complexes between protein and small molecules should be able to predict affinities similar to the experimental ones, to distinguish between good poses, close to the crystallographic binding site, and bad ones, and to separate binder from non-binder molecules. Some of these tasks may be poorly performed because the scores attributed are approximate.

In the docking program AutoDock Vina [124] ΔG_b (equation 1.4) is approximated by the following scoring function (E^{dock}):

$$E^{dock} = \frac{U^{dock}_{noncov}}{1 + 0.0585 N_{tor}} \tag{1.18}$$

$$U^{dock}_{noncov} = \sum_{i<j} -0.0356 e^{-(d_{ij}/0.5)^2} - 0.00516 e^{-[(d_{ij}-3)/2]^2} + 0.84 U_{cl} - 0.0351 U_{hyd} - 0.587 U_{hb} \tag{1.19}$$

$$d_{ij} = r_{ij} - W_i - W_j \tag{1.20}$$

$$U_{cl} = \begin{cases} d_{ij}^2 & \text{if} \quad d_{ij} < 0 \\ 0 & \text{if} \quad d_{ij} \geq 0 \end{cases} \tag{1.21}$$

$$U_{hyd} = \begin{cases} 1 & \text{if} \quad d_{ij} < 0.5\text{Å} \\ 0 & \text{if} \quad d_{ij} > 1.5\text{Å} \end{cases} \tag{1.22}$$

$$U_{hb} = \begin{cases} 1 & \text{if} \quad d_{ij} < -0.7\text{Å} \\ 0 & \text{if} \quad d_{ij} > 0 \end{cases} \tag{1.23}$$

where N_{tor} is the number of ligand rotatable bonds and U_{noncov}^{dock} is the sum of noncovalent interactions in docking, represented by energetic contributions from steric clashes (first three terms of equation 1.19), hydrophobic interactions (U_{hyd}) and hydrogen bonds (U_{hb}) between ligand and protein. r_{ij} is the distance between atoms i and j and W is the van der Waals radius. The coefficients multiplying each energetic contribution to estimate U_{noncov}^{dock} in equation 1.19 were obtained by parametrization of the equation using ligand-protein complexes with experimental ΔG_b values determined. U_{cl}, U_{hyd} and U_{hb} vary linearly as a function of d_{ij} between the extreme values of d_{ij} in equations 1.21, 1.22 and 1.23.

The scoring function, E^{dock}, contains many approximations to represent ΔH_b and ΔS_b in equation 1.4. ΔS_b is represented by N_{tor}. Restrictions to the ligand translation and rotation due to binding, reduction in the number of protein configurations due to conformational selection and increase in the number of solvent configurations available due to release of water molecules interacting with protein or ligand after binding can also contribute to ΔS_b. However, such terms are not considered in equation 1.18.

Moreover, ΔH_b is represented by U_{noncov}^{dock} (equation 1.19), which contains terms to describe van der Waals interactions and hydrogen bonds in the bound state only. Changes in covalent interactions, such as bonds or dihedrals in the ligand or in the protein, in noncovalent intramolecular interactions or in electrostatic interactions due to binding may have significant contributions to ΔH_b. These terms are not taken into consideration in the scoring function presented in equation 1.18.

Therefore, keeping the protein rigid and neglecting contributions to ΔH_b and ΔS_b in the scoring function contribute to the imprecision of molecular docking. These issues will be addressed in chapter 2.

1.3.2 Molecular dynamics simulations

Over the past years, structural biology provided atomic-resolution structures of proteins and macromolecular complexes as big as virus capsids [125]. However, such structures are static. Proteins are flexible in solution (section 1.3.1.1) and their motions allow them to perform functions such as cell signaling and catalysis. MD simulations [126] are

used to model the motions and conformations accessible to proteins, revealing microscopic details of how proteins are able to perform their functions.

MD simulations provide trajectories of the system coordinates along time using molecular mechanics or Newton's law of motion:

$$\vec{F_i} = m_i\vec{a_i} \tag{1.24}$$

where $\vec{F_i}$ is the force acting over atom i, m_i is the mass and $\vec{a_i}$ is the acceleration. The force acting over every atom is calculated from the potential energy. The length of the trajectory, or the number of times the equation 1.24 will be integrated, depends on the timescale of the phenomena of interest.

The main challenges in performing MD simulations of biomolecules are to do an accurate description of the potential energy of the system (section 1.3.2.1) and achieve reasonable configurational sampling (section 1.3.2.2), or obtaining the correct populations of the microstates and states of the system. These challenges will be presented in the next sections.

1.3.2.1 Potential energy

In molecular mechanics the potential energy (U_{pot}) of the system is usually described using force fields. However, the use of force fields to describe biomolecules presents some challenges and approximations [127,128]. Ideally, the potential energy of microscopic systems should be described by quantum mechanics equations, but solving these equations presents high computational costs for molecules as large as proteins. The parameters to describe covalent and noncovalent energies are usually available for amino acids only. So, if a protein contains a metal center or is bound to a small molecule, parameters to describe the covalent and noncovalent interaction energies of the metal center or molecule must be derived. Moreover, atoms are represented with a fixed point charge. So, it is not possible to represent polarization or charge transfer [127, 128]. As metal ions have charges and coordination numbers that depend on the environment, a force field representation is usually poor for such ions, because charges and bonds are usually fix during the simulation.

The force field contains terms to describe covalent (U_{cov}) and noncovalent (U_{noncov}) interactions:

$$U_{pot} = U_{cov} + U_{noncov} \tag{1.25}$$

The covalent interactions are given by the sum of the terms corresponding to bond (U_{bond}), angle (U_{angle}), dihedral (U_{dih}) and improper dihedral (U_{imp}) energies [129]:

$$U_{cov} = U_{bond} + U_{angle} + U_{dih} + U_{imp} \tag{1.26}$$

Bond and angle energies are usually approximated by harmonic functions [129]:

$$U_{bond} \approx \sum_{bond} \frac{1}{2} k_b (b - b_0)^2 \tag{1.27}$$

$$U_{angle} \approx \sum_{ang} \frac{1}{2} k_\theta (\theta - \theta_0)^2 \tag{1.28}$$

where k_b and k_θ are force constants, b is the length of the bond between two atoms, θ is the angle between three atoms, and b_0 and θ_0 are the equilibrium values. The dihedral energy surface may have multiple energy minima, so it is better approximated by a periodic function [129]:

$$U_{dih} \approx \sum_{dih} \frac{1}{2} k_d [1 + cos(n_d \phi - \delta_d)] \tag{1.29}$$

where k_d is a force constant, n_d represents the periodicity of the angle, δ_d represents the phase of the angle and ϕ is the angle of the dihedral. The same equation can be used for the energy of improper dihedrals, which describe out-of-plane deviations.

The harmonic potential (equation 1.27) can be replaced by a Morse potential (U_{Morse}) to describe bond energies when simulation of covalent bond rupture is desired [130]:

$$U_{Morse} = \sum_{bond} D_M [1 - exp(-\beta_M (b - b_0))]^2 \tag{1.30}$$

where D_M is the depth of the potential well and β_M is the steepness of the well. For increasing $(b - b_0)$ values the harmonic potential gives high energies, forcing the system to stay close to the equilibrium value b_0. On the other hand, the Morse potential gives lower energies than the harmonic potential for increasing $(b - b_0)$ values, allowing bond

stretching and rupture during the simulation (figure 9). It should be noted that the use of a Morse potential to represent covalent bond rupture is also an approximation. Covalent bond rupture involves changes in the electronic structure, changes of partial charges and polarization effects. However, such changes and effects are not represented when a Morse potential is used.

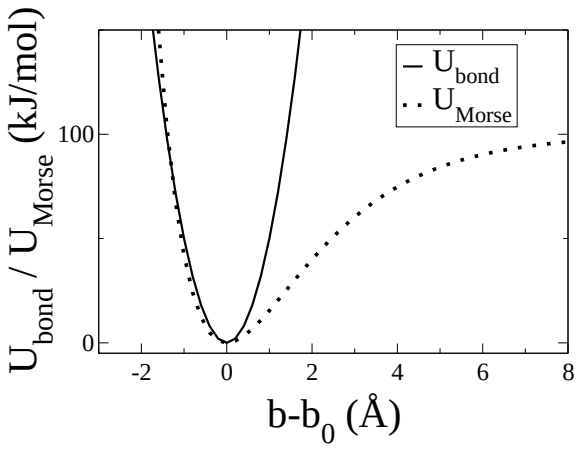


Figure 9 – Potential energies of a bond described by an harmonic (U_{bond}, equation 1.27) or by a Morse potential (U_{Morse}, equation 1.30) as a function of the difference between bond length (b) and equilibrium bond length (b_0).

Noncovalent interactions are given by the sum of electrostatic (U_{elec}) and van der Waals (U_{vdW}) terms [129]:

$$U_{noncov} = U_{elec} + U_{vdW} \tag{1.31}$$

Noncovalent interactions are usually modeled by pair-wise potentials. The calculation of the electrostatic energy (U_{elec}) is based on the Coulomb law [129]:

$$U_{elec} = k_e \sum_{i<j} \frac{q_i q_j}{r_{ij}} \tag{1.32}$$

where k_e is a constant that depends on the dielectric permissivity of the medium, q_i and q_j are the partial charges of atoms i and j and r_{ij} is the distance between these atoms. The calculation of the van der Waals energy (U_{vdW}) is approximated by the Lennard-Jones function [129]:

$$U_{vdW} \approx \sum_{i<j} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.33}$$

where ϵ_{ij} is the depth of the potential well describing the interaction between atoms i and j and σ_{ij} is the distance at which the potential reaches its minimum. The term $1/r_{ij}^{12}$ is related to interactions of electron clouds close to each other, leading to repulsion between the atoms, while the term $1/r_{ij}^{6}$ is related to the dispersion energy due to correlated fluctuations in the charge distributions of the two atoms, leading to attraction between them [129].

The equilibrium terms and force constants of equations 1.27, 1.28 and 1.29, the atomic charges, σ and ϵ values of equations 1.32 and 1.33 and the equations 1.27 to 1.29, 1.32 and 1.33 compose the force field. Parameters of the force field are usually obtained from quantum–mechanical calculations or from fitting to reproduce quantum–mechanical calculations or experimental observables such as liquid densities, heats of vaporization or protein crystal structures [131–134].

Besides the approximations presented above, the solvent can be represented in an implicit manner, by using equations to model the average interaction energy of the solvent with the solutes in the system. The use of implicit solvation reduces the computational cost, as the forces and motions of explicit water molecules do not need to be computed. Moreover, the relaxation of water is instantaneous for every solute configuration, reducing the amount of computational effort required to obtain reasonable configurational sampling (section 1.3.2.2). However, the use of implicit solvation also has disadvantages. For instance, it is not possible to represent hydrogen bonds between solute and solvent.

Noncovalent interaction energies between the solute and the implicit solvent (G_{sol}) are given by [135]:

$$G_{sol} \approx G_{GB} + G_{NP} + G_{cav} \tag{1.34}$$

where G_{GB} represents the free energy of polarization according to the generalized Born approximation, G_{NP} represents the nonpolar free energy of interaction between solute and implicit solvent and G_{cav} is the energy required to build a cavity for the solute in the solvent, including the work to reorganize solvent molecules around the solute and the work against the solvent pressure to create the cavity [135]. The non-electrostatic term of

equation 1.34 can be calculated as [136]:

$$G_{NP} + G_{cav} = \eta SASA \tag{1.35}$$

where $SASA$ is the solute solvent accessible surface area and η is a constant. G_{GB} is obtained by the generalized Born approximation. The formulation given by Still *et al.* [136] is used in many simulation programs:

$$G_{GB} = -\frac{1}{2}\left(1 - \frac{1}{\zeta}\right)\sum_{i \leq j}\frac{q_i q_j}{f(r_{ij}, a_{ij})} \tag{1.36}$$

where ζ is the medium dielectric constant, and $a_{ij} = (a_i a_j)^{1/2}$, where a_i and a_j are the Born radii of atoms i and j. $f(r_{ij}, a_{ij}) = (r_{ij}^2 + a_{ij}^2 e^{-B})$, where $B = r_{ij}^2/(2a_{ij})^2$. Due to the functional form of $f(r_{ij}, a_{ij})$, G_{GB} results in the Born model, which estimates the free energy of polarization of a spherical charge, when i=j and in the sum of the expressions of the Coulomb and Born models when two charges are far apart [136].

Equation 1.24 may be modified to incorporate the effects of friction and collisions between water and solute molecules in the propagation of the system when implicit solvation is employed. These effects are incorporated by stochastic or Langevin dynamics [137]:

$$m_i \vec{a_i} = -m_i \gamma_i \vec{v_i} + \vec{F_i} + R_i \tag{1.37}$$

where $\vec{v_i}$ is the velocity, γ_i is the friction constant and R_i is a noise process, which models the effect of random collisions between water and solute.

1.3.2.2 Configurational sampling

It is considered that good sampling of molecular simulations is achieved when the simulated configurations are obtained with the same weights or populations observed experimentally. In equilibrium conditions the relative populations of the configurations accessible to the system are given by the Boltzmann distribution [138, 139]:

$$\rho(x_c) \propto exp[-\beta U_{pot}(x_c)] \tag{1.38}$$

where $\rho(x_c)$ is the probability density or population of configuration x_c. Therefore, the more favorable U_{pot} is for a configuration, the higher is the population of this configuration.

In experiments with many units of one molecule in solution, $\rho(x_c)$ is equal to the fraction of molecules in configuration x_c in one time point. However, MD simulations are usually performed for one unit of one molecule in solution to keep computational costs low. In this case $\rho(x_c)$ is equal to the fraction of time the molecule was observed in configuration x_c during the simulation. The assumption that time averages, as those of MD simulations, can reproduce ensemble averages, as those of experiments, is known as the ergodic theorem [140]. The population of a state is given by the sum of the populations of the configurations that belong to this state [138]:

$$P(x_s) = \int_{V_A} \rho(x_c)dx_c \propto \int_{V_A} exp[-\beta U_{pot}(x_c)]dx_c \quad (1.39)$$

where $P(x_s)$ is the probability or population of state x_s and V_A comprises all the configurations that belong to state x_s. So, MD simulations should be long enough to guarantee that all configurations of the states of interest were visited multiple times, such that reasonable $\rho(x_c)$ and $P(x_s)$ values can be estimated. However, the length of MD simulations is limited by the system size and the computational resources available.

Biochemical phenomena such as protein-ligand binding (section 1.1.1) and forced protein unfolding (section 1.1.2) are slow for the timescales usually reached by MD simulations. Ligand binding and unbinding are infrequent events which usually take milliseconds or more to happen due to large dwell times (τ_{dt}, equation 1.11). AFM experiments are usually performed at pulling velocities (v_c, equation 1.13) of 10^{-6} m/s, requiring milliseconds to lead to unfolding of all the protein units in a polyprotein. On the other hand, conventional MD simulations are limited to the microsecond timescale [9, 11]. Therefore, methods or approximations to enhance configurational sampling are necessary to simulate these phenomena.

1.3.3 Enhanced sampling methods

Configurational sampling may be enhanced by increasing the computational time spent in regions of interest (sections 1.3.3.1 and 1.3.3.2) or by speeding up the occurrence of conformational transitions in the system (section 1.3.3.3). The next sections describe such methods and approximations used here to enhance configurational sampling.

1.3.3.1 Linear interaction energy

Linear interaction energy (LIE) [141] is an approach to estimate binding affinities (section 1.1.1). Traditional computational methods to estimate affinities, such as free energy perturbation (FEP) [142] and thermodynamic integration (TI) [143], require multiple simulations of points along a computational pathway connecting the end-points of the binding process. LIE can be considered an approach to increase configurational sampling when compared to FEP and TI because it focus the computational effort in the regions of interest, the bound and unbound states of the ligand. Due to this focused computational effort, the LIE approach is able to estimate affinities at a lower computational cost compared to FEP and TI.

The LIE approach estimates affinities by assuming a linear response of the intermolecular interactions. Affinities are predicted (ΔG_b^{LIE}) using energy contributions obtained from MD simulations of the ligand free in solvent and bound to the protein [141]:

$$\Delta G_b^{LIE} = \alpha_{LIE}(\langle U_{vdW}^{LP}\rangle - \langle U_{vdW}^{L}\rangle) + \beta_{LIE}(\langle U_{elec}^{LP}\rangle - \langle U_{elec}^{L}\rangle) \quad (1.40)$$

where $\langle\cdots\rangle$ represents a configurational average and U^{LP} and U^{L} are the interaction energies between the ligand and its environment when the ligand is in the bound and unbound states, respectively. The differences of average interactions are multiplied by coefficients derived from the linear response assumption (β_{LIE}=0.5) [144] or obtained by calibration of equation 1.40 to reproduce experimental affinities (α_{LIE}). Variations of equation 1.40 have been used, such as obtaining the value of β_{LIE} by calibration, including a free coefficient to account for contributions not included in U_{vdW} and U_{elec} or including additional terms that may contribute for binding, such as changes in the solvent accessible surface area or in the intramolecular energies of the ligand and of the protein [68,145,146].

The LIE approach has been applied successfully to predict affinities for different ligand-protein complexes [37,67,68,121,145–150]. For instance, a LIE equation with four coefficients parametrized for HIV reverse transcriptase resulted in an average deviation between experimental and estimated affinities of 1.3 kcal/mol for 57 inhibitors [67]. Another LIE equation with three coefficients parametrized for the same protein resulted in

average deviations of 0.8 kcal/mol for 39 inhibitors [68]. LIE equations were employed by our group to predict binding affinities for complexes between phosphatase and its inhibitors [121] and in my master's thesis to predict binding affinities between T4 lysozyme mutants and small molecules [37].

One of the main limitations of LIE is the poor transferability of the coefficients among different proteins. Coefficients of LIE equations usually predict affinities that resemble the experimental ones for complexes of the specific protein used to calibrate them only. Attempts to increase the transferability of the coefficients were proposed [55, 151], such as adapting them by the number of hydrogen bonds the ligand can make or by the ligand or binding site relative polarities. This issue will be addressed in chapter 2.

1.3.3.2 Weighted ensemble

The weighted ensemble (WE) method [152, 153] enhances sampling of infrequent biochemical phenomena. It resembles the LIE approach (section 1.3.3.1), since it also enhances sampling by increasing the computational effort in the regions of interest. However, in the WE method the regions of interest are those of low probability. Such regions are usually associated with transition configurations of conformational changes, which have unfavorable potential energies and, therefore, low probabilities (equation 1.38). One consequence of focusing computational effort in low probability regions is the reduction of dwell times (τ_{dt}, equation 1.11), which usually account for most of the time necessary to observe a single infrequent event.

In the WE method a progress coordinate that describes the infrequent biochemical phenomena, such as the distance between two atoms or groups, is defined and divided into bins. A group of trajectories of the system in an initial state is propagated by MD simulations and receive initial equal weights or probabilities. Every τ steps, the group of trajectories is resampled by evaluating each bin occupancy. Trajectories may be replicated or pruned with a proper weight attribution to keep a given number of trajectories per bin, once a bin has been visited. For instance, if one of the initial trajectories reached a new unvisited bin, and a number of 4 trajectories per bin was set up initially, this trajectory

is split in 4 and each of the new trajectories receives 1/4 of the weight of the mother trajectory. Thus, sampling in bins of low probability is enhanced (figure 10). However, if a bin has more than 4 trajectories, the exceeding trajectories are removed and their weights are divided among the remaining trajectories of the bin. This reduces the computational effort spent in bins of high probability. The cycle of propagation and resampling steps is repeated until state populations are converged or, in other words, do not change with increasing simulation time. In the end a group of trajectories is created with accurate weights.

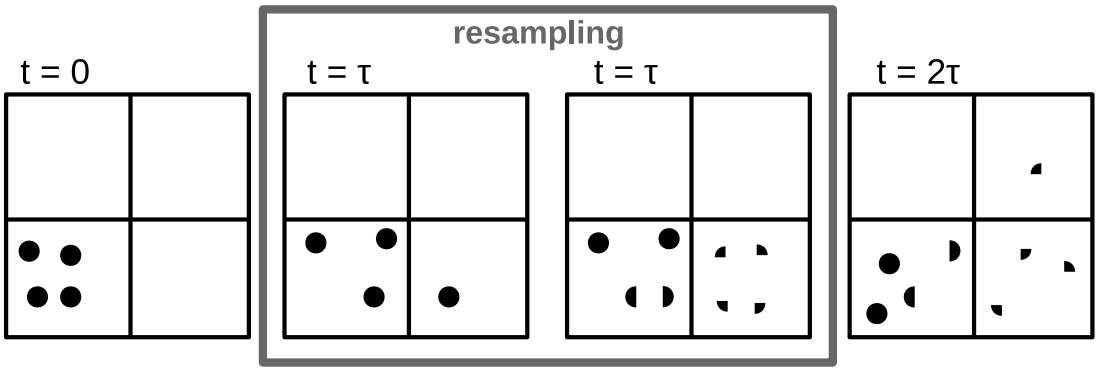


Figure 10 – Weighted ensemble method. In this example trajectories are replicated or merged every τ steps to keep 4 trajectories (circles) per bin (squares). One of the trajectories reached a new unvisited bin. So, in the resampling step, this trajectory is split in 4 and each of the new trajectories receives 1/4 of the weight of the mother trajectory (quarter circles).

Transition rates and state populations can be estimated from a set of trajectories obtained from a WE procedure. The population of a state is given by the sum of weights of the trajectories belonging to the bins corresponding to this state. If the trajectories arriving at the target state B are immediately fed back into the initial state A during the WE procedure, the transition rate from A to B (k_{AB}) can be estimated as the sum of probability fluxes into B [154]:

$$k_{AB} = \sum_{j \neq B} f_{jB} \tag{1.41}$$

where f_{jB} is the probability flux, or probability per unit time, from bin j to the bins of the state B and j includes all the bins, except those which define the state B. The definition of states A and B can be adjusted to allow the use of k_{AB} values to estimate k_{on} and k_{off} values (section 1.1.1).

The WE method has been applied to study pathways and kinetic rates of many biochemical phenomena such as protein and peptide conformational transitions [155–157], protein unfolding [158], protein-peptide binding [159], protein-protein binding [160] and protein-ligand unbinding [88, 161].

The main limitations of the WE method are the generation of correlated trajectories and the dependence on a progress coordinate to describe the infrequent biochemical phenomena [11, 153]. Due to the trajectory splitting and pruning scheme to keep a given number of trajectories per bin, an ensemble of trajectories sharing part of their history is generated, leading to correlation among trajectories [11, 153]. The progress coordinate should include the slowest degrees of freedom in the infrequent biochemical phenomena. Therefore, some knowledge of the phenomena is required to define the progress coordinate. If one of the slow degrees of freedom is not included in the progress coordinate, reasonable sampling of all the important configurations may not be achieved.

Methods that add an artificial term to the potential energy of the system, thus reducing the free energy barrier for state transitions, have also been used to enhance sampling of infrequent biochemical phenomena [60, 162]. The advantage of the WE approach over these methods is that it does not change the potential energy, therefore avoiding perturbations in the group of transition configurations and in the mechanism of state transitions.

1.3.3.3 Steered molecular dynamics

In steered molecular dynamics (SMD) simulations [32, 163] a term ($U_{add}[\xi(t)]$) is added to the potential energy (U_{pot}) to force the system to leave the initial state and reach the desired state:

$$U_{SMD} = U_{pot} + U_{add}[\xi(t)] \tag{1.42}$$

where U_{SMD} is the new potential energy of the system. $U_{add}[\xi(t)]$ depends on the progress coordinate ξ, which can be the distance between two groups. $U_{add}[\xi(t)]$ usually has the

form of an harmonic potential of force constant k_p:

$$U_{add}[\xi(t)] = \frac{k_p}{2}[\xi(t) - \xi_0(t)]^2 \quad (1.43)$$

where $\xi(t)$ and $\xi_0(t)$ are the current and reference values of the progress coordinate, respectively. $\xi_0(t)$ changes in time according to the pulling velocity (v_p):

$$\xi_0(t) = \xi(0) + v_p t \quad (1.44)$$

SMD is usually employed to model forced protein unfolding (section 1.1.2) due to the similarity between $U_{add}[\xi(t)]$ and the combination of stage and cantilever in AFM experiments. $U_{add}[\xi(t)]$ and the stage are moved with constant pulling velocity, leading to increasing distances between a pulled group and a reference group and forced unfolding of the protein units of a polyprotein. Moreover, forced protein unfolding by SMD produces force-extension curves similar to the ones of AFM. Pulling forces are obtained by the derivative of $-U_{add}[\xi(t)]$ (equation 1.43) in respect to ξ, resulting in an equation similar to equation 1.12.

In SMD simulations enhanced sampling is achieved by the use of high pulling velocities, which are usually orders of magnitude faster than those of AFM experiments and speed up the occurrence of conformational transitions. Due to the use of high pulling velocities full unfolding of a polyprotein, which is achieved in milliseconds in AFM experiments, can be obtained in nanoseconds, a timescale affordable in MD simulations. Moreover, the use of high pulling velocities results in simulations with low computational cost. Thus, tens or hundreds of SMD simulations can be performed, allowing the estimation of average properties such as average unfolding forces ($\bar{F}_{unf}$) and contour length increments (ΔLc).

The use of much faster pulling velocities in SMD requires care in the comparison of the results from SMD simulations and AFM experiments. As average unfolding forces depend on the pulling velocity, it is not possible to compare them directly. An indirect comparison is possible by fitting the force spectrum to the microscopic model presented before (section 1.1.2), which is valid for both intermediate and fast pulling velocities regimes.

SMD simulations provided microscopic details of forced unfolding experiments for many proteins [164–173]. For instance, SMD simulations revealed the molecular basis for the plateau phase seen in fibrinogen force-extension curves [170] and that the mechanical stability of the titin I91 domain is due to contacts between beta-strand pairs [164, 166, 167, 169, 171]. These SMD simulations were used to model proteins that unfold due to disruption of noncovalent interactions only. Despite the many AFM experiments of forced protein unfolding where disruption of covalent interactions is involved [98–103, 174], SMD simulations have not been used to model such experiments because classical force fields (section 1.3.2.1) are unable to represent the rearrangement of electronic structure involved in bond dissociation. This issue will be addressed in chapter 4.

1.4 Aims

1.4.1 Prediction of affinities for protein-small molecule complexes

Molecular docking (section 1.3.1) is a computational method often used for rational drug design. However, it presents two major approximations that can be sources of error. One of them is treating the protein as rigid (section 1.3.1.1) and the other is using an approximate scoring function (section 1.3.1.2).

One of the aims of this thesis was to develop a computational method to predict binding affinities (section 1.1.1) with better accuracy and including protein flexibility in docking. T4 lysozyme mutants L99A and L99A/M102Q (section 1.2.1), HIV-1 reverse transcriptase (section 1.2.2) and human FKBP12 (section 1.2.3) were used as model systems. Docking was performed using a group of protein configurations obtained from MD simulations (section 1.3.2) to include protein flexibility. The scoring function was replaced by a LIE equation (section 1.3.3.1), which focus the computational effort in the bound and unbound states of the ligand, thus predicting affinities at lower computational cost than other methods. Coefficients of the LIE equation were adapted by the ligand or binding site relative polarities to increase their transferability among different model systems.

1.4.2 Pathways for protein-small molecule unbinding

The binding kinetics (section 1.1.1) of T4 lysozyme mutants (section 1.2.1) is not fully understood. The engineered binding site of these mutants is hidden from solvent and openings on the protein surface for ligand escape are small. Knowledge about the pathways for a ligand to dissociate from the binding site can help in the prediction of kinetic rates. However, pathways for ligand exit from the buried binding site of T4 lysozyme mutants and the associated protein conformational adjustments have not been fully resolved.

Another aim of this thesis was to determine pathways for benzene exit from T4 lysozyme L99A mutant and the associated protein conformational changes. MD simulations (section 1.3.2) were combined with the WE approach (section 1.3.3.2) to enhance sampling of infrequent unbinding events.

1.4.3 Forced protein unfolding

AFM experiments (section 1.1.2) revealed information about rubredoxin (section 1.2.4) forced unfolding and mechanical stability. However, the microscopic details of the forced unfolding mechanism have not been fully resolved.

The last aim of this thesis was to determine the microscopic mechanism of forced unfolding of rubredoxin. Full unfolding of rubredoxin involves rupture of Fe-S covalent bonds. Here, covalent bond cleavage was allowed by replacing an harmonic potential (equation 1.27) by a Morse potential (equation 1.30) to represent Fe-S bonds. SMD simulations (section 1.3.3.3), which mimic AFM experiments, were combined to high pulling velocities to enhance sampling of unfolding events.

2 Ligand-receptor affinities computed by an adapted linear interaction model for continuum electrostatics and by protein conformational averaging

Ariane Nunes-Alves and Guilherme Menegon Arantes

Department of Biochemistry, Instituto de Química, Universidade de São Paulo, SP, Brazil

Reprinted with permission from Nunes-Alves, A.; Arantes, G. M. Ligand-receptor affinities computed by an adapted linear interaction model for continuum electrostatics and by protein conformational averaging. *J. Chem. Inf. Model.*, v. 54, p. 2309-2319, 2014.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

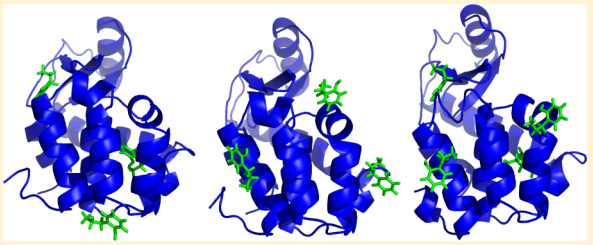
Ligand−Receptor Affinities Computed by an Adapted Linear Interaction Model for Continuum Electrostatics and by Protein Conformational Averaging

Ariane Nunes-Alves and Guilherme Menegon Arantes*

Department of Biochemistry, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil

Supporting Information

ABSTRACT: Accurate calculations of free energies involved in small-molecule binding to a receptor are challenging. Interactions between ligand, receptor, and solvent molecules have to be described precisely, and a large number of conformational microstates has to be sampled, particularly for ligand binding to a flexible protein. Linear interaction energy models are computationally efficient methods that have found considerable success in the prediction of binding free energies. Here, we parametrize a linear interaction model for implicit solvation with coefficients adapted by ligand and binding site relative polarities in order to predict ligand binding free energies. Results obtained for a diverse series of ligands suggest that the model has good predictive power and transferability. We also apply implicit ligand theory and propose approximations to average contributions of multiple ligand−receptor poses built from a protein conformational ensemble and find that exponential averages require proper energy discrimination between plausible binding poses and false-positives (*i.e.*, decoys). The linear interaction model and the averaging procedures presented can be applied independently of each other and of the method used to obtain the receptor structural representation.



1. INTRODUCTION

Prediction of binding affinities between small-molecule ligands and protein receptors has both fundamental and applied importance.[1] In practice, this is a very challenging task[2] because the ligand functional or bound configurations have a small energy difference from the huge amount of alternative ligand unbound configurations.[3] The number and strength of contributions in the ligand bound and unbound states are similar. Consequently, intermolecular interactions have to be evaluated with accuracies much better than 1 kcal mol^{-1} to discriminate the small energy gap between the two states.[3,4] In addition, a huge number of configurations has to be generated and their energy calculated to sample the important conformational microstates of the molecular system.[3,5,6] The number of configurations to be sampled will increase if the protein or the ligand has a more flexible structure and if their binding pose is unknown or not unique.[2,7]

Despite the challenges, there has been enormous progress in the prediction of binding free energies, and several methods have been proposed to tackle the problem.[1,8,9] In one hand, the application of detailed all-atom force fields, molecular dynamics (MD) simulations (or related approaches), and rigorous free energy estimators[10−13] have found impressive agreement with experimental affinities;[14−17] but, given the high computational costs associated, these methods have been successfully applied mainly to less flexible proteins and ligands for which binding sites are known or easy to determine.[18] The high computational costs still prohibit these rigorous methods from being applied in screenings of large ligand sets. On the other hand, molecular docking[19−21] employs approximate descriptions of intermolecular interactions usually parametrized against empirical data and efficient conformational search methods to generate binding poses,[22,23] rank or enrich ligand sets,[24,25] and determine ligand affinities.[2,26] However, docking has many documented failures[27,23,28] which may be due to severe approximations in the calculation of interactions and lack of transferability for ligands or receptors not included in the method parametrization as well as to insufficient conformational sampling.

Another family of methods shows accuracy and computational ease in between the two approaches just mentioned. They are called linear interaction energy (LIE) models[29−32] because a linear response of the intermolecular interactions[33] is assumed in the estimation of binding free energies by the equation

$$\Delta G_{LIE} = \alpha\Delta\langle V_{vdW}^{l-e}\rangle + \beta\Delta\langle V_{elet}^{l-e}\rangle + \gamma \quad (1)$$

where a force field description of intermolecular van der Waals (*vdW*) and electrostatic (*elet*) interactions between ligand and its environment (V^{l-e}) is employed. The difference (Δ) of ensemble averaged ($\langle\cdots\rangle$) interactions between the ligand free state (when environment is the solvent only) and bound state (when environment is the solvated protein complex) is

Received: May 19, 2014
Published: July 30, 2014

multiplied by coefficients derived from the linear response assumption (β) or fit to empirical data (α and γ).[32,34]

LIE models have been applied successfully to predict affinities for a range of ligand−receptor complexes.[32,35−38] However, in many of these applications, the LIE models were specifically parametrized to the system studied. In order to increase the model transferability, Hansson et al. proposed the adaptation of coefficients to ligand properties (e.g., the number of possible hydrogen bonds).[39] Recently, Linder et al. suggested an adaptative LIE model where coefficients in eq 1 are adjusted by the relative polarities of the ligand and of the binding cavity achieving accuracy and model transferability.[40]

To increase computational efficiency and to avoid the sometimes slow convergence of explicit solvent contributions[41,42] in eq 1, continuum electrostatics descriptions of solvation[43−46] have been used in LIE models.[36,41,47−49] Here, we propose and describe the necessary parametrization of LIE models that combine an implicit solvent description with adaptative coefficients[40] to predict binding affinities. Local configurational sampling of ligand−receptor complexes usually done by molecular dynamics simulations is substituted by more economic molecular docking and geometry optimizations.[21,36,47]

The methods mentioned so far rely their predictions on one initial receptor structure, typically obtained from X-ray crystallography. During conformational search in molecular docking, the receptor structure is maintained rigid, maybe allowing for side-chain rotations or smoothened interactions.[50−52] In methods applying ensemble averages, protein configurations near the initial structure are visited in relatively short MD simulations; but, for flexible receptors, sufficient sampling of protein motions will be difficult to achieve in both approaches. A possible solution in those cases is to start the search or averaging from a conformational ensemble, i.e., from multiple representations of the receptor structure.[6,7,53,54]

Several approaches, mostly related to docking, are now used to predict binding poses and affinities from receptor conformational ensembles.[22,55−59] Usually a dominant pose and dominant state approximation is applied.[57−59] This means that the binding free energy or the related docking score for a given ligand−receptor pair is estimated from the most favorable pose (only one) found after evaluating several complexes obtained from the different receptor structures in the ensemble. This approximation should be appropriate for the level of accuracy expected in docking, but it dismisses important contributions such as multiple binding poses, receptor reorganization energy and thermal fluctuations, and the related entropic effects. Thus, it may be useful to average contributions obtained from an ensemble of ligand−receptor poses.[60,61]

Based on the implicit ligand theory recently developed,[61] two ensemble averages can be defined for the calculation of binding free energy between a ligand and a receptor represented by a conformational ensemble embedded in an implicit solvent. The first average concerns with the ligand configurational distribution that may be obtained for interaction with one given *rigid* receptor structure. It has been called the binding potential of mean-force, B, and may be estimated by an exponential mean

$$B = -k_B T \ln \frac{1}{P} \sum_{i=1}^{P} \exp\left(-\frac{\Psi_i}{k_B T} \right) \quad (2)$$

where k_B is the Boltzmann constant, T is the temperature, P is the number of ligand configurations sampled, and Ψ_i is the *implicit solvent-mediated* interaction energy for the ith ligand pose.[61]

The second ensemble average accounts for the receptor configurational distribution. Similarly, it may be estimated by an exponential mean, leading to an expression for the binding free energy

$$\Delta G_{ave} = -k_B T \ln \frac{1}{N} \sum_{n=1}^{N} \exp\left(-\frac{B_n}{k_B T} \right) + \Delta G_{\xi} \quad (3)$$

where B_n is the binding potential of mean-force (eq 2) for the nth receptor configuration out of a total of N configurations sampled.[61] ΔG_{ξ} represents a correction to standard concentration due to restriction of the volume sampled by the ligand.

