Free Energy-Based Computational Methods for the Study of Protein-Peptide Binding Equilibria

Emilio Gallicchio

Abstract This chapter discusses the theory and application of physics based free energy methods to estimate protein-peptide binding free energies. It presents a statistical mechanics formulation of molecular binding, which is then specialized in three methodologies: (i) alchemical absolute binding free energy estimation with implicit solvation, (ii) alchemical relative binding free energy estimation with explicit solvation, and (iii) potential of mean force binding free energy estimation. Case studies of protein-peptide binding application taken from the recent literature are discussed for each method.

Short title: Free Energy-Based Computational Methods

Keywords: Free Energy, Binding Free Energy, Equilibrium Binding Constant, Alchemical Perturbation, Potential of Mean Force, Protein-Peptide Binding Modeling, Molecular Dynamics, Molecular Recognition, Statistical Mechanics

Emilio Gallicchio
Department of Chemistry, Brooklyn College of the City University of New York, and Ph.D. Program in Biochemistry and Ph.D. Program in Chemistry at The Graduate Center of the City University of New York, New York, NY e-mail: egallicchio@brooklyn.cuny.edu

1 Introduction

Peptide and peptide-derived molecules are widely used to target protein-protein interactions for medicinal purposes and basic biological research. In-silico models play an increasingly significant role in the study of protein-peptide interactions. As excellently reviewed elsewhere,[1, 2, 3] computational methods for studying protein-peptide interactions have evolved on somewhat separate tracks from those used for small molecule-protein interactions. These differences are partly due to the greater flexibility and size of peptides and their tendency to interact with proteins through many relatively weak interactions. Nevertheless, because the same fundamental physical forces regulate all molecular recognition phenomena, it is helpful to relate computational models under a standard set of principles.

This chapter is devoted to a class of physics-based free energy methods considered the most accurate and detailed for modeling the thermodynamics of molecular binding equilibria. These methods model the interactions between molecules as well as their motion at the atomic level. We derive each method discussed from a well-established statistical mechanics theory of non-covalent molecular association. The chapter attempts to demystify the theory and the seemingly arcane formulas and computational procedures used in the field and point out the specific features of the methods that make them more or less suitable for studying protein-peptide interactions.

There is an implicit acknowledgment here that an understanding of these methodologies and how to select and apply them appropriately cannot be accomplished fully without referring to the underlying theory. The treatment employed here requires only a basic familiarity with concepts of statistics (probability distributions, averages, marginalization) and of classical statistical thermodynamics (classical partition functions and their manipulations, and their relationship with the free energy).

After presenting the theory and methods, we then illustrate their applications by discussing three case studies. We hope that this format will help convey the characteristics and relationships between the various methodologies and the fundamental principles on which they are based.

2 Statistical Mechanics Formulation

In this section, we derive and discuss a statistical mechanics theory of molecular binding. The concepts and the formulas expressed here will be used later to rationalize the specific computational methods and practices used in the case studies reviewed in Section 3.

We attempt to use unambiguous notation throughout, but sometimes we adopt a simplified notation to unclutter the equations. For example, in intermediate formulas, we often omit limits of integration and Jacobian factors for curvilinear coordinates when they do not affect the form and interpretation of the final result. In some places, we use function arguments to distinguish two functions. For example, we might denote the ligand and receptor's potential energy functions with the same symbol U, as $U(x_L)$ and $U(x_R)$, even though they are different mathematical functions.

2.1 The Standard Free Energy of Binding

We will consider here the reversible non-covalent binding equilibrium between receptor molecules R and ligand molecules L to form a complex RL in an ideal solution:

$$\mathrm{R} + \mathrm{L} \rightleftharpoons \mathrm{RL}, \tag{1}$$

with the dimensionless equilibrium constant

$$K_b = \left[\frac{[\mathrm{RL}]/C^\circ}{([\mathrm{R}]/C^\circ)([\mathrm{L}]/C^\circ)} \right]_{\mathrm{eq}}, \tag{2}$$

where [...] are concentrations, C° is the standard state concentration (conventionally set as 1 M or 1 molecule/1668 Å^3), and the "eq" subscript states that all concentrations are evaluated at equilibrium. The Gibb's molar standard binding free energy, which is the main objective of the computational models of binding discussed here, is defined as

$$\Delta G_b^\circ = -k_{\mathrm{B}} T \ln K_b \tag{3}$$

where k_{B} is Boltzmann's constant and T is the temperature in the Kelvin scale (in the following we will assume constant temperature pressure conditions).

Implicit in this quasi-chemical description of the binding equilibrium is the idea that the separated species in solution R and L, as well as the complex RL, are defined in some way. In an experimental setting, the apparatus used to measure equilibrium concentrations provides a working definition of the species. The nature of the experimental reporter used to monitor the formation of the complex is of particular relevance.[4] The change of a spectroscopic signal, as in NMR and UV/VIS fluorescence assays,[5] likely probes a set of conformations of the complex in which specific groups of the receptor and the ligand are in contact. Hence, different spectroscopic reporters would, in general, yield different estimates of the standard free energy of binding.[6] Spectroscopic reporters stand in contrast to experimental reporters, such as those in calorimetric, surface plasmon resonance (SPR), amplified luminescent proximity (AlphaScreen), and equilibrium dialysis binding assays, that probe unspecific molecular association.[4, 7, 8, 9, 6] Here, we focus mainly on computational models that define the complex using structural means–typically specific distances and angles between groups of atoms [10]–and are therefore more suitable to describe measurements of binding constants with specific spectroscopic experimental reporters.