Here, we propose approximations to these two exponential averages in order to account for multiple binding poses and for protein conformation flexibility. In the next section, we provide details of the training and test sets and the methodology used to calibrate adapted LIE models for continuum solvation. System setup, generation of receptor−ligand poses, and definition of the force fields employed are also described. Both Results and Discussion are divided in two parts. First, we report the construction and performance of the adapted LIE models and the procedures for averaging contributions from an ensemble of ligand−receptor poses. Then, we analyze the accuracy and shortcomings of the proposed LIE model as well as the applicability of the ligand−receptor conformational averaging.

2. COMPUTATIONAL METHODS

Calibration and tests of the proposed approximations were conducted using bacteriophage T4 lysozyme mutants L99A[62] and L99A/M102Q,[17,63] HIV-1 reverse transcriptase (HIVRT), and human FK506 binding protein 12 (FKBP) as model systems (Table 1). These proteins were chosen based on the availability of experimental structures and binding affinities. The ligand set varies in size from fragment-like small molecules which bind T4 lysozyme to lead-like molecules which bind FKBP. Only neutral ligands were considered (Table S3 and Figure S1 in the Supporting Information, SI).

Table 1. Proteins Included in the Training and Test Sets and Ranges of Binding Affinities of Associated Ligands (in kcal mol^{-1})

protein[a]	PDB code	ligands	$-\Delta G_{exp}$
L99A	3DMV	benzene derivatives	4.5−6.7
M102Q	1LI3	benzene derivatives	4.3−5.8
HIVRT	1RT1	HIV1−HIV6	4.9−11.8
FKBP	1FKG	FKB1−FKB5	7.8−11.2

[a]T4 lysozyme mutants L99A and M102Q, HIV-1 reverse transcriptase (HIVRT), and human FK506 binding protein 12 (FKBP).

2.1. Parametrization of the Model. Following Linder et al.,[40] system-derived descriptors are used here to scale coefficients in the linear interaction models. Ligand (π) and cavity or binding site (η) relative polarities were given by the ratio PSA/SA, where SA represents the ligand or cavity total surface area and PSA represents the area of its subset of polar atoms (Table S3 and Table S4). Ligand surface area was obtained from the "3V" server,[64] and ligand polar surface area

was calculated using the approach of Ertl et al.[65] Cavity area was obtained from the *SA* of residues in contact with (or less than 4 Å from) the ligand. Thus, each binding pose has a characteristic η. Protein carbonyl C, O, N, and H bound to O and N were assigned as polar atoms.[66]

Coefficients in the LIE equations were obtained by an optimization procedure that minimized deviations between calculated and experimental affinities for a training set of 10 T4 lysozyme ligands, 3 HIVRT ligands, and 10 T4 lysozyme false-positive poses, as indicated in Table S5. The test set used to check the performance of the parametrized equations was composed of a different set of 15 T4 lysozyme binders, 10 T4 lysozyme false-positive poses, 9 T4 lysozyme nonbinders, 3 HIVRT ligands, and 5 FKBP ligands, as indicated in Table 2. Optimization was carried out with a combination of genetic (GA)[67] and simplex[68] algorithms as previously described.[69] A population of 10 individuals with each coefficient represented by 12 bits was used in the GA. Coefficients in eq 6 could vary between [0,10] for k_S, [−20,20] for k_6, and [−10,10] for the others k_i (eqs 6 and S1−S3). Populations evolved for 10^6 generations. Simplex optimization was carried from the best GA individuals until the difference in deviations between successive generations was smaller than 10^{-6} kcal mol^{-1}.

2.2. Construction of Receptor−Ligand Structures. Protein structures retrieved from the Protein Databank (PDB) were used after removal of water and other crystallization molecules. Incomplete side chains were built with the WHATIF server.[70] Hydrogens were constructed using the GROMACS PDB parser[71] for proteins and Babel 2.2[72] for ligands. Ligand geometry was optimized using Gaussian[73] with the AM1[74] potential if holo crystal structures were unavailable.

To train and test the LIE models, holo structures were taken from the PDB when available. Otherwise, the most favorable binding pose obtained from docking the ligand to an apo structure was used (Table S3). These poses were compared to known crystal structures of congeneric ligands bound to the same protein to confirm the docked ligand was complexed in a plausible binding mode.

Unstable or artificial poses of known binders may be generated in docking due to inaccuracies in the scoring functions.[27,28,75] Such artificial poses, here called false-positives, were used as a decoy set to assist in the parametrization of the LIE equations. Assuming the ligand will occupy a site different from the known crystallographic site, false-positive poses were obtained by docking ligands to apo crystal structures using a grid excluding the known binding site. Selected poses were submitted to energy minimization, careful heating up ramps, and 10−20 ns explicit solvent molecular dynamics simulations as described below. False-positive poses were retained only if the ligand spent more than 20% of the trajectory dissociated from the protein. Ligand-protein dissociation was monitored by the ligand solvent accessible surface area (SASA).

Tentative configurational ensembles were generated for apo T4 lysozyme L99A and M102Q mutants, HIVRT bound to ligand HIV1 and FKBP bound to ligand FKB1 (Table 1). Ensembles generated from apo HIVRT and FKBP could not be used for docking due to large conformational changes which occluded the binding sites (see Discussion for further details). Receptor structures were submitted to energy minimization, and implicit solvent molecular dynamics simulations were run for 160−235 ns. For each protein, an ensemble was constructed by 50 configurations (excluding the ligand in the case of HIVRT and FKBP) collected along trajectories at regular time intervals (3−4 ns) after stabilization of C_α root mean-squared deviation (RMSD). For each configuration in an ensemble, 20 docking poses were generated resulting in a total of 1000 ligand-bound structures for each protein−ligand pair.

Dockings to crystal structures were performed with AutoDock 4.0[50] with its genetic algorithm search run with 150 individuals for 27,000 generations maximum. Dockings to

Table 2. Binding Free Energies (in kcal mol^{-1}) Experimentally Measured and Estimated by Eq 6 for the Ligand *Test Set*

ligand	ΔG_{exp}[a]	ΔG_{ALICE}
L99A		
n-butylbenzene[b]	−6.7	−6.3
propylbenzene	−6.5	−5.7
ethylbenzene[b]	−5.7	−5.1
toluene	−5.5	−3.7
benzene[b]	−5.2	−3.0
3-ethyltoluene	−5.1	−5.5
meta-xylene	−4.7	−4.5
2-ethyltoluene	−4.5	−4.8
propylbenzene	>−2.0	−1.4
ortho-xylene (A)	>−2.0	0.7
toluene	>−2.0	−1.7
4-ethyltoluene	>−2.0	−2.1
benzene	>−2.0	−0.8
3-methylpyrrole	>−2.0	−2.6
phenol	>−2.0	−3.1
1,3,5-trimethylbenzene	>−2.0	−4.4
cyclohexane	>−2.0	−2.6
2-fluoroaniline	>−2.0	−2.8
M102Q		
(phenylamino)acetonitrile[b]	−5.8	−4.8
toluene	−5.2	−4.1
3-methylpyrrole	−5.2	−2.9
thieno[3,2-b]thiophene[b]	−4.9	−3.5
2-ethylphenol[b]	−4.8	−5.0
catechol[b]	−4.4	−2.5
2-ethoxyphenol[b]	−4.3	−4.9
thieno[3,2-b]thiophene	>−2.0	−0.3
(phenylamino)acetonitrile	>−2.0	−0.4
catechol	>−2.0	−2.4
2-propylphenol (A)	>−2.0	−1.7
2-ethoxyphenol	>−2.0	−0.7
phenylhydrazine	>−2.0	−3.0
2-methoxyphenol[b]	>−2.0	−4.5
4-vinylpyridine	>−2.0	−4.2
N-(*o*-tolyl)cyanoformamide	>−2.0	−5.0
HIVRT		
HIV3	−8.1	−10.5
HIV4	−10.6	−9.5
HIV5	−6.4	−7.4
FKBP		
FKB1[b]	−11.0	−10.9
FKB2[b]	−11.2	−11.4
FKB3	−7.8	−7.2
FKB4	−8.5	−7.8
FKB5	−9.6	−9.9

[a]Repeated from Table S3. [b]Holo structure taken from the PDB. False-positive poses of binder molecules are underlined. These poses and nonbinders were assumed to have $\Delta G_{exp} > -2.0$ kcal mol^{-1}. The label (A) represents different false-positive poses of the same ligand.

dx.doi.org/10.1021/ci500301s | J. Chem. Inf. Model. 2014, 54, 2309−2319

the configurational ensemble were done with AutoDock Vina[21] setting the exhaustiveness level to 8. Conformational search options were chosen in order to thoroughly search for the possible docking poses in a given protein structure. Grids with 0.375 Å spacing and 60 to 80 points were centered in the known binding sites. Protein structures were kept frozen, but bond torsions were allowed in ligands. Typically, T4 lysozyme ligands had 0−4 torsions activated, HIVRT ligands had 3−12 torsions, and FKBP ligands had 6−13 torsions (Table S2). The correction of the restricted volume sampled by the ligand to the standard concentration (1 M)[60] in eq 3 was calculated from the average volume of the grid used for docking, 2.7×10^4 Å^3.

2.3. Protein Force Field and Simulation Details. Energy contributions for the linear interaction models (eqs 6 and S1−S3) were obtained after geometry optimizations of protein−ligand complexes in implicit solvent using the conjugate gradient approach (T4 lysozyme) or the BFGS algorithm (HIVRT and FKBP).[68] Free protein and ligand contributions were obtained without the ligand or protein, respectively, but using the same geometry of the complex. The GBr^6 method[43] was used to calculate the solvent polarization free energies, G_{GB} in eq 6.

GROMACS 4.5[71] was used for all protein geometry optimizations and MDs. Dynamics were carried out at 300 K with a 2 fs time-step, and covalent bonds were constrained with LINCS.[76] Proteins were represented by the OPLS-AA force field.[77]

In explicit solvent simulations, structures were solvated in a dodecahedral box with edges at least 8 Å far from the protein. The SPC/E potential[78] was employed for water, and chloride ions were added to neutralize the charge of the systems. Periodic boundary conditions were activated. The velocity rescale method[79] was used to control the temperature at 300 K, and pressure control at 1 bar was applied with the Parrinello−Rahman method.[80] PME[81] was used to treat long-range electrostatics, and a switched potential (cutoffs 0.8, 1.2 nm) was used to treat van der Waals interactions. Before production MD, systems were heated in cycles of short 20 ps simulations with gradual temperature increase (10 K, 50 K, 100 K, 200 K, and 300 K) and reduction of position restraints over heavy atoms (240 kcal nm^{-2}, 120 kcal nm^{-2}, 24 kcal nm^{-2}, 2 kcal nm^{-2}, and 0).

In implicit solvent simulations, the generalized Born (GB) approximation was used.[44] The OBC model was used to estimate Born radii,[45] and the nonpolar contribution was calculated as in Schaefer et al.[46] with a surface tension of 5.4 cal mol^{-1} Å^{-2} for all atoms. MDs were run with a leapfrog stochastic dynamics integrator, with a friction coefficient $\tau = 10$ ps^{-1}.

2.4. Ligand Force Field. Topologies for ligands were built manually based on the OPLS-AA force field. Bonding, Lennard-Jones, and implicit solvation parameters unavailable for certain atom types in OPLS-AA were approximated from similar chemical functions. Parameters for dihedral angles of the thymine ring in HIVRT ligands were taken from the AMBER99 force field[82,83] OPLS-AA partial charges were used for nonpolar ligands or for ligands with one polar group. For ligands with more than one polar group, partial charges were recalculated with AM1-CM2.[74,84] For small ligands, partial charges for the whole molecule were recalculated. For the bulky FKBP and HIVRT ligands, the molecule was divided in fragments, and those with more than one polar group had their partial charges recalculated. For HIVRT ligands containing sulfur, the partial charges were recalculated with HF/6-31G*.

For all ligands, the partial charges used here resulted in total and component dipole moments in good agreement with a quantum mechanical (QM) reference (HF/6-31G*, Table S1). For instance, OPLS-AA partial charges were used for 4-vinylpyridine resulting in a total dipole moment $\mu = 2.7$ D which is in good agreement with the QM reference $\mu = 2.6$ D. Another example is 2-fluoroaniline which has two polar groups. Its partial charges were recalculated as described above and resulted in $\mu = 2.0$ D which is in good agreement with the QM reference, $\mu = 1.9$ D.

All ligand topologies are available online[85] or from the authors upon request.

2.5. Approximations to Implicit Ligand Theory. Given that only configurations with favorable interaction energies will contribute significantly to the exponential average in eq 2, here we use ligand docking to quickly generate ligand−receptor poses with favorable interactions for a rigid receptor conformation and approximate $\Psi \approx \Delta G_{int}$. Thus, eq 2 leads to

$$B_E = -k_B T \ln \frac{1}{P} \sum_{i=1}^{P} \exp\left(- \frac{\Delta G_{\mathrm{int},i}}{k_B T} \right) \quad (4)$$

where ΔG_{int} is an intrinsic binding free energy used to estimate the stability of a given ligand binding pose (see Results, section 3.1). As docking does not generate an equilibrium distribution of ligand−receptor configurations, application of eq 4 is approximate. Substitution of interaction energies (Ψ in eq 2) for an intrinsic binding free energy parametrized against experimental data may partially correct inaccuracies in the docking energy function and introduce entropic contributions.

In the limit that one individual sample dominates the exponential average in eq 4, the dominant pose approximation may be used

$$B_D = \min_i (\Delta G_{\mathrm{int},i}) \quad (5)$$

where only a single intrinsic free energy of binding contribution is used for each rigid receptor structure.

For the receptor ensemble average, eq 3 is used with B_n calculated by either eqs 4 or 5. A dominant state approximation may also be invoked where only a single receptor configuration [min$_n$ (B$_n$)] is used.[61]

Values of $N = 50$ and $\Delta G_\xi = 0.9$ kcal mol^{-1} were employed here (see section 2.2). A maximum of $P = 20$ complex configurations were drawn from docking a ligand to each rigid receptor configuration. Thus, a maximum of 1000 binding poses were used in eq 3. For the calculation of B_E in eq 4, poses with intrinsic free energies less favorable by 2.0 kcal mol^{-1} than the most stable pose were discarded, effectively leading to $1 \leq P \leq 20$ (see Table S7).

3. RESULTS

3.1. Parametrization and Performance of LIE Models. The first goal here was to obtain an accurate yet computationally efficient free energy function to estimate the stability of a given binding pose obtained for a small-molecule and a given receptor configuration. This function was called an intrinsic binding free energy (ΔG_{int}).

Several equations based on LIE models previously proposed for implicit[41,47,49] and explicit[29,32,40,86] solvents were tested. A combination of implicit solvent and geometry optimization of

ligand−receptor complexes proved reasonably accurate and computationally fast. As indicated below by comparisons of errors observed between LIE models parametrized here and previously available, the following adapted linear interaction model for continuum electrostatics (ALICE) gave the best results

$$\Delta G_{int} \approx \Delta G_{ALICE} = k_1(2 - \eta - \pi)V^c_{vdW} + k_2(\eta + \pi)V^c_{elet} + k_3\eta(G^c_{GB} - G^p_{GB}) + k_4\pi G^l_{GB} + k_5\Delta SASA^l + k_6 \quad (6)$$

where solvent polarization free energy (G_{GB}), van der Waals (V_{vdW}), and electrostatic (V_{elet}) potentials were calculated for optimized geometries of the complex (c), protein (p), and ligand (l) species. $\Delta SASA^l$ is the difference in SASA between bound and free ligand. The processes of ligand insertion in solution and in the receptor cavity were assumed to be fully decoupled so that the 6 LIE coefficients were independent.

Energy contributions and relative polarity descriptors for ligand (π) and receptor cavity (η) are given in the Supporting Information for all ligands (Table S3 and Table S4). Parameters obtained after optimization of eq 6 against the training set described above were $k_1 = 0.09$, $k_2 = 0.31$, $k_3 = 1.16$, $k_4 = -2.85$, $k_5 = 0.017$ kcal mol^{-1} Å^{-2}, and $k_6 = 3.36$ kcal mol^{-1} (Table S6).

Binding free energies estimated with eq 6 are shown in Figure 1, Table 2, and Table S5. For the 42 small-molecule

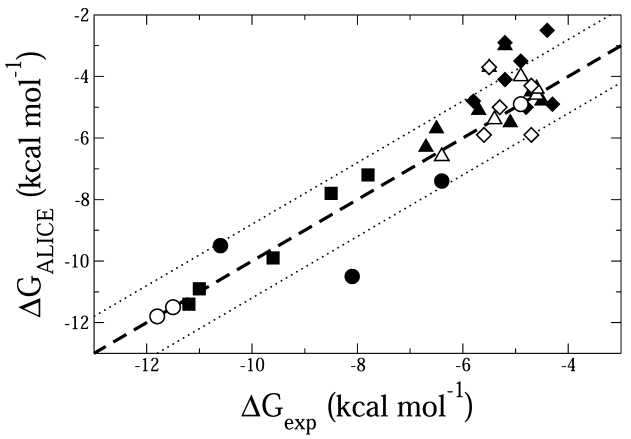


Figure 1. Binding free energies estimated by eq 6. T4 lysozyme L99A ligands are shown as triangles (▲), M102Q ligands are lozenges (◆), HIVRT ligands are circles (●), and FKBP ligands are squares (■). Ligands in the training set are shown as empty symbols, and ligands in the test set are shown as filled symbols. Dashed and dotted lines indicate $y = x \pm 1.2$ kcal mol^{-1}.

complexes in the test set, the RMSD between calculated (ΔG_{ALICE}) and experimental affinities is 1.2 kcal mol^{-1}, the coefficient of determination (R^2)[87] is 0.8, and the maximum error (E_{max}) is 3.0 kcal mol^{-1} (Table S6), which drops to 2.4 kcal mol^{-1} if only binder molecules are considered. All false-positive poses were properly recognized, but many nonbinders were not.

In order to compare the results of eq 6 with the Vina docking energy function,[21] ligands in our test set were docked with Vina to their respective protein crystal structure (native docking) or to a congeneric holo structure if a native one was not available. The binding pose given by the most favorable score in Vina was chosen for comparisons. Error analysis shows that binding free energies calculated with eq 6 resulted in smaller deviations from experiment than those estimated with Vina (RMSD = 1.1 or 1.7 kcal mol^{-1}, E_{max} = 2.2 or 4.3 kcal mol^{-1}, and R^2 = 0.7 or 0.4, respectively). In particular, the performance of our ALICE model is significantly better than Vina if only FKBP ligands are considered (RMSD = 0.6 or 2.6 kcal mol^{-1}, E_{max} = 1.0 or 4.3 kcal mol^{-1}, and R^2 = 0.8 or −2.8, respectively).

It is instructive to describe some of the other LIE models parametrized and tested here. An implicit solvent LIE equation equivalent to the formulation given by Su et al.[49] but with the cavity and van der Waals free energies of solvation condensed to one nonpolar contribution[46] (eq S1 in the SI) resulted in a RMSD of 2.2 kcal mol for the test set (E_{max} = 5.6 kcal mol^{-1}). Affinities estimated with an adaptative version of the same model (eq S2) resulted in a RMSD of 1.6 kcal mol^{-1} (E_{max} = 4.6 kcal mol^{-1}). An adapted LIE eq (eq S3) in which k_6 was scaled by $(1-\eta)$ presented a RMSD of 1.4 kcal mol^{-1} (E_{max} = 2.8 kcal mol^{-1}). The number of outliers found for predictions with this last model was, however, 50% larger than found with eq 6 (Table S6).

Adapted LIE models with the same energetic contributions but with different combinations of the polarity descriptors were also tested, but eq 6 is the most accurate model. A similar result was observed by Linder et al.[40] for adapted LIE models in explicit solvent.

3.2. Averaging Multiple Ligand and Receptor Configurations. The second goal of this study was to test procedures and approximations based on the implicit ligand theory[61] to average the intrinsic free energies calculated for an ensemble of ligand−receptor complexes. Three combinations of ligand pose and receptor configuration averages were tested: In ΔG_{EE}, eq 4 is used to average the ligand poses and to calculate the binding potential of mean-force for each receptor structure, and eq 3 is used to average the receptor configurational distribution. In ΔG_{DE}, eq 5 is used to calculate the binding potential of mean-force, and eq 3 is used to average the receptor distribution. Finally in ΔG_{DD}, eq 5 is again used to calculate the binding potential of mean-force, and a dominant state approximation is used for the receptor distribution (see section 2.5).

Table 3 shows results obtained by the averaging procedures for the full ligand set (previously divided in training and test sets). Error analysis in comparison to experimental affinities is shown in Table 4.

The highest deviations observed in Table 3 are due to the L99A ligands benzene, toluene, and 1,3,5-trimethylbenzene and to the M102Q ligands 2-fluoroaniline, toluene, 3-methylpyrrole, thieno[3.2-b]thiophene, and *N*-(*o*-tolyl)cyanoformamide. All of these also show high ΔG_{ALICE} deviations. In order to isolate contributions of the averaging procedures from inaccuracies in the intrinsic free energy function, all ALICE outliers, i.e., the ligands cited above and catechol, 2-methoxyphenol, 4-vinylpyridine, and HIV3, were removed from the error analysis.

Deviations calculated for this ligand set show slightly smaller RMSDs and determination coefficients closer to one when going from the exponential averages (ΔG_{EE}) to the dominant pose (ΔG_{DE}) and state (ΔG_{DD}) approximations. However, the maximum errors (E_{max}) are higher for ΔG_{DD} due to overstabilization of HIVRT and FKBP ligands.

It is useful to analyze errors for each receptor separately. For T4 lysozyme mutants, the dominant pose and state approximation results in smaller deviations than the exponential averaging procedures. In fact, ΔG_{DD} shows a RMSD smaller than that observed for ΔG_{ALICE} for L99A ligands (Table 2 and Table S6) suggesting that receptor conformational selection

Table 3. Binding Free Energies (in kcal mol^{-1}) Experimentally Measured and Estimated by the Averaging Procedures Described in the Text for the Full Ligand Set

ligand	ΔG_{exp}[a]	ΔG_{EE}	ΔG_{DE}	ΔG_{DD}
L99A				
isobutylbenzene	−6.4	−4.1	−4.4	−5.6
4-ethyltoluene	−5.4	−3.7	−4.1	−4.7
para-xylene	−4.6	−3.1	−3.5	−4.1
indole	−4.9	−2.7	−3.2	−3.8
ortho-xylene	−4.6	−3.1	−3.6	−4.2
n-butylbenzene	−6.7	−4.4	−4.7	−5.7
propylbenzene	−6.5	−3.8	−4.2	−5.1
ethylbenzene	−5.7	−3.1	−3.7	−4.1
toluene	−5.5	−2.5	−3.1	−3.5
benzene	−5.2	−1.7	−2.3	−3.0
3-ethyltoluene	−5.1	−3.7	−4.1	−4.8
meta-xylene	−4.7	−3.1	−3.6	−4.0
2-ethyltoluene	−4.5	−3.7	−4.1	−4.8
3-methylpyrrole	>−2.0	−1.3	−2.0	−2.3
phenol	>−2.0	−1.5	−2.1	−2.6
1,3,5-trimethylbenzene	>−2.0	−3.6	−3.9	−4.9
cyclohexane	>−2.0	−1.9	−2.4	−2.7
2-fluoroaniline	>−2.0	−1.7	−2.3	−2.8
M102Q				
2-fluoroaniline	−5.5	−1.5	−2.1	−2.8
5-chloro-2-methylphenol	−5.3	−2.6	−3.2	−4.0
benzyl acetate	−4.7	−4.4	−4.5	−5.6
ortho-cresol	−4.7	−2.1	−2.6	−3.2
2-propylphenol	−5.6	−3.5	−3.8	−4.7
(phenylamino)acetonitrile	−5.8	−3.2	−3.4	−4.4
toluene	−5.2	−2.3	−2.6	−3.1
3-methylpyrrole	−5.2	−0.9	−1.6	−2.0
thieno[3.2-b]thiophene	−4.9	−1.7	−1.9	−2.7
2-ethylphenol	−4.8	−2.7	−3.2	−3.9
catechol	−4.4	−2.8	−3.0	−4.5
2-ethoxyphenol	−4.3	−2.6	−3.1	−3.9
phenylhydrazine	>−2.0	−1.6	−2.3	−3.2
2-methoxyphenol	>−2.0	−2.1	−2.7	−3.3
4-vinylpyridine	>−2.0	−2.3	−2.7	−3.2
N-(*o*-tolyl)cyanoformamide	>−2.0	−1.8	−2.4	−3.6
HIVRT				
HIV1	−11.5	−10.4	−11.1	−12.2
HIV2	−4.9	−4.6	−5.2	−6.3
HIV3	−8.1	−8.6	−9.2	−10.0
HIV4	−10.6	−8.7	−9.2	−10.6
HIV5	−6.4	−7.5	−8.2	−9.8
HIV6	−11.8	−10.3	−10.5	−12.0
FKBP				
FKB1	−11.0	−11.9	−12.6	−13.8
FKB2	−11.2	−11.7	−12.2	−13.3
FKB3	−7.8	−7.4	−8.2	−9.3
FKB4	−8.5	−7.6	−8.2	−9.4
FKB5	−9.6	−10.7	−11.2	−12.8

[a]Repeated from Table S3. Nonbinders were assumed to have ΔG_{exp} > −2.0 kcal mol^{-1}.

Table 4. Error Analysis of the Binding Free Energies Calculated by the Averaging Procedures Proposed for Different Ligand Sets

	ΔG_{EE}	ΔG_{DE}	ΔG_{DD}
Full Ligand Set (Table 3)[a]			
RMSD	1.7	1.5	1.3
E_{max}	2.9	2.6	3.4
R^2	0.6	0.7	0.8
L99A Ligands[a]			
RMSD	1.7	1.4	0.7
E_{max}	2.7	2.3	1.5
R^2	0.0	0.3	0.8
M102Q Ligands[a]			
RMSD	2.2	1.8	1.3
E_{max}	2.9	2.6	2.1
R^2	−3.0	−1.8	−0.4
HIVRT Ligands			
RMSD	1.2	1.2	1.7
E_{max}	1.9	1.8	3.4
R^2	0.8	0.8	0.6
FKBP Ligands			
RMSD	0.8	1.1	2.3
E_{max}	1.1	1.7	3.2
R^2	0.6	0.3	−1.9

[a]Ligands with a ΔG_{ALICE} deviation from the experimental affinity higher than one RMSD (1.2 kcal mol^{-1}, Table 2) were removed from the error analysis. Deviations (in kcal mol^{-1}, except for R^2) were calculated for each set in comparison to experimental affinities.

contributes to the calculation of binding free energies even for the small and hydrophobic L99A ligands and for the relatively rigid T4 lysozyme engineered cavity. Results for the M102Q mutant show higher deviations which in part may be due to the higher inaccuracies in the ALICE model for this receptor (Table 2).

For HIVRT and specially for FKBP ligands, the opposite trend is observed. The best predictions were obtained for exponential averaging of both pose and receptor configurations (ΔG_{EE}). The dominant approximations overstabilize the binding free energies because some binding poses are as much as 3.5 kcal mol^{-1} more stable than the experimental free energy.

The relative rankings of binding affinities among the HIVRT and the FKBP ligands are recovered with all averaging procedures, except for the two ligands with the most favorable affinities in each receptor. However, experimental and calculated differences between these two ligands are smaller than 0.3 kcal mol^{-1}.

If the Vina docking energy function is used to approximate Ψ in eq 2, the free energies predicted show significative disagreement with experiment for all ligand sets. For example, ΔG_{EE}^{Vina} calculated for the FKBP ligands give RMSD = 3.8 kcal mol^{-1}, E_{max} = 4.6 kcal mol^{-1}, and R^2 = −7.0 suggesting that the docking energy function will give meaningless results if used to approximate the solvent-mediated interaction energy in eq 2. This result can be traced to the Vina inability to discriminate false-positive poses. Almost all poses generated from docking were on average used to calculate B_E^{Vina}. On the other hand, less than half of the generated poses were on average used to calculate B_E^{ALICE} for the same set of ligand−receptor poses (Figure S2 and Table S7).

4. DISCUSSION

4.1. ALICE Model Contributions, Performance, and Limitations. Several LIE equations with different definitions of the nonpolar solvation contribution and with different combinations of polarity descriptors were described and tested here. The best predictions for a test set composed of 42 small-

molecule complexes of 4 different receptors were obtained with eq 6 with 6 adjustable parameters. None of the ligands in the test set were used in the parametrization training set. Deviations observed with this ALICE model and with previously proposed LIE models are similar. For instance, the implicit solvent LIE model proposed by Su et al. has 4 adjustable parameters and resulted in a RMSD of 1.3 kcal mol^{-1} with R^2 = 0.62 for a set of 57 HIVRT ligands (including the 6 HIVRT ligands used here).[49] The implicit solvent LIE model with 2 adjustable parameters proposed by Kolb et al. showed a RMSD of 1.6 kcal mol^{-1} with a correlation coefficient of 0.52 for a set of 128 EGFR kinase ligands.[36] The LIE model proposed by Wall et al. has 3 adjustable parameters and showed a RMSD of 1.6 kcal mol^{-1} and a correlation coefficient of 0.62 for a set of 15 neuraminidase inhibitors.[35] Finally, the adapted LIE model for explicit solvent proposed by Linder et al. has 3 adjustable parameters and resulted in a mean absolute deviation (similar to a RMSD) of 1.6 kcal mol^{-1}, E_{max} = 3.4 kcal mol^{-1}, and R^2 = 0.72 for a diverse set of 38 ligands and their respective 16 receptors.[40] Deviations reported for these four LIE models were obtained with the same ligands (or with a congeneric set of ligands) used for training the models. Still, the ALICE model proposed has the smallest RMSD and the determination coefficient closest to 1.

In order to analyze the energy contributions included in the ALICE model, it should be noted that implicit solvation is not pairwise decomposable in general. Consequently, the splitting of solute−solvent interactions necessary for LIE calculations is not unique. The solvent polarization energy in the bound ligand state is given by $(G^{c}_{GB} - G^{c'}_{GB})$ where the initial state, c', indicates a complex with ligand charges turned off. As discussed by Su et al.,[49] the initial state is approximated here to the free protein (p). With the GBr^6 method,[43] the electrostatic polarization energy calculated for the training set changes by 0.2−0.7 kcal mol^{-1} between these two initial state definitions. This polarization response is scaled by the cavity polarity descriptor in eq 6. A combination with the ligand polarity did not result in better predictions.

Nonpolar solvent contributions from the free ligand and the receptor complex were replaced by a simple $\Delta SASA$ term without loss of accuracy. Previous work suggested that the constant term in a LIE equation, γ in eq 1 or k_6 in eq 6, may be related to binding site hydrophobicity[32,34] or nonpolar surface area.[88] An ALICE model (eq S3 in the SI) in which the constant term is scaled by the nonpolar cavity surface $(1-\eta)$ was tested, but this modification also did not result in better performance.

The comparable accuracy obtained here for different LIE models (eq 6 and eqs S1−S3) suggests the exact form of a LIE equation is less important given a proper parametrization is conducted. Significant departure from theoretical values is observed for the parameters obtained here. For instance, $k_2(\eta+\pi)$, equivalent to the parameter β in previous LIE models (eq 1), ranged from 0.03 to 0.27. This is below the theoretical value of β = 0.5.[29,49] It is expected that parameter values will mutually compensate model assumptions and inaccuracies in the solvent model, molecular mechanical potentials, etc. Thus, the parametrized LIE equations presented here may be cast as linear free energy relationships which coefficients are only bounded by the linear response theory.[30]

Applications of these LIE models depend, however, on their transferability for receptors and ligands not included in the training set used to parametrize the equations. Here coefficients were scaled by ligand and cavity polarities in order to increase model transferability.[40] Eq 6 correctly predicts affinities for ligands which receptors were either included (T4 lysozyme and HIVRT) or not (FKBP) in the training set. The sensitivity of ΔG_{ALICE} on the η and π descriptor values is small. Variations of ∼0.2 kcal mol^{-1} were observed when descriptor values were scaled by ±20%. Predictions for other receptors and ligands should have similar accuracy, but an extensive test of transferability is left for future studies.

The computational efficiency observed for eq 6 suggests it can be used to predict affinities for large ligand sets. For instance, our ALICE model could be used instead of the scoring or energy functions currently employed in molecular docking. To this end, ligand−receptor poses would have to be generated by the conformational search procedures found in docking[21,50] or by another method such as mining-minima.[89] Ligand and protein topologies containing connectivity, force field parameters, and polarity descriptors would have to be available or built. Although cumbersome when manually done, this process can be made fairly automatic.[64,66,90]

In order to improve the ALICE model proposed, it may be useful to analyze the highest deviations found. The following ligands are described as eq 6 outliers since they show deviations larger than one RMSD: benzene, toluene, 1,3,5-trimethylbenzene (L99A), 2-fluoroaniline, catechol, 3-methylpyrrole, thieno[3,2-*b*]thiophene, 4-vinylpiridine, 2-methoxyphenol, *N*-(*o*-tolyl)cyanoformamide (M102Q), and HIV3. All but 2-fluoroaniline belong to the test set, and all T4 lysozyme outliers have underestimated free energies. Most of these ligands are also outliers for eqs S1−S3 (see Table S6).

The binding affinity increases for L99A ligands upon addition of linear methylene units, as seen for benzene to toluene and up to *n*-butylbenzene. The experimental free energy difference upon methylene addition in this series is 0.2−0.3 kcal mol^{-1} but between ethylbenzene and propylbenzene, which is 0.8 kcal mol^{-1}. Although eq 6 incorrectly predicts a small stability to benzene and toluene, appropriate affinities are predicted upon increasing the number of methylene units. This observation suggests a slightly unbalanced description of the nonpolar contributions involving aliphatic and aromatic carbons. An atom-type dependent surface tension, k_5 in eq 6, could amend this problem.

The hydrophobic ligand 1,3,5-trimethylbenzene should interact more favorably with the L99A nonpolar engineered binding site than with water, as suggested by the free energy calculated with the ALICE model. Docking suggests that there is enough room to accommodate this relatively bulky ligand in the L99A cavity (Table S7). However, the experimental free energy shows that 1,3,5-trimethylbenzene is a L99A nonbinder. As T4 lysozyme must show some breathing or opening movement to allow the entrance or exit of ligands from the engineered cavity,[91] we speculate that 1,3,5-trimethylbenzene is a kinetic nonbinder and that there may not be a low energy pathway allowing its entrance into the L99A cavity.

The highest number of outliers were found for M102Q ligands. Possibly electrostatic interactions were not described or sampled correctly in the LIE models tested. Yet no correlation was found between outliers and significant flaws in the description of their dipole moments (Table S1) or the availability of experimental ligand−receptor structures. The lack of holo crystal structures for some ligands and the possibility that receptor structures used for the calculation of ΔG_{ALICE} are not representative of complexes observed in

aqueous solution may contribute for inaccuracies in the prediction of intrinsic binding free energies. This appears to be the case for catechol as discussed in the next section.

4.2. Analysis of the Proposed Conformational Averaging Schemes. Ensembles for all target receptors were initially built with apo protein representations. For the T4 lysozyme mutants, 3 structures out of 50 in the L99A ensemble and 11 structures out of 50 for the M102Q ensemble had their binding cavity fully blocked by side-chains rotations. Docking to these receptor structures did not yield complexes with ligand inside the engineered binding site even for small ligands such as toluene. This problem was more pronounced for the HIVRT and FKBP bulky ligands as none of the receptor structures obtained from apo molecular dynamics after equilibration were able to accommodate ligands in its crystallographic binding site. Thus, molecular dynamics obtained from holo structures were used to generate the HIVRT and FKBP ensembles.

Affinities were consistently underestimated for T4 lysozyme mutants by all three combinations of ensemble averages tested. Although ΔG_{DD} shows a small RMSD and a favorable error analysis for L99A ligands, the dominant approximation that counts only the most favorable pose underestimates affinities for almost all T4 lysozyme ligands but catechol (see below). This tendency suggests that the tentative apo M102Q ensemble used here systematically degrades the structural representation in comparison to the crystal structures. As noted, side-chains rotations block the binding cavity even for the relatively rigid engineered site in T4 lysozyme. Apo protein ensembles have to be used carefully and possibly enlarged or modified[92] to accommodate ligand binding.

The only notable exception is catechol binding to M102Q. Although the ALICE free energy function predicts unstable binding for catechol to the crystal (holo) configuration, favorable binding is predicted by ΔG_{DD}. Thus, receptor configurational sampling is important for the prediction of catechol binding, and the M102Q crystal configuration may have a low contribution to the binding affinity. In fact, it has been shown that catechol has at least two binding modes[93] and that enhanced sampling is required to compute its binding free energy correctly.[18]

Increasing the number of poses that contribute to exponential averaging results in less favorable binding free energies for all ligands tested. This "dilution" effect has two possible causes. One is artificially related to the discrimination of poses that should contribute to the ensemble averages. It depends on the discriminatory quality of the intrinsic free energy function. The second cause has a physical origin as the macroscopically measured affinity may be an average of several binding poses, some of which will have higher intrinsic free energies of binding.

Free energies calculated for the T4 lysozyme mutants decrease up to 50% on going from ΔG_{DD} to ΔG_{EE}. The dilution effect is more pronounced for T4 lysozyme because the intrinsic free energy differences between ligand poses inside or outside the engineered binding site is small (<2.0 kcal mol^{-1}). Consequently, it is harder to discriminate poses that should contribute to the ensemble averages from decoy poses. In fact, for all T4 lysozyme ligands, the average number of poses included in the calculation of the binding potential of mean-force (eq 4) is higher than the average number of poses inside the receptor binding site (Table S7).

The importance of pose discrimination on ensemble averages is also illustrated when the Vina energy function is used as an approximation instead of ΔG_{ALICE} (eq 6). The exponentially averaged free energies obtained in this case are significantly less favorable than the corresponding dominant approximation (ΔG_{DD}^{Vina}) for the same set of ligand−receptor poses.

For HIVRT and FKBP ligands, pose discrimination with the ALICE model is rather accurate. Although the absolute values of binding free energies are higher, the predictions with the ALICE intrinsic free energy function show smaller deviations for both receptors (Table 2). The better discrimination is reflected by the average number of poses used to calculate B_E which is closer to the average number of poses found inside the receptor binding site (Table S7).

Thus, for HIVRT and FKBP ligands, the decrease of calculated free energies on going from ΔG_{DD} to ΔG_{EE} does not appear artifactual but a physical effect due to averaging a distribution of ligand−receptor conformations. Induced fit is expected to change the receptor conformational distribution upon ligand complexation. Indeed, some of the configurations found for the complexes with the receptor ensemble show more favorable intrinsic free energies of binding. This also explains the overstabilization observed for ΔG_{DD} predictions.

Receptor reorganization free energy will be accounted for if the ensemble is canonically distributed and wide enough to represent the relevant receptor motions. If ligand binding has a large reorganization free energy, the associated receptor motions will have high free energy costs, and the number or fraction of binding-competent configurations in the ensemble will be small. Consequently, a ligand complex formed with such rare receptor configurations will only contribute significatively to the final exponential average if it has a highly favorable intrinsic binding free energy. On the other hand, if the reorganization free energy for ligand binding is small, the fraction of binding-competent configurations in the ensemble will be large, and a reasonable number of ligand complexes with these popular receptor configurations will contribute to the average even if their intrinsic binding free energies are not highly favorable.

We do not investigate here what is the appropriate size of the ensemble to account for the reorganization energy correctly, but 50 configurations appear to be insufficient, specially if an apo structure is used to generate the ensemble.[92] When a holo structure is used to generate a tentative ensemble, the full receptor reorganization energy may not be accounted for, but the relative contribution upon a series of congeneric ligands can be retrieved.

Finally, which of the three averaging procedures should be applied? The answer depends on the receptor ensemble and the intrinsic free energy function. For receptor ensembles with unknown or biased distributions, with insufficient sampling and for free energy functions that cannot discriminate correct binding poses from decoys, the dominant pose and state approximation (ΔG_{DD}) should be applied. For more flexible receptors and upon increasing the quality of the conformational ensemble, the state exponential average (ΔG_{DE}) might be included. The exponential average for multiple ligand configurations (ΔG_{EE}) should be employed only when a calibrated and discriminatory intrinsic free energy function is available. This is probably the case for the ALICE model proposed (eq 6) with lead-like ligands.

5. CONCLUSIONS

We have parametrized an adapted linear interaction energy model that employs a simple combination of energy

minimization of ligand-protein geometries with implicit solvation. This model is able to retrospectively predict binding affinities for different receptors with accuracy similar to other LIE models which employ more expensive molecular dynamics simulations to sample configurations and more detailed solvent models.[29−31,40,49] LIE models using geometry optimization and implicit solvation have already been successfully applied by other authors.[36,47]

Conformational sampling is divided and approximated in several steps here. Solvent degrees of freedom, in particular the important dielectric response, are treated implicitly with a continuum electrostatic model. Ligand internal torsion and relative orientation in receptor complexes are evaluated within the docking algorithm when poses are generated. Finally, receptor conformations are sampled from a configurational ensemble, which is generated once and repeatedly used for preparing ligand-protein complexes for a series of congeneric ligands. It should be noted that the proposed ALICE model can be used to predict binding affinities given either a single receptor structure or a conformational distribution.

For the prediction of binding to flexible protein targets, it is useful to represent the receptor structure by conformational ensembles.[6,7,53,56,57] Based on the implicit ligand theory,[61] averaging procedures were proposed and tested to estimate affinities for ligand binding to four different receptors with structures represented by tentative conformational ensembles. The scheme proposed is computationally efficient and could be applied to average contributions of $\sim 10^6$ ligand−receptor poses for each ligand tested.

In principle, to be useful in the prediction of ligand binding, a receptor conformational ensemble should follow the Boltzmann distribution and describe all motions or structural rearrangements relevant for small-molecule complexation. Here, tentative ensembles were built without much attention to statistical distribution or to sampling of relevant motions by running simple molecular dynamics in implicit solvent. Consequently, results obtained by averaging contributions from these receptor ensembles do not aim to reproduce experimental affinities and should be analyzed only qualitatively. It should be noted that there is no consensus on how to obtain conformational ensembles that are suitable to predict ligand binding.[7,53,54,92]

Nevertheless, we find that good discrimination between binding poses and decoys is essential to calculate accurate binding affinities, particularly when contributions from several putative binding poses and different receptor configurations are (exponentially) averaged. Approximations used here such as sampling by docking instead of a statistical distribution or innacuracies in the intrinsic free energy function may contribute to the difficulty in distinguishing binding poses from decoys. For lead-like ligands, we found that the ALICE model proposed is able to discriminate poses resulting in binding free energy predictions in good agreement with experiment.

■ ASSOCIATED CONTENT

Supporting Information

Equations S1−S3, figures with HIVRT and FKBP ligand structures and an example of ligand pose distribution for T4 lysozyme, and tables with dipole moments, number of torsions activated in docking, experimental energies, descriptors and LIE contributions for all ligands considered, training set predictions, error analysis, and number of poses used in averaging procedures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail:garantes@iq.usp.br.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Funding from FAPESP (projects 11/04354-6 and 12/02501-4) and CNPq (project 141950/2013-7) are gratefully acknowledged.

■ REFERENCES

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813−1818.
(2) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein−Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* **2014**, *54*, 944−955.
(3) Fleishman, S.; Baker, D. Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell* **2012**, *149*, 262−273.
(4) Faver, J. C.; Yang, W.; Merz, K. M. The Effects of Computational Modeling Errors on the Estimation of Statistical Mechanical Variables. *J. Chem. Theory Comput.* **2012**, *8*, 3769−3776.
(5) Zuckerman, D. M. Equilibrium Sampling in Biomolecular Simulations. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.
(6) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Bockmann, R. A. Predicting Free Energy Changes Using Structural Ensembles. *Nat. Methods* **2009**, *6*, 3−4.
(7) Arantes, G. M. Flexibility and Inhibitor Binding in Cdc25 Phosphatases. *Proteins* **2010**, *78*, 3017−3032.
(8) Gilson, M. K.; Zhou, H. X. Calculation of Protein−Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21−42.
(9) Guvench, O.; MacKerell, A. D., Jr. Computational Evaluation of Protein−Small Molecule Binding. *Curr. Opin. Struct. Biol.* **2009**, *19*, 56−61.
(10) Deng, Y.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234−2246.
(11) Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Basic Ingredients of Free Energy Calculations: a Review. *J. Comput. Chem.* **2010**, *31*, 1569−1582.
(12) Gallicchio, E.; Levy, R. M. Advances in All Atom Sampling Methods for Modeling Protein−Ligand Binding Affinities. *Curr. Opin. Struct. Biol.* **2011**, *21*, 161−166.
(13) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical Free Energy Methods for Drug Discovery: Progress and Challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150−160.
(14) Rizzo, R. C.; Tirado-Rives, J.; Jorgensen, W. L. Estimation of Binding Affinities for HEPT and Nevirapine Analogues with HIV−1 Reverse Transcriptase Via Monte Carlo Simulations. *J. Med. Chem.* **2001**, *44*, 145−154.
(15) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* **2007**, *371*, 1118−1134.
(16) Jayachandran, G.; Shirts, M. R.; Park, S.; Pande, V. S. Parallelized−Over−Parts Computation of Absolute Binding Free Energy with Docking and Molecular Dynamics. *J. Chem. Phys.* **2006**, *125*, 084901.
(17) Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K. Predicting Ligand Binding Affinity with Alchemical Free Energy Methods in a Polar Model Binding Site. *J. Mol. Biol.* **2009**, *394*, 747−763.

(18) Mobley, D. L.; Chodera, J. D.; Dill, K. A. On the Use of Orientational Restraints and Symmetry Corrections in Alchemical Free Energy Calculations. *J. Chem. Phys.* **2006**, *125*, 084902.
(19) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(20) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein−Ligand Docking: Current Status and Future Challenges. *Proteins* **2006**, *65*, 15−26.
(21) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(22) Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H.; Wang, S.; Carlson, H. A. Small Molecule Inhibitors of the MDM2−P53 Interaction Discovered by Ensemble−Based Receptor Models. *J. Am. Chem. Soc.* **2007**, *129*, 12809−12814.
(23) Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32*, 742−755.
(24) Keenan, S. M.; Geyer, J. A.; Welsh, W. J.; Prigge, S. T.; Waters, N. C. Rational Inhibitor Design and Iterative Screening in the Identification of Selective Plasmodial Cyclin Dependent Kinase Inhibitors. *Comb. Chem. High Throughput Screening* **2005**, *8*, 27−38.
(25) Bisson, W. H.; Cheltsov, A. V.; Bruey-Sedano, N.; Lin, B.; Chen, J.; Goldberger, N.; May, L. T.; Christopoulos, A.; Dalton, J. T.; Sexton, P. M.; Zhang, X.-K.; Abagyan, R. Discovery of Antiandrogen Activity of Nonsteroidal Scaffolds of Marketed Drugs. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 11927−11932.
(26) Frimurer, T. M.; Peters, G. H.; Iversen, L. F.; Andersen, H. S.; Møller, N. P. H.; Olsen, O. H. Ligand−Induced Conformational Changes: Improved Predictions of Ligand Binding Conformations and Affinities. *Biophys. J.* **2003**, *84*, 2273−2281.
(27) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(28) Kim, R.; Skolnick, J. Assessment of Programs for Ligand Binding Affinity Prediction. *J. Comput. Chem.* **2008**, *29*, 1316−1331.
(29) Åqvist, J.; Medina, C.; Sammuelsson, J.-E. A New Method for Predicting Binding Affinity in Computer−Aided Drug Design. *Protein Eng.* **1994**, *7*, 385−391.
(30) Carlson, H. A.; Jorgensen, W. L. An Extended Linear Response Method for Determining Free Energies of Hydration. *J. Phys. Chem.* **1995**, *99*, 10667−10673.
(31) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Enthalpy−Entropy and Cavity Decomposition of Alkane Hydration Free Energies: Numerical Results and Implications for Theories of Hydrophobic Solvation. *J. Phys. Chem. B* **2000**, *104*, 6271−6285.
(32) Åqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand Binding Affinities from MD Simulations. *Acc. Chem. Res.* **2002**, *35*, 358−365.
(33) King, G.; Warshel, A. Investigation of the Free Energy Functions for Electron Transfer Reactions. *J. Chem. Phys.* **1990**, *93*, 8682−8692.
(34) Almlof, M.; Brandsdal, B. O.; Åqvist, J. Binding Affinity Prediction with Different Force Fields: Examination of the Linear Interaction Energy Method. *J. Comput. Chem.* **2004**, *25*, 1242−1254.
(35) Wall, I. D.; Leach, A. R.; Salt, D. W.; Ford, M. G.; Essex, J. W. Binding Constants of Neuraminidase Inhibitors: an Investigation of the Linear Interaction Energy Method. *J. Med. Chem.* **1999**, *42*, 5142−5152.
(36) Kolb, P.; Huang, D.; Dey, F.; Caflisch, A. Discovery of Kinase Inhibitors by High−Throughput Docking and Scoring Based on a Transferable Linear Interaction Energy Model. *J. Med. Chem.* **2008**, *51*, 1179−1188.
(37) Stjernschantz, E.; Marelius, J.; Medina, C.; Jacobsson, M.; Vermeulen, N. P. E.; Oostenbrink, C. Are Automated Molecular Dynamics Simulations and Binding Free Energy Calculations Realistic Tools in Lead Optimization? An Evaluation of the Linear Interaction Energy (LIE) Method. *J. Chem. Inf. Model.* **2006**, *46*, 1972−1983.
(38) de Amorim, H. L. N.; Caceres, R. A.; Netz, P. A. Linear Interaction Energy (LIE) Method in Lead Discovery and Optimization. *Curr. Drug Targets* **2008**, *9*, 1100−1105.
(39) Hansson, T.; Marelius, J.; Åqvist, J. Ligand Binding Affinity Prediction by Linear Interaction Energy Methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27−35.
(40) Linder, M.; Ranganathan, A.; Brinck, T. Adapted Linear Interaction Energy": a Structure−Based LIE Parametrization for Fast Prediction of Protein−Ligand Affinities. *J. Chem. Theory Comput.* **2013**, *9*, 1230−1239.
(41) Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. New Linear Interaction Method for Binding Affinity Calculations Using a Continuum Solvent Model. *J. Phys. Chem. B* **2001**, *105*, 10388−10397.
(42) Alves, A. F. N. M.Sc. thesis, Instituto de Química, Universidade de São Paulo, São Paulo, 2013. Available at http://www.teses.usp.br/teses/disponiveis/46/46131/tde-08052013-144801/ (accessed June 20, 2014).
(43) Tjong, H.; Zhou, H.-X. GBr^6: A Parameterization−Free, Accurate, Analytical Generalized Born Method. *J. Phys. Chem. B* **2007**, *111*, 3055−3061.
(44) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(45) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large−Scale Conformational Changes with a Modified Generalized Born Model. *Proteins* **2004**, *55*, 383−394.
(46) Schaefer, M.; Bartels, C.; Karplus, M. Solution Conformations and Thermodynamics of Structured Peptides: Molecular Dynamics Simulation with an Implicit Solvation Model. *J. Mol. Biol.* **1998**, *284*, 835−848.
(47) Huang, D.; Caflisch, A. Efficient Evaluation of Binding Free Energy Using Continuum Electrostatic Solvation. *J. Med. Chem.* **2004**, *47*, 5791−5797.
(48) Carlsson, J.; Ander, M.; Nervall, M.; Åqvist, J. Continuum Solvation Models in the Linear Interaction Energy Method. *J. Phys. Chem. B* **2006**, *110*, 12034−12041.
(49) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. Linear Interaction Energy (LIE) Models for Ligand Binding in Implicit Solvent: Theory and Application to the Binding of NNRTIs to HIV−1 Reverse Transcriptase. *J. Chem. Theory Comput.* **2007**, *3*, 256−277.
(50) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge−Based Desolvation. *J. Comput. Chem.* **2007**, *28*, 1145−1152.
(51) Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.
(52) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in Molecular Recognition: the Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 45−55.
(53) Wong, C. F.; Kua, J.; Zhang, Y.; Straatsma, T.; McCammon, J. A. Molecular Docking of Balanol to Dynamics Snapshots of Protein Kinase A. *Proteins* **2005**, *61*, 850−858.
(54) Mamonov, A. B.; Bhatt, D.; Cashman, D. J.; Ding, Y.; Zuckerman, D. M. General Library−Based Monte Carlo Technique Enables Equilibrium Sampling of Semi−Atomistic Protein Models. *J. Phys. Chem. B* **2009**, *113*, 10891−10904.
(55) Kallblad, P.; Mancera, R. L.; Todorov, N. P. Assessment of Multiple Binding Modes in Ligand−Protein Docking. *J. Med. Chem.* **2004**, *47*, 3334−3337.
(56) Novoa, E. M.; de Pouplana, L. R.; Barril, X.; Orozco, M. Ensemble Docking from Homology Models. *J. Chem. Theory Comput.* **2010**, *6*, 2547−2557.
(57) Totrov, M.; Abagyan, R. Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178−184.

(58) B-Rao, C.; Subramanian, J.; Sharma, S. D. Managing Protein Flexibility in Docking and Its Applications. *Drug Discovery Today* **2009**, *14*, 394−400.
(59) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target Flexibility: an Emerging Consideration in Drug Discovery and Design. *J. Med. Chem.* **2008**, *51*, 6237−6255.
(60) Gallicchio, E.; Lapelosa, M.; Levy, R. M. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein−Ligand Binding Affinities. *J. Chem. Theory Comput.* **2010**, *6*, 2961−2977.
(61) Minh, D. D. L. Implicit Ligand Theory: Rigorous Binding Free Energies and Thermodynamic Expectations from Molecular Docking. *J. Chem. Phys.* **2012**, *137*, 104106.
(62) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. A Cavity−Containing Mutant of T4 Lysozyme Is Stabilized by Buried Benzene. *Nature* **1992**, *355*, 371−373.
(63) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A Model Binding Site for Testing Scoring Functions in Molecular Docking. *J. Mol. Biol.* **2002**, *322*, 339−355.
(64) Voss, N. R.; Gerstein, M.; Steitz, T. A.; Moore, P. B. The Geometry of the Ribosomal Polypeptide Exit Tunnel. *J. Mol. Biol.* **2006**, *360*, 893−906.
(65) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area As a Sum of Fragment−Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(66) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein−Ligand Interactions. *J. Mol. Biol.* **2003**, *326*, 607−620.
(67) Carroll, D. L. *Genetic Algorithm Driver, Version 1.7*; 1998.
(68) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN 77: the Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992.
(69) Arantes, G. M.; Loos, M. Specific Parametrisation of a Hybrid Potential to Simulate Reactions in Phosphatases. *Phys. Chem. Chem. Phys.* **2006**, *8*, 347−353.
(70) Chinea, G.; Padron, G.; Hooft, R. W. W.; Sander, C.; Vriend, G. The Use of Position−Specific Rotamers in Model−Building by Homology. *Proteins* **1995**, *23*, 415−421.
(71) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a High−Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845−854.
(72) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an Open Chemical Toolbox. *J. Cheminf.* **2011**, DOI: doi:10.1186/1758-2946-3-33.
(73) Frisch, M. J. et al. *Gaussian 09, Revision A.1*; Gaussian, Inc.: Wallingford, CT, 2009.
(74) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(75) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the Development of Universal, Fast and Highly Accurate Docking/scoring Methods: a Long Way to Go. *Br. J. Pharmacol.* **2008**, *153*, S7−S26.
(76) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: a Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(77) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All−Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(78) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(79) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(80) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: a New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(81) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: an N×log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(82) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(83) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(84) Li, J.; Xing, J.; Cramer, C. J.; Truhlar, D. G. Accurate Dipole Moments from Hartree−Fock Calculations by Means of Class IV Charges. *J. Chem. Phys.* **1999**, *111*, 885−892.
(85) *Ligand Topologies*. http://gaznevada.iq.usp.br/wp-content/uploads/opls_site.tar.gz (accessed May 2014).
(86) Jones-Hertzog, D. K.; Jorgensen, W. L. Binding Affinities for Sulfonamide Inhibitors with Human Thrombin Using Monte Carlo Simulations with a Linear Response Method. *J. Med. Chem.* **1997**, *40*, 1539−1549.
(87) Zar, J. H. *Biostatistical Analysis*, 4th ed.; Prentice Hall: NJ, 1999.
(88) Reynolds, J. A.; Gilbert, D. B.; Tanford, C. Empirical Correlation Between Hydrophobic Free Energy and Aqueous Cavity Surface Area. *Proc. Natl. Acad. Sci. U. S. A.* **1974**, *71*, 2925−2927.
(89) Chen, W.; Gilson, M. K.; Webb, S. P.; Potter, M. J. Modeling Protein−Ligand Binding by Mining Minima. *J. Chem. Theory Comput.* **2010**, *6*, 3540−3557.
(90) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr. CHARMM General Force Field: a Force Field for Drug−like Molecules Compatible with the CHARMM All−Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671−690.
(91) Mulder, F. A. A.; Hon, B.; Muhandiram, D. R.; Dahlquist, F. W.; Kay, L. E. Flexibility and Ligand Exchange in a Buried Cavity Mutant of T4 Lysozyme Studied by Multinuclear NMR. *Biochemistry* **2000**, *39*, 12614−12622.
(92) Seeliger, D.; de Groot, B. L. Conformational Transitions upon Ligand Binding: Holo−Structure Prediction from Apo Conformations. *PLoS Comput. Biol.* **2010**, *6*, e1000634.
(93) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for Docking. *J. Med. Chem.* **2005**, *48*, 3714−3728.

Ligand-receptor affinities computed by an adapted linear interaction model for continuum electrostatics and by protein conformational averaging

Ariane Nunes-Alves and Guilherme Menegon Arantes

Department of Biochemistry, Instituto de Química, Universidade de São Paulo,

Av. Prof. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil

E-mail: garantes@iq.usp.br

Supporting Information

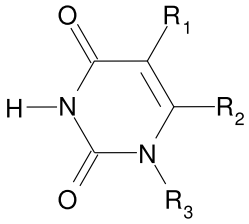
Along with eq. 6 in the main text, the following LIE equations were parametrized and tested:

$$\Delta G_{int} \approx k_1 V^c_{vdW} + k_2 V^c_{elet} + k_3 (G^c_{GB} - G^p_{GB}) + k_4 G^l_{GB} + k_7 (G^c_{NP} - G^p_{NP}) + k_8 G^l_{NP} + k_6 \quad \text{(S1)}$$

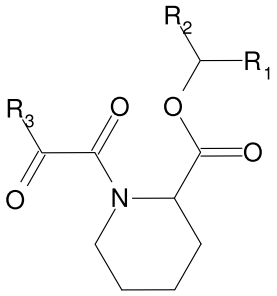
$$\approx k_1 (2 - \eta - \pi) V^c_{vdW} + k_2 (\eta + \pi) V^c_{elet} + k_3 (G^c_{GB} - G^p_{GB}) + k_4 \pi G^l_{GB} + k_7 (G^c_{NP} - G^p_{NP}) + k_8 \pi G^l_{NP} + k_6 \quad \text{(S2)}$$

$$\approx k_1 (2 - \eta - \pi) V^c_{vdW} + k_2 (\eta + \pi) V^c_{elet} + k_3 \eta (G^c_{GB} - G^p_{GB}) + k_4 \pi G^l_{GB} + k_5 \Delta SASA^l + k_6 (1 - \eta) \quad \text{(S3)}$$

where G_{NP} is the implicit solvent nonpolar free energy[46] for the complex (c), protein (p) and ligand (l) species, and the remaining terms were defined in eq. 6. Coefficients are given in Table S6.



	R_1	R_2	R_3
HIV1:	i-Pr	CH_2Ph	$CH_2OCH_2CH_3$
HIV2:	Me	SPh	H
HIV3:	Me	SPh	Bu
HIV4:	Et	SPh	$CH_2OCH_2CH_3$
HIV5:	Me	CH_2Ph	$CH_2OCH_2CH_2OH$
HIV6:	i-Pr	SPh-3,5-di-Me	$CH_2OCH_2CH_2OH$



	R_1	R_2	R_3
FKB1:	Ph	CH_2CH_2Ph	tert-Pentyl
FKB2:	Cy	CH_2CH_2Ph	tert-Pentyl
FKB3:	H	Me	Cy
FKB4:	H	Me	tert-Pentyl
FKB5:	H	CH_2CH_2Ph	tert-Pentyl

Figure S1: Structures of HIVRT and FKBP ligands employed in the data set. Please, note that an extra methylene unit was misplaced on FKB3 and FKB4 structures given in a previous reference.[94]

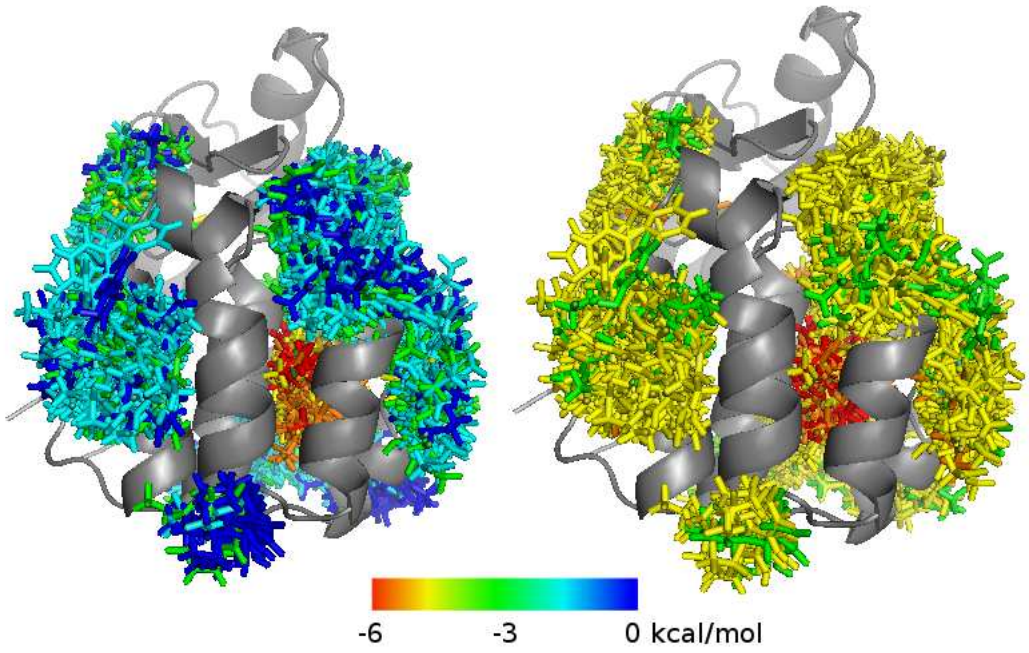


Figure S2: Superposition of poses for 2-propylphenol bound to M102Q obtained by docking to the tentative configurational ensemble. Only one protein structure is shown in gray cartoon. The ligand is shown in sticks and is colored according to ALICE (left) and Vina (right) intrinsic binding free energy. The crystalographic binding site is populated with ligands shown in red and orange.

Table S1: Dipole moments and respective components (in Debye) for selected ligands.

ligand	dipole - FF[a]				dipole - QM[b]			
	total	x	y	z	total	x	y	z
M102Q								
2-fluoroaniline	2.0	1.3	0.9	1.2	1.9	1.4	0.6	1.1
catechol	2.1	-1.7	-1.2	0.0	2.1	-1.7	-1.3	0.0
benzyl acetate	4.3	1.3	-4.1	0.3	4.8	1.4	-4.6	0.3
2-ethoxyphenol	2.0	1.0	1.4	1.1	2.5	1.1	1.9	1.1
phenylhydrazine	0.9	0.8	-0.4	0.1	1.4	1.3	-0.6	0.2
2-methoxyphenol	2.1	1.4	1.0	-1.2	2.4	2.0	0.8	-1.2
3-methylpyrrole	1.4	-1.3	-0.1	-0.5	1.7	-1.4	-0.1	-0.9
4-vinylpyridine	2.7	-2.6	-0.6	-0.1	2.6	-2.6	-0.4	-0.1
N-(o-tolyl) cyanoformamide	4.9	-3.2	2.8	2.4	5.1	-3.4	2.9	2.5
HIVRT								
HIV1	2.6	-1.6	0.0	2.1	4.2	-2.9	-0.1	3.0
HIV2	4.1	-3.3	-1.3	2.1	6.9	-4.8	-2.7	4.1
HIV3	5.1	-5.1	0.9	0.3	5.4	-5.3	1.2	0.5
HIV4	4.0	4.0	0.0	0.1	4.2	4.2	-0.4	0.0
HIV5	4.5	2.9	-1.1	3.3	5.2	3.9	-1.5	3.1
HIV6	3.3	-3.0	0.4	1.4	4.6	-4.4	1.0	1.0
FKBP								
FKB1	3.9	-1.3	0.7	3.5	4.6	-0.7	0.8	4.5
FKB2	4.5	0.6	2.3	3.8	4.6	0.7	2.9	3.4
FKB3	4.2	2.4	0.6	-3.4	4.8	2.2	0.7	-4.2
FKB4	4.5	3.4	-2.0	2.3	4.3	3.5	-2.0	1.5
FKB5	4.3	0.4	-0.4	-4.3	4.1	1.6	-0.6	-3.7

[a] Force field parametrized here. [b] Quantum mechanical reference (HF/6-31G*).

Table S2: Number of torsions activated in docking experiments for all ligands tested.