In practice, the association between a peptide ligand and a protein receptor is also often monitored by indirect biochemical means, such as enzymatic inhibition[11] or pull-down assays,[12] that are only indirectly related to the equilibrium binding constant of the ligand-receptor complex. The computational models' ability to reproduce or explain this type of data is expected to be semi-quantitative at best, as it would be a correlation between experimental binding constants and activity data.

While ambiguities in relating molecular computer simulations to experimental biophysical data of molecular binding exist for any molecular complex, the issue is explicitly discussed here because it is expected to be particularly widespread for the study of the interactions involving peptides, which are generally more flexible than most small-molecule drug compounds and engage protein receptors over a large binding surface in a variety of binding modes. It is useful to keep these issues in mind when designing a computational model and the answers that one can reasonably extract from it. Computational modeling can be a valuable tool when used judiciously by exploiting its strengths while managing its unavoidable limitations.

2.2 Statistical Mechanics Theory of Non-Covalent Molecular Binding

Under the assumptions above, Gilson et al.[13] derived a statistical mechanics expression for the binding constant [Eq. (2)] which, with a few reasonable approximations (discussed below), can be written as:[14]

$$K_b = \frac{C^\circ}{8\pi^2} \frac{z_{RL}}{z_R z_L}, \tag{4}$$

where z_i is the intramolecular configurational partition function of one molecules of species i in solution.

A full derivation of Eq. (4) is beyond the scope of this chapter. However, it is briefly outlined here to introduce the notation. Eq. (4) is derived by writing the molar standard binding free energy as the difference of the standard chemical potentials of the complex and those of the receptor and ligand

$$\Delta G_b^\circ = \mu_{RL}^\circ - \mu_R^\circ - \mu_L^\circ \tag{5}$$

and employing the McMillan-Mayer expression for the standard chemical potential of a solute in an ideal solution[15]

$$\mu_i^\circ = -k_B T \ln \frac{\phi_i}{\Lambda_i^3 C^\circ} \tag{6}$$

where ϕ_i the internal canonical molecular partition function of solute i in solution and Λ_i is the thermal De Broglie wavelength of the center of mass of the solute. The internal molecular partition function includes only the internal degrees of freedom of the solute obtained after separating the translational degrees of freedom of the molecular center of mass. Furthermore, the solute's internal canonical molecular partition function in solution is understood in the context of the concept of the solvent potential of mean force,[16] in which the solvent degrees are averaged out.

While a quantum-mechanical treatment is required in general, adopting a classical expression for the molecular partition function is appropriate for the present discussion limited to non-covalent molecular association equilibria, which do not involve the formation or breaking of chemical bonds. The internal canonical molecular partition function is written as

$$\phi_i = \frac{8\pi^2 z_i}{\prod_j \lambda_j^3} \tag{7}$$

where the denominator comes from the integration over the momenta,[1] the factor of $8\pi^2$ comes from the integration over the orientational degrees of freedom of the solute,[2] and z_i is the vibrational molecular configurational partition function

$$z_i = \int d\boldsymbol{x}_i e^{-\beta\Psi_i(\boldsymbol{x}_i)} \tag{8}$$

where $\beta = 1/(k_B T)$ is the inverse temperature, $\boldsymbol{x}_i$ denotes the collection of the vibrational degrees of freedom of solute i and

$$\Psi_i(\boldsymbol{x}_i) = U_i(\boldsymbol{x}_i) + W_i(\boldsymbol{x}_i) \tag{9}$$

is the effective potential energy of a specific configuration of the solute in solution. The effective potential energy is given by the sum of the intramolecular potential energy $U_i(\boldsymbol{x}_i)$ and the solvent of potential of mean force $W_i(\boldsymbol{x}_i)$, defined as the free energy of solvation of the solute kept rigid in configuration $\boldsymbol{x}_i$.[3] In the present notation,

$$e^{-\beta W_i(\boldsymbol{x}_i)} = \frac{1}{Z_N} \int d\boldsymbol{r}_v^N \, e^{-\beta U_i(\boldsymbol{r}_v^N, \boldsymbol{x}_i)} \tag{10}$$

where $\boldsymbol{r}_v^N$ denotes the collection of degrees of freedom of N solvent molecules, $U_i(\boldsymbol{r}_v^N, \boldsymbol{x}_i)$ is the potential energy of the mixture of N solvent molecules and one solute molecule i, and Z_N is the configurational partition function of the pure solvent,[4] expressed as the integral in Eq. (10) but without the solute.

---

[1] The details are omitted since the the contributions from momenta cancel out in this classical treatment.

[2] We assume that the orientational degrees of freedom can be separated from the vibrational degrees of freedom without significant loss of accuracy. This is generally an excellent approximation at moderate temperatures.

[3] It should be noted that the solvent potential of mean force formalism does not introduce new assumptions or approximations than the ones already adopted. In this context, it is only a convenient notation aid. We will discuss later implicit solvation models which approximate the solvent potential of mean force.

[4] The notation can be easily extended to solvent mixtures including ions and co-solvents.

Eq. (4) is obtained by inserting Eq. (6) for each species, using Eqs. (7–10), into Eq. (5) noticing that the kinetic energy factors cancel out, and finally inverting Eq. (3).

The definition of the vibrational configurational partition function of the complex, z_{RL}, receives special consideration in this theory.[13] In the complex, the translational and orientational degrees of freedom of the ligand are represented by the internal degrees of freedom of the complex that specify the position and orientation of the ligand with respect to a coordinate system attached to the receptor.[10] Furthermore, the integration along these coordinates is limited to some specified range of configurational space that encodes our structural definition of what constitutes a valid configuration of a ligand "bound" to the receptor (see the discussion in Section 2.1). The structural definition of the bound complex is a necessary and somewhat arbitrary input of the theory.[13, 4, 14, 7] Without it, the free energy of the bound complex relative to the unbound state is undefined and, consequently, the standard binding free energy and the binding constant would also be undefined in this theory. It is customary to represent the bound region of the complex by an indicator function $I(\zeta_L)$, where ζ_L represents the collection of the six coordinates[5] that specify the position and orientation of the ligand relative to the receptor.[10][6] The indicator function is set to 1 if the position and orientation of the ligand is such that receptor and ligand are considered bound and zero otherwise so that z_{RL} can be written as

$$z_{RL} = \int d\boldsymbol{x}_R d\boldsymbol{x}_L d\zeta_L I(\zeta_L) e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)} \tag{11}$$

[5] Three translations and three orientations for a non-linear ligand.