L99A and M102Q ligands			
isobutylbenzene	4	cyclohexane	0
4-ethyltoluene	3	2-fluoroaniline	1
benzene	0	5-chloro-2-methylphenol	2
indole	0	benzyl acetate	4
ortho-xylene	2	ortho-cresol	2
n-butylbenzene	4	catechol	2
propylbenzene	3	(phenylamino)acetonitrile	2
ethylbenzene	2	2-propylphenol	4
toluene	1	toluene	1
3-ethyltoluene	3	thieno[3,2-b]thiophene	0
meta-xylene	2	2-ethylphenol	3
para-xylene	2	2-ethoxyphenol	4
2-ethyltoluene	3	phenylhydrazine	2
3-methylpyrrole	1	2-methoxyphenol	3
phenol	1	4-vinylpyridine	1
1,3,5-trimethylbenzene	3	N-(o-tolyl) cyanoformamide	2
FKBP ligands		HIVRT ligands	
FKB1	13	HIV1	9
FKB2	13	HIV2	3
FKB3	6	HIV3	7
FKB4	10	HIV4	8
FKB5	12	HIV5	8
		HIV6	12

Table S3: PDB code for holo structures, descriptors (η and π), surface area variations (in Å^2), experimental binding free energies and energetic contributions (both in kcal mol^{-1}) used for the LIE models.

ligand	PDB code[a]	ΔG^b_{exp}	η	π	$\Delta SASA^l$	V^c_{vdW}	V^c_{elet}	$G^c_{GB} - G^p_{GB}$	G^l_{GB}
				L99A					
isobutylbenzene	184L[95]	-6.4[96]	0.13	0.00	-287	-31.0	-1.3	1.0	-0.4
benzene	1L83[62]	-5.2[96]	0.13	0.00	-198	-17.7	-2.0	0.5	-0.5
indole	185L[95]	-4.9[96]	0.14	0.10	-228	-22.6	-5.9	0.6	-1.5
ortho-xylene	188L[95]	-4.6[96]	0.13	0.00	-215	-24.7	-2.3	0.9	-0.7
4-ethyltoluene	-	-5.4[96]	0.15	0.00	-271	-25.7	-2.1	1.0	-0.6
ethylbenzene	1NHB[95]	-5.7[96]	0.13	0.00	-251	-25.1	-1.8	0.8	-0.5
para-xylene	187L[95]	-4.6[96]	0.11	0.00	-228	-24.5	-1.4	0.6	-0.6
n-butylbenzene	186L[95]	-6.7[96]	0.12	0.00	-274	-30.4	-1.9	1.2	-0.5
meta-xylene	-	-4.7[96]	0.17	0.00	-245	-22.5	-2.7	0.9	-0.7
propylbenzene	-	-6.5[96]	0.13	0.00	-271	-26.9	-1.6	0.9	-0.5
2-ethyltoluene	-	-4.5[96]	0.17	0.00	-252	-24.0	-1.5	0.8	-0.6
3-ethyltoluene	-	-5.1[96]	0.14	0.00	-269	-25.8	-2.4	1.1	-0.6
toluene	-	-5.5[96]	0.18	0.00	-231	-19.3	-1.3	0.4	-0.6
phenol	-	>-2.0[96]	0.14	0.16	-219	-18.8	-6.8	0.2	-1.6
cyclohexane	-	>-2.0[96]	0.18	0.00	-183	-18.6	-0.2	0.8	0.1
1,3,5-trimethylbenzene	-	>-2.0[96]	0.17	0.00	-222	-24.6	-2.7	1.0	-0.7
2-fluoroaniline	-	>-2.0[63]	0.22	0.19	-229	-18.7	-3.8	0.2	-1.5
3-methylpyrrole	-	>-2.0[63]	0.17	0.12	-198	-18.5	-2.2	0.3	-1.3
				M102Q					
ortho-cresol	3HT6[17]	-4.7[17]	0.21	0.13	-230	-23.9	-7.4	0.6	-1.2
2-fluoroaniline	1LGW[63]	-5.5[63]	0.20	0.19	-230	-20.5	-10.4	1.2	-1.5
catechol	1XEP[93]	-4.4[17]	0.22	0.30	-215	-22.9	-2.6	-0.5	-1.6
5-chloro-2--methylphenol	3HT8[17]	-5.3[17]	0.22	0.12	-268	-24.1	-7.1	0.0	-1.6
benzyl acetate	3HUK[17]	-4.7[17]	0.19	0.14	-278	-32.3	-2.9	0.0	-1.6
(phenylamino)-acetonitrile	2RBN[97]	-5.8[17]	0.18	0.20	-264	-26.5	-12.0	1.1	-2.4
thieno[3,2-b]thiophene	3HUQ[17]	-4.9[17]	0.23	0.38	-244	-24.6	-2.9	0.6	-0.7
2-ethoxyphenol	3HU8[17]	-4.3[17]	0.19	0.16	-265	-29.4	-2.8	0.0	-2.0
2-propylphenol	3HTB[17]	-5.6[17]	0.18	0.11	-279	-30.1	-4.6	0.2	-1.5
2-ethylphenol	3HT7[17]	-4.8[17]	0.19	0.12	-255	-26.6	-5.4	0.1	-1.5
3-methylpyrrole	-	-5.2[63]	0.19	0.12	-198	-18.6	-7.4	1.0	-1.3
toluene	-	-5.2[63]	0.22	0.00	-234	-22.3	-1.8	0.7	-0.6
2-methoxyphenol	3HT9[17]	>-2.0[17]	0.16	0.18	-251	-26.2	-3.9	0.1	-1.4

ligand	PDB code[a]	ΔG^{b}_{exp}	η	π	$\Delta SASA^{l}$	V^{c}_{vdW}	V^{c}_{elet}	$G^{c}_{GB} - G^{p}_{GB}$	G^{l}_{GB}
4-vinylpyridine	-	>-2.0[17]	0.22	0.08	-239	-24.1	1.0	-0.6	-1.2
phenylhydrazine	-	>-2.0[17]	0.22	0.25	-213	-22.1	-9.5	1.1	-2.0
N-(o-tolyl) cyanoformamide	-	>-2.0[98]	0.21	0.27	-265	-30.9	-10.6	1.1	-2.2
				HIVRT					
HIV1	1RT1[99]	-11.5[100]	0.24	0.18	-425	-59.7	-11.1	-3.1	-6.1
HIV2	-	-4.9[101]	0.24	0.34	-319	-44.9	-6.9	-2.5	-4.9
HIV3	-	-8.1[101]	0.24	0.24	-442	-54.9	-15.3	-2.0	-5.7
HIV4	-	-10.6[101]	0.24	0.26	-404	-56.9	-13.7	-2.7	-6.1
HIV5	-	-6.4[102]	0.23	0.27	-397	-50.6	-18.7	-4.4	-8.8
HIV6	-	-11.8[103]	0.23	0.28	-473	-60.9	-30.5	-1.8	-7.9
				FKBP					
FKB1	1FKG[104]	-11.0[104]	0.39	0.14	-389	-49.7	-15.9	0.2	-3.5
FKB2	1FKH[104]	-11.2[104]	0.35	0.14	-399	-51.6	-21.5	2.1	-3.5
FKB3	-	-7.8[104]	0.36	0.21	-317	-38.7	-13.2	0.6	-3.0
FKB4	-	-8.5[104]	0.36	0.21	-328	-39.1	-14.9	0.6	-3.0
FKB5	-	-9.6[104]	0.36	0.17	-376	-50.4	-12.7	0.4	-3.2

[a] If PDB code is not given, a holo structure was not available and the ligand-receptor configuration was obtained from docking to the receptor structure in PDB 1NHB (L99A), 3HT6 (M102Q), 1RT1 (HIVRT) or 1FKG (FKBP). [b] Superscripts give the original references from where experimental affinities were retrieved. Non-binders are shown in bold and were assumed to have $\Delta G_{exp} > -2.0$ kcal mol^{-1}.

Table S4: Adaptative descriptor (η), surface area variations (in Å^2) and energetic contributions (in kcal mol^{-1}) for false-positive poses.

ligand[a]	η	$\Delta SASA^l$	V^c_{vdW}	V^c_{elet}	$(G^c_{GB}$ - $G^p_{GB})$	G^l_{GB}
			L99A			
indole	0.50	-158	-18.0	-3.6	0.4	-1.5
isobutylbenzene	0.46	-129	-14.4	-7.4	1.6	-0.4
benzene	0.47	-149	-13.4	0.2	0.3	-0.5
4-ethyltoluene (A)	0.40	-175	-17.8	1.5	-0.3	-0.6
4-ethyltoluene (B)	0.56	-195	-17.5	-2.7	0.8	-0.6
ortho-xylene (A)	0.51	-149	-15.9	1.4	0.2	-0.7
ortho-xylene (B)	0.51	-66	-11.2	-2.5	0.6	-0.7
propylbenzene	0.52	-171	-17.7	0.7	0.6	-0.5
toluene	0.48	-185	-16.4	2.3	0.0	-0.6
para-xylene	0.55	-96	-12.5	-2.7	0.5	-0.6
			M102Q			
catechol	0.47	-147	-11.4	-15.9	0.6	-1.6
2-fluoroaniline (A)	0.53	-168	-12.3	-9.6	0.7	-1.5
2-fluoroaniline (B)	0.43	-118	-8.1	-10.6	0.4	-1.5
ortho-cresol	0.48	-136	-9.6	-17.6	1.9	-1.2
5-chloro-2-methylphenol	0.48	-123	-12.0	-6.0	-0.2	-1.6
thieno[3,2-b]thiophene	0.51	-138	-13.3	-3.7	0.4	-0.7
(phenylamino)acetonitrile	0.30	-113	-12.5	-12.1	0.9	-2.4
2-propylphenol (A)	0.33	-157	-15.2	-7.3	-0.4	-1.5
2-propylphenol (B)	0.57	-146	-17.1	-2.7	-0.7	-1.5
2-ethoxyphenol	0.63	-172	-19.4	4.4	-1.4	-2.0

[a] A and B represent different poses for the same ligand.

Table S5: Binding free energies (in kcal mol^{-1}) experimentally measured and estimated by eq. 6 for ligands in the *training set*.

ligand	ΔG^{a}_{exp}	ΔG_{ALICE}	ligand	ΔG^{a}_{exp}	ΔG_{ALICE}
L99A			M102Q		
isobutylbenzene[b]	-6.4	-6.6	2-fluoroaniline[b]	-5.5	-3.7
4-ethyltoluene	-5.4	-5.4	5-chloro-2-methylphenol[b]	-5.3	-5.0
para-xylene[b]	-4.6	-4.6	benzyl acetate[b]	-4.7	-5.9
indole[b]	-4.9	-4.0	ortho-cresol[b]	-4.7	-4.3
ortho-xylene[b]	-4.6	-4.4	2-propylphenol[b]	-5.6	-5.9
indole	>-2.0	-1.6	2-propylphenol (B)	>-2.0	-2.1
isobutylbenzene	>-2.0	-1.0	2-fluoroaniline (A)	>-2.0	-1.8
para-xylene	>-2.0	0.0	2-fluoroaniline (B)	>-2.0	-0.7
4-ethyltoluene	>-2.0	-2.1	ortho-cresol	>-2.0	-1.9
ortho-xylene (B)	>-2.0	-1.0	5-chloro-2-methylphenol	>-2.0	-0.9
HIVRT					
HIV1[b]	-11.5	-11.5			
HIV2	-4.9	-4.9			
HIV6	-11.8	-11.8			

[a]repeated from Table S3. [b]Complexes with holo crystal structure available. False-positive poses are shown with underline. A and B represent different poses for the same ligand.

Table S6: Optimized coefficients and error analysis for the LIE models considered here.

LIE model	eq. 6	eq. S1	eq. S2	eq. S3
Coefficients				
k_1	0.09	0.30	0.13	0.09
k_2	0.31	0.01	0.22	0.15
k_3	1.16	-1.70	-0.12	2.02
k_4	-2.85	0.11	-4.39	-2.30
k_5^a	0.017			0.012
k_6^b	3.36	-4.04	2.61	2.22
k_7		5.23	5.49	
k_8		3.30	-3.55	
Deviations for training set complexes				
RMSD	0.5	0.8	0.7	0.8
E_{max}	1.8	2.8	1.9	2.1
R^2	1.0	0.9	0.9	0.9
Deviations for test set complexes				
RMSD	1.2	2.0	1.6	1.4
E_{max}	3.0	5.6	4.6	2.8
R^2	0.8	0.5	0.7	0.8
Outliers in the test set[c]				
binders	6	8	6	9
non-binders	4	5	5	6

[a] kcal mol^{-1} Å^{-2}. [b] kcal mol^{-1}. All other coefficients are dimensionless. [c] Outlier ligands show deviations larger than one RMSD. Deviations (in kcal mol^{-1}, except for R^2) from experimental affinities were calculated for the 42 complexes in our test set (see Table 2).

Table S7: Average ($\pm$ standard deviation) number of poses found inside the crystalographic binding site and used to calculate B_E (P in eq. 4) with the ALICE and Vina intrinsic free energy functions.

ligand	binding site	B_E^{ALICE}	B_E^{Vina}	ligand	binding site	B_E^{ALICE}	B_E^{Vina}
L99A				M102Q			
isobutylbenzene	2±2	10±7	19±4	2-fluoroaniline	4±3	13±4	17±4
4-ethyltoluene	2±2	9±7	18±5	5-chloro-2--methylphenol	2±2	12±6	19±3
benzene	3±2	9±5	15±6	benzyl acetate	1±1	10±7	19±3
indole	3±2	9±6	17±5	ortho-cresol	4±3	11±5	17±5
ortho-xylene	5±3	9±5	18±4	catechol	3±3	9±5	19±2
n-butylbenzene	2±2	9±7	17±5	(phenylamino)-acetonitrile	1±1	7±6	19±4
propylbenzene	2±2	7±6	17±5	2-propylphenol	2±2	10±7	19±4
ethylbenzene	3±2	6±5	15±5	toluene	3±2	9±6	14±6
toluene	4±2	7±5	15±6	3-methylpyrrole	2±2	12±5	19±2
3-ethyltoluene	3±3	9±7	18±5	thieno[3,2-b]thiophene	1±2	13±8	20±0
meta-xylene	4±3	10±6	18±3	2-ethylphenol	3±3	11±6	18±4
para-xylene	3±2	9±6	18±4	2-ethoxyphenol	1±2	12±6	20±1
2-ethyltoluene	3±3	9±6	18±5	**phenylhydrazine**	2±2	10±5	18±4
3-methylpyrrole	4±2	11±3	20±0	**2-methoxyphenol**	3±3	11±5	19±2
phenol	4±2	13±3	18±4	**4-vinylpyridine**	2±2	10±7	19±2
1,3,5-trimethylbenzene	3±3	13±6	20±2	**N-(o-tolyl) cyanoformamide**	0±1	13±5	20±0
cyclohexane	6±2	10±5	18±4				
2-fluoroaniline	5±3	14±3	19±3	HIVRT			
FKBP				HIV1	9±3	6±2	8±4
FKB1	5±3	9±4	20±1	HIV2	12±3	12±3	13±4
FKB2	5±4	9±4	20±1	HIV3	9±4	8±3	8±4
FKB3	8±3	10±3	19±2	HIV4	9±4	7±3	8±4
FKB4	9±4	10±4	19±2	HIV5	11±4	8±3	10±5
FKB5	7±4	8±4	20±1	HIV6	4±3	4±2	15±6

Non-binders are shown in bold. Averages and standard deviations were calculated for $N = 50$ receptor structures. To determine the poses inside the crystalographic binding site, a histogram of minimum distances between ligand and binding site was built for the whole ensemble. The histogram showed a bimodal distribution for all ligands and the distance separating the two modes was taken as a cutoff distance to count ligands bound to the crystalographic site.

SI References

(94) Lapelosa, M.; Gallicchio, E.; Levy, R. M. Conformational Transitions and Convergence of Absolute Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 47–60

(95) Morton, A.; Matthews, B. W. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry* **1995**, *34*, 8576–8588

(96) Morton, A.; Baase, W. A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry* **1995**, *34*, 8564–8575

(97) Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring Docking Hit Lists for Model Cavity Sites: Predictions and Experimental Testing. *J. Mol. Biol.* **2008**, *377*, 914–934

(98) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a Flexible–Receptor Docking Algorithm in a Model Binding Site. *J. Mol. Biol.* **2004**, *337*, 1161–1182

(99) Hopkins, A. L.; Ren, J.; Esnouf, R. M.; Willcox, B. E.; Jones, E. Y.; Ross, C.; Miyasaka, T.; Walker, R. T.; Tanaka, H.; Stammers, D. K.; Stuart, D. I. Complexes of HIV–1 Reverse Transcriptase with Inhibitors of the HEPT Series Reveal Conformational Changes Relevant to the Design of Potent Non–Nucleoside Inhibitors. *J. Med. Chem.* **1996**, *39*, 1589–1600

(100) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Inouye, N.; Baba, M.; Shigeta, S.; Walker, R. T.; Clercq, E. D.; Miyasaka, T. Synthesis and Antiviral Activity of 6–Benzyl Analogs of 1–[(2–Hydroxyethoxy)methyl]–6–(phenylthio)thymine (HEPT) As Potent and Selective Anti–HIV–1 Agents. *J. Med. Chem.* **1995**, *38*, 2860–2865

(101) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R.; De Clercq, E.; Miyasaka, T. Synthesis and Antiviral Activity of Deoxy Analogs

of 1–[(2–Hydroxyethoxy)methyl]–6–(phenylthio)thymine (HEPT) As Potent and Selective Anti–HIV–1 Agents. *J. Med. Chem.* **1992**, *35*, 4713–4719

(102) Tanaka, H.; Baba, M.; Hayakawa, H.; Sakamaki, T.; Miyasaka, T.; Ubasawa, M.; Takashima, H.; Sekiya, K.; Nitta, I.; Shigeta, S.; Walker, R. T.; Balzarini, J.; De Clercq, E. A New Class of HIV–1 Specific 6–Substituted Acyclouridine Derivatives: Synthesis and Anti–HIV–1 Activity of 5– or 6–Substituted Analogs of 1–[(2–Hydroxyethoxy)methyl]–6–(phenylthio)thymine (HEPT). *J. Med. Chem.* **1991**, *34*, 349–357

(103) Tanaka, H.; Takashima, H.; Ubasawa, M.; Sekiya, K.; Nitta, I.; Baba, M.; Shigeta, S.; Walker, R. T.; De Clercq, E.; Miyasaka, T. Structure–Activity Relationships of 1–[(2–Hydroxyethoxy)methyl]–6–(phenylthio)thymine Analogs: Effect of Substitutions at the C–6 Phenyl Ring and at the C–5 Position on Anti–HIV–1 Activity. *J. Med. Chem.* **1992**, *35*, 337–345

(104) Holt, D. A.; Luengo, J. I.; Yamashita, D. S.; Oh, H. J.; Konialian, A. L.; Yen, H. K.; Rozamus, L. W.; Brandt, M.; Bossard, M. J.; Levy, M. A.; Eggleston, D. S.; Liang, J.; Schultz, L. W.; Stout, T. J.; Clardy, J. Design, Synthesis, and Kinetic Evaluation of High–Affinity FKBP Ligands and the X–Ray Crystal Structures of Their Complexes with FKBP12. *J. Am. Chem. Soc.* **1993**, *115*, 9925–9938

3 Small molecule escapes from inside T4 lysozyme by multiple pathways

Ariane Nunes-Alves[a], Daniel M. Zuckerman[b] and Guilherme Menegon Arantes[a]

[a]Department of Biochemistry, Instituto de Química, Universidade de São Paulo, SP, Brazil

[b]Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, Portland, OR, US

Small molecule escapes from inside T4 lysozyme by multiple pathways

Ariane Nunes-Alves,[†] Daniel M. Zuckerman,[*,‡] and Guilherme Menegon Arantes[*,†]

Department of Biochemistry, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil, and Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, 2730 SW Moody Avenue, 97239, Portland, OR, US

E-mail: zuckermd@ohsu.edu; garantes@iq.usp.br

Abstract

The T4 lysozyme L99A mutant is often used as a model system to study small molecule binding to proteins, but pathways for ligand entry and egress from the buried binding site and the associated protein conformational changes have not been fully resolved. Here, molecular dynamics simulations were employed to model benzene exit from its binding cavity using the weighted ensemble (WE) approach to enhance sampling of low-probability unbinding trajectories. Independent WE simulations revealed four pathways for benzene exit which correspond to transient tunnels spontaneously formed in previous simulations of *apo* T4 lysozyme. Thus, benzene unbinding occurs through multiple pathways partially created by intrinsic protein structural fluctuations. Motions of several α-helices and side chains were involved in ligand escape from metastable microstates. WE simulations also provided preliminary estimates of rate constants for each exit pathway. These results complement previous works and provide a semi-quantitative characterization of pathway heterogeneity for binding of small molecules to proteins.

[*]To whom correspondence should be addressed
[†]Universidade de São Paulo
[‡]Oregon Health & Science University

1 Introduction

Binding of small molecules to proteins, a fundamental process in cellular metabolism, is also explored pharmacologically to treat a large number of diseases. Drug development strategies usually try to improve ligand affinity to a target protein, based on the structure and interactions observed on bound ligand-protein complexes. But binding kinetics and ligand residence time may also determine the physiological response and efficiency of a drug.[1–4]

Kinetic, affinity and structural information on protein-ligand association may be obtained from the ensemble of transition pathways[5–8] for the binding process. Each pathway describes the evolution of conformational and orientational degrees of freedom of ligand and protein that leads from an unbound to a bound configuration. As expected from path-ensemble symmetry,[9] the reverse bound-to-unbound process should be described by the same pathway ensemble.

Computer simulations have been of great help to reveal possible pathways and kinetic information of protein-ligand binding.[10–31] For example, Kubas and co-workers[27] simulated pathways for molecular O_2 binding to a hydrogenase and proposed mutations along the binding protein tunnels that slowed down the access of small molecules, thus reducing the hydrogenase inhibition caused by O_2. Casasnovas and co-workers[28] simulated dissociation of a potent pyrazol inhibitor from p38 MAP kinase and proposed congeneric ligands that could spend longer residence times when bound to the same kinase.

However, sampling of pathway ensembles for ligand-protein (un)binding is difficult because these are rare events on the timescales usually reached in molecular dynamics (MD) simulations. Methods such as weighted ensemble (WE),[32,33] milestoning[34,35] and transition interface sampling[36] are valuable because they enhance sampling of pathways by increasing computational time spent in conformational regions with low visiting probability, without introducing bias on the simulated potential energy.[37,38] The WE method has been applied to study pathways and kinetics for protein conformational transitions,[39–41] host-guest association,[42] protein-peptide association[43] and protein-ligand dissociation.[4,22,31]

One of the proteins most often used to study association with small molecules is the T4 lysozyme (T4L),[44–46] the structure of which is shown in figure 1. The T4L L99A mutation creates a hydrophobic cavity in the protein C-terminal domain with a volume of $\approx$150 Å^3. This site may accommodate gases such as xeon or O_2 and small molecules such as benzene and its

nonpolar derivatives.[47–50] The engineered cavity is dehydrated and shows little structural variation upon ligand complexation. Due to this simplicity and ease of experimental manipulation, more than 700 T4L complexes with small molecules have been characterized and used extensively to study structure-affinity relationships[46] and to compare with computational predictions of binding pose and affinity.[51–54]

Crystal structures of T4L L99A mutant bound to benzene and a congeneric series of ligands[44,47] show the binding cavity is buried without any clear tunnel or channel connecting to the protein surface (figure 1). Thus, T4L must undergo structural fluctuations or conformational "breathing" to allow ligand excursion into the binding cavity. The kinetics of benzene entry and exit from T4L L99A mutant have been estimated from nuclear magnetic resonance (NMR) spectroscopy (k_{off}=950 s^{-1} for benzene at 303 K),[55] implying that the conformational motions involved in benzene escape should be observed in a millisecond timescale.

Several studies have addressed the possible pathways and the structural transitions used by small molecules to escape from inside T4L L99A. A weakly populated (or "excited") state involving motion of the flexible T4L helix F (figure 1) was observed from relaxation-dispersion NMR measurements and suggested to be responsible for benzene exit.[56] Later structural determination of this weakly populated state showed that the binding cavity remains inaccessible,[57] but a recent metadynamics simulation showed that the conformational transition from the crystal structure to the weakly populated state forms a transient protein tunnel along which benzene can egress from T4L.[58] An accelerated MD simulation was able to identify a different route for benzene entry[59] and two recent conventional MD simulations identified three other pathways for O_2 unbinding[50] and five putative tunnels formed transiently in *apo*[60] T4L L99A that might accommodate small molecule excursions to the binding cavity.

In order to resolve the number and nature of ligand exit pathways and the conformational transitions involved in T4L L99A mutant, the WE approach is employed here as a discovery tool to sample trajectories for benzene exit. The ensemble of unbinding pathways obtained shows that benzene escapes via four different protein tunnels, all of which formed spontaneously in *apo* T4L L99A in a previous study which did not sample explicit ligand binding or unbinding.[60] We find that motions in helix F, as previously suggested,[46,55,56] but more importantly in helices C, H and J, and side chains of several residues are involved in ligand escape. Rates were estimated

for each pathway and allow a preliminary quantification of their contribution to the overall rate constant.

2 Computational methods

2.1 Structural model and molecular dynamics simulations

The structure of T4L L99A mutant bound to benzene was obtained from the PDB structure 1L83[44] after removal of water and crystallization molecules. Hydrogens were constructed using the GROMACS PDB parser.[61]

Interactions were described by the CHARMM36 force field.[62–64] All simulations employed implicit solvation in the generalized Born surface area (GB/SA) form.[65] This should be a reasonable approximation as the binding cavity in T4L L99A is dehydrated. The OBC method was used to estimate Born radii[66] and the nonpolar contribution was calculated as in Schaefer *et al.*[67] with a surface tension of 5.4 cal mol^{-1} Å^{-2} for all atoms.

Molecular dynamics trajectory segments in WE simulations were run using GROMACS 4.5.[61] Dynamics were carried out at 300 K or at 400 K, with a 2 fs time-step, a leapfrog stochastic dynamics integrator and a friction of 10 ps^{-1}. Covalent hydrogen bonds were constrained with LINCS.[68] Structures were saved every 2 ps for analysis. The total aggregate simulation time as ≈ 29 μs.

2.2 Weighted ensemble simulations and rate estimation

The WE algorithm[32,33] was used to enhance sampling of benzene unbinding events. Briefly, in the WE method a progress coordinate is defined and divided into bins to describe the process of interest. A group of trajectories of the simulated system is propagated from an initial state with initially equal weights. Every τ steps, the occupancy of each coordinate bin is reevaluated and trajectories may be replicated or merged with a proper weight attribution to maintain a given number of trajectories per bin. This number of trajectories and the binning scheme to divide the progress coordinate do not affect the results of simulations, but affect the efficiency of the WE method in sampling rare events.[69]

WESTPA software package[70] was used to manage trajectory splitting and merging. The

initial state was defined as benzene bound to T4L as in the crystallographic position and the target or final state was reached when benzene had a solvent accessible surface area (SASA) higher than 60 % of its maximum SASA. This definition was enough to see benzene reaching the protein surface. But, as discussed below, this definition is *not* equivalent to complete unbinding as measured in experiments. Resampling frequencies (τ) of 10 ps (for simulations run at 400 K) or 2 ps (for runs at 300 K) were used. The number of trajectories per bin varied from 4 to 5 according to the progress coordinate used. WE simulations were run in a non-equilibrium steady-state scheme, with trajectories being recycled back to the initial state once they reached the target state. WE simulations were run for 4000 iterations, resulting in a maximum trajectory length of 8 ns. More details about the WE method are given in the Supporting Information (Table S2).

Time-windowed rate constants ($k_{\Delta t}$) were estimated from the non-equilibrium steady-state set of trajectories obtained from WE simulations:[71]

$$k_{\Delta t}(t) = \frac{\Delta p}{\Delta t} \tag{1}$$

where Δp is the sum of probabilities or weights of the trajectories that reached the target state in the time interval $t - \Delta t$ to t. Transition rates from the bound to the unbound state (dissociation rate constants) were calculated using a Δt of 2 ns.

2.3 Definition of progress coordinates for WE simulations

Two sets of WE simulations with different progress coordinates were carried out: an exploratory set to find possible routes for benzene unbinding; and a production set to estimate rate constants and determine the protein conformational transitions involved in ligand escape, for each unbinding route found.

In the exploratory simulations, two one-dimensional progress coordinates were used: the distance between benzene and binding site center-of-mass (COM); and the root mean squared deviation (RMSD) between the current benzene position and the bound position found in the crystal structure. The binding site was defined by the Cα atoms of L84, V87, R95, A98, A99, V111, L118, N122, A129 and L133 in the artificial binding cavity. Only benzene carbon atoms were considered for RMSD calculations and structures were previously aligned by superimposing

the binding site Cαs just defined. The twelve possible symmetric images of benzene in each configuration were compared to the reference crystal structure and the lowest RMSD value was adopted.

The exploratory progress coordinates were partitioned into small bins along coordinate regions where benzene movements were more restricted, as observed in initial simulations. Larger bins were employed along regions where benzene could diffuse more freely. Unbound states were defined by a distance or RMSD higher than 2.0 nm, in agreement with the SASA criteria described above. Simulations run at high temperature (400 K) for 150 iterations, resulting in a maximum trajectory length of 1.5 ns. Bin boundaries and further details on the progress coordinates are given in table S2.

Based on initial pathways found in the exploratory simulations, production simulations were set up by using Voronoi bins[22,31] to map each of four pathways separately. In a Voronoi mapping, a set of centers is specified and all points in configurational space are attributed to bins for which center they are closest to, according to a given distance criteria. Here, a Voronoi center was manually selected as the structure of a ligand-protein complex along unbinding trajectories obtained in the exploratory set of simulations. A total of 25, 23, 24 and 26 Voronoi centers were used to sample trajectories for the paths denoted below by the colors blue, orange, pink and cyan, respectively. Simulations run at 300K for 4000 iterations, resulting in a maximum trajectory length of 8 ns.

In the production set of simulations, a three-dimensional progress coordinate was employed for a small subset of Voronoi bins based on two atomic pair distances. The distances were chosen to delineate small protein conformational transitions observed in the exploratory simulations, as shown on table S1 (see the Supporting Information). In order to avoid a combinatorial increase in the total number of bins and WE trajectories simulated, the distance dimensions were included in a nested way to only two Voronoi bins for each of the four paths.

Conformational transitions involved in ligand escape from metastable or long-lived microstates along the unbinding pathways were obtained from analysis of the transition structure observed immediately after the Voronoi bin with highest lifetime in a trajectory was unoccupied. Further details for this analysis are described in the Supporting Information (section S5.2)

3 Results

3.1 Identification of pathways for ligand escape

In the exploratory set of simulations run at 400 K, four exit pathways were found for benzene unbinding from T4L, as shown in Figure 2 and denoted by color code. In the blue pathway the ligand egresses from the buried binding site by passing through helices C and F. Benzene mainly transits through helices C and D in the orange pathway, helices F and I in the pink pathway and helices H and J in the cyan pathway.

Table 1 shows that the four pathways were found when RMSD was used as progress coordinate, while three of these paths were visited when the distance from the binding site was used as progress coordinate. The four pathways found at 400 K could also be sampled at simulations performed at 300 K using a Voronoi mapping of the unbinding process.

The Voronoi bin map used as progress coordinate in the production set of simulations allowed sampling of trajectories for a specific pathway, thus facilitating the calculation of rate constants and identification of protein conformational transitions involved in ligand unbinding. Six independent WE simulations were carried out for each set of Voronoi bins defined (table 1), leading to a total of 11237, 16062, 15777 and 12642 unbinding events through the blue, orange, pink and cyan pathways, respectively.

Table 1 shows that WE simulations with a progress coordinate defined from Voronoi bins for the blue pathway also sampled unbinding events for the orange pathway, and vice-versa. This overlap for two different Voronoi mappings is due to a similar route followed by the ligand in the first half of the orange and blue pathways. Unbinding trajectories were reassigned to the correct pathway (blue or orange) before further analysis by adding two new Voronoi centers to the blue pathway (total of 27 centers) and one new center to the orange pathway (total of 24 centers), as detailed in the Supporting Information (section S5.1). Thus, the correct estimation of rate constants and detection of conformational transitions was possible for each pathway.

3.2 Protein structural transitions in long-lived microstates

Considering that small-molecule ligands such as benzene are clearly buried inside a hydrophobic cavity in T4L L99A holo crystal structures[44,49] (figure 1), conformational transitions or some

"breathing" of the protein structure are expected to allow ligand escape to the protein surface.

Transition structures involved in ligand escape from metastable or long-lived microstates along the unbinding pathways were analyzed. Table 2 shows fractions of unbinding trajectories with transition structures that displayed side chain rotations involved in ligand unbinding. Fractions are similar for different reference structures, showing that the identification of side chains is independent of the reference. Rotations in Y88 and I78 side chains are the most often observed for the blue and orange pathways. These side chains move away from the binding site, allowing the ligand to depart from the buried cavity. Rotations of W126, R154 and V111 side chains in the cyan pathway are related to motions in helices H, J and F, respectively. Motions of the other side chains listed on table 2 allow the ligand to exit the bound state and reach the protein surface. Figure 4 shows some of the side chains involved in exit pathways.

Backbone fluctuation is also involved in benzene unbinding. Significant displacements in helix C in transition structures were observed in fractions of 0.35 and 0.65 of unbinding trajectories via blue and orange pathways, respectively. This helix moves away from the binding site and facilitates ligand exit from the buried cavity. The orange path also involved displacements of helix D in one fifth of the transition structures. One third of trajectories in the cyan pathway presented significant fluctuations in the backbone of helices H and J, increasing the distance between these helices and allowing benzene passage. On the other hand, only 15% and 12% of trajectories for the blue and pink pathways, respectively, presented transition structures with displacements in helix F. Fluctuations along this helix F were previously suggested as essential displacements for small molecule unbinding from T4L.[55,57,58]

3.3 Rate constants for each unbinding pathway

Preliminary estimates for dissociation rate constants were calculated as a function of simulation time for each exit pathway in order to quantify their relative contributions to the overall ligand escape rate constant, as shown in Figure 3. However, the relative contributions of each pathway can not be clearly distinguished, as their rank order changes over simulation time. Note that the probability fluxes, upon which the rates are estimated, are expected to increase with simulation time until they reach steady values. Additionally, the present estimates are dominated by only one or two individual WE simulations (figure S1). Thus, further sampling appears to be required

to enable confident distinctions among the pathways.

The experimental rate constant for benzene dissociation (k_{off}) from T4L L99A mutant has been determined[55] by lineshape analysis of protein NMR as k_{off} =950 s^{-1}. The experimental value is close to the rate calculated for the pink and cyan pathways (at 8 ns in figure 3) but lower than the rates estimated for the blue and orange pathways (figure 3). Rates calculated from the WE simulations are expected to be higher than the experimental value. Here a less strict definition of unbound state was used (section 4), leading to benzene unbinding up to the protein surface rather than complete dissociation.

4 Discussion

4.1 Exit pathways and comparison to previous T4L studies

Four pathways were found here for benzene egress from the binding cavity buried in T4L L99A mutant. Each escape route was characterized by a preliminary rate constant and a different set of protein structural transitions.

Interestingly, a previous long timescale (30μs) conventional MD simulation of *apo* T4L L99A identified 5 protein tunnels, without sampling explicit ligand entry or egress, which could be used by small molecules to access the buried cavity.[60] Four of these tunnels, named D/F/G (tunnel through helices D, F and G), C/D, F/G/H and H/J, correspond respectively to the blue, orange, pink and cyan pathways found here. An additional D/G tunnel was also identified, but the authors suggested this D/G and the H/J tunnels would be too narrow to accommodate the passage of benzene.[60] The WE simulations shown here were not able to find benzene escape through tunnel D/G, but identified benzene transit via tunnel H/J (cyan pathway) with a slower rate and indeed a narrower tunnel than escape through the other pathways.

Other previous simulations also studied small molecule (un)binding to T4L.[50,58,59] Three routes were found for molecular O_2 dissociation in conventional MD simulations.[50] One route matches the cyan path (H/J tunnel) found here for benzene transit. The other two O_2 routes correspond to the tunnel D/G mentioned above and to another tunnel found between helices D, E, G, H and J. Both these two last tunnels were not observed here for benzene transit. Two MD simulations combined with enhanced sampling methods separately identified benzene

transit from the binding site through the blue[59] and pink[58] pathways, with similar protein structural transitions to the ones observed here. Ligand excursion through the orange pathway was observed for the first time here.

The two previous simulations showing benzene (un)binding to T4L were able to find only one ligand transit pathway each[58,59] probably due to the enhanced sampling procedure employed. Miao *et al.* used a Gaussian bias on the potential energy[59] that may not efficiently increase sampling of processes determined by entropic barriers such as required to sample multiple entry pathways. Wang *et al.* used metadynamics[72] with a path progress coordinate designed to sample the protein transitions involved in displacements of T4L residue F114 and helix F.[57,58] These transitions are associated with benzene unbinding via the pink pathway which was thus the only ligand exit pathway they observed.[58]

Taken together these results show that small molecules such as benzene unbind from T4L L99A through multiple and kinetically competitive pathways that correspond to ligand passage through transient tunnels spontaneously opened by intrinsic fluctuations of the protein structure, observed in the absence of any ligand. This is reminiscent of a conformational selection mechanism,[73,74] but for protein tunnels permitting small molecule transit.

It should be noted the results shown here are the first to our knowledge to find together and consistently the four pathways for small molecule egress from T4L, by explicitly collecting exit trajectories instead of observing protein structural fluctuations that may lead to putative tunnels. The exploratory set of simulations presented here was carried out in 2016 without any hint of the pathways explored in the pink, orange and cyan paths. At that time, previous works[50,58,60] were not yet published or known to us. We were only aware of the tunnel involved in the blue path,[59] although no information about it was used during our exploratory simulations. Remarkably, model composition (solvation, for instance) and force field descriptions differ significantly between our and previous studies.[50,58–60] Nevertheless, a consistent set of ligand egress tunnels and pathways is found.

Pathways for ligand entry and exit have also been investigated in other proteins, containing binding sites with variable surface exposure. Multiple protein tunnels are usually available for small gaseous molecules to diffuse into binding sites. For instance, hydrogenases may react with molecular O_2 and three entry routes for gaseous molecules have been characterized.[12,14,27] On

the other hand, a dominant route usually emerges for proteins with exposed binding sites and bulky ligands. This is the case for unbinding of a drug-like inhibitor from p38 MAP kinase.[28] Binding to T4L lies somewhere in between and is similar to cytochrome P450 which binds organic molecules larger than gases but smaller than bulky drug-like ligands in a buried cavity via 3 or 4 different tunnels.[10,17,18,21,26,29]

4.2 Structural analysis of egress pathways

The analysis of transition structures presented on section 3.2 will only find protein movements related to benzene escape from metastable microstates (or long-lived Voronoi bins). Thus, conformational transitions involved in short-lived microstates have not been analyzed. It is noticeable, however, that no specific conformational gate such as a single side chain or helix displacement is clearly dominant for each pathway. Instead, benzene exit depends on a combination of intrinsic protein motions. Nonetheless, some side chains were found to be important for ligand egress and point mutations on T4L could be suggested in order to modulate the dissociation rate constant. Residues Y88, I78 (blue and orange pathways) or F114 (pink pathway) could be exchanged for a less bulky residue such as alanine to speed up benzene exit.

Crystallographic[44,45,49] and NMR[56,57] data has shown that helix F is the most flexible element nearby the artificial cavity in T4L mutants. Thus, it has been suggested[46,55,75] that displacements along this helix are involved in ligand unbinding. The WE simulations suggest that, although observed, movements on helix F are not determinant or ubiquitous among unbinding pathways. Instead, transitions in helices C, H and J are more important for ligand transit. These motions might lead to conformational states which are too weakly populated to be detected experimentally.

4.3 Methodological limitations

It may be concluded that using the ligand RMSD as a progress coordinate in combination with the WE method and high-temperature MD simulation is an efficient and reliable procedure to explore ligand exit pathways from buried cavities in proteins. Ligand distance to the binding site COM lacks information on ligand orientation and proved a less reliable progress coordinate as one pathway was not found when this metric was used.

Increasing the number of dimensions in the progress coordinate may reduce the waiting time in long-lived microstates due to increment of the number of trajectories in the WE procedure or to enhancement of protein conformational transitions that facilitate ligand transit. The additional distance dimensions used here as progress coordinates (table S1) were later confirmed to be involved in ligand unbinding (table 2).

Many unbinding trajectories are observed with a maximum length of 8 ns in our WE simulations. Since the WE approach preserves the dynamics while enhance sampling, the transit time for ligand diffusion and for the necessary protein conformational transitions has similar (or smaller) length, although the waiting time seems to be much longer. This is in agreement with repeated observations of opening of protein tunnels during the 30 μs MD simulation of *apo* T4L.[60]

5 Conclusions

Pathways for benzene exit from the buried binding cavity of T4L L99A mutant and the associated protein conformational changes were characterized here with the WE simulation method. For the first time, four separate ligand unbinding processes were observed in a single study. The four pathways found here are also in agreement with ligand transit routes observed in previous holo T4L simulations.[50,58,59] Conformational transitions in several side chains, most notably Y88, I78 and F114, as well as displacements in helices C, D, F, H and J are involved in benzene unbinding. Our study also provided preliminary estimates of the escape rates for the different pathways, although further sampling is required to narrow statistical uncertainties.

The present study used the WE method in two stages: an exploratory or discovery mode using naive progress coordinates to find possible pathways, followed by a production stage with more focused sampling. Combining the WE method with a RMSD progress coordinate and high temperature MD simulations appears as a plausible approach to find multiple (un)binding pathways for small molecules. Production simulations using Voronoi bins may then be obtained to better characterize the dissociation rates and the conformational transitions involved in ligand transit.

The four protein pathways found for benzene egress may also be used for larger binders such as meta-xylene and N-butylbenzene.[46] Similarly, it is also worth investigating if hydrophobic

molecules such as 1,3,5-trimethylbenzene, that still could fit inside the binding cavity, experimentally do not bind T4L L99A mutant[46] due to a very slow entry kinetics.[54] Such bulkier molecules may require wider protein tunnels to access the binding cavity, which WE could help to rule out if extremely low rates were found.

Acknowledgement

A.N.A. acknowledges Lillian Chong, Adam Pratt, Rory Donovan and Ernesto Suarez for discussions and technical help during a recent visit to the University of Pittsburgh, and Murilo Teixeira for suggestions on the manuscript. Funding from FAPESP (projects 2014/17008-7, 2014/21900-2, 2015/19912-5 and 2016/24096-5) and from NIH Grant GM115805 is gratefully acknowledged. Computing resources were in part provided by the University of Pittsburgh Center for Research Computing.

References

(1) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. *Nat. Rev. Drug Discov.* **2006**, *5*, 730–739.

(2) Copeland, R. A. *Nat. Rev. Drug Discov.* **2016**, *15*, 87–95.

(3) Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H.; Segala, E.; Amaral, M.; Guo, D.; Andres, D.; Georgi, V.; Stoddart, L. A.; Hill, S.; Cooke, R. M.; Graaf, C. D.; Leurs, R.; Frech, M.; Wade, R. C.; de Lange, E. C. M.; IJzerman, A. P.; Müller-Fahrnow, A.; Ecker, G. F. *Drug Discov. Today* **2017**, *22*, 896–911.

(4) Tang, Z.; Roberts, C. C.; Chang, C. A. *Front. Biosci.* **2017**, *22*, 960–981.

(5) Dellago, C.; Bolhuis, P. G.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 9236–9245.

(6) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(7) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.

(8) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18043–18048.

(9) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2011**, *7*, 2520–2527.

(10) Winn, P. J.; Lüdemann, S. K.; Gauges, R.; Lounnas, V.; Wade, R. C. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5361–5366.

(11) Martínez, L.; Sonoda, M. T.; Webb, P.; Baxter, J. D.; Skaf, M. S.; Polikarpov, I. *Biophys. J.* **2005**, *89*, 2011–2023.

(12) Cohen, J.; Kim, K.; King, P.; Seibert, M.; Schulten, K. *Structure* **2005**, *13*, 1321–1329.

(13) Buch, I.; Giorgino, T.; Fabritiis, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10184–10189.

(14) Wang, P. H.; Best, R. B.; Blumberger, J. *J. Am. Chem. Soc.* **2011**, *133*, 3548–3556.

(15) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *J. Am. Chem. Soc.* **2011**, *133*, 9181–9183.

(16) Huang, D.; Caflisch, A. *PLoS Comput. Biol.* **2011**, *7*, e1002002.

(17) Cojocaru, V.; Winn, P. J.; Wade, R. C. *Curr. Drug Metab.* **2012**, *13*, 143–154.

(18) Yu, X.; Cojocaru, V.; Wade, R. C. *Biotechnol. Appl. Biochem.* **2013**, *60*, 134–145.

(19) Bisha, I.; Rodriguez, A.; Laio, A.; Magistrato, A. *PLoS Comput. Biol.* **2014**, *10*, 1–8.

(20) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E386–E391.

(21) Yu, X.; Nandekar, P.; Mustafa, G.; Cojocaru, V.; Lepesheva, G. I.; Wade, R. C. *Biochim. Biophys. Acta* **2016**, *1860*, 67–78.

(22) Dickson, A.; Lotz, S. D. *J. Phys. Chem. B* **2016**, *120*, 5377–5385.

(23) Teo, I.; Mayne, C. G.; Schulten, K.; Lelièvre, T. *J. Chem. Theory Comput.* **2016**, *12*, 2983–2989.

(24) Paloncýová, M.; Navrátilová, V.; Berka, K.; Laio, A.; Otyepka, M. *J. Chem. Theory Comput.* **2016**, *12*, 2101–2109.

(25) Rydzewski, J.; Nowak, W. *Phys. Life Rev.* **2017**, *In press*.

(26) Rydzewski, J.; Nowak, W. *Sci. Rep.* **2017**, *7*, 7736.

(27) Kubas, A.; Orain, C.; De Sancho, D.; Saujet, L.; Sensi, M.; Gauquelin, C.; Meynial-Salles, I.; Soucaille, P.; Bottin, H.; Baffert, C.; Fourmond, V.; Best, R. B.; Blumberger, J.; Léger, C. *Nat. Chem.* **2017**, *9*, 88–95.

(28) Casasnovas, R.; Limongelli, V.; Tiwary, P.; Carloni, P.; Parrinello, M. *J. Am. Chem. Soc.* **2017**, *139*, 4780–4788.

(29) Magistrato, A.; Sgrignani, J.; Krause, R.; Cavalli, A. *J. Phys. Chem. Lett.* **2017**, *8*, 2036–2042.

(30) Tiwary, P.; Mondal, J.; Berne, B. J. *Sci. Adv.* **2017**, *3*, e1700014.

(31) Dickson, A.; Lotz, S. D. *Biophys. J.* **2017**, *112*, 620–629.

(32) Huber, G. A.; Kim, S. *Biophys. J.* **1996**, *70*, 97–110.

(33) Zuckerman, D. M.; Chong, L. T. *Annu. Rev. Biophys.* **2017**, *46*, 43–57.

(34) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880–10889.

(35) Votapka, L. W.; Amaro, R. E. *PLoS Comput. Biol.* **2015**, *11*, e1004381.

(36) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762–7774.

(37) Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41–62.

(38) Chong, L. T.; Saglam, A. S.; Zuckerman, D. M. *Curr. Opin. Struct. Biol.* **2017**, *43*, 88–94.

(39) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 3527–3539.

(40) Adelman, J. L.; Dale, A. L.; Zwier, M. C.; Bhatt, D.; Chong, L. T.; Zuckerman, D. M.; Grabe, M. *Biophys. J.* **2011**, *101*, 2399–2407.

(41) Suárez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2014**, *10*, 2658–2667.

(42) Zwier, M. C.; Kaus, J. W.; Chong, L. T. *J. Chem. Theory Comput.* **2011**, *7*, 1189–1197.

(43) Zwier, M. C.; Pratt, A. J.; Adelman, J. L.; Kaus, J. W.; Zuckerman, D. M.; Chong, L. T. *J. Phys. Chem. Lett.* **2016**, *7*, 3440–3445.

(44) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. *Nature* **1992**, *355*, 371–373.

(45) Morton, A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8576–8588.

(46) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. *Protein Sci.* **2010**, *19*, 631–641.

(47) Morton, A.; Baase, W. A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8564–8575.

(48) Mulder, F. A. A.; Hon, B.; Muhandiram, D. R.; Dahlquist, F. W.; Kay, L. E. *Biochemistry* **2000**, *39*, 12614–12622.

(49) Merski, M.; Fischer, M.; Balius, T. E.; Eidam, O.; Shoichet, B. K. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 5039–5044.

(50) Kitahara, R.; Yoshimura, Y.; Xue, M.; Kameda, T.; Mulder, F. A. *Sci. Rep.* **2016**, *6*, 20534.

(51) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. *J. Mol. Biol.* **2002**, *322*, 339–355.

(52) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *J. Mol. Biol.* **2007**, *371*, 1118–1134.

(53) Wang, K.; Chodera, J. D.; Yang, Y.; Shirts, M. R. *J. Comput. Aided Mol. Des.* **2013**, *27*, 989–1007.

(54) Nunes-Alves, A.; Arantes, G. M. *J. Chem. Inf. Model.* **2014**, *54*, 2309–2319.

(55) Feher, V. A.; Baldwin, E. P.; Dahlquist, F. W. *Nat. Struct. Biol.* **1996**, *3*, 516–521.

(56) Mulder, F. A. A.; Mittermaier, A.; Hon, B.; Dahlquist, F. W.; Kay, L. E. *Nat. Struct. Biol.* **2001**, *8*, 932–935.

(57) Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E. *Nature* **2011**, *477*, 111–114.

(58) Wang, Y.; Papaleo, E.; Lindorff-Larsen, K. *eLife* **2016**, *5*, e17505.

(59) Miao, Y.; Feher, V. A.; McCammon, J. A. *J. Chem. Theory Comput.* **2015**, *11*, 3584–3595.

(60) Schiffer, J. M.; Feher, V. A.; Malmstrom, R. D.; Sida, R.; Amaro, R. E. *Biophys. J.* **2016**, *111*, 1631–1640.

(61) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845–854.

(62) 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. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(63) MacKerell, A. D.; Feig, M.; Brooks, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(64) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(65) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(66) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.

(67) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835–848.

(68) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(69) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *132*, 054107.

(70) Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; Suárez, E.; Lettieri, S.; Wang, D. W.; Grabe, M.; Zuckerman, D. M.; Chong, L. T. *J. Chem. Theory Comput.* **2015**, *11*, 800–809.

(71) Bhatt, D.; Zhang, B. W.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *133*, 014110.

(72) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(73) Weikl, T. R.; Paul, F. *Protein Sci.* **2014**, *23*, 1508–1518.

(74) Gianni, S.; Dogan, J.; Jemth, P. *Biophys. Chem.* **2014**, *189*, 33–39.

(75) Vallurupalli, P.; Chakrabarti, N.; Pomès, R.; Kay, L. E. *Chem. Sci.* **2016**, *7*, 3602–3613.

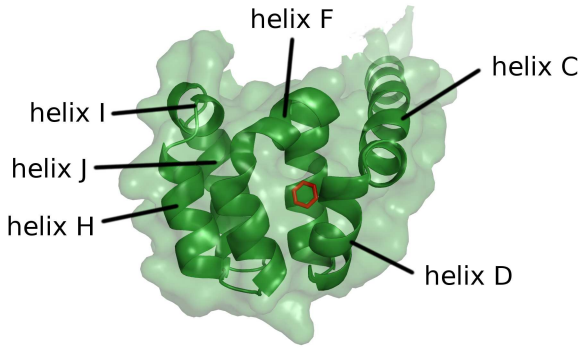
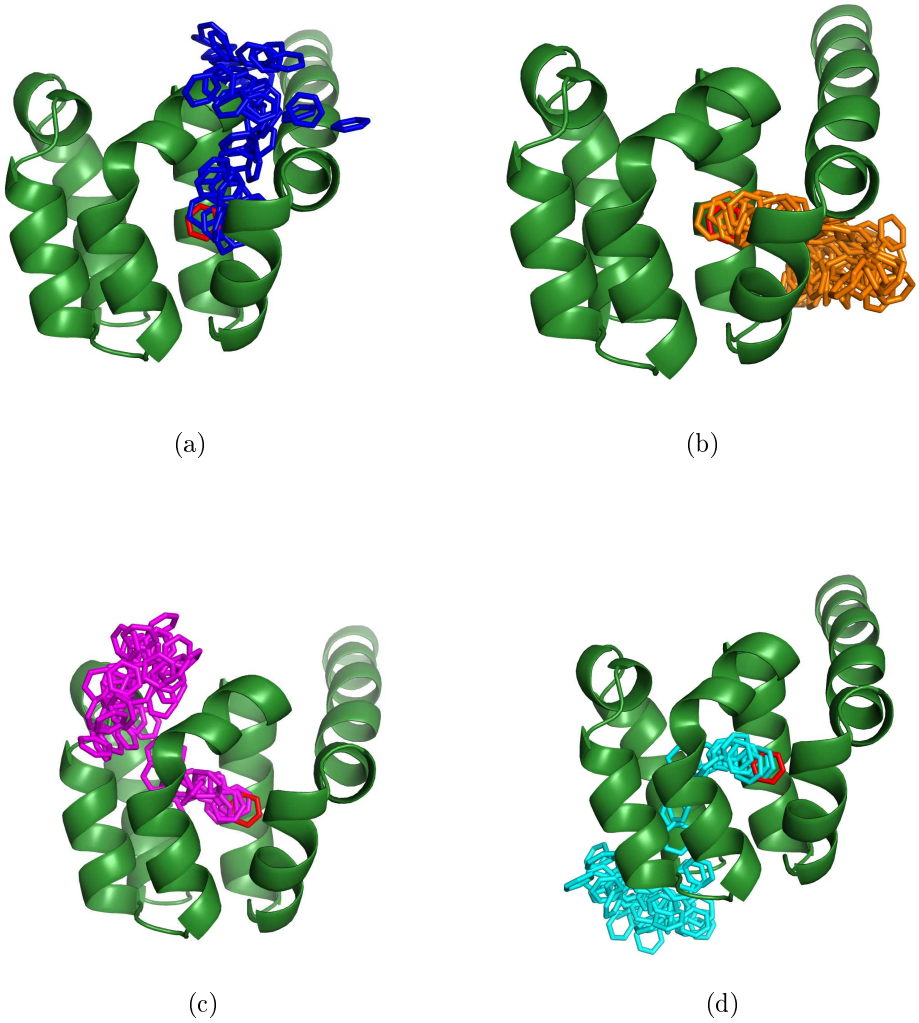


Figure 1: Crystal structure of the C-terminal domain of T4 lysozyme L99A mutant bound to benzene (red). The protein is represented with its molecular surface (green transparency), showing the ligand is fully buried.



(a) (b)

(c) (d)

Figure 2: Four pathways were found for benzene egress from the buried binding site in T4 lysozyme. Blue (a), orange (b), pink (c) and cyan (d) colors show benzene positions sampled during one WE simulation. Each color represents a different exit pathway. Benzene position in the crystal structure is shown in red sticks. Only the T4L C-terminal domain is shown, but the complete protein was used in all simulations.

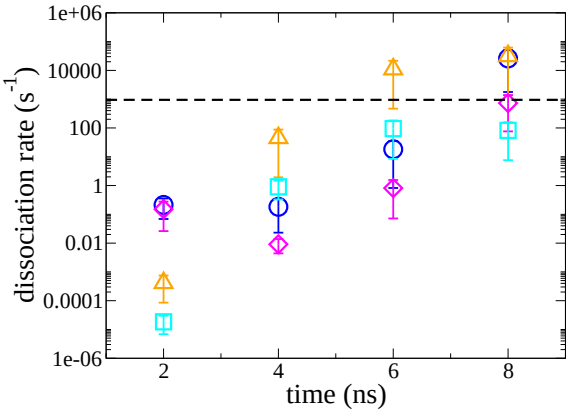
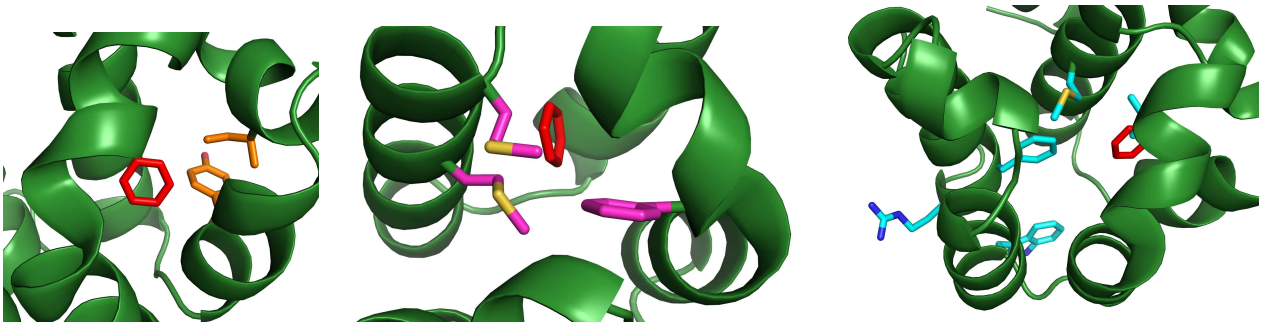


Figure 3: Average rate constant estimated for each pathway of benzene unbinding from T4L. Symbols: blue pathway - circle; orange pathway - triangle; pink pathway - diamond; cyan pathway - square. Bars indicate the standard error. The dotted line indicates the experimental rate constant. Rates were averaged every 2 ns for 6 WE simulations (12 WE simulations for the blue and orange pathways).



(a) (b) (c)

Figure 4: Residue side chains involved in benzene unbinding. (a) Y88 and I78 (orange sticks) contribute to blue and orange pathways. (b) F114, M102 and M106 (pink) contribute to pink path. (c) M102, F153, R154, W126 and V111 (cyan) contribute to cyan path. Benzene position in the crystal structure is shown in red sticks.

Table 1: Sampling of exit pathways by WE simulations performed with distance to cavity COM (dist), ligand RMSD, and Voronoi bins (V) progress coordinates in different temperatures. Total refers to the total amount of WE simulations run.

	400 K		300 K					
pathway	**dist**	**RMSD**	$\mathbf{V}_{blue}$		$\mathbf{V}_{orange}$		$\mathbf{V}_{pink}$	$\mathbf{V}_{cyan}$
			before[a]	after	before	after		
blue	4	2	6	6	6	0	0	0
orange	3	2	6	0	6	6	0	0
pink	2	5	0	0	0	0	6	0
cyan	0	2	0	0	0	0	0	6
total	4	5	6		6		6	6

[a]before and after post-production reassignment step, as explained in the main text.

Table 2: Fraction of trajectories in WE simulations showing residue side chain rotation associated with benzene transit from the microstate (Voronoi bins) with longest lifetime along an unbinding event. Reference structures were collect either 100 ps or 10 ps before the transition structure. Data obtained from 6 WE simulations for each pathway.

blue			**orange**		
side chain	**100 ps**	**10 ps**	**side chain**	**100 ps**	**10 ps**
I78	0.78	0.67	Y88	0.55	0.69
Y88	0.52	0.48	I78	0.57	0.49
pink			**cyan**		
side chain	**100 ps**	**10 ps**	**side chain**	**100 ps**	**10 ps**
F114	0.43	0.50	M102	0.36	0.45
M102	0.42	0.25	F153	0.59	0.35
M106	0.35	0.25	R154	0.29	0.31
			W126	0.26	0.23
			V111	0.26	0.22

Small molecule escapes from inside T4 lysozyme by multiple pathways

Ariane Nunes-Alves[1], Daniel M. Zuckerman[2] and Guilherme Menegon Arantes[1]

[1]*Department of Biochemistry, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil*

[2]*Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, 2730 SW Moody Avenue, 97239, Portland, OR, US*

Supporting Information

S5.1 Reassignment of the unbinding trajectories to the correct pathways

Visual inspection of the trajectories showed that WE simulations with a progress coordinate defined from Voronoi bins for the blue pathway also sampled unbinding events for the orange pathway, and vice-versa. A post-production step was performed to properly separate unbinding events and reassign to the correct exit pathway. The unbinding event was reassigned to the pathway which contained the Voronoi center closest to the unbound state configuration. This was possible by using 27 Voronoi centers (instead of 25 used in the WE production runs) to the blue pathway and 24 (instead of 23) centers to the orange pathway. Reassignement allowed proper separation of pathways and correct estimation of rate constants and conformational changes for each pathway.

S5.2 Identification of protein conformational transitions involved in ligand unbinding

Conformational changes allowing ligand egress from T4L were analyzed. Metastable microstates were associated to Voronoi bins with high lifetimes and the protein transitions involved in ligand progression along the unbinding pathway were found by:

- identification of the metastable microstate with the Voronoi bin with highest lifetime for each successful unbinding trajectory;

- collection of the structure in the frame immediately after the metastable state was unoccupied (transition structure);
- construction of a list of the side chains contacting benzene in the transition structure, using a distance cutoff of 0.5 nm;
- computation of the (heavy atom) RMSD between transition structure and reference structure for each side chain, after structure alignment using the binding site C_{alpha} defined in section 2.2;
- selection of 3 side chains with highest RMSD.

Reference structures were collected either 100 ps or 10 ps before the transition structure in the same trajectory.

Analysis revealed that movement of helices C, D, F, H and J could also be involved in ligand unbinding. This was quantified by:

- identification of the metastable microstate and collection of transition structure for each successful unbinding trajectory;
- check if one of the side chains of helices C (residues 69-81), D (residues 83-90) F (residues 107-114), H (residues 126-135) or J (residues 143-155) was contacting benzene in the transition structure, using a cutoff of 0.5 nm;
- check the helix displacement.

The criteria for helix displacement was based on average atom-pair distance distributions shown in figure S2. Helix C was considered displaced when the average distance was higher than 1.01 nm, helix D when higher than 1.02 nm, helix F when higher than 1.25 nm and helices H and J when higher than 0.67 nm.

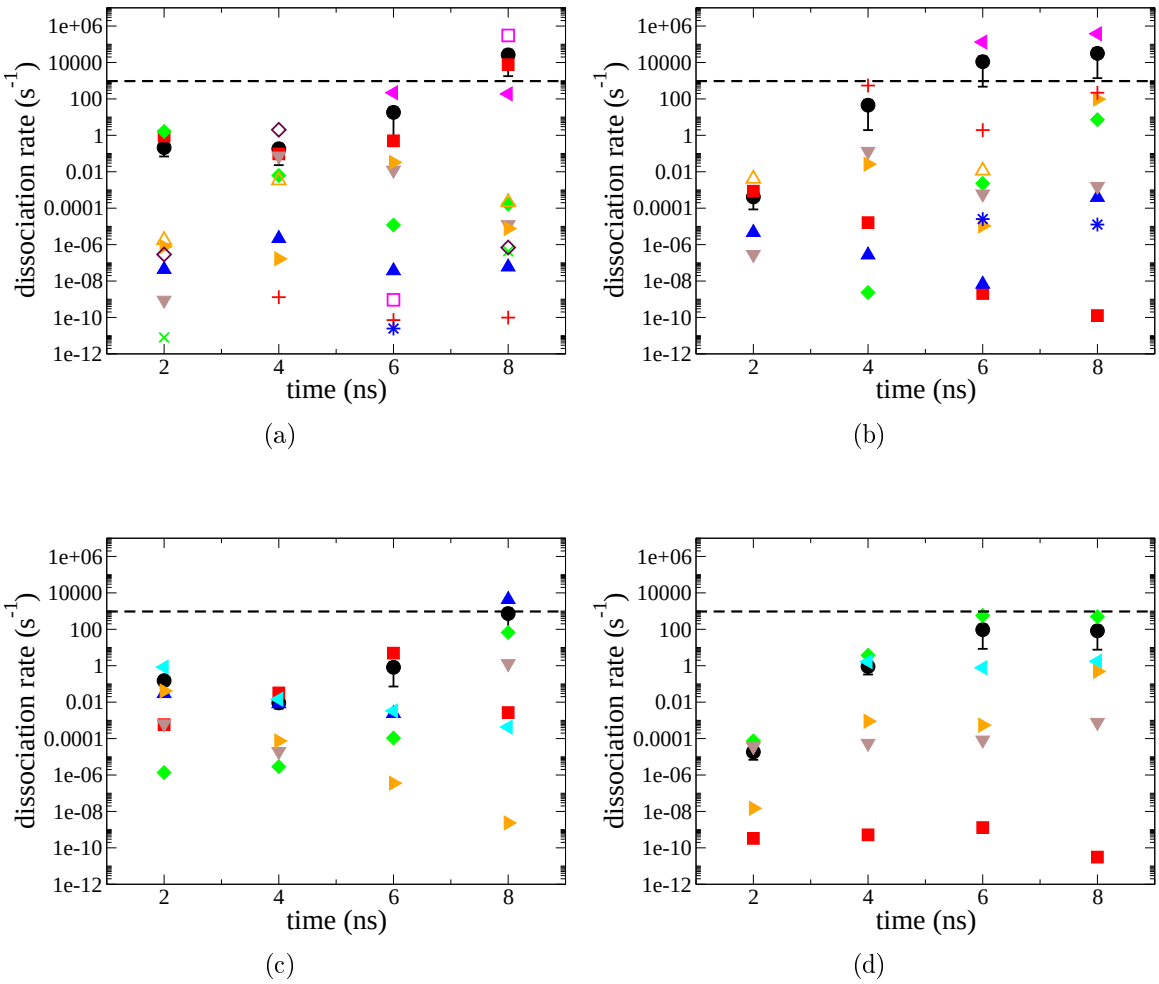


Figure S1: Average and individual rate constant estimated for each pathway of benzene unbinding from T4L. (a) Blue pathway, (b) orange pathway, (c) pink pathway and (d) cyan pathway. Black circles indicate averages, remaining symbols indicate individual estimates from one WE simulation. Bars represent the standard error in the averages, which are dominated by larger values. Dotted lines indicate the experimental rate constant.

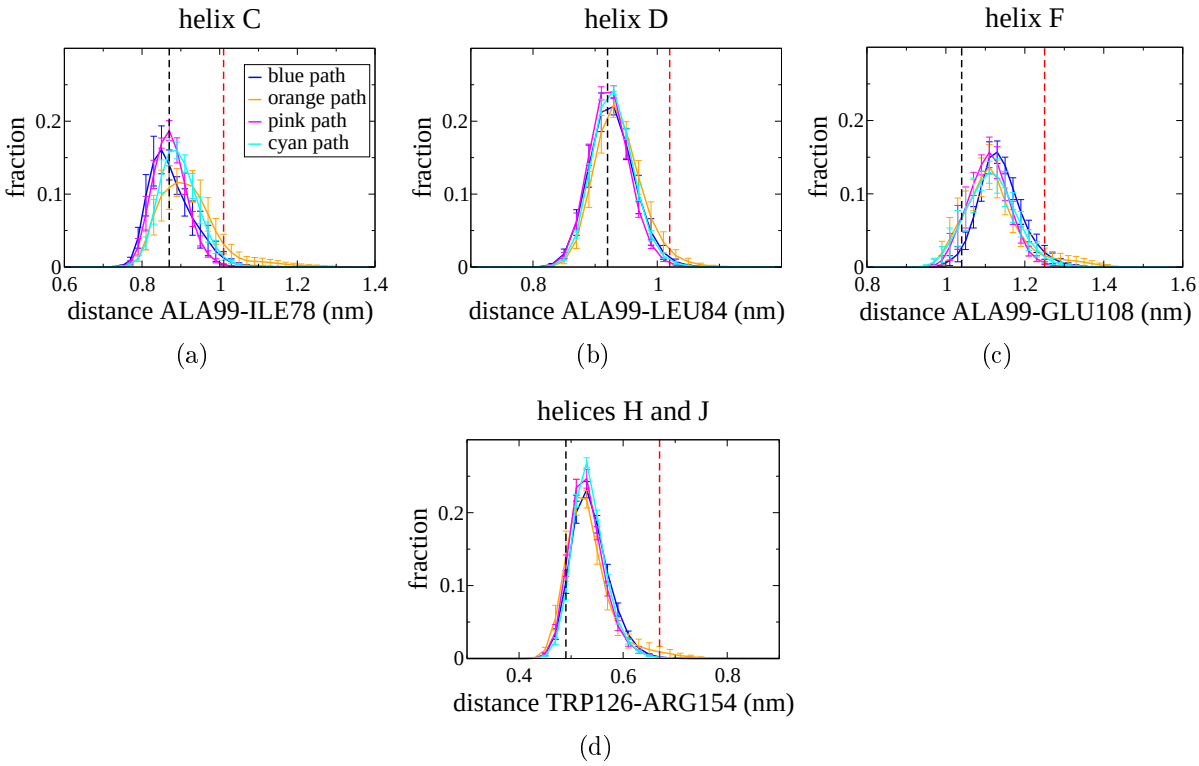


Figure S2: Average distance distributions from 4 to 8 ns of six WE simulations for each pathway (color coded as shown in the insert). Bars represent the standard error in the averages. Black and red lines represent the distance in the crystal structure and the distance criteria for helix displacement, respectively.

Table S1: Definition of additional progress coordinates for benzene unbinding. Two atom-pair distances were selected for each pathway. The bin boundaries (in nm) and the Voronoi bins where the progress coordinate was added are also shown. Voronoi bins are numbered in order of increasing distance from the binding site.

pathway	**distance**	**bin boundaries**	**Voronoi bins**
blue	CZ-Y88 - CA-A99	0.70, 0.83, 0.95	1, 2
	CA-L84 - CA-A99	0.90, 0.97, 1.05	5, 6
orange	CZ-Y88 - CA-A99	0.70, 0.83, 0.95	1, 2
	CB-Y88 - CB-I78	0.50, 0.58, 0.65	7, 8
pink	CB-V111 - CA-A99	0.80, 0.85, 0.90	2, 3
	CE-M102 - CA-A146	0.70, 0.77, 0.85	7, 11
cyan	CG-L133 - CA-S117	0.62, 0.72, 0.80	8, 9
	CD-R154 - CA-W126	0.60, 0.70, 0.80	19, 20

Table S2: Parameters used to run WE simulations for different progress coordinates (PC). WE sim.: number of WE simulations run, traj./bin: number of trajectories per bin, τ: resampling frequency (in ps), iterations: number of resampling steps of one WE simulation, max. length: maximum length of one WE simulation (in ns, calculated as τ*iterations), bins: amount of bins of the progress coordinate, aggreg. time: aggregate simulation time of the progress coordinate (in ns, calculated as (WE sim.)*(traj./bin)*(max. length)*bins).

PC	**WE sim.**	**traj./bin**	τ	**iterations**	**max. length**	**bins**	**aggreg. time**
dist	4	5	10	150	1.5	12	360
RMSD	5	5	10	150	1.5	17	637.5
V_{blue}	6	4	2	4000	8	37	7104
V_{orange}	6	4	2	4000	8	35	6720
V_{pink}	6	4	2	4000	8	36	6912
V_{cyan}	6	4	2	4000	8	38	7296

Bin boundaries for distance PC (in nm): 0.30, 0.35, 0.38, 0.41, 0.50, 0.55, 0.60, 0.65, 0.80, 1.60, 1.68. Boundaries for RMSD PC (nm): 0.15, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.70, 0.80, 0.90, 1.00, 1.10, 1.50, 1.90.

4 Mechanical unfolding of macromolecules coupled to bond dissociation

Ariane Nunes-Alves and Guilherme Menegon Arantes

Department of Biochemistry, Instituto de Química, Universidade de São Paulo, SP, Brazil

Manuscript submitted to *J. Chem. Theory Comput.*

Mechanical unfolding of macromolecules coupled to bond dissociation

Ariane Nunes-Alves and Guilherme Menegon Arantes*

Department of Biochemistry, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil

E-mail: garantes@iq.usp.br

Abstract

Single molecule force spectroscopy has become a powerful tool to investigate molecular mechanisms in biophysics and materials science. In particular, the new field of polymer mechanochemistry has emerged to study how tension may induce chemical reactions in a macromolecule. A rich example is the mechanical unfolding of the metalloprotein rubredoxin coupled to dissociation of iron-sulfur bonds that has recently been studied in detail by atomic force microscopy. Here, we present a simple molecular model composed of a classical all-atom force field description, implicit solvation and steered molecular dynamics simulation to describe the mechanical properties and mechanism of forced unfolding coupled to covalent bond dissociation of macromolecules. We apply this model and test it extensively to simulate forced rubredoxin unfolding, and dissect the sensitivity of calculated mechanical properties with model parameters. The model provides a detailed molecular explanation of experimental observables such as force-extension profiles and contour length increments. Changing the points of force application along the macromolecule results in different unfolding mechanisms, characterized by disruption of hydrogen bonds and secondary protein structure, and determines the degree of solvent access to the reactive center. We expect this molecular model will be broadly applicable to simulate (bio)polymer mechanochemistry.