[6] The specific choice of the ζ_L coordinates is arbitrary as long as they do not couple directly or indirectly the intramolecular coordinates of the receptor or the ligand.

2.3 The Binding Free Energy Formula

Since the direct evaluation of partition functions is not generally feasible, Eq. (4) is not amenable to direct computation. One strategy is to transform it into an average over the conformational ensemble in which receptor and ligand are uncoupled. To do so, we reorganize the integration variables in the numerator so that they match exactly those in the denominator. First, define

$$\int d\zeta_L I(\zeta_L) = V_{\text{site}}\Omega_{\text{site}} \tag{12}$$

which measures the spatial (V_{site}) and angular (Ω_{site}) extent of the bound state of the complex when receptor and ligand are uncoupled. [7] Then, multiply and divide Eq. (4) by Eq. (12) by keeping the integral form in the denominator and the integrated form in the numerator. The result is

$$K_b = C^\circ V_{\text{site}} \frac{\Omega_{\text{site}}}{8\pi^2} \langle e^{-\beta u} \rangle_0 \tag{13}$$

where

$$\langle e^{-\beta u} \rangle_0 = \int d\boldsymbol{x}_R d\boldsymbol{x}_L d\zeta_L e^{-\beta u(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)} \rho_0(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L) \tag{14}$$

is the ensemble average of the Boltzmann weight of the *effective binding energy* u of defined as the difference in effective potential energies of the complex in the specified configuration and of that of the separated receptor and ligand without changing their internal configurations

[7] Eq. (12)is is colloquially referred to as the volume of the receptor binding site. The notation used here suggests that translational and orientational components are not coupled in the definition of $I(\zeta_L)$. The present treatment is still valid if this is not the case, except that in this case the value of the integral of the indicator function is not written as the product of spatial and orientational components. Finally, $\Omega_{\text{site}} = 8\pi^2$ if the definition of the bound complex does not involve orientational coordinates, that is when only the position of the ligand is used to judge whether it is bound to the receptor.

$$u(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L) = \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L) - \Psi_R(\boldsymbol{x}_R) - \Psi_L(\boldsymbol{x}_L) \tag{15}$$

with the normalized probability density function

$$\rho_0(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L) = \frac{I(\zeta_L) e^{-\beta\Psi_R(\boldsymbol{x}_R)} e^{-\beta\Psi_L(\boldsymbol{x}_L)}}{\int d\boldsymbol{x}_R d\boldsymbol{x}_L d\zeta_L I(\zeta_L) e^{-\beta\Psi_R(\boldsymbol{x}_R)} e^{-\beta\Psi_L(\boldsymbol{x}_L)}} \tag{16}$$

which corresponds to an unphysical state of the complex in which the ligand is bound to the receptor (the density is zero unless $I(\zeta_L) = 1$) but it does not interact with it (the potential function lacks receptor-ligand coupling terms). We will hereafter refer to this state as the *decoupled* state of the complex. Conversely, the *coupled* state of the complex is the physical state in which the bound ligand and the receptor interact through the $\Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)$ potential function.

Inserting Eq. (13) into Eq. (3) yields the following expression for the standard free energy of binding

$$\Delta G_b^\circ = \Delta G_{b,\mathrm{id}}^\circ + \Delta G_b \tag{17}$$

where

$$\Delta G_{b,\mathrm{id}}^\circ = -k_B T \ln C^\circ V_{\mathrm{site}} - k_B T \ln \frac{\Omega_{\mathrm{site}}}{8\pi^2} \tag{18}$$

is the *ideal* component of the standard free energy of binding, corresponding to the reversible work for transferring a ligand from an ideal solution at concentration C° to the binding site region in the absence of ligand-receptor interactions, and

$$\Delta G_b = -k_B T \ln \langle e^{-\beta u} \rangle_0 \tag{19}$$

is the *excess* component of the standard free energy of binding, corresponding to the reversible work for turning on the receptor-ligand interactions while the ligand is sequestered within the binding site region of the receptor. The goal of the computational models discussed in this chapter is the estimation of the excess free energy

of binding. The ideal component is generally computed analytically by integration of the expression that defines the indicator function of the bound complex.

Eq. (19) provides, in principle, a computational route to evaluate the binding free energy. The process is often called *alchemical* because it is unrealizable in Nature. Never the less it produces estimates that can be compared to experimental measurements. It instructs to (i) obtain a sample of Boltzmann's-distributed conformations of the complex in the uncoupled state (by molecular dynamics, typically), (ii) evaluate the binding energy function u [Eq. (15)] for each sample by turning on without conformational rearrangements the coupling between ligand and receptor, and finally (iii) find the average of the Boltzmann weight $\exp(-\beta u)$. While straightforward, this process is numerically ill-conditioned, and it fails for all but the simplest systems. This problem arises because atoms of the ligand and the receptor are very likely to clash when uncoupled. Consequently, the binding energy u is large and positive, and $\exp(-\beta u)$ is negligibly small for the vast majority of samples. Effectively, the sampling process generates mostly zeros, and the average is dominated by the very rare cases when, by chance, ligand and receptor do not clash and are primed to form favorable interactions even in the absence of such interactions.

To appreciate more quantitatively the severity of this numerical problem, let's rewrite the ensemble average in Eq. (19) as a statistical average

$$\langle e^{-\beta u} \rangle_0 = \int_{-\infty}^{+\infty} du\, e^{-\beta u} p_0(u) \tag{20}$$

where $p_0(u)$ is the probability density distribution of the binding energy in the uncoupled state. As shown for example in Fig. (1) for the complex between 3-iodotoluene and the L99A mutant of T4-lysozyme,[17] $p_0(u)$ (in green) is greatest for large and positive values of the binding energy. For this system, the probability of finding a conformation for which the integrand of Eq. (20) is significant (the red curve) is six or more orders of magnitude smaller than the probability of occurrence

of conformations with atomic clashes. It would take a prohibitively large number of independent samples of the decoupled ensemble to obtain a sufficiently large subset at favorable binding energies to estimate the binding free energy with any precision. Effectively, the binding free energy is dominated by the low binding energy tail of $p_0(u)$, which is difficult to estimate precisely and which is greatly amplified by the exponential term in Eq. (20). [8]

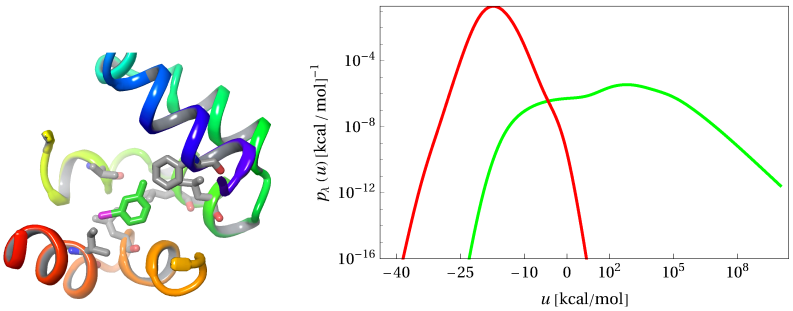

Fig. 1 The probability density $p_0(u)$ (right, green curve) of the binding energy for the alchemical uncoupled state ($\lambda = 0$) of the the complex between 3-iodotoluene and the L99A mutant of T4-lysozyme (left) at 300 K.[17] The red curve is $p_1(u)$, the probability density of the binding energy in the coupled ensemble ($\lambda = 1$), which is proportional to the integrand in Eq. (20) $\exp(-\beta u)\, p_0(u)$[18]. Note that the y-axis and the positive x-axis are in a logarithmic scale.

The 3-iodobenzene/T4-lysozyme complex illustrated in Fig. (1) is a rather simple system. The severity of the numerical problem is far greater for ligand peptides, significantly larger and more flexible than a small molecule. Random placement of a peptide molecule in the protein receptor site will almost inevitably result in conformations with atomic clashes that do not contribute significantly to the binding free energy. Moreover, peptides can assume a large variety of conformations when decoupled from the receptor, with only a small fraction of them compatible with binding, thereby further reducing the probability of generating useful bound conformations.

[8] It is tempting to try to address the clashes caused by coupling by reversing the process by decoupling. However, an equilibrium thermodynamic process like this one must be reversible, so the process's direction is irrelevant.

In practice, various strategies ranging from stratification (break up the binding process by introducing appropriate intermediate states) to importance sampling (preferential sampling of bound states) have been devised to overcome the numerical problems in alchemical free energy averages. Some of these strategies will be discussed in the case studies later in this chapter. While often very useful, applying these advanced strategies to protein-peptide complexes remains very challenging, as reflected in the paucity of successful alchemical absolute binding free energy calculations for protein-peptide complexes reported in the literature.

2.3.1 The Double-Decoupling Method

Eq. (19) is not directly applicable to the calculation of binding free energies unless the solvent potential of mean force, $W_i(\boldsymbol{x}_i)$, or a suitable implicit solvent approximation for it, is available for the ligand, the receptor, and their complex. The solvent potential of mean force is required for conformational sampling and the evaluation of effective binding energies for each sample using Eqs. (15) and (9).

The alternative is to employ an explicit representation of the solvent. The relevant partition functions include integrating the solutes' internal degrees of freedom and the degrees of freedom of the solvent molecules. The result is a binding free energy formulation known as double-decoupling[13] involving two exponential averages of the same form as Eq. (19), one for coupling the ligand from vacuum to the solvated receptor and another for coupling the ligand to the pure solvent. These two processes, the second of which is related to the solvation of the ligand, are part of a thermodynamic cycle that brings the ligand from the solvent bulk to the solvated receptor through an intermediate state in which the ligand is in vacuum (Fig. 2).

The double-decoupling method is regarded as the leading computational model for calculating protein-small molecule binding free energies. However, due to their sizes,

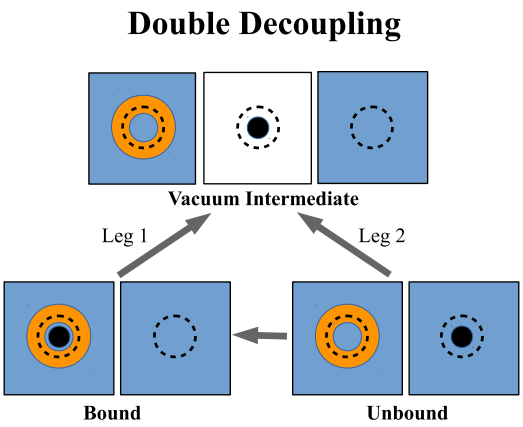

Fig. 2 Schematic illustration of the thermodynamic cycle of the double decoupling method for the calculation of the binding free energy between a molecular receptor (orange doughnut) and a ligand (black circle). The dashed circle within the receptor represents the binding site region. The blue boxes represent the solvent. The bound and unbound end states are transformed to a common intermediate state in which the ligand is in vacuum (white). The excess binding free energy is the difference of the free energy changes of the two legs, $\Delta G_b = \Delta G_2 - \Delta G_1$.