*To whom correspondence should be addressed

1 Introduction

The stability and denaturation kinetics of macromolecules are fundamental properties relevant to both natural polymers and designed materials. Besides the more traditional thermal and chemical unfolding techniques based on ensemble-averaged observations, polymer stability has been recently probed by mechanical manipulation at the single-molecule level, particularly by atomic force microscopy (AFM) and optical tweezers methods.[1–4]

These single-molecule force spectroscopy techniques are often applied to unfold polymers composed of repetitions of several folded units, leading to measured force-extension curves with a regular saw-tooth pattern.[5] Increments in contour length between force peaks can be used as fingerprints to assign the specific macromolecular region unfolded under tension and attribute its stability to the distribution of measured peak forces.[6] Further information on the polymer unfolding kinetics can be obtained from experiments run at different pulling rates which depict the dependency of unfolding force on pulling velocity, also called the (dynamic) force spectrum of the material.[5,7–9]

Although studies have been performed most often in macromolecules that mechanically unfold only due to disruption of non-covalent interactions, there is an increasing number of studies reporting polymer mechanochemical activation: Si–O bond dissociation in stretched polydimethylsiloxane,[10] poly(tetrahydrofuran) detachment from a silver(I)-N-carbene complex,[11,12] Au–S bond rupture in the gold binding protein GolB[13] and iron-catechol complex dissociation.[14] Another rich example is the mechanical unfolding of the metalloprotein rubredoxin coupled to dissociation of ferric-thiolate (Fe–S) covalent bonds that has been extensively studied by AFM[15–20] and electronic structure calculations.[20–22]

Rubredoxin is a simple iron-sulfur protein with four cysteine side-chains S_γ bound to just one Fe atom in a tetrahedral orientation (Fig. 1). This FeS center is buried in the polymer interior and complete rubredoxin unfolding requires exposure to solvent and rupture of at least two of the four Fe–S bonds. Rubredoxin polyproteins, or polymerical constructions composed of repetitions of protein units, have been prepared by chemical cross-linking[23] or genetic engineering.[24] In the former case, protein units are connected through a pair of additional cysteine residues introduced in the rubredoxin sequence by point mutations. Polyproteins with different topologies have been prepared using this procedure and their mechanical unfolding rate and

force distributions were shown by AFM to depend on the points of force application along the rubredoxin sequence.[17]

Steered molecular dynamics (SMD) simulations[25,26] have been of great aid in the interpretation of mechanical unfolding experiments by providing a molecular picture and mechanistic details. For instance, SMD simulations revealed the molecular basis for the plateau phase seen in fibrinogen force-extension curves[27] or that the mechanical stability of the titin I91 domain is due to contacts between β-strand pairs.[28,29]

Covalent mechanochemistry[30,31] and its effect on the stability of macromolecules[32,33] can not be modeled with SMD employing classical force fields. These phenomena have traditionally been studied by quantum-chemical methods, where a small molecule model system containing the disrupted bond is simulated by constrained geometry optimization[20,21,34–38] or by *ab initio* molecular dynamics.[31,38] Both approaches are computationally expensive and may not sample enough reactive events or orthogonal degrees of freedom. Such model and sampling limitations may hamper assignment of complete unfolding mechanisms and determination of rupture force spectra, distributions and the influence of non-reactive but structurally important regions on the activated reaction center. Reactive force fields have recently been introduced as computationally efficient descriptions of reactive systems.[39] But, to our knowledge, this method has not yet been applied to simulate force spectroscopy.

Here we introduce a simple molecular mechanical method to simulate forced unfolding of macromolecules coupled to covalent bond dissociation. The method may be used to depict the unfolding mechanism of a complete polymer with good sampling statistics of tens to hundreds of unfolding events. In the following Methods section, we describe the energy model composed of an empirical all-atom force field for non-reactive atoms and a Morse potential to describe the reactive covalent bonds. In the Results section, the model is applied to simulate the forced unfolding of rubredoxin in several polyprotein constructions and compared to a collection of AFM experiments previously available on equivalent systems. We discuss the sensitivity of mechanical properties to model parameters and the different mechanisms for complete unfolding obtained. We conclude the presented method is appropriate to simulate force spectroscopy and mechanochemical activation of macromolecules in detail.

2 Computational methods

2.1 Setup of molecular models

Two polymer models were used here: a single protein rubredoxin unit and a polyprotein with three units connected in line. These models are referred by $[A,B]_n$, where A and B indicate the position of point mutation to Cys in each rubredoxin unit that allow polyprotein connection, and $n = 1, 3$ indicates the number of units in each model. The crystal structure of ferric rubredoxin from *Pyrococcus furiosus* (PDB code 1BRF[40]) was used for both models. Mutations were introduced manually using PyMOL.[41] In the polyprotein, rubredoxin units were connected in the N-to-C orientation by a 1,2-diethoxyethane linker bound to S_γ of the mutated Cys residues in each unit, as shown schematically in Figure 2. In AFM experiments, polyproteins are composed of 3 to 6 rubredoxin units connected by maleimide-thiol cross-linking chemistry that covalently attaches the mutated Cys residues in each unit by a linker molecule similar to 1,2-diethoxyethane, but with unknown (probably random) relative orientation.[15]

Forced unfolding of $[RD1,49]_{1,3}$ and $[RD15,49]_{1,3}$ models begins with pre-dissociation of β-strands 1-3 (Fig. 1), as suggested by the observation of pre-peaks in experimental AFM force-extension profiles.[17] This was confirmed here by running exploratory pulling simulations (section 2.3) and observing rupture of hydrogen bonds between β-strands 1-3. Models $[RD1,49]_{1,3}$ and $[RD15,49]_{1,3}$ used here for the remaining results were built from an initial configuration with this β-strand already dissociated. For the other polyproteins, a pre-dissociation of the rubredoxin β-sheet was not observed experimentally[17] neither modelled here.

2.2 Force field and simulation details

Macromolecular interactions were described by the all-atom CHARMM27 empirical force field.[42,43] For the FeS center, covalent parameters for angles and dihedrals using the CHARMM functional form and Lennard–Jones parameters (ϵ_{Fe} =0.048 kJ/mol and σ_{Fe} =0.38 nm) were taken from previous work.[44] These covalent parameters are similar to other values proposed for rubredoxin.[45,46] Partial charges $q_{Fe} = 1.04$, $q_{S\gamma} = -0.45$ and $q_{C\beta} = -0.18$ for Cys were based on Mulliken population analysis of previous quantum chemical calculations on FeS mimetic compounds.[21,22] All unfolding simulations employed implicit solvation in the generalized Born

surface area (GB/SA) form.[47] This approximation was introduced to decrease the computational costs and allow extensive sampling of unfolding trajectories. The Still method was used to estimate Born radii,[48] and the nonpolar contribution was given by an uniform surface tension of 2.1 kJ $mol^{-1}nm^{-2}$.[49] Atomic radius for Fe was taken as 0.20 nm by comparison to other radii.

The four Fe–S bonds found in rubredoxin were represented by a Morse potential[50] to model bond dissociation during pulling simulations. Morse parameters were adjusted to quantum chemical calculations at the density functional level (DFT)[51] for the rate-limiting step of Fe–S dissociation in water for the $Fe(SCH_3)_4^-$ mimetic compound (Figure 5a in a previous publication[20]). The equilibrium bond distance set to b_0=0.23 nm was obtained from the isolated $Fe(SCH_3)_4^-$ optimized geometry. Steepness β=30 nm^{-1} and depth D_e=90 kJ/mol were adjusted to the position and energy barrier of the mimetic transition state for Fe–S bond dissociation.[20]

Initial single rubredoxin and polyprotein models were submitted to geometry optimization and molecular dynamics (MD) simulation during ∼200 ns with position restraints (50 kJ mol^{-1} nm^{-2}) applied to the first N–terminal and the last C–terminal mutated Cys residue. Configurations used to start the pulling simulations were collected at regular time intervals (4-10 ns) after stabilization of Cα root mean-squared deviation (RMSD).

Potential energy curves for Fe–S bond dissociation in the rubredoxin model were obtained by geometry optimization with the BFGS algorithm.[52,53] The Fe–S bond was scanned by restraining the Fe–S distance of Cys41 by a harmonic potential with force constant $k = 10^5$ kJ mol^{-1} nm^{-2}.

GROMACS 4.5[54] was used for all simulations and for building hydrogen atoms on the protein models. Dynamics were carried out at 300 K with a 2 fs time-step, with a leapfrog stochastic dynamics integrator and a friction τ =10 ps^{-1}. Covalent hydrogen bonds were constrained with LINCS.[55]

2.3 Forced unfolding simulations

Forced unfolding trajectories were obtained using SMD simulations.[25] A time-dependent harmonic potential ($V[\xi]$) is added to the system energy function to mimic protein pulling by the

cantilever or pulling tip on force spectroscopy experiments:

$$V[\xi(t)] \quad = \quad \frac{k_p}{2}[\xi(t) - \xi_0(t)]^2 \tag{1}$$

where the reference value of the progress coordinate, ξ_0, changes linearly in time:

$$\xi_0(t) = \xi(0) + v_p t \tag{2}$$

Simulations with constant velocity were performed with $v_p = 10^{-1}$ m/s and pulling force constant $k_p = 83$ pN/nm. Structures and forces were saved every 5 ps for analysis.

The progress coordinate ξ was defined as the distance between the Cα in a reference center and in a pulling center. When pulling by the C–terminal, the reference center was set to the first N–terminal mutated Cys residue and the pulling center was set to the last C–terminal mutated Cys. For example, in model $[\mathrm{RD15,49}]_3$ the reference center was C15 in the first rubredoxin unit and the pulling center was C49 in the third rubredoxin unit. Reference and pulling centers were exchanged when pulling by the N–terminal.

SMD simulations were initialized with a different random seed for stochastic dynamics and a different initial structure (section 2.2). Between 35 to 40 unfolding simulations were obtained for each polyprotein model, resulting in N=105-120 complete unfolding events. A Fe–S bond with distance higher than 0.37 nm was considered broken. After each bond rupture, the simulation was paused and the associated angle and dihedral contributions were removed from the system topology. The trajectory was then continued from the same geometry.

Secondary structural content along trajectories was analyzed by counting the involved hydrogen bonds. The sheets between $\beta-$strands 1-2 and 1-3 (Fig. 1) were considered formed when at least two of the hydrogen bonds found in the crystal structure were present.

2.4 Calculation of contour lengths

Polyprotein simulations were used for calculation of contour length increments after a force peak or unfolding event. The increment is $\Delta\mathrm{Lc} = \mathrm{Lc(u)} - \mathrm{Lc(f)}$, where Lc is the contour length or maximum extension between the anchoring points in the folded (f) and unfolded (u) protein. Contour lengths will be labeled by a superscript indicating values obtained from sequences and

crystal structures (PDB), SMD simulations or AFM experiments. Anchoring points are the points of force application along the polypeptide and are defined as the aminoacids enclosing the protein region unfolded in the corresponding peak in force-extension profiles[17] (see Table 2 for models studied here).

As in previous works,[15,17] Lc(f)^{PDB} was calculated as the through-space distance between the Cα of anchoring points in the crystal structure. On the other hand, Lc(u)^{PDB} was estimated as $(n-1)*0.365$ nm, where n is the number of aminoacids between anchoring points and 0.365 nm is the average contribution in length per aminoacid.[56]

Lc(f)^{SMD} was calculated as the through space distance between the Cα of anchoring points one frame before the first Fe–S bond rupture, preceeding protein unfolding indicated by the respective force peak in the force-extension profiles. Lc(u)^{SMD} was calculated as the distance between the same Cα in the frame before the next force peak is achieved. In AFM experiments, ΔLc^{AFM} is obtained by fitting the unfolding peaks in force-extension profiles to the worm like chain (WLC) model[57] and calculating the difference between fitted Lc values from successive peaks.

3 Results

3.1 Analysis of simulation parameters

A sensitivity analysis of calculated mechanical properties due to variations in simulation parameters is presented on this section for the rubredoxin model [RD1,49]$_1$, except when noted. Only one type of parameter was varied for each set of simulations shown below while the others remained with values given on the Methods section.

Table 1 shows that average rupture forces strongly depend on the Morse potential depth and steepness, and on partial charges for FeS atoms, particularly when more polarized charges are used. Average forces are clearly more sensitive to changes in the Morse depth. Increasing the depth by 1.5 fold increases the forces by almost 6 fold. More polarized charges lead to lower rupture forces probably due to stabilization by the dielectric solvent when the FeS center is exposed.

Potential energy profiles for Fe–S bond rupture are affected by the same force field param-

eters (D_e, β and $q_{Fe/S}$ in Fig. S1, in the Supporting Information). This is expected given the derivative relation between dissociation energies and forces. The Morse steepness determines both the profile inclination and the Fe–S distance for bond dissociation.

On the other hand, Lennard–Jones σ_{Fe} and ϵ_{Fe} parameters, and atomic radius for Fe used in GB calculations have a small influence on potential energy curves for Fe–S bond dissociation (Fig. S2), except when low GB radius and σ were tested which lead to premature Fe–S dissociation before force application. Values for these parameters were chosen based on previous calibrations[21,44] to avoid strong interactions between Fe and negative side-chains, and an unbalanced solvation contribution.

Figure 3 shows that the pulling velocity (v_p, equation 2) dramatically changes the distribution of rupture forces for the Fe–S bond. For very high velocities (v_p=10 m/s), the distribution depends on the pulling direction. Decreasing the velocity removes this dependency, but only velocities down to $v_p \leq$0.1 m/s result in a force distribution with the expected bell shape[26,58] and a standard deviation compatible with the experimental one ($\sim$150 pN).[15]

Average rupture forces obtained with a pulling velocity of v_p=0.1 m/s are similar between model $[RD1,49]_1$ ($\bar{F}$=758 $\pm$ 127 pN, N=11 unfolding events) and model $[RD1,49]_3$ pulled by either the C–terminal ($\bar{F}$=738 $\pm$ 93 pN, N=69) or by the N–terminal ($\bar{F}$=744 $\pm$ 86 pN, N=33). Mechanistic details and the sequence of structural changes for complete unfolding (section 3.3) are also equivalent between models $[RD1,49]_1$ and $[RD1,49]_3$.

The pulling force constant (k_p) does not significantly affect the rupture forces (Table S1), but controls the shape of force-extension curves (Fig. S3). High force constants result in distinct peaks for each Fe–S bond rupture and hence, multiple peaks for a complete protein unfolding event. One peak followed by a continuos relaxation curve, as observed experimentally,[15,17] is obtained from simulation only when decreasing the force constant to about twice the experimental value (40 pN/nm).[15]

Figure 4 shows that the protein extension measured from termini Cα distances follows the harmonic pulling potential or reference coordinate (ξ_0, equation 2) only when a high pulling force constant is used. When a low force constant similar to the experimental value is used, the protein extension does not strictly follow the reference pulling coordinate. Figure 4 also shows the maximum force in force-extension profiles is reached just before the first Fe–S bond

rupture.

Slower pulling velocities and lower force constants considerably increase the amount of computer time necessary for simulating protein unfolding. For instance, complete unfolding of the $[RD1,49]_1$ model with a pulling velocity $v_p = 10^{-2}$ m/s takes about 3 weeks of wall-clock time on modern Intel Xeon processors using 4 cores, which is the highest number of cores we could sustain a decent parallel scalability for such a molecular model with a relatively small number of particles.

3.2 Comparison between SMD and AFM data

A comparison between unfolding forces observed in SMD simulations and AFM experiments is not straightforward as simulations have to be conducted in much higher pulling velocities than AFM, except in (rare) experiments when high speeds are realized.[7,59] For rubredoxin unfolding, there is a gap of almost 4 orders of magnitude between our slowest pulling simulation and the fastest measurement,[17] as shown for the force spectrum in Figure 5.

However, the dependency of unfolding forces on pulling velocities has been studied by theory in detail.[26,58,60–63] In Fig. 5, we also show adjustment of the full microscopic theoretical model derived by Hummer and Szabo[58] to fit SMD and AFM force spectra for rubredoxin unfolding. The fit quality is rather good with a mean deviation of 22 pN between the model line and data points. The theoretical model predicts an unfolding barrier position $\Delta x^{\ddagger}$=0.17 nm and a spontaneous unfolding rate k_0=0.22 s^{-1}, in good agreement with the barrier position and spontaneous rate obtained by fitting the AFM data alone either with the same theoretical model ($\Delta x^{\ddagger}$=0.11 nm and k_0=1.0 s^{-1}) or with a simplified phenomenological theory[61] ($\Delta x^{\ddagger}$=0.11 nm and k_0=0.15 s^{-1}).[17]

Additionally, by assuming the microscopic theoretical model[58] is valid, the force field parameters (for instance, Morse β and D_e as shown in Figure S4) can be refined by comparison of the experimental AFM force spectra with SMD simulations obtained with different sets of parameters.

Comparison between simulated and AFM contour length increments in Table 2 shows excellent agreement for polyproteins [RD15,49] and [RD15,35], and good agreement for polyprotein [RD1,49], considering one standard deviation and the two experimental results available.[15,17]

Simulated and AFM increments only disagree for [RD1,35].

It should be noted that the procedure based on WLC fits to obtain ΔLc from AFM data and that based on Cα distances described in Section 2.4 give equivalent results when applied to the same set of simulated force-extension profiles. For example, the same set of N=8 simulated unfolding events for model $[RD1,49]_3$ resulted in ΔLc^{SMD} =12.2$\pm$0.5 nm when using Cα distances and ΔLc^{SMD} =12.5$\pm$0.7 nm when using WLC fits. Equivalent results are observed for other rubredoxin polyproteins.

Table 2 shows that contour length increments expected from crystal structures (ΔLc^{PDB}) are overestimated in comparison with the increments obtained from simulations (ΔLc^{SMD}) for all polyprotein models, except $[RD1,49]_3$. Closer inspection indicates that $Lc(u)^{PDB}$ and $Lc(u)^{SMD}$ are in very good agreement. Thus, discrepancies in ΔLc increments are due to underestimated $Lc(f)^{PDB}$ values, based on the unperturbed crystal structure.

3.3 Microscopic mechanism for forced macromolecular unfolding

Pulling simulations of polyprotein models were also used to describe the detailed mechanism of rubredoxin mechanical unfolding. In the beginning of all simulations, linker and protein regions connecting rubredoxin units are first extended without much perturbation in the rest of the molecular structure (time t <50 ns in Fig. 4b). Then, tension starts to build up and little extension is gained until the Fe–S bonds are broken (50< t <250 ns, in Fig. 4b). During this second phase, structural fluctuations are relatively small in the regions under tension and in between anchoring points (see these regions depicted in Fig. 2).

Table 3 shows the stability of secondary structures and salt bridges found in rubredoxin along pulling simulations. Contacts between $\beta-$strands 1-2 are preserved before Fe–S bond rupture in all $[RD1,49]_3$ and $[RD1,35]_3$ simulations, but are disrupted in all $[RD15,49]_3$ and $[RD15,35]_3$ simulations. Hydrogen bonds between $\beta-$strands 1-3 are preserved before Fe–S bond rupture in half of the simulations for $[RD15,35]_3$ and for the majority of $[RD1,35]_3$ simulations. The stability of salt bridges A1-E14 and K6-E49 follows the same qualitative pattern, whereas less stable. Notice that the wild-type salt bridge between residues 6-49 is not present in $[RD1,49]_3$ and $[RD15,49]_3$ due to the E49C mutation.

Exploratory pulling simulations showed that contacts between $\beta-$strands 1-3 are disrupted

early for models $[RD1,49]_3$ and $[RD15,49]_3$, during the first extension phase described above. Low force peaks corresponding to dissociation of hydrogen bonds between $\beta-$strands 1-3 are observed in the simulated force-extension profiles only for these models ($[RD1,49]_3$ and $[RD15,49]_3$, data not shown), in agreement with experimental AFM data.[17]

Cooperativity in the stability of salt bridges and secondary structures before Fe–S bond rupture also depends on the polyprotein connectivity. For instance, in model $[RD15,35]_3$ the survival of interactions overlap for the salt bridge A1-E14 and the hydrogen bonds holding together $\beta-$strands 1-2, meaning that these contacts are disrupted simultaneously (Fig. S5). But in model $[RD15,49]_3$, the salt bridge is disrupted first, followed sequentially by disruption of hydrogen bonds in $\beta-$strands 1-2. When salt bridges and secondary structures are disrupted before Fe–S bond dissociation, disruption is observed at the first half of the simulation time necessary to break the Fe–S bond (Fig. S5).

Hydrogen bonds are found in the rubredoxin crystal structure (Fig. 1) between S_γ in the FeS center and backbone amides. S_γCys41 makes one hydrogen bond which is broken before Fe–S bond rupture during all pulling simulations for models $[RD1,49]_3$ and $[RD15,49]_3$. Tension is applied to the Fe–S_γCys41 bond on these models. S_γCys8 also makes one hydrogen bond and it is equivalently disrupted in $[RD15,35]_3$ and $[RD15,49]_3$. On the other hand, S_γCys5 and S_γCys38 make two hydrogen bonds each and these two are preserved before Fe–S bond rupture in all pulling simulations, except for 20% of the $[RD1,35]_3$ trajectories. Consequently, solvent access is considerably higher near S_γCys41 in $[RD1,49]_3$ and $[RD15,49]_3$, and near S_γCys8 in $[RD15,35]_3$ and $[RD15,49]_3$. For polyprotein $[RD1,35]_3$, solvent access is similar for both S_γCys5 and S_γCys38.

4 Discussion

We have presented a molecular model to study the mechanical unfolding of a macromolecule coupled to covalent bond rupture and shown its application to simulate the forced unfolding of the rubredoxin metalloprotein.

A sensitivity analysis of calculated properties shows that parameters for the reactive center describing partial atomic charges and the Morse potential for bond dissociation have to be carefully calibrated as they have a large influence on rupture forces (Table 1 and Fig. S1). Here,

these parameters were adjusted to quantum chemical calculations at the DFT level performed on an isolated mimetic molecule and further refined in comparison to experimental data. In particular, the force spectra can be used to distinguish which parameter set better fits a microscopy theoretical model in comparison to AFM data (Fig. S4). Less sensitive parameters such as those describing Lennard-Jones interactions can be retrieved from force fields previously parametrized for equilibrium properties.

This model for bond dissociation and the proposed approach for calibration are only viable if the chemical bonds disrupted upon mechanical unfolding are previously known or at least suggested. If two or more bonds can dissociate, each of them would have to be calibrated to a characteristic set of Morse parameters, as often done in reactive force fields.[39] Here the four Fe–S bonds found in rubredoxin were proposed to dissociate and treated with the same Morse potential. Of course it is expected that the force field description for the remaining non-reactive atoms in the macromolecule can properly model their interactions.[42,43]

The major limitation of our energy model is the classical description of the disrupted bond. The lack of an electronic structure or quantum chemical description prevents the inclusion of charge and spin reorganization effects in the simulation. These effects are important to model reactions involving organic and metal centers[21] and to discern details of reaction mechanisms. Thus, we are unable to use the rubredoxin simulations shown here to distinguish the order that each of the four possible Fe–S bonds are broken to complete protein unfolding.

The effect of the pulling force constant (k_p) on the simulated force-extension curves can be rationalized. When a high constant is used, protein extension strictly follows the harmonic pulling reference coordinate (Fig. 4a) and samples only local forces, similar to the drift regime proposed before.[26] Complete unfolding is observed after multiple bond rupture steps, leading to multiple force peaks in force-extension curves. When a low force constant is used, the protein extension can fluctuate more and does not strictly follow the harmonic pulling coordinate (Fig. 4b), similar to an activated regime.[26] The reaction coordinate may also sample thermal forces and more tension may accumulate. Complete unfolding is observed in one step, leading to one force peak in the force-extension curves, which corresponds to the first Fe–S bond rupture, quickly followed by disruption of other Fe–S bonds.

The force spectrum obtained from SMD and AFM data on rubredoxin unfolding was well

adjusted here by the full microscopic model proposed by Hummer and Szabo[58] that assumes the rupture energy is described by a cusp potential. Although similar theoretical models based on smooth potentials have also been proposed,[63] their adjustment to force spectra obtained at high speeds did not lead to significant changes on the derived barrier position and unfolding rate.[7] This suggests some independence of adjusted parameters on the exact form of the rupture potential. Thus, the microscopic model[58] can be employed to fit the force spectrum when unfolding is coupled to covalent bond dissociation as modelled here by a smooth Morse potential.

Simulated and experimental contour length increments are in very good agreement for three of the four polyproteins studied here. Contour lengths before unfolding [Lc(f)] have traditionally been estimated from through-space distances between anchoring points on the macromolecular crystal structure. This may not be a reliable practice as our comparison of $\mathrm{Lc(f)}^{PDB}$ and $\mathrm{Lc(f)}^{SMD}$ suggests. The coupling of protein units and force application before bond rupture perturb the structure of each unit and hence, the distance between anchoring points (Table 2). On the other hand, construction and simulation of polyprotein models requires more labor than inspecting the crystal structures.

Notice the excellent agreement between contour lengths after unfolding Lc(u) obtained from SMD simulations and estimated from the formula $(n-1)*0.365$ nm, where n is the number of aminoacids between anchoring points (Table 2). Here, $(n-1)$ is used in the formula instead of n adopted before[15,17] because the combined extension of the two residues in anchoring points will contribute only one length unit to Lc(u).

Cavagnero *et al.*[64] proposed a mechanism for thermal denaturation of rubredoxin in three steps: rubredoxin first loses part of the secondary structure; Fe–S bonds are broken, iron is released and more secondary structure is lost; and the hydrophobic core is exposed leading to the unfolded state. This is based on data obtained by several optical spectroscopy methods at low pH, when Glu14 and Glu49 side chains should be protonated and their respective salt bridges be broken.[65] Nevertheless, this proposed mechanism is roughly the sequence we obtain from mechanical unfolding simulations in polyproteins $[\mathrm{RD15,35}]_3$ and $[\mathrm{RD1,49}]_3$. Solvent exposure of the hydrophobic core varies for each polyprotein model, but it is also correlated with disruption of secondary structure.

The mechanism of macromolecular unfolding simulated here clearly depends on the points

of force application along the polymer. Salt-bridges and hydrogen bonds holding secondary structures together may be completely disrupted before covalent bond dissociation, as seen in rubredoxin model $[RD15,49]_3$, or almost entirely preserved, as seen in model $[RD1,35]_3$. The fraction of these intramolecular contacts present before Fe–S bond rupture in the four polyprotein models tested here increases in the order $[RD15,49]_3 < [RD15,35]_3 < [RD1,49]_3 < [RD1,35]_3$. This is approximately the same order found for the average rupture force and the reverse order found for the intrinsic rate of unfolding k_0 on AFM measurements,[17] suggesting that the stability of intramolecular non-covalent contacts plays a role on macromolecules subject to mechanochemical activation.

Solvent exposure of the macromolecular interior and water access to the reactive FeS center in rubredoxin are controlled by partial protein unfolding, with disruption of secondary structures and of native hydrogen bonds between S_γ and backbone amides. Our simulations show that water penetration is higher near S_γCys41 and S_γCys8, suggesting the respective Fe–S bonds would be more reactive than the other two. In fact, it has been shown by quantum-chemical calculations that water substitution leads to faster Fe–S bond cleavage in rubredoxin models.[20]

Thus, the mechanical anisotropy previously observed in rubredoxin polyproteins[17] may not be only due to differences in the intrinsic stabilities among the four Fe–SCys bonds,[15,17] but also because of the variable degree of solvent access to the FeS center found here between different polyproteins.

Two types of mechanism for Fe–S bond rupture in mechanical unfolding of rubredoxin have been observed for a structural variant of rubredoxin.[18] A concurrent process, where multiple Fe–S bonds rupture simultaneously was observed in 80% of AFM force-extension profiles. And a sequential mechanism, where rupture of different Fe–S bonds can be individually distinguished, was observed in the other 20% of AFM profiles. Observation of simultaneous processes depends on the time resolution of measurements, which is in the order of μs or slower for the mentioned AFM experiments. In our simulations, bond ruptures in the same protein unit are separated by tens of ps. Therefore, within a μs time window, bond rupture occurs simultaneously in all simulations. It is possible that the sequential process observed in the AFM experiments is due to the protein construction containing an extra unnatural loop elongation in the studied rubredoxin variant.[18]

We conclude that the molecular mechanical model presented here may be applied to study the forced unfolding of macromolecules coupled to covalent bond rupture. Empirical force field parameters for bond dissociation can be obtained from quantum-chemical calculations on model compounds and further refined in comparison to experimental force spectra. SMD simulations revealed the mechanism of macromolecular unfolding and the sequence that intramolecular contacts are disrupted for four different polyproteins. Solvent penetration near the reactive center may be determinant for the mechanical stability of each polymer.

The simulation methods presented here are not limited to proteins. They can be applied to simulate forced unfolding and mechanochemical activation of any macromolecule for which an appropriate model of the molecular structure and a set of cleavable covalent bonds are known.

Given the lack of an electronic structure description of the reactive center in our classical model, the detailed mechanism and sequence of Fe–S bond rupture in the stretched rubredoxin models could not be analyzed. These would require a hybrid quantum chemical/molecular mechanical energy model which we have been investigating in our laboratory.[21] Nevertheless, unfolding trajectories extensively sampled here with a computationally cheap method may be of great value as initial reactive conformations in future studies.

Acknowledgement

We acknowledge funding from FAPESP (projects 2014/17008-7, 2014/21900-2 and 2016/24096-5) and CNPq (projects 141950/2013-7 and 306133/2015-6).

References

(1) Fisher, T. E.; Oberhauser, A. F.; Carrion-Vazquez, M.; Marszalek, P. E.; Fernandez, J. M. *Trends Biochem. Sci.* **1999**, *24*, 379–384.

(2) Neuman, K. C.; Nagy, A. *Nat. Methods* **2008**, *5*, 491–505.

(3) Passeri, D.; Rossi, M.; Tamburri, E.; Terranova, M. L. *Anal. Bioanal. Chem.* **2013**, *405*, 1463–1478.

(4) Kilpatrick, J. I.; Revenko, I.; Rodriguez, B. J. *Adv. Healthc. Mater.* **2015**, *4*, 2456–2474.

(5) Schönfelder, J.; Sancho, D. D.; Perez-Jimenez, R. *J. Mol. Biol.* **2016**, *428*, 4245–4257.

(6) Franco, I.; Ratner, M. A.; Schatz, G. C. In *Nano and cell mechanics: fundamentals and frontiers*; Espinosa, H. D., Bao, G., Eds.; John Wiley & Sons: Chichester, UK, 2013; Chapter 14, pp 359–388.

(7) Rico, F.; Gonzalez, L.; Casuso, I.; Puig-Vidal, M.; Scheuring, S. *Science* **2013**, *342*, 741–743.

(8) Ando, T.; Uchihashi, T.; Scheuring, S. *Chem. Rev.* **2014**, *114*, 3120–3188.

(9) Rajendran, A.; Endo, M.; Sugiyama, H. *Chem. Rev.* **2014**, *114*, 1493–1520.

(10) Schwaderer, P.; Funk, E.; Achenbach, F.; Weis, J.; Bräuchle, C.; Michaelis, J. *Langmuir* **2008**, *24*, 1343–1349.

(11) Karthikeyan, S.; Potisek, S. L.; Piermattei, A.; Sijbesma, R. P. *J. Am. Chem. Soc.* **2008**, *130*, 14968–14969.

(12) Piermattei, A.; Karthikeyan, S.; Sijbesma, R. P. *Nat. Chem.* **2009**, *1*, 133–137.

(13) Wei, W.; Sun, Y.; Zhu, M.; Liu, X.; Sun, P.; Wang, F.; Gui, Q.; Meng, W.; Cao, Y.; Zhao, J. *J. Am. Chem. Soc.* **2015**, *137*, 15358–15361.

(14) Li, Y.; Wen, J.; Qin, M.; Cao, Y.; Ma, H.; Wang, W. *ACS Biomater. Sci. Eng.* **2017**, *3*, 979–989.

(15) Zheng, P.; Li, H. *J. Am. Chem. Soc.* **2011**, *133*, 6791–6798.

(16) Zheng, P.; Takayama, S.-I. J.; Mauk, A. G.; Li, H. *J. Am. Chem. Soc.* **2012**, *134*, 4124–4131.

(17) Zheng, P.; Chou, C.-C.; Guo, Y.; Wang, Y.; Li, H. *J. Am. Chem. Soc.* **2013**, *135*, 17783–17792.

(18) Zheng, P.; Takayama, S.-I. J.; Mauk, A. G.; Li, H. *J. Am. Chem. Soc.* **2013**, *135*, 7992–8000.

(19) Zheng, P.; Wang, Y.; Li, H. *Angew. Chem. Int. Ed. Engl.* **2014**, *53*, 14060–14063.

(20) Zheng, P.; Arantes, G. M.; Field, M. J.; Li, H. *Nat. Commun.* **2015**, *6*, 7569.

(21) Arantes, G. M.; Bhattacharjee, A.; Field, M. J. *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 8144–8146.

(22) Arantes, G. M.; Field, M. J. *J. Phys. Chem. A* **2015**, *119*, 10084–10090.

(23) Zheng, P.; Cao, Y.; Li, H. *Langmuir* **2011**, *27*, 5713–5718.

(24) Li, Y. D.; Lamour, G.; Gsponer, J.; Zheng, P.; Li, H. *Biophys. J.* **2012**, *103*, 2361–2368.

(25) Grubmuller, H.; Heymann, B.; Tavan, P. *Science* **1996**, *271*, 997–999.

(26) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568–1581.

(27) Lim, B. B. C.; Lee, E. H.; Sotomayor, M.; Schulten, K. *Structure* **2008**, *16*, 449–459.

(28) Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K. *Biophys. J.* **1998**, *75*, 662–671.

(29) Lee, E. H.; Hsin, J.; Sotomayor, M.; Comellas, G.; Schulten, K. *Structure* **2009**, *17*, 1295–1306.

(30) Liang, J.; Fernández, J. M. *ACS Nano* **2009**, *3*, 1628–1645.

(31) Ribas-Arino, J.; Marx, D. *Chem. Rev.* **2012**, *112*, 5412–5487.

(32) Groote, R.; Jakobs, R. T. M.; Sijbesma, R. P. *Polym. Chem.* **2013**, *4*, 4846–4859.

(33) Makarov, D. E. *J. Chem. Phys.* **2016**, *144*, 030901.

(34) Lupton, E. M.; Nonnenberg, C.; Frank, I.; Achenbach, F.; Weis, J.; Bräuchle, C. *Chem. Phys. Lett.* **2005**, *414*, 132–137.

(35) Barros, T. C.; Yunes, S.; Menegon, G.; Nome, F.; Chaimovich, H.; Politi, M. J.; Dias, L. G.; Cuccovia, I. M. *J. Chem. Soc. Perkin Trans. 2* **2001**, *12*, 2342–2350.

(36) Groote, R.; Szyja, B. M.; Pidko, E. A.; Hensen, E. J. M.; Sijbesma, R. P. *Macromolecules* **2011**, *44*, 9187–9195.

(37) Kochhar, G. S.; Heverly-Coulson, G. S.; Mosey, N. J. In *Polymer mechanochemistry*; Boulatov, R., Ed.; Springer International Publishing, 2015; pp 37–96.

(38) Stauch, T.; Dreuw, A. *Chem. Rev.* **2016**, *116*, 14137–14180.