it is not generally applicable to peptides. It is presented here because it forms the basis for the relative binding free energy method employed in the case study of Section 3.2. To see why double-decoupling is not readily applicable to peptides, consider, for example, the first leg in Fig. (2), which is the inverse of the coupling of the peptide to the solvated receptor. For the same reasons outlined above concerning Eq. (17), it would be very challenging to compute the free energy of this process because, in addition to the many atomic clashes with the receptor atoms, the uncoupled peptide will also clash with solvent molecules that would be present in the binding site. Similar challenges would exist for the hydration leg.

The double-decoupling formula is derived from the statistical mechanics theory outlined in Section 2.2 by first inserting the definition of the solvent potential of mean force [Eq. (10)] in each of the configurational partition functions in Eq. (4) and then multiplying and dividing by the configurational partition function of the ligand in vacuum $Z_{0,L}$ to obtain:

$$K_b = \frac{C^\circ}{8\pi^2} \frac{Z_{N,RL}}{Z_{N,R} Z_{0,L}} \frac{Z_N Z_{0,L}}{Z_{N,L}} \tag{21}$$

where $Z_{N,i}$ is the configurational partition function of a system with N solvent molecules with one molecule of species i whose position and orientation, like in Section 2.2, is fixed. So for example

$$Z_{N,RL} = \int d\boldsymbol{x}_R d\boldsymbol{x}_L d\zeta_L I(\zeta_L) d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L, \boldsymbol{r}_v^N)} \tag{22}$$

where $U(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L, \boldsymbol{r}_v^N)$ is the potential energy function of a system with N solvent molecules containing the receptor-ligand complex RL in the configuration specified by the internal degrees of freedom $\boldsymbol{x}_R, \boldsymbol{x}_L$, and ζ_L. $Z_{0,L}$ represents the configurational partition function of the ligand in vacuum.

The reciprocal of the last term in Eq. (21) can be written as

$$\frac{Z_{N,L}}{Z_N Z_{0,L}} = \frac{\int d\boldsymbol{x}_L d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_L, \boldsymbol{r}_v^N)}}{\int d\boldsymbol{x}_L d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_L)} e^{-\beta U(\boldsymbol{r}_v^N)}} = \langle e^{-\beta u_L} \rangle_{N+L} = e^{\beta \Delta G_2} \tag{23}$$

where $u_L = U(\boldsymbol{x}_L, \boldsymbol{r}_v^N) - U(\boldsymbol{x}_L) - U(\boldsymbol{r}_v^N)$ is the instantaneous change in potential energy for bringing the ligand from vacuum to solution and $\langle \ldots \rangle_{N+L}$ indicates the ensemble average over pure solvent and the ligand in vacuum. As indicated in Eq. (23), this term is related to the solvation free energy of the ligand[9] or the opposite process of leg 2 in Figure 2.

The ratio of partition functions corresponding to the complex in Eq. (21) is converted to an average by multiplying and dividing by $V_{\text{site}}\Omega_{\text{site}}$ as done earlier to derive Eq. (13)

$$\frac{Z_{N,RL}}{Z_{N,R} Z_{0,L}} = V_{\text{site}}\Omega_{\text{site}} \langle e^{-\beta u_{RL}} \rangle_{N,R+L} = V_{\text{site}}\Omega_{\text{site}} e^{\beta \Delta G_1} \tag{24}$$

[9] Specifically, the free energy of a solute in a fixed position and orientation in vacuum to a fixed position and orientation in solution; a quantity also known as the solvation free energy in the Ben-Naim standard state.[19, 20]

where $u_{RL} = U(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L, \boldsymbol{r}_v^N) - U(\boldsymbol{x}_R, \boldsymbol{r}_v^N) - U(\boldsymbol{x}_L)$ is the instantaneous change in potential energy for bringing the ligand from vacuum to a position and orientation ζ_L relative to receptor in a solution containing the receptor, and $\langle \ldots \rangle_{N,R+L}$, similarly to Eq. (19), indicates the ensemble average over the uncoupled ensemble in which the ligand is bound to the receptor ($I(\zeta_L) = 1$) but it does not interact with either the receptor nor the solvent. As indicated in Eq. (24) this ensemble average gives the free energy of the inverse of leg 1 in Figure 2. Combining Eqs. (23), (24), (21), (17), (18) and (3) we finally arrive at the double-decoupling expression for the excess binding free energy:

$$\Delta G_b = \Delta G_2 - \Delta G_1 \tag{25}$$

as illustrated in Figure 2.

Note that the free energy formula for each leg is in the same form of an exponential average [Eq. (24)] of the alchemical potential energy change as the direct binding free energy formula we derived in Section 2.3. Thus, similar considerations apply for each leg of double-decoupling. In each case, the formula instructs to obtain samples of configurations of either the systems with the ligand in solution or the ligand in the solvated receptor in their decoupled ensembles. It then instructs to average over the set of samples the Boltmann's weight of the potential energy change for turning on the coupling between the ligand and the environment without conformational rearrangements. Here too, each leg's averaging process is expected to be numerically ill-conditioned (see, for example, Figure 1) and not generally applicable directly in molecular simulations. Some numerical approaches to this problem are illustrated in the Case Studies section of this chapter.

2.4 The Potential of Mean Force Method

In this section we derive a non-alchemical formulation of the statistical mechanics expression (4) which leads to the *potential of mean force* formula for the of binding constant.