(39) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; van Duin, A. C. T. *NPJ Comput. Mater.* **2016**, *2*, 15011.

(40) Bau, R.; Rees, D. C.; Kurtz, D. M.; Scott, R. A.; Huang, H.; Adams, M. W. W.; Eidsness, M. K. *J. Biol. Inorg. Chem.* **1998**, *3*, 484–493.

(41) Schrödinger, LLC, PyMOL, The PyMOL Molecular Graphics System, Version 1.2r2.

(42) 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. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(43) MacKerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(44) Chang, C. H.; Kim, K. *J. Chem. Theory Comput.* **2009**, *5*, 1137–1145.

(45) Carvalho, A. T. P.; Teixeira, A. F. S.; Ramos, M. J. *J. Comput. Chem.* **2013**, *34*, 1540–1548.

(46) Yelle, R. B.; Park, N.-S.; Ichiye, T. *Proteins* **1995**, *22*, 154–167.

(47) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(48) Qui, D.; Shenkin, P.; Hollinger, F.; Still, W. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(49) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835–848.

(50) Morse, P. M. *Phys. Rev.* **1929**, *34*, 57–64.

(51) Parr, R. G.; Yang, W. *Density-functional theory of atoms and molecules*, 1st ed.; Oxford University Press, 1996.

(52) Byrd, R. H.; Lu, P.; Nocedal, J. *SIAM J. Scientif. Statistic. Comput.* **1995**, *16*, 1190–1208.

(53) Zhu, C.; Byrd, R. H.; Nocedal, J. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560.

(54) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845–854.

(55) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(56) Dietz, H.; Rief, M. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1244–1247.

(57) Marko, J. F.; Siggia, E. D. *Macromolecules* **1995**, *28*, 8759–8770.

(58) Hummer, G.; Szabo, A. *Biophys. J.* **2003**, *85*, 5–15.

(59) Yu, H.; Siewny, M. G. W.; Edwards, D. T.; Sanders, A. W.; Perkins, T. T. *Science* **2017**, *355*, 945–950.

(60) Bell, G. *Science* **1978**, *200*, 618–627.

(61) Evans, E.; Ritchie, K. *Biophys. J.* **1997**, *72*, 1541–1555.

(62) Rief, M.; Fernandez, J. M.; Gaub, H. E. *Phys. Rev. Lett.* **1998**, *81*, 4764–4767.

(63) Dudko, O. K.; Hummer, G.; Szabo, A. *Phys. Rev. Lett.* **2006**, *96*, 108101.

(64) Cavagnero, S.; Zhou, Z. H.; Adams, M. W. W.; Chan, S. I. *Biochemistry* **1998**, *37*, 3377–3385.

(65) Cavagnero, S.; Zhou, Z. H.; Adams, M. W. W.; Chan, S. I. *Biochemistry* **1995**, *34*, 9865–9873.

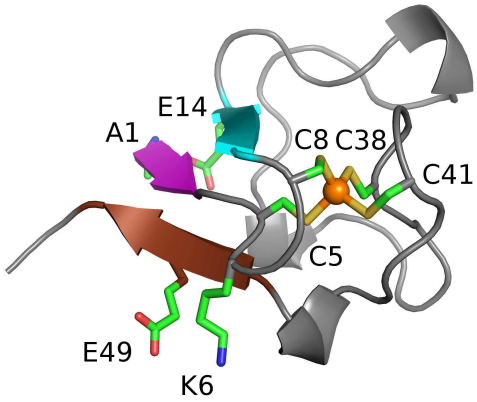


Figure 1: Rubredoxin crystal structure from *P. furiosus*.[40] Fe is shown in orange, Cys bound to Fe and residues in salt bridges are shown as sticks. β-strands 1, 2 and 3 are indicated in pink, cyan and brown respectively.

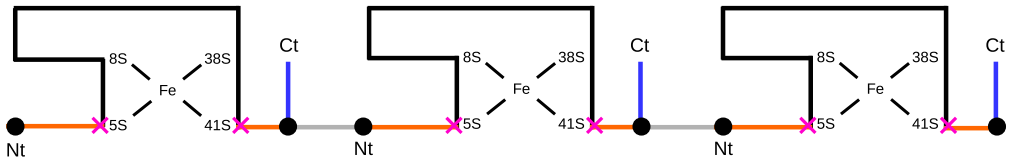


Figure 2: Schematic model of the polyprotein $[RD1,49]_3$ used in the simulations. Rubredoxin units are connected by a linker (in gray). Black circles indicate the points of mutation (residues 1 and 49), pink crosses indicate the anchoring points (residues 5 and 41), orange lines indicate the regions under tension, blue lines indicate the regions outside the points of force application, black lines indicate the regions in between anchoring points, Nt and Ct indicate the protein terminals. Positions of Fe and native S_γ centers are also indicated.

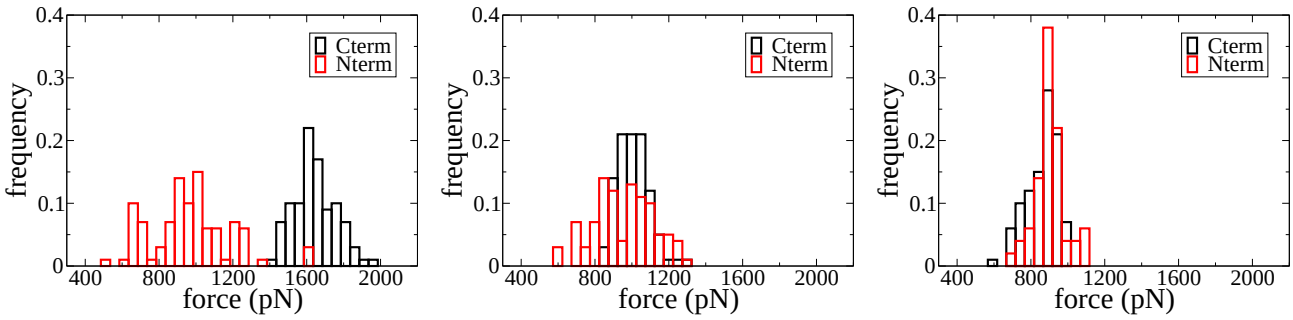
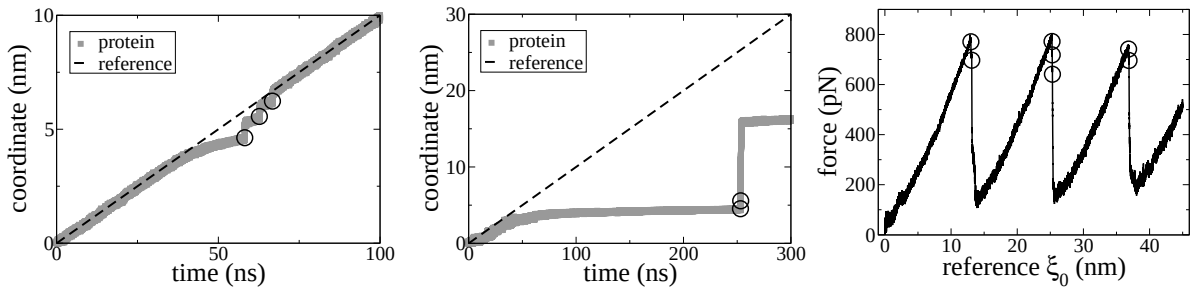


Figure 3: Distribution of rupture forces obtained for N=70-100 simulations with pulling velocities $v_p = 10$ m/s (panel **a**), $v_p = 1$ m/s (panel **b**) and $v_p = 10^{-1}$ m/s (panel **c**) of model $[RD1,49]_1$.



(a) (b) (c)

Figure 4: Time evolution of the reference progress coordinate (ξ_0, equation 2) and the protein extension coordinate for model $[\mathrm{RD1,49}]_1$ obtained with a pulling force constant $k_p = 1667$ pN/nm (panel **a**) and $k_p = 83$ pN/nm (panel **b**). One representative force-extension curve simulated with $k_p = 83$ pN/nm for model $[\mathrm{RD1,49}]_3$ is shown in panel **c**. Circles indicate Fe–S bond rupture.

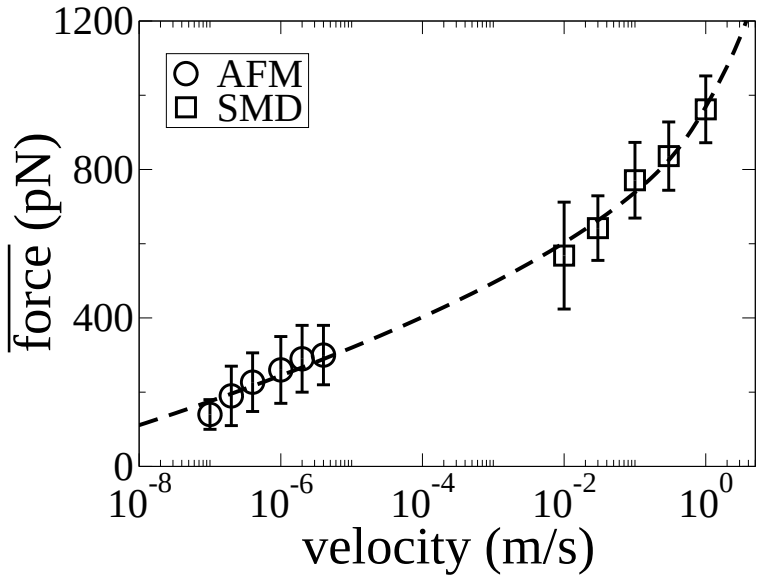


Figure 5: Force spectra or average rupture forces obtained at different pulling velocities for AFM experiments[17] and SMD simulations (N=20) of model $[\mathrm{RD1,49}]_1$. The dashed line shows the full microscopic model derived by Hummer and Szabo[58] fitted to both AFM and SMD data. Error bars in SMD points indicate a standard deviation.

Table 1: Average rupture forces ($\bar{F}$, average ± standard deviation in pN) calculated for the first Fe–S bond rupture in $[RD1,49]_1$ (N=10 simulations) with different values of depth (D_e, kJ/mol) and steepness (β, nm^{-1}) of the Morse potential, and partial charges q_{Fe} and q_S assigned to the FeS center. Only one type of parameter was changed in each set of simulations.

D_e	$\bar{F}$	β	$\bar{F}$	q_{Fe}	q_S	$\bar{F}$
70	415 ± 78	15	348 ± 60	1.44	-0.55	525 ± 42
90	758 ± 127	20	530 ± 35	1.04	-0.45	758 ± 127
130	1632 ± 57	30	758 ± 127	0.64	-0.35	995 ± 75
170	2437 ± 118	40	1007 ± 88	0.24	-0.25	1163 ± 65

Table 2: Contour lengths (Lc, average ± standard deviation in nm) before (f) and after (u) polyprotein unfolding and increments (ΔLc) obtained from the crystal structure (Lc^{PDB}), AFM experiments (Lc^{AFM})[17] and SMD simulations (Lc^{SMD}, N=70-80 values). Residue number of anchoring points is also indicated.

model	ΔLc^{PDB}	ΔLc^{SMD}	ΔLc^{AFM}		
$[RD1,49]_3$	12.2	12.0 ± 0.1	13.0 ± 0.8[17] and 12.6 ± 1.3[15]		
$[RD15,49]_3$	7.4	6.0 ± 0.1	6.4 ± 0.7		
$[RD15,35]_3$	5.3	2.7 ± 0.1	2.7 ± 0.4		
$[RD1,35]_3$	10.4	9.0 ± 0.1	11.1 ± 1.2		
	$Lc(f)^{PDB}$	$Lc(f)^{SMD}$	$Lc(u)^{PDB}$	$Lc(u)^{SMD}$	anchoring points
$[RD1,49]_3$	0.9	1.0 ± 0.1	13.1	13.0 ± 0.1	5, 41
$[RD15,49]_3$	2.1	3.3 ± 0.1	9.5	9.3 ± 0.1	15, 41
$[RD15,35]_3$	2.0	4.5 ± 0.1	7.3	7.2 ± 0.1	15, 35
$[RD1,35]_3$	2.0	3.3 ± 0.1	12.4	12.3 ± 0.1	1, 35

Table 3: Fraction of proteins with salt bridges and secondary structure preserved before the first Fe–S bond rupture along pulling simulations (N=105-120 unfolding events).

model	salt bridges		β-sheet	
	A1-E14[a]	K6-E49	strands 1-2	strands 1-3
$[RD1,49]_3$	0.34	-	1.00	-
$[RD15,49]_3$	0.00	-	0.00	-
$[RD15,35]_3$	0.00	0.37	0.00	0.52
$[RD1,35]_3$	0.38	0.67	1.00	0.81

[a] C1-E14 in $[RD1,49]_3$ and $[RD1,35]_3$. Residue and β-sheet numbering and positions are shown in Figure 1. Salt bridge K6-E49 and the contact between β-strands 1-3 were absent by construction in models $[RD1,49]_3$ and $[RD15,49]_3$.

Supporting Information

Mechanical unfolding of macromolecules coupled to bond dissociation

Ariane Nunes-Alves and Guilherme Menegon Arantes

Department of Biochemistry, Instituto de Química, Universidade de São Paulo,

Av. Prof. Lineu Prestes 748, 05508-900, São Paulo, SP, Brazil

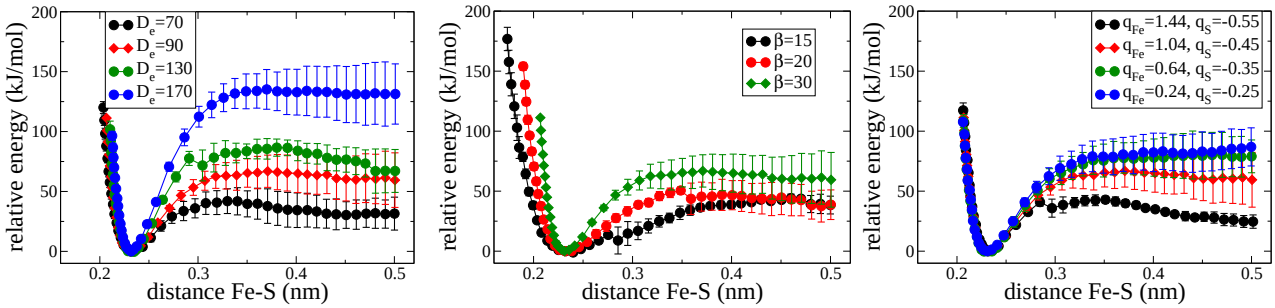
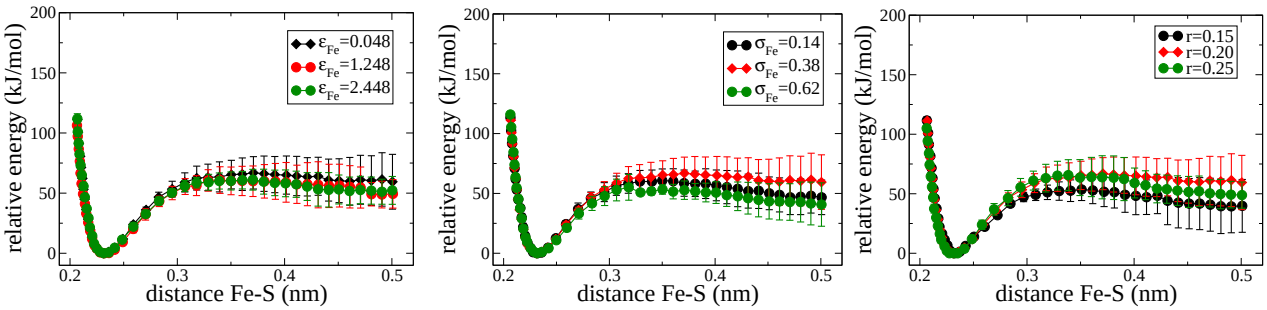


Figure S1: Potential energy profiles of Fe–SCys41 bond dissociation obtained for different Morse depths D_e (in kJ/mol, panel **a**), Morse steepness β (in nm^{-1}, panel **b**) and partial charges assigned for Fe and Cys S_γ (panel **c**). Diamond symbols correspond to values used for unfolding simulations. Error bars indicate one standard deviation for N=3 different starting geometries.



(a) (b) (c)

Figure S2: Potential energy profiles of Fe–SCys41 bond dissociation obtained for different Lennard–Jones parameters ϵ_{Fe} (kJ/mol, panel **a**) and σ_{Fe} (in nm, panel **b**) and atomic radius for GB calculations for Fe (in nm, panel **c**). Diamond symbols correspond to values used for unfolding simulations. Error bars indicate one standard deviation for N=3 different starting geometries.

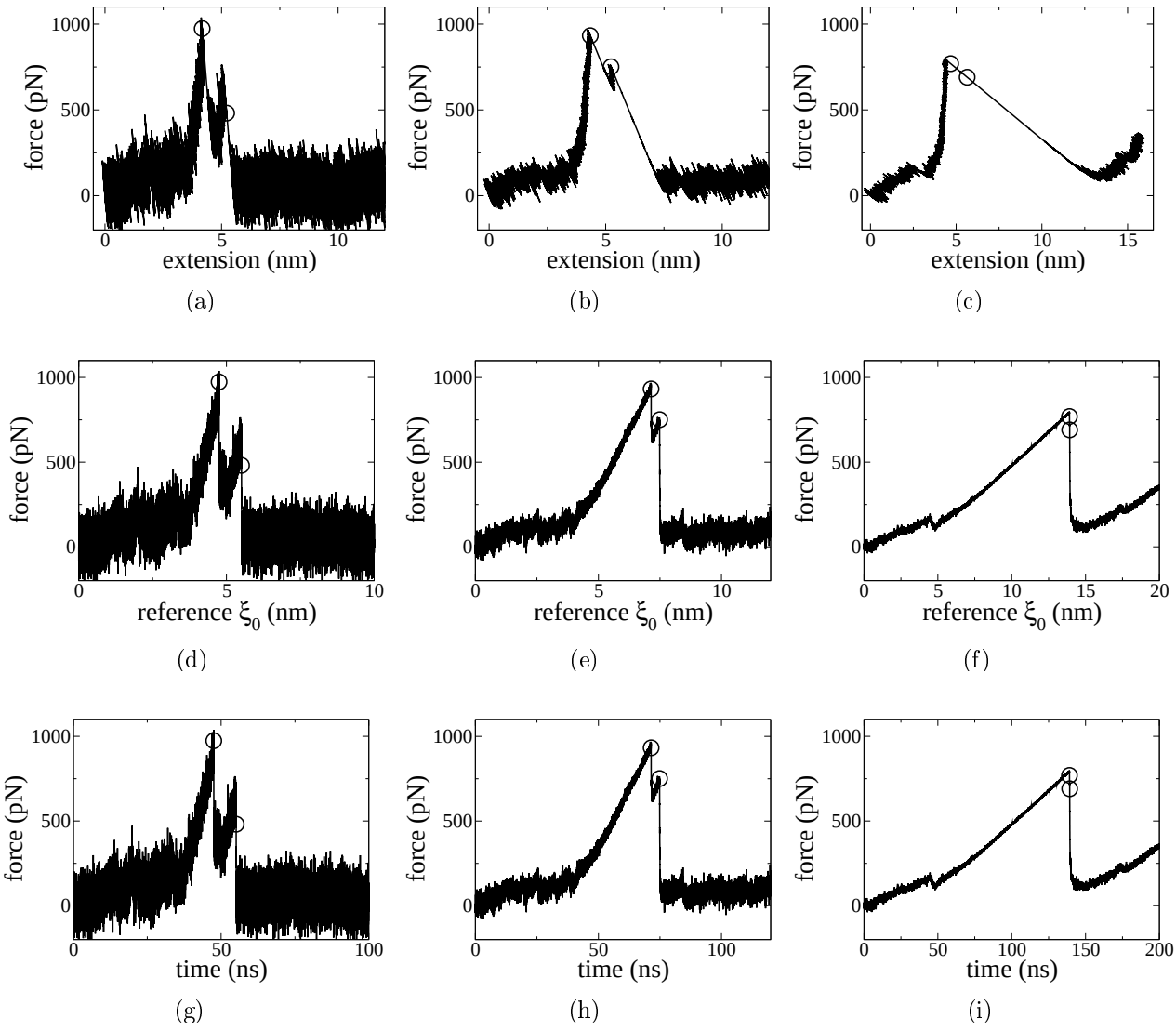


Figure S3: Force-extension curves obtained for model $[\mathrm{RD1,49}]_1$ with a pulling force constant $k_p = 1667$ pN/nm (panels **a**, **d** and **g**), $k_p = 333$ pN/nm (panels **b**, **e** and **h**) and $k_p = 83$ pN/nm (panels **c**, **f** and **i**). Upper panels give the true protein extension as measured from distances of Cα in the reference and pulling centers, middle panels give the reference coordinate (ξ_0, equation 2) and lower panels give the simulation time. Circles indicate Fe–S bond rupture.

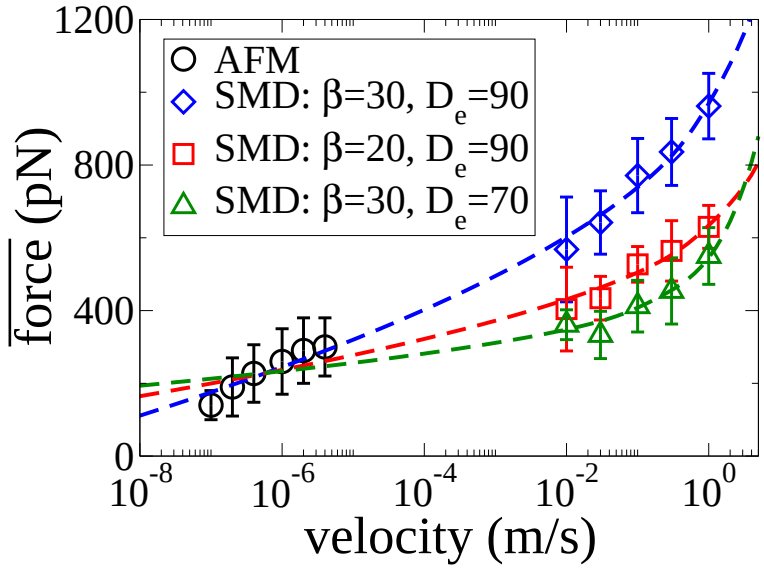


Figure S4: Force spectra or average rupture forces obtained at different pulling velocities for AFM experiments[17] and SMD simulations (N=20) of model $[RD1,49]_1$. Simulations were carried out with different values of Morse depths D_e (in kJ/mol) and Morse steepness β (in nm^{-1}). Dashed line show the full microscopic model derived by Hummer and Szabo[58] fitted to both AFM and SMD data. Error bars in SMD points indicate a standard deviation. The mean deviation between the model line and data points is 22 pN, 30 pN and 35 pN, respectively for parameter sets (β=30 nm^{-1}, D_e=90 kJ/mol – diamond symbol), (β=20 nm^{-1}, D_e=90 kJ/mol – square) and (β=30 nm^{-1}, D_e=70 kJ/mol – triangle).

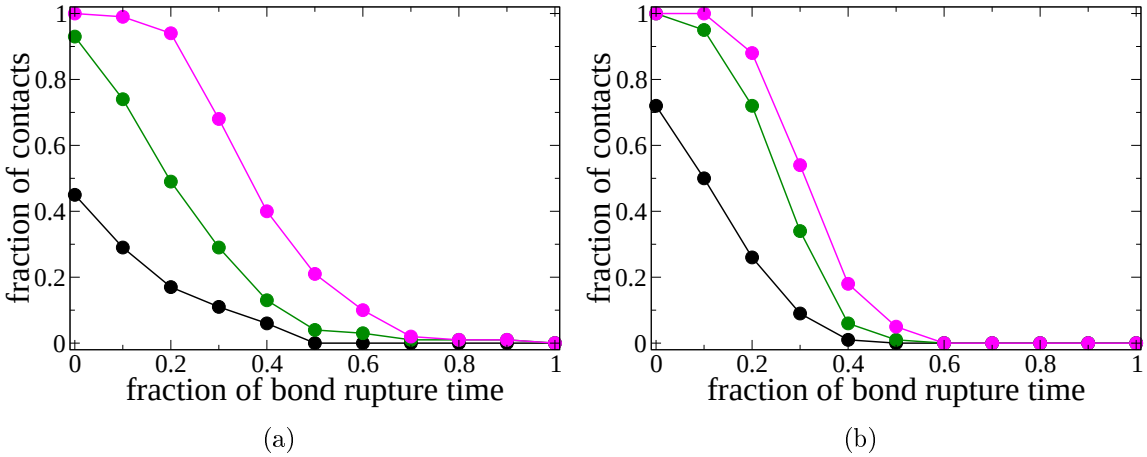


Figure S5: Relative time evolution of interaction survival (or fraction of contacts present) before the first Fe–S bond dissociation for pulling of models $[RD15,49]_3$ (panel **a**) and $[RD15,35]_3$ (panel **b**) with N=105-120 unfolding events. Salt bridge A1-E14 (black) and hydrogen bonds describing β-sheet 1-2 Trp3-HN···O-Tyr12 (green) and Cys5-HN···O-Tyr10 (pink) are shown.

Table S1: Average rupture forces ($\bar{F}$, average ± standard deviation in pN) calculated for the first Fe–S bond rupture in $[RD1,49]_1$ (N=10 simulations) with different pulling force constants (k_p in pN/nm).

k_p	F
1667	762 ± 101
333	831 ± 126
167	792 ± 65
83	758 ± 127
42	700 ± 63

5 Conclusion

Biochemical phenomena such as ligand-protein binding (section 1.1.1) and forced protein unfolding (section 1.1.2) are slow for the timescales usually reached by conventional MD simulations. These phenomena were studied in this thesis by combining MD simulations with methods or approximations to enhance configurational sampling.

In chapter 2 the LIE method (section 1.3.3.1) was used to predict binding affinities for ligand-protein complexes. LIE can be considered an approach to increase configurational sampling when compared to traditional computational methods to predict affinities because it focus the computational effort in the regions of interest in the ligand-protein binding process, the bound and unbound states of the ligand. Affinities predicted by LIE using ligand-protein complexes from crystal structures were similar to or better than those predicted by the docking scoring function and other LIE equations. One of the main limitations of LIE is the poor transferability of the coefficients to scale interaction energies among different model systems. Here transferability was increased by the use of coefficients adapted by the ligand or binding site relative polarities.

Major approximations in molecular docking (section 1.3.1) were improved. The LIE equation parametrized using complexes from crystal structures was used to replace the approximate scoring function (section 1.3.1.2) in the prediction of binding affinities. Moreover, the lack of protein flexibility (section 1.3.1.1) was overcome by the use of a group of protein configurations obtained from MD simulations to perform docking. Different averaging schemes were tested to obtain overall affinities for the ligand-protein complexes obtained from docking, revealing that many complex configurations contribute to the affinities of ligands for flexible proteins such as FKBP12 and HIV-1 reverse transcriptase, while affinities estimated for rigid proteins such as T4 lysozyme mutants are dominated by one complex configuration.

In chapter 3 the WE approach (section 1.3.3.2) was combined with MD simulations (section 1.3.2) to enhance sampling of benzene unbinding from the buried binding

site of T4 lysozyme. Enhanced sampling of infrequent unbinding events was achieved by increasing the computational effort in low probability regions of pre-defined progress coordinates.

Pathways for ligand exit and the protein conformational changes involved in ligand unbinding had not been fully resolved. Here, an exploratory set of simulations using the distance between the ligand and the binding site or the root mean squared deviation from the ligand in the crystal structure as progress coordinates revealed four possible pathways for benzene exit from the buried binding site of T4 lysozyme. Then, a production set of simulations using Voronoi bin mapping as progress coordinate was used to allow sampling of trajectories for a specific path, thus facilitating the calculation of rate constants and identification of protein conformational transitions involved in ligand unbinding for each specific path. The simulations suggest that motions in alpha helices C, H and J are more important for ligand transit than motions in alpha helix F, which was expected to be involved in ligand unbinding. Moreover, the simulations also suggest that motions in residues Y88 and I78 are important for ligand unbinding. Dissociation rates calculated from simulations did not converge, so further sampling is required to enable confident distinctions among the rates for the different pathways.

In chapter 4 SMD simulations (section 1.3.3.3) were used to model the forced unfolding of rubredoxin. Enhanced sampling of slow unfolding events was achieved by the use of pulling velocities of 10^{-1} m/s, which are much faster than typical pulling velocities of AFM experiments, of 10^{-6} m/s. As average unfolding forces depend on the pulling velocity, it was not possible to compare forces from AFM experiments and simulations directly. Here an indirect comparison was made by fitting the force spectrum to the microscopic model presented (section 1.1.2).

Full unfolding of rubredoxin involves covalent bond rupture. Here the full unfolding process was simulated by SMD simulations coupled to a classical description of bond dissociation, which was achieved using a Morse potential (equation 1.30). Parameters for this Morse potential were not available in the force field and were obtained by adjustment to quantum chemical calculations for Fe–S dissociation performed in our group. Parameters

to describe the FeS center in rubredoxin were not available in the force field either, and were obtained from previous calibrations of other authors and of our group or adjusted to avoid strong interactions between Fe and negative side-chains and an unbalanced solvation contribution. Moreover, an extensive analysis of the sensitivity of the mechanical properties of rubredoxin to the velocity (v_p) and force constant (k_p) of $U_{add}[\xi(t)]$ was performed. The v_p and k_p values were chosen to allow the reproduction of AFM results at a reasonable computational cost.

Simulations showed microscopic details of the forced unfolding of rubredoxin, revealing that changing the points of force application along rubredoxin results in different unfolding mechanisms. Contacts between beta-strands 1-2 and between beta-strands 1-3 are preserved before Fe-S bond rupture in most simulations of the rubredoxin mutated in positions 1 and 35, which is also the rubredoxin mutant with a slower k_{unf} value. On the other hand, such contacts are partially or completely ruptured before Fe-S bond rupture in most simulations of the other three rubredoxin mutants.

SMD simulations provided ΔLc values in better agreement with ΔLc^{AFM} values than estimates from crystal structures, ΔLc^{PDB} values, for rubredoxins mutated in positions 15 and 49 and in positions 15 and 35. This may have happened because perturbations in the protein structure before the force peak is achieved were considered in the simulations. ΔLc values from simulations and crystal structures were similar for rubredoxin mutated in positions 1 and 49. However, such values differed for rubredoxin mutated in positions 1 and 35, and better agreement with ΔLc^{AFM} values was achieved by the ΔLc^{PDB} values, indicating that improvements in the modeling of this mutant are needed.

Therefore, enhanced sampling methods, such as LIE, WE and SMD used here, allow modeling of biochemical phenomena that happen in the millisecond timescale or slower, even if MD simulations, which reach the microsecond timescale, are employed.

6 References

1 NOBEL FOUNDATION. *http://www.nobelprize.org*. 2017. (accessed in Feb. 2017).

2 KOTECHA, A. et al. Structure–based energetics of protein interfaces guides foot–and–mouth disease virus vaccine design. *Nat. Struct. Mol. Biol.*, v. 22, p. 788–794, 2015.

3 GOHLKE, H.; KLEBE, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed. Engl.*, v. 41, p. 2644–2676, 2002.

4 EYRING, H. The activated complex in chemical reactions. *J. Chem. Phys.*, v. 3, p. 107–115, 1935.

5 ROMANOWSKA, J. et al. Computational approaches for studying drug binding kinetics. In: KESERü, G. M.; SWINNEY, D. C. (Ed.). *Thermodynamics and kinetics of drug binding*. [S.l.]: Wiley-VCH Verlag GmbH & Co. KGaA, 2015. cap. 11, p. 211–235.

6 COPELAND, R. A.; POMPLIANO, D. L.; MEEK, T. D. Drug–target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.*, v. 5, p. 730–739, 2006.

7 COPELAND, R. A. The drug-target residence time model: a 10-year retrospective. *Nat. Rev. Drug Discov.*, v. 15, p. 87–95, 2016.

8 SCHUETZ, D. A. et al. Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug Discov. Today*, v. 22, p. 896–911, 2017.

9 TANG, Z.; ROBERTS, C. C.; CHANG, C. A. Understanding ligand-receptor non-covalent binding kinetics using molecular modeling. *Front. Biosci.*, v. 22, p. 960–981, 2017.

10 ZWIER, M. C.; KAUS, J. W.; CHONG, L. T. Efficient explicit–solvent molecular dynamics simulations of molecular association kinetics: methane/methane, Na+/Cl-, methane/benzene, and K+/18–crown–6 ether. *J. Chem. Theory Comput.*, v. 7, p. 1189–1197, 2011.

11 CHONG, L. T.; SAGLAM, A. S.; ZUCKERMAN, D. M. Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struct. Biol.*, v. 43, p. 88–94, 2017.

12 ZHUANG, X.; RIEF, M. Single-molecule folding. *Curr. Opin. Struct. Biol.*, v. 13, p. 88–97, 2003.

13 BINNIG, G.; QUATE, C. F.; GERBER, C. Atomic force microscope. *Phys. Rev. Lett.*, v. 56, p. 930–933, 1986.

14 OESTERHELT, F. et al. Unfolding pathways of individual bacteriorhodopsins. *Science*, v. 288, p. 143–146, 2000.

15 MÜLLER, D. J. et al. Stability of bacteriorhodopsin alpha-helices and loops analyzed by single-molecule force spectroscopy. *Biophys. J.*, v. 83, p. 3578–3588, 2002.

16 YU, H. et al. Hidden dynamics in the unfolding of individual bacteriorhodopsin proteins. *Science*, v. 355, p. 945–950, 2017.

17 RIEF, M. et al. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, v. 276, n. 5315, p. 1109–1112, 1997.

18 CARRION-VAZQUEZ, M. et al. Mechanical and chemical unfolding of a single protein: a comparison. *Proc. Natl. Acad. Sci. U. S. A.*, v. 96, p. 3694–3699, 1999.

19 MARSZALEK, P. E. et al. Mechanical unfolding intermediates in titin modules. *Nature*, v. 402, p. 100–103, 1999.

20 LI, H. et al. Reverse engineering of the giant muscle protein titin. *Nature*, v. 418, p. 998–1002, 2002.

21 FOWLER, S. B. et al. Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. *J. Mol. Biol.*, v. 322, p. 841–849, 2002.

22 LINKE, W. A.; GRÜTZNER, A. Pulling single molecules of titin by AFM - recent advances and physiological implications. *Pflügers Arch.*, v. 456, p. 101–115, 2008.

23 FISHER, T. E. et al. The study of protein mechanics with the atomic force microscope. *Trends Biochem. Sci.*, v. 24, p. 379–384, 1999.

24 FRANCO, I.; RATNER, M. A.; SCHATZ, G. C. Single-molecule pulling: phenomenology and interpretation. In: ESPINOSA, H. D.; BAO, G. (Ed.). *Nano and cell mechanics: fundamentals and frontiers*. [S.l.]: John Wiley & Sons, 2013. cap. 14, p. 359–388.

25 LI, Y. D. et al. The molecular mechanism underlying mechanical anisotropy of the protein GB1. *Biophys. J.*, v. 103, p. 2361–2368, 2012.