Using the definition of the internal configurational partition function of the complex in Eq. (11) and the analogous ones for the receptor and ligand, Eq. (4) is written as

$$K_b = \frac{C^\circ}{8\pi^2} \frac{\int d\boldsymbol{x}_R d\boldsymbol{x}_L d\zeta_L I(\zeta_L) e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)}}{\int d\boldsymbol{x}_R d\boldsymbol{x}_L e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L^*)}} \tag{26}$$

where we have written the product $z_R z_L$ of the separated receptor and ligand as the partition function of a single system in which the ligand is placed in an arbitrary position ζ_L^* sufficiently removed from the receptor so that it does not interact with it. Eq. (26) is then written as

$$K_b = \frac{C^\circ}{8\pi^2} \int_{\text{site}} d\zeta_L e^{-\beta \Delta F(\zeta_L)} \tag{27}$$

where the integration is within the binding site region where $I(\zeta_L) \neq 0$, and the potential of mean force (PMF) function is defined as

$$e^{-\beta \Delta F(\zeta_L)} = \frac{\int d\boldsymbol{x}_R d\boldsymbol{x}_L e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)}}{\int d\boldsymbol{x}_R d\boldsymbol{x}_L e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L^*)}} \tag{28}$$

where $\Delta F(\zeta_L)$ is the value of the PMF at ζ_L relative to the value far away from the receptor. With this definition the PMF is zero at any point far away from the receptor.

The PMF as defined corresponds to the probability density of $p(\zeta_L)$ of finding the ligand in the orientation and position ζ_L relative to the receptor:

$$p(\zeta_L) = \frac{\int d\boldsymbol{x}_R d\boldsymbol{x}_L e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)}}{\int d\boldsymbol{x}_R d\boldsymbol{x}_L d\zeta_L' e^{-\beta \Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)}} = \langle \delta(\zeta_L' - \zeta_L) \rangle \tag{29}$$

so that

$$\Delta F(\zeta_L) = -k_B T \ln \frac{p(\zeta_L)}{p(\zeta_L^*)} \tag{30}$$

The potential of mean force expression (27) formally instructs to map out the probability density (29) to observe the ligand around the receptor in orientation and position ζ_L, including far away from the receptor and within the binding site region, and to then integrate it within the binding site region to obtain the binding constant using Eq. (27).

Some comments are in order. First, the PMF function can be obtained in the solvent of potential of mean force formulation as suggested by Eq. (28) or using an explicit representation of the solvent by inserting the definitions of the effective potential energy Ψ and of the solvent of potential of mean force (10) into Eq. (28)

$$e^{-\beta \Delta F(\zeta_L)} = \frac{\int d\boldsymbol{x}_R d\boldsymbol{x}_L d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_R, \boldsymbol{x}_L, \boldsymbol{r}_v^N, \zeta_L)}}{\int d\boldsymbol{x}_R d\boldsymbol{x}_L d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_R, \boldsymbol{x}_L, d\boldsymbol{r}_v^N, \zeta_L^*)}} \tag{31}$$

It is evident therefore that the PMF is obtained by monitoring the probability of occurrence of the ligand at ζ_L whether an implicit or explicit description of the solvent is used.

Secondly, the potential of mean force formula for the binding constant (27) does not require knowledge of the probability density $p(\zeta_L)$ everywhere around the receptor. It requires it only within the binding site region and at one arbitrary point ζ_L^* far away from the receptor in the solvent bulk to compute $\Delta F(\zeta_L)$ from Eq. (30). The latter is a fundamental point. It is not sufficient to study the distribution of placements of the ligand in the binding site o compute the binding free energy. We also require the probability of finding the ligand in the binding site relative to finding it somewhere in the solvent bulk. In practice, the PMF is obtained in a volume that includes both the binding site and positions far away from the receptor to connect the two regions in a statistical sense.[21, 22, 23]

Finally, the PMF is rarely obtained over all six degrees of freedom of ζ_L (three positions and three orientations). In practice, the PMF is collected only along some of the dimensions by averaging over the others. The averaging procedure is formally described by marginalization of $p(\zeta_L)$. For example, to obtain the probability of the position $\boldsymbol{r}_L$ of the ligand regardless of its orientation we integrate $p(\zeta_L) = p(\boldsymbol{r}_L, \theta_1, \psi_1, \psi_2)$ over the three Euler angles θ_1, ψ_1, and ψ_2

$$p(\boldsymbol{r}_L) = \int d(\cos\theta_1) d\psi_1 d\psi_2 p(\boldsymbol{r}_L, \theta_1, \psi_1, \psi_2) \tag{32}$$

In the bulk, the ligand distribution does not depend on the orientation and we get

$$p(\boldsymbol{r}_L^*) = \int d(\cos\theta_1) d\psi_1 d\psi_2 p(\boldsymbol{r}_L^*, \theta_1, \psi_1, \psi_2) = 8\pi^2 p(\zeta_L^*) \tag{33}$$

Next, integrate Eq. (27) over θ_1, ψ_1, and ψ_2, assuming that the binding site definition does not depend on orientations, and expressing $e^{-\beta\Delta F(\zeta_L)}$ as $p(\zeta_L)/p(\zeta_L^*)$, to obtain

$$K_b = \frac{C^\circ}{8\pi^2} \int_{\text{site}} d\boldsymbol{r}_L p(\boldsymbol{r}_L)/p(\zeta_L^*) = K_b = C^\circ \int_{\text{site}} d\boldsymbol{r}_L e^{-\beta\Delta F(\boldsymbol{r}_L)} \tag{34}$$

where

$$\Delta F(\boldsymbol{r}_L) = -k_B T \ln \frac{p(\boldsymbol{r}_L)}{p(\boldsymbol{r}_L^*)} \tag{35}$$

and we have used Eqs. (32) and (33). The implementation of Eq. (34) requires the PMF with respect to the position of the ligand regardless of its orientation.

3 Case Studies of Applications of Free Energy Methods to Protein-Peptide Binding Free Energy Estimation

In this section, we review some applications of the free energy methods derived from the statistical mechanics theory of non-covalent molecular binding introduced in Section 2.2 to the study of protein-peptide binding phenomena. We will focus in particular on theoretical and methodological aspects that will be introduced and discussed as needed. The following case studies are far from an exhaustive representation of the literature in the field. They have been selected primarily to illustrate the application of the theory and methods presented in Section 2. We also do not attempt to review each work exhaustively.

3.1 Binding of Cyclic Peptides to HIV Integrase with the Single-Decoupling Method and Implicit Solvation

As part of the infection cycle, HIV inserts its genome into a human chromosome. The HIV integrase (IN) enzyme responsible for this process is recruited to the nuclear chromatin by the human lens epithelium-derived growth factor (LEDGF) transcriptional coactivator.[24] There have been significant attempts[8, 25, 26, 27] to develop therapies against HIV based on disrupting the interaction of LEDGF with HIV IN, which occurs at the so-called LEDGF binding domain of integrase (Fig. 3). The study of the interaction of LEDGF and LEDGF-derived synthetic peptides with HIV-IN has provided useful insights for competitive inhibitors' design.[28, 29] As an example, Figure 3 illustrates the crystal structure of the LEDGF binding domain of the HIV IN dimer complexed with a cyclic peptide.[29]

Building upon an earlier successful application of alchemical binding free energy calculations of small-molecule inhibitors targeting the LEDGF/HIV IN

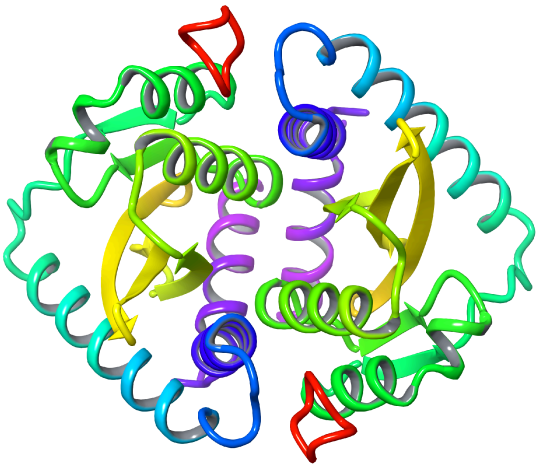

Fig. 3 The 3AVN crystal structure of the dimer of the LEDGF binding domain of HIV integrase (multi-color ribbons) bound to SHKIDNLD cyclic peptides (red tube).[29]

interaction,[30] Kilburg and Gallicchio[31] modeled the binding free energies between HIV IN and of five of the thirteen cyclic peptides assayed by Rhodes et al.[29] The alchemical binding free energy study by Kilburg and Gallicchio recapitulated the trends observed in the experimental assays and identified the specific structural and energetic signatures responsible for favorable binding. Conversely, the calculations provided explanations for the lack of binding observed for two sequences for which structural information is not available.

The study by Kilburg & Gallicchio[31] remains one of a few examples of the successful application of alchemical free energy methods to the computation of the absolute binding free energies of protein-peptide complexes. This was made possible by employing a implementation of Eq. (17) which was first reported under the name of Binding Energy Distribution Analysis Method (BEDAM).[18, 32] as part of the IMPACT molecular simulation program.[33] The latest implementation

as a plugin of the OpenMM molecular dynamics library[34] has been named the Single-Decoupling Method (SDM),[17][10] a name chosen to better place it in the same theoretical context as the Double-Decoupling Method (DDM)[13] discussed in Section 2.3.1. In the following, we will use the latter name to refer to both implementations. SDM has been used in two studies involving protein-peptide binding to date.[35, 31]

The implementation of Eq. (17) requires the averaging of the Boltzmann weight of the effective binding energy in Eq. (15), which in turn requires the specification of the intramolecular potential energy $U_i(\boldsymbol{x}_i)$ and the solvent potential of mean force $W_i(\boldsymbol{x}_i)$ for each configuration $\boldsymbol{x}_i$ of the molecular species involved. The former is available from a molecular mechanics force field (OPLS-AA[36] in the applications discussed here) while the solvent potential of mean force is approximated by an implicit solvent model.[16] SDM employs the Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent model[37, 38] which is now maintained as an OpenMM plugin.[39][11]

3.1.1 Alchemical Pathways and Stratification

We use this case study to illustrate the very general concept of an *alchemical pathway* and the idea of performing conformational sampling along the pathway to improve the convergence characteristics of the basic binding free energy formula [Eq. (19)]. This technique, commonly known in the field as *stratification* is used in many free energy estimation problems.[40]

As discussed in Section 2.3, Eq. (19) is not directly applicable in numerical simulations because, fundamentally, the coupled and uncoupled ensembles preferentially visit distinct regions of conformational space (see Figure 1 for example). The free

[10] `github.com/rajatkrpal/openmm_sdm_plugin`
[11] `github.com/egallicc/openmm_agbnp_plugin`

energy, however, is a thermodynamic state function, and it should be possible to compute it as the sum of free energy changes over a series of intermediate states, each sufficiently similar to its neighbors so that free energy estimation formulas such as Eq. (19) among these are numerically well-behaved.[41, 42] [12] The intermediate so-called *alchemical states* are generally implemented by means of an alchemical progress parameter λ that tunes the system's potential energy function such that $\lambda = 0$ corresponds to the initial state and $\lambda = 1$ corresponds to the final state. A simple–but not necessarily the most efficient[17, 43] choice–is a linear interpolating function of the form

$$U_\lambda(x) = U_0(x) + \lambda u(x) \tag{36}$$

where $U_0(x)$ is the potential energy function that describes the initial state and $u(x) = U_1(x) - U_0(x)$, where $U_1(x)$ is the potential function of the final state, is the *perturbation potential*. The progress parameter λ and the specific parameterization of the alchemical potential is said to define an *alchemical path* that connects, in a thermodynamic sense, the initial and final states.