26 ZHENG, P.; CAO, Y.; LI, H. Facile method of constructing polyproteins for single-molecule force spectroscopy studies. *Langmuir*, v. 27, p. 5713–5718, 2011.

27 MARKO, J. F.; SIGGIA, E. D. Stretching DNA. *Macromolecules*, v. 28, p. 8759–8770, 1995.

28 RICO, F. et al. High-speed force spectroscopy unfolds titin at the velocity of molecular dynamics simulations. *Science*, v. 342, p. 741–743, 2013.

29 ANDO, T.; UCHIHASHI, T.; SCHEURING, S. Filming biomolecular processes by high-speed atomic force microscopy. *Chem. Rev.*, v. 114, p. 3120–3188, 2014.

30 BELL, G. Models for the specific adhesion of cells to cells. *Science*, v. 200, p. 618–627, 1978.

31 EVANS, E.; RITCHIE, K. Dynamic strength of molecular adhesion bonds. *Biophys. J.*, v. 72, p. 1541–1555, 1997.

32 IZRAILEV, S. et al. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys. J.*, v. 72, p. 1568–1581, 1997.

33 RIEF, M.; FERNANDEZ, J. M.; GAUB, H. E. Elastically coupled two-level systems as a model for biopolymer extensibility. *Phys. Rev. Lett.*, v. 81, p. 4764–4767, 1998.

34 HUMMER, G.; SZABO, A. Kinetics from nonequilibrium single–molecule pulling experiments. *Biophys. J.*, v. 85, p. 5–15, 2003.

35 DUDKO, O. K.; HUMMER, G.; SZABO, A. Intrinsic rates and activation free energies from single–molecule pulling experiments. *Phys. Rev. Lett.*, v. 96, p. 108101, 2006.

36 BAASE, W. A. et al. Lessons from the lysozyme of phage T4. *Protein Sci.*, v. 19, p. 631–641, 2010.

37 ALVES, A. F. N. *Um método computacional para estimar afinidades entre proteínas flexíveis e pequenos ligantes*. Thesis (Master) — Instituto de Química, Universidade de São Paulo, São Paulo, 2013.

38 POTEETE, A. R.; HARDY, L. W. Genetic analysis of bacteriophage T4 lysozyme structure and function. *J. Bacteriol.*, v. 176, p. 6783–6788, 1994.

39 GRüTTER, M. G. et al. Structural studies of mutants of the lysozyme of bacteriophage T4. The temperature–sensitive mutant protein Thr157→Ile. *J. Mol. Biol.*, v. 197, p. 315–329, 1987.

40 ALBER, T. et al. Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. *Nature*, v. 330, p. 41–46, 1987.

41 MATSUMURA, M.; BECKTEL, W. J.; MATTHEWS, B. W. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature*, v. 334, p. 406–410, 1988.

42 MATTHEWS, B. W.; REMINGTON, S. J. The three dimensional structure of the lysozyme from bacteriophage T4. *Proc. Natl. Acad. Sci. U. S. A.*, v. 71, p. 4178–4182, 1974.

43 ERIKSSON, A. E. et al. A cavity–containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature*, v. 355, p. 371–373, 1992.

44 QUILLIN, M. L. et al. Size versus polarizability in protein-ligand interactions: binding of noble gases within engineered cavities in phage T4 lysozyme. *J. Mol. Biol.*, v. 302, p. 955–977, 2000.

45 WEI, B. Q. et al. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, v. 322, p. 339–355, 2002.

46 MORTON, A.; BAASE, W. A.; MATTHEWS, B. W. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. *Biochemistry*, v. 34, p. 8564–8575, 1995.

47 MORTON, A.; MATTHEWS, B. W. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: linkage of dynamics and structural plasticity. *Biochemistry*, v. 34, p. 8576–8588, 1995.

48 GRAVES, A. P.; BRENK, R.; SHOICHET, B. K. Decoys for docking. *J. Med. Chem.*, v. 48, p. 3714–3728, 2005.

49 DENG, Y.; ROUX, B. Calculation of standard binding free energies: aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.*, v. 2, p. 1255–1273, 2006.

50 MOBLEY, D. L. et al. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, v. 371, p. 1118–1134, 2007.

51 GRAVES, A. P. et al. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J. Mol. Biol.*, v. 377, p. 914–934, 2008.

52 BOYCE, S. E. et al. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.*, v. 394, p. 747–763, 2009.

53 JIANG, W.; ROUX, B. Free energy perturbation Hamiltonian replica–exchange molecular dynamics (FEP/H-REMD) for absolute ligand binding free energy calculations. *J. Chem. Theory Comp.*, v. 6, p. 2559–2565, 2010.

54 GALLICCHIO, E.; LAPELOSA, M.; LEVY, R. M. Binding energy distribution analysis method (BEDAM) for estimation of protein-ligand binding affinities. *J. Chem. Theory Comput.*, v. 6, p. 2961–2977, 2010.

55 LINDER, M.; RANGANATHAN, A.; BRINCK, T. "Adapted linear interaction energy": a structure-based LIE parametrization for fast prediction of protein-ligand affinities. *J. Chem. Theory Comput.*, v. 9, p. 1230–1239, 2013.

56 WANG, K. et al. Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics. *J. Comput. Aided Mol. Des.*, v. 27, p. 989–1007, 2013.

57 MERSKI, M. et al. Homologous ligands accommodated by discrete conformations of a buried cavity. *Proc. Natl. Acad. Sci. U. S. A.*, v. 112, p. 5039–5044, 2015.

58 FEHER, V. A.; BALDWIN, E. P.; DAHLQUIST, F. W. Access of ligands to cavities within the core of a protein is rapid. *Nat. Struct. Biol.*, v. 3, p. 516–521, 1996.

59 SCHIFFER, J. M. et al. Capturing invisible motions in the transition from ground to rare excited states of T4 lysozyme L99A. *Biophys. J.*, v. 111, p. 1631–1640, 2016.

60 MIAO, Y.; FEHER, V. A.; MCCAMMON, J. A. Gaussian accelerated Molecular Dynamics: unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.*, v. 11, p. 3584–3595, 2015.

61 WANG, Y.; PAPALEO, E.; LINDORFF-LARSEN, K. Mapping transiently formed and sparsely populated conformations on a complex energy landscape. *eLife*, v. 5, p. e17505, 2016.

62 KITAHARA, R. et al. Detecting O_2 binding sites in protein cavities. *Sci. Rep.*, v. 6, p. 20534, 2016.

63 MULDER, F. A. A. et al. Studying excited states of proteins by NMR spectroscopy. *Nat. Struct. Biol.*, v. 8, p. 932–935, 2001.

64 BOUVIGNIES, G. et al. Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature*, v. 477, p. 111–114, 2011.

65 FRANKEL, A. D.; YOUNG, J. A. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.*, v. 67, p. 1–25, 1998.

66 MENéNDEZ-ARIAS, L.; SEBASTIáN-MARTíN, A.; ÁLVAREZ, M. Viral reverse transcriptases. *Virus Res.*, v. 234, p. 153–176, 2017.

67 SU, Y. et al. Linear interaction energy (LIE) models for ligand binding in implicit solvent: theory and application to the binding of NNRTIs to HIV–1 reverse transcriptase. *J. Chem. Theory Comput.*, v. 3, p. 256–277, 2007.

68 CARLSSON, J.; BOUKHARTA, L.; ÅQVIST, J. Combining docking, molecular dynamics and the linear interaction energy method to predict binding modes and affinities for non-nucleoside inhibitors to HIV-1 reverse transcriptase. *J. Med. Chem.*, v. 51, p. 2648–2656, 2008.

69 NICHOLS, S. E. et al. Predictive power of molecular dynamics receptor structures in virtual screening. *J. Chem. Inf. Model.*, v. 51, p. 1439–1446, 2011.

70 SANTOS, L. H.; FERREIRA, R. S.; CAFFARENA, E. R. Computational drug design strategies applied to the modelling of human immunodeficiency virus-1 reverse transcriptase inhibitors. *Mem. Inst. Oswaldo Cruz*, v. 110, p. 847–864, 2015.

71 TANAKA, H. et al. A new class of HIV–1 specific 6–substituted acyclouridine derivatives: synthesis and anti–HIV–1 activity of 5– or 6–substituted analogs of 1–[(2–hydroxyethoxy)methyl]–6–(phenylthio)thymine (HEPT). *J. Med. Chem.*, v. 34, p. 349–357, 1991.

72 TANAKA, H. et al. Structure–activity relationships of 1–[(2–hydroxyethoxy)methyl]–6–(phenylthio)thymine analogs: effect of substitutions at the C–6 phenyl ring and at the C–5 position on anti–HIV–1 activity. *J. Med. Chem.*, v. 35, p. 337–345, 1992.

73 TANAKA, H. et al. Synthesis and antiviral activity of deoxy analogs of 1–[(2–hydroxyethoxy)methyl]–6–(phenylthio)thymine (HEPT) as potent and selective anti–HIV–1 agents. *J. Med. Chem.*, v. 35, p. 4713–4719, 1992.

74 TANAKA, H. et al. Synthesis and antiviral activity of 6–benzyl analogs of 1–[(2–hydroxyethoxy)methyl]–6–(phenylthio)thymine (HEPT) as potent and selective anti–HIV–1 agents. *J. Med. Chem.*, v. 38, p. 2860–2865, 1995.

75 REN, J. et al. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat. Struct. Biol.*, v. 2, p. 293–302, 1995.

76 HOPKINS, A. L. et al. Complexes of HIV–1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non–nucleoside inhibitors. *J. Med. Chem.*, v. 39, p. 1589–1600, 1996.

77 ESNOUF, R. M. et al. Unique features in the structure of the complex between HIV-1 reverse transcriptase and the bis(heteroaryl)piperazine (BHAP) U-90152 explain resistance mutations for this nonnucleoside inhibitor. *Proc. Natl. Acad. Sci. U. S. A.*, v. 94, p. 3984–3989, 1997.

78 HSIOU, Y. et al. Structures of Tyr188Leu mutant and wild-type HIV-1 reverse transcriptase complexed with the non-nucleoside inhibitor HBY 097: inhibitor flexibility is a useful design feature for reducing drug resistance. *J. Mol. Biol.*, v. 284, p. 313–323, 1998.

79 BONNER, J. M.; BOULIANNE, G. L. Diverse structures, functions and uses of FK506 binding proteins. *Cell. Signal.*, v. 38, p. 97–105, 2017.

80 SWANSON, J. M. J.; HENCHMAN, R. H.; MCCAMMON, J. A. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.*, v. 86, p. 67–74, 2004.

81 LEE, M. S.; OLSON, M. A. Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. *Biophys. J.*, v. 90, p. 864–877, 2006.

82 YTREBERG, F. M. Absolute FKBP binding affinities obtained via nonequilibrium unbinding simulations. *J. Chem. Phys.*, v. 130, p. 164906, 2009.

83 HOLT, D. A. et al. Design, synthesis, and kinetic evaluation of high–affinity FKBP ligands and the X–ray crystal structures of their complexes with FKBP12. *J. Am. Chem. Soc.*, v. 115, p. 9925–9938, 1993.

84 BURKHARD, P.; TAYLOR, P.; WALKINSHAW, M. D. X-ray structures of small ligand-FKBP complexes provide an estimate for hydrophobic interaction energies. *J. Mol. Biol.*, v. 295, p. 953–962, 2000.

85 SZEP, S. et al. Structural coupling between FKBP12 and buried water. *Proteins*, v. 74, p. 603–611, 2009.

86 MUSTAFI, S. M. et al. Analysing the visible conformational substates of the FK506-binding protein FKBP12. *Biochem. J.*, v. 453, p. 371–380, 2013.

87 HUANG, D.; CAFLISCH, A. The free energy landscape of small molecule unbinding. *PLoS Comput. Biol.*, v. 7, p. e1002002, 2011.

88 DICKSON, A.; LOTZ, S. D. Ligand release pathways obtained with WExplore: residence times and mechanisms. *J. Phys. Chem. B*, v. 120, p. 5377–5385, 2016.

89 BAU, R. et al. Crystal structure of rubredoxin from *Pyrococcus furiosus* at 0.95 Angstroms resolution, and the structures of N–terminal methionine and formylmethionine variants of Pf Rd. Contributions of N–terminal interactions to thermostability. *J. Biol. Inorg. Chem.*, v. 3, p. 484–493, 1998.

90 JENNEY, F. E. et al. Anaerobic microbes: oxygen detoxification without superoxide dismutase. *Science*, v. 286, p. 306–309, 1999.

91 KLUMP, H. H.; ADAMS, M. W. W.; ROBB, F. T. Life in the pressure cooker: the thermal unfolding of proteins from hyperthermophiles. *Pure & Appl. Chem.*, v. 66, p. 485–489, 1994.

92 HILLER, R. et al. Stability and dynamics in a hyperthermophilic protein with melting temperature close to 200 °C. *Proc. Natl. Acad. Sci. U. S. A.*, v. 94, p. 11329–11332, 1997.

93 CAVAGNERO, S. et al. Response of rubredoxin from *Pyrococcus furiosus* to environmental changes: implications for the origin of hyperthermostability. *Biochemistry*, v. 34, p. 9865–9873, 1995.

94 LAZARIDIS, T.; LEE, I.; KARPLUS, M. Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. *Protein Sci.*, v. 6, p. 2589–2605, 1997.

95 CAVAGNERO, S. et al. Kinetic role of electrostatic interactions in the unfolding of hyperthermophilic and mesophilic rubredoxins. *Biochemistry*, v. 37, p. 3369–3376, 1998.

96 CAVAGNERO, S. et al. Unfolding mechanism of rubredoxin from *Pyrococcus furiosus*. *Biochemistry*, v. 37, p. 3377–3385, 1998.

97 PRAKASH, S.; SUNDD, M.; GUPTASARMA, P. The key to the extraordinary thermal stability of *P. furiosus* holo-rubredoxin: iron binding-guided packing of a core aromatic cluster responsible for high kinetic stability of the native structure. *PLoS One*, v. 9, p. e89703, 2014.

98 ZHENG, P.; LI, H. Highly covalent ferric–thiolate bonds exhibit surprisingly low mechanical stability. *J. Am. Chem. Soc.*, v. 133, p. 6791–6798, 2011.

99 ZHENG, P. et al. Hydrogen bond strength modulates the mechanical strength of ferric–thiolate bonds in rubredoxin. *J. Am. Chem. Soc.*, v. 134, p. 4124–4131, 2012.

100 ZHENG, P. et al. Single molecule force spectroscopy reveals the molecular mechanical anisotropy of the FeS_4 metal center in rubredoxin. *J. Am. Chem. Soc.*, v. 135, p. 17783–17792, 2013.

101 ZHENG, P. et al. Single molecule force spectroscopy reveals that iron is released from the active site of rubredoxin by a stochastic mechanism. *J. Am. Chem. Soc.*, v. 135, p. 7992–8000, 2013.

102 ZHENG, P.; WANG, Y.; LI, H. Reversible unfolding–refolding of rubredoxin: a single–molecule force spectroscopy study. *Angew. Chem. Int. Ed. Engl.*, v. 53, p. 14060–14063, 2014.

103 ZHENG, P. et al. Force–induced chemical reactions on the metal centre in a single metalloprotein molecule. *Nat. Commun.*, v. 6, p. 7569, 2015.

104 ARANTES, G. M.; BHATTACHARJEE, A.; FIELD, M. J. Homolytic cleavage of Fe–S bonds in rubredoxin under mechanical stress. *Angew. Chem. Int. Ed. Engl.*, v. 52, p. 8144–8146, 2013.

105 ARANTES, G. M.; FIELD, M. J. Ferric-thiolate bond dissociation studied with electronic structure calculations. *J. Phys. Chem. A*, v. 119, p. 10084–10090, 2015.

106 KUNTZ, I. D. et al. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.*, v. 161, p. 269–288, 1982.

107 GUEDES, I. A.; MAGALHÃES, C. S. de; DARDENNE, L. E. Receptor–ligand molecular docking. *Biophys. Rev.*, v. 6, p. 75–87, 2014.

108 KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, v. 3, p. 935–949, 2004.

109 BOWMAN, A. L. et al. Small molecule inhibitors of the MDM2–p53 interaction discovered by ensemble–based receptor models. *J. Am. Chem. Soc.*, v. 129, p. 12809–12814, 2007.

110 COELHO-CERQUEIRA, E. et al. Beyond topoisomerase inhibition: antitumor 1,4-naphthoquinones as potential inhibitors of human monoamine oxidase. *Chem. Biol. Drug. Des.*, v. 83, p. 401–410, 2014.

111 BISSON, W. H. et al. Discovery of antiandrogen activity of nonsteroidal scaffolds of marketed drugs. *Proc. Natl. Acad. Sci. U. S. A.*, v. 104, p. 11927–11932, 2007.

112 PAULI, I. et al. Discovery of new inhibitors of *Mycobacterium tuberculosis* InhA enzyme using virtual screening and a 3D-pharmacophore-based approach. *J. Chem. Inf. Model.*, v. 53, p. 2390–2401, 2013.

113 FERREIRA, L. G. et al. Molecular docking and structure-based drug design strategies. *Molecules*, v. 20, p. 13384–13421, 2015.

114 MALTAROLLO, V. G. et al. Structure-based virtual screening and discovery of new PPARδ/γ dual agonist and PPARδ and γ agonists. *PLoS One*, v. 10, p. e0118790, 2015.

115 IRWIN, J. J.; SHOICHET, B. K. Docking screens for novel ligands conferring new biology. *J. Med. Chem.*, v. 59, p. 4103–4120, 2016.

116 JIANG, F.; KIM, S.-H. “Soft docking": matching of molecular surface cubes. *J. Mol. Biol.*, v. 219, p. 79–102, 1991.

117 LEACH, A. R. Ligand docking to proteins with discrete side–chain flexibility. *J. Mol. Biol.*, v. 235, p. 345–356, 1994.

118 MORRIS, G. M. et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.*, v. 30, p. 2785–2791, 2009.

119 BROUGHTON, H. B. A method for including protein flexibility in protein–ligand docking: improving tools for database mining and virtual screening. *J. Mol. Graphics Model.*, v. 18, p. 247–257, 2000.

120 CARLSON, H. A. et al. Developing a dynamic pharmacophore model for HIV–1 integrase. *J. Med. Chem.*, v. 43, p. 2100–2114, 2000.

121 ARANTES, G. M. Flexibility and inhibitor binding in Cdc25 phosphatases. *Proteins*, v. 78, p. 3017–3032, 2010.

122 KNEGTEL, R. M. A.; KUNTZ, I. D.; OSHIRO, C. M. Molecular docking to ensembles of protein sctructures. *J. Mol. Biol.*, v. 266, p. 424–440, 1997.

123 AMARO, R. E.; LI, W. W. Emerging methods for ensemble-based virtual screening. *Curr. Top. Med. Chem.*, v. 10, p. 3–13, 2010.

124 TROTT, O.; OLSON, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, v. 31, p. 455–461, 2010.

125 ZHAO, G. et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, v. 497, p. 643–646, 2013.

126 MCCAMMON, J. A.; GELIN, B. R.; KARPLUS, M. Dynamics of folded proteins. *Nature*, v. 267, p. 585–590, 1977.

127 STONE, A. J. Intermolecular potentials. *Science*, v. 321, p. 787–789, 2008.

128 BEST, R. B. Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol.*, v. 22, p. 52–61, 2012.

129 FIELD, M. J. *A practical introduction to the simulation of molecular systems.* 2nd. ed. Cambridge: Cambridge University Press, 2007.

130 MORSE, P. M. Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys. Rev.*, v. 34, p. 57–64, 1929.

131 MACKERELL, A. D. et al. All–atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, v. 102, p. 3586–3616, 1998.

132 MACKERELL, A. D.; FEIG, M.; BROOKS, C. L. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.*, v. 126, p. 698–699, 2004.

133 BEST, R. B. et al. Optimization of the additive CHARMM all–atom protein force field targeting improved sampling of the backbone ϕ, ψ and side–chain $\chi 1$ and $\chi 2$ dihedral angles. *J. Chem. Theory Comput.*, v. 8, p. 3257–3273, 2012.

134 MAIER, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.*, v. 11, p. 3696–3713, 2015.

135 LEACH, A. R. *Molecular modelling: principles and applications.* 2nd. ed. Harlow: Prentice–Hall, 2001.

136 STILL, W. C. et al. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, v. 112, p. 6127–6129, 1990.

137 TOMé, T.; OLIVEIRA, M. J. de. *Dinâmica estocástica e irreversibilidade.* 2nd. ed. São Paulo: Edusp, 2014.

138 ZUCKERMAN, D. M. *Statistical physics of biomolecules: an introduction.* 1st. ed. [S.l.]: CRC Press, 2010.

139 ZUCKERMAN, D. M. Equilibrium sampling in biomolecular simulations. *Annu. Rev. Biophys.*, v. 40, p. 41–62, 2011.

140 MOORE, C. C. Ergodic theorem, ergodic theory, and statistical mechanics. *Proc. Natl. Acad. Sci. U. S. A.*, v. 112, p. 1907–1911, 2015.

141 ÅQVIST, J.; MEDINA, C.; SAMMUELSSON, J.-E. A new method for predicting binding affinity in computer–aided drug design. *Protein Eng.*, v. 7, p. 385–391, 1994.

142 ZWANZIG, R. W. High–temperature equation of state by a perturbation method. I. nonpolar gases. *J. Chem. Phys.*, v. 22, p. 1420–1426, 1954.

143 KIRKWOOD, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, v. 3, p. 300–313, 1935.

144 WARSHEL, A.; RUSSELL, S. T. Calculations of electrostatic interactions in biological systems and in solutions. *Q. Rev. Biophys.*, v. 17, p. 283–422, 1984.

145 WALL, I. D. et al. Binding constants of neuraminidase inhibitors: an investigation of the linear interaction energy method. *J. Med. Chem.*, v. 42, p. 5142–5152, 1999.

146 ÅQVIST, J.; LUZHKOV, V. B.; BRANDSDAL, B. O. Ligand binding affinities from MD simulations. *Acc. Chem. Res.*, v. 35, p. 358–365, 2002.

147 KOLB, P. et al. Discovery of kinase inhibitors by high–throughput docking and scoring based on a transferable linear interaction energy model. *J. Med. Chem*, v. 51, p. 1179–1188, 2008.

148 AMORIM, H. L. N. de; CACERES, R. A.; NETZ, P. A. Linear interaction energy (LIE) method in lead discovery and optimization. *Curr. Drug Targets*, v. 9, p. 1100–1105, 2008.

149 VALIENTE, P. A. et al. New parameterization approaches of the LIE method to improve free energy calculations of PlmII-inhibitors complexes. *J. Comput. Chem.*, v. 31, p. 2723–2734, 2010.

150 PEREIRA, E. G.; MOREIRA, M. Â. M.; CAFFARENA, E. R. Molecular interactions of c-ABL mutants in complex with imatinib/nilotinib: a computational study using linear interaction energy (LIE) calculations. *J. Mol. Model.*, v. 18, p. 4333–4341, 2012.

151 HANSSON, T.; MARELIUS, J.; ÅQVIST, J. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.–Aided Mol. Des.*, v. 12, p. 27–35, 1998.

152 HUBER, G. A.; KIM, S. Weighted–ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.*, v. 70, p. 97–110, 1996.

153 ZUCKERMAN, D. M.; CHONG, L. T. Weighted ensemble simulation: review of methodology, applications, and software. *Annu. Rev. Biophys.*, v. 46, p. 43–57, 2017.

154 BHATT, D.; ZHANG, B. W.; ZUCKERMAN, D. M. Steady–state simulations using weighted ensemble path sampling. *J. Chem. Phys.*, v. 133, p. 014110, 2010.

155 BHATT, D.; ZUCKERMAN, D. M. Heterogeneous path ensembles for conformational transitions in semiatomistic models of adenylate kinase. *J. Chem. Theory Comput.*, v. 6, p. 3527–3539, 2010.

156 ADELMAN, J. L. et al. Simulations of the alternating access mechanism of the sodium symporter Mhp1. *Biophys. J.*, v. 101, p. 2399–2407, 2011.

157 SUáREZ, E. et al. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *J. Chem. Theory Comput.*, v. 10, p. 2658–2667, 2014.

158 DICKSON, A.; III, C. L. B. WExplore: hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm. *J. Phys. Chem. B*, v. 118, p. 3532–3542, 2014.

159 ZWIER, M. C. et al. Efficient atomistic simulation of pathways and calculation of rate constants for a protein-peptide binding process: application to the MDM2 protein and an intrinsically disordered p53 peptide. *J. Phys. Chem. Lett.*, v. 7, p. 3440–3445, 2016.

160 SAGLAM, A. S.; CHONG, L. T. Highly efficient computation of the basal kon using direct simulation of protein-protein association with flexible molecular models. *J. Phys. Chem. B*, v. 120, p. 117–122, 2016.

161 DICKSON, A.; LOTZ, S. D. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by WExplore. *Biophys. J.*, v. 112, p. 620–629, 2017.

162 LAIO, A.; PARRINELLO, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.*, v. 99, p. 12562–12566, 2002.

163 GRUBMULLER, H.; HEYMANN, B.; TAVAN, P. Ligand binding: molecular mechanics calculation of the streptavidin–biotin rupture force. *Science*, v. 271, p. 997–999, 1996.

164 LU, H. et al. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys. J.*, v. 75, p. 662–671, 1998.

165 LU, H.; SCHULTEN, K. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins*, v. 35, p. 453–463, 1999.

166 LU, H.; SCHULTEN, K. The key event in force-induced unfolding of titin's immunoglobulin domains. *Biophys. J.*, v. 79, p. 51–65, 2000.

167 ISRALEWITZ, B.; GAO, M.; SCHULTEN, K. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.*, v. 11, p. 224–230, 2001.

168 LORENZO, A. C.; CAFFARENA, E. R. Elastic properties, Young's modulus determination and structural stability of the tropocollagen molecule: a computational study by steered molecular dynamics. *J. Biomech.*, v. 38, p. 1527–1533, 2005.

169 SOTOMAYOR, M.; SCHULTEN, K. Single-molecule experiments in vitro and in silico. *Science*, v. 316, p. 1144–1148, 2007.

170 LIM, B. B. C. et al. Molecular basis of fibrin clot elasticity. *Structure*, v. 16, p. 449–459, 2008.

171 LEE, E. H. et al. Discovery through the computational microscope. *Structure*, v. 17, p. 1295–1306, 2009.

172 PEPŁOWSKI, L. et al. Molecular jamming - The cystine slipknot mechanical clamp in all-atom simulations. *J. Chem. Phys.*, v. 134, p. 085102, 2011.

173 HE, C. et al. Mechanically untying a protein slipknot: multiple pathways revealed by force spectroscopy and steered molecular dynamics simulations. *J. Am. Chem. Soc.*, v. 134, p. 10428–10435, 2012.

174 WEI, W. et al. Structural insights and the surprisingly low mechanical stability of the Au-S bond in the gold-specific protein GolB. *J. Am. Chem. Soc.*, v. 137, p. 15358–15361, 2015.

Attachments

Curriculum Vitae

Ariane Ferreira Nunes Alves
PhD student
Institute of Chemistry
Universidade de São Paulo - Brazil

Work address:
Department of Biochemistry, Institute of Chemistry
Universidade de São Paulo
Av. Prof. Lineu Prestes, 748, room 752
CEP 05508-000, São Paulo, SP, Brazil
Phone number: +55 11 3091-9044
E-mail: anunesalves@usp.br
researchgate.net/profile/Ariane_Nunes-Alves

1. Formal education

Ph.D. Biochemistry, Universidade de São Paulo, Brazil (2013-expected October 2017).
M.Sc. Biochemistry, Universidade de São Paulo, Brazil (2011-2013).
B.Sc. Biology, Universidade de São Paulo, Brazil (2006-2010).

2. Research activities

Ph.D. student, Universidade de São Paulo, Brazil (2013-expected October 2017)
Advisor: Dr. Guilherme Menegon Arantes, Dept. of Biochemistry, Institute of Chemistry
Activities: understand the molecular mechanism of rubredoxin forced unfolding in atomic force microscopy experiments using steered molecular dynamics simulations, sample exit pathways of ligands from T4 lysozyme binding site using weighted ensemble simulations, develop a computational method based on linear interaction energy to estimate affinities for protein-ligand complexes obtained from crystal structures or docking.

Visiting Ph.D. student, University of Pittsburgh, USA (2016, 6 months)
Advisor: Dr. Daniel M. Zuckerman, Dept. of Computational and Systems Biology, School of Medicine
Activities: sample exit pathways of ligands from T4 lysozyme binding site using weighted ensemble simulations.

M.Sc. student, Universidade de São Paulo, Brazil (2011-2013)
Advisor: Dr. Guilherme Menegon Arantes, Dept. of Biochemistry, Institute of Chemistry
Activities: develop a computational method based on linear interaction energy to estimate

affinities for protein-ligand complexes obtained from crystal structures or docking.

Undergraduate research, Universidade de São Paulo, Brazil (2007-2010)
Advisor: Dr. Alexander Henning Ulrich, Dept. of Biochemistry, Institute of Chemistry
Activities: study the pharmacological activity of tobacco nitrosamines over acetylcholine receptors using the patch-clamp technique.

Undergraduate research, Cornell University, USA (2009, 3 months)
Advisor: Dr. George P. Hess, Dept. of Molecular Biology and Genetics
Activities: study the pharmacological activity of the tobacco nitrosamine N-nitrosonornicotine over acetylcholine receptors using single-channel kinetics.

3. Awards

- Travel award from the Biophysical Society (2014)
- Travel award from the Institute of Chemistry, Universidade de São Paulo (2013)
- Best poster presentation in the area of computational biochemistry at the XLII Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology (2013)
- Cone-sul symposium travel award from the Brazilian Society for Biochemistry and Molecular Biology (2013)
- Travel award from the Society for Research on Nicotine and Tobacco - Europe (2010)
- Travel award from the Institute of Biosciences, Universidade de São Paulo (2010)
- Best poster presentation in the area of ion channels and transporters at the XXIV Annual Meeting of the Federation of Societies for Experimental Biology (2009)

4. Articles published in scientific journals

1. **Nunes-Alves, A.**; Arantes, G.M. Ligand-receptor affinities computed by an adapted linear interaction model for continuum electrostatics and by protein conformational averaging. *J. Chem. Inf. Model.*, 54: 2309-2319, 2014.
2. **Nunes-Alves, A.**; Nery, A.A.; Ulrich, H. Tobacco nitrosamine N-nitrosonornicotine as inhibitor of neuronal nicotinic acetylcholine receptors. *J. Mol. Neurosci.*, 49: 52-61, 2013.

5. Invited talks

1. Computing the affinity between small ligands and flexible proteins. Universidade de São Paulo, Institute of Chemistry, Department of Biochemistry, São Paulo, Brazil (2015).

6. Lectures and oral presentations

1. How many pathways are available for a ligand to exit a protein binding site? Seminars of the Graduate Students and Postdoctorates of the Department of Biochemistry, Institute of Chemistry, Universidade de São Paulo, São Paulo, Brazil (2017).

2. Weighted ensemble of pathways for ligand unbinding from T4 lysozyme. 61^{st} Annual Meeting of the Biophysical Society, New Orleans, USA (2017).

3. Ligand-receptor affinities computed by an adapted linear interaction model and by protein conformational averaging. Hands-on Workshop on Computational Biophysics, Pittsburgh, USA (2016).

4. Weighted ensemble of pathways for ligand unbinding. 8^{th} Annual Retreat of the Department of Computational and Systems Biology of the University of Pittsburgh, Roanoke, USA (2016).

5. The role of protein flexibility in the binding of small molecules. XLII Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology, Foz do Iguaçu, Brazil (2013).

6. Role of protein flexibility in the binding of small ligands. Hands-on Course - Coarse Grain Methods for Biomolecular Simulations, Montevideo, Uruguay (2011).

7. Tobacco nitrosamine N-nitrosonornicotine as inhibitor of the nicotinic acetylcholine receptor subtype $\alpha 3\beta 4$. XXV Annual Meeting of the Federation of Societies for Experimental Biology, Águas de Lindóia, Brazil (2010).

8. Tobacco nitrosamine N-nitrosonornicotine as inhibitor of nicotinic acetylcholine receptors. XXIV Annual Meeting of the Federation of Societies for Experimental Biology, Águas de Lindóia, Brazil (2009).

7. Abstracts published in annals of events

1. **Nunes-Alves, A.**; Arantes, G.M. Forced unfolding coupled to covalent bond cleavage investigated by steered molecular dynamics. Gordon Research Conference on Computational Chemistry, Girona, Spain (2016).

2. **Nunes-Alves, A.**; Arantes, G.M. Metalloprotein mechanical unfolding investigated by steered molecular dynamics. III Brazilian School for Molecular Modeling, Santo André, Brazil (2015).

3. **Nunes-Alves, A.**; Arantes, G.M. A computational method to estimate affinities between flexible proteins and small ligands. 58^{th} Annual Meeting of the Biophysical Society, San Francisco, USA (2014).

4. **Nunes-Alves, A.**; Arantes, G.M. The role of protein flexibility in the binding of small molecules. XLII Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology, Foz do Iguaçu, Brazil (2013).

5. **Nunes-Alves, A.**; Arantes, G.M. Computational method to cluster complexes of small ligands and flexible proteins. XLI Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology, Foz do Iguaçu, Brazil (2012).

6. **Nunes-Alves, A.**; Arantes, G.M. A procedure to cluster complexes of small ligands and flexible proteins. XVI Brazilian Symposium for Theoretical Chemistry, Ouro Preto, Brazil (2011).

7. **Nunes-Alves, A.**; Nery, A.A.; Ulrich, H. Tobacco nitrosamine N-nitrosonornicotine as inhibitor of rat nicotinic acetylcholine receptors. 12^{th} Annual Meeting of the Society for Research on Nicotine and Tobacco - Europe, Bath, England (2010).

8. **Nunes-Alves, A.**; Nery, A.A.; Ulrich, H. Tobacco nitrosamine N-nitrosonornicotine as inhibitor of the nicotinic acetylcholine receptor subtype $\alpha 3\beta 4$. XXV Annual Meeting of the Federation of Societies for Experimental Biology, Águas de Lindóia, Brazil (2010).

9. **Nunes-Alves, A.**; Nery, A.A.; Ulrich, H. Tobacco nitrosamine N-nitrosonornicotine as inhibitor of nicotinic acetylcholine receptors. XXIV Annual Meeting of the Federation of Societies for Experimental Biology, Águas de Lindóia, Brazil (2009).

10. **Nunes-Alves, A.**; Nery, A.A.; Ulrich, H. Pharmacological activity of the tobacco nitrosamine N-nitrosonornicotine on nicotinic acetylcholine receptors. XXXVIII Annual Meeting of the Brazilian Society for Biochemistry and Molecular Biology, Águas de Lindóia, Brazil (2009).

11. **Nunes-Alves, A.**; Ulrich, H. Pharmacological activity of tobacco nitrosamines 4-(methylnitrosamino)-1-(3-pyridil)-1-butanone and N-nitrosonornicotine. I Neurolatam (I Congress ibro/larc of Neuroscience for Latin America, Caribbean and Iberian Peninsula), Búzios, Brazil (2008).