The specific alchemical potential energy function adopted by Kilburg & Gallicchio[31] to study peptide binding is, in the notation of Section 2.3,

$$\Psi_\lambda(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L) = \Psi(\boldsymbol{x}_R) + \Psi(\boldsymbol{x}_L) + \lambda u(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L) \tag{37}$$

where the first term on the r.h.s. is the potential energy function of the decoupled ensemble (corresponding to $U_0(x)$ in Eq. (36)) and the binding energy function u is defined by Eq. (15). [13] It is straightforward to see that Ψ_λ at $\lambda = 1$ is the potential energy function of the coupled state. An *alchemical binding free energy*

[12] This concept has since evolved into rigorous statistical interpretations and numerical algorithms, some of which are discussed later in this section.

[13] To improve convergence, Kilburg & Gallicchio actually used a *soft-core* form of the binding energy function.[44, 17] Soft-core functions are critical aspects of alchemical binding free energy calculations.

profile, $\Delta G(\lambda)$, along the thermodynamic path is defined, which corresponds to the free energy of the intermediate alchemical state at λ relative to the uncoupled state ($\lambda = 0$)[18]

$$\Delta G(\lambda) = -k_B T \ln\langle e^{-\beta\lambda u}\rangle_0 \tag{38}$$

which is Eq. (19) with u replaced with λu, the perturbation energy at the alchemical state at λ. By definition, the excess free energy of binding (19) is the difference between the end points of the alchemical binding free energy profile

$$\Delta G_b = \Delta G(\lambda = 1) - \Delta G(\lambda = 0) \tag{39}$$

In Kilburg & Gallicchio's study, the alchemical path was subdivided into 26 intermediate states mostly linearly spaced between 0 and 1, except the region near $\lambda = 0$, which required more closely spaced points. Conformational sampling was conducted at each λ-state by molecular dynamics (MD)[14] using the alchemical potential energy function (37). The binding energy function (15) and its gradients were evaluated at each MD time step by first evaluating the potential energy of the complex $\Psi_{RL}(\boldsymbol{x}_R, \boldsymbol{x}_L, \zeta_L)$ and then displacing the peptide in the implicit solvent medium at a large distance away from the protein receptor to evaluate the potential energy $\Psi_R(\boldsymbol{x}_R) + \Psi_L(\boldsymbol{x}_L)$ without protein-peptide interactions.[15] Samples of the decoupled energy $\Psi_0 = \Psi_R(\boldsymbol{x}_R) + \Psi_L(\boldsymbol{x}_L)$ and of the binding energy u were saved at each alchemical state at regular intervals. As discussed in Section 3.1.3, these are the inputs for the estimation of the binding free energy profile and of the excess binding free energy through Eq. (39).

[14] Specifically by replica-exchange molecular dynamics in temperature and λ space as described in Section 3.1.2)

[15] The ligand displacement approach to compute the alchemical potential energy was made necessary by the many-body nature of the implicit solvation model. As briefly discussed in Section 3.2, with pairwise decomposable potentials it is more common that λ is is integrated into the calculation of individual interatomic interaction energies.

3.1.2 Replica Exchange Conformational Sampling

Stratification implies that an alchemical binding free energy calculation is commonly carried out as a collection of molecular simulations, each with a different alchemical potential energy function [Eq. (36)] at a series of values of the alchemical progress parameter λ. The accuracy of alchemical free energy calculations depends heavily on the conformational sampling's quality at each λ-state. In this context, the conformational sampling's challenge is to generate a diverse set of configurations distributed according to Boltzmann's distribution for the given temperature and potential energy function. It is not sufficient, like in molecular docking, to propose a set of low-energy configurations. The configurations should also appear according to their probability of occurrence. Conformational sampling in alchemical simulations is carried out by Monte Carlo and, more often, Molecular Dynamics (MD). MD conformational sampling is limited by the slow time-scales of biomolecules' motion, and a host of advanced conformational sampling algorithms have been devised to overcome it.[45] Kilburg & Gallicchio employed two-dimensional replica-exchange conformational sampling in temperature and alchemical spaces.[46, 31]

It is useful to consider separately the problem of sampling intermolecular degrees of freedom (the position and orientation of the ligand relative to the receptor, denoted by ζ_L above) from the sampling of intramolecular degrees of freedom (the individual conformations of the peptide and the receptor, denoted by $\boldsymbol{x}_L$ and $\boldsymbol{x}_R$). The first problem is related to the simulation algorithm's ability to explore all relevant binding modes of the protein-receptor complex for fixed receptor and peptide conformations. Missing the most stable binding mode would, of course, underestimate the binding affinity. The sampling of intermolecular degrees of freedom is straightforward near the decoupled state ($\lambda \simeq 0$) where protein-peptide interactions are weak, and the peptide can nearly freely translate and rotate within the binding site volume. In contrast, because of receptor-peptide interactions, rotations and translations are severely

hindered near the coupled state ($\lambda \simeq 1$) where the peptide visits alternative binding modes only very rarely. Therefore, one solution to this problem is to make it so that the MD thread evolves the system in conformational space as well as λ space. In this way, new binding modes are formed when λ is small and, if they are sufficiently stable, they will be retained when the MD thread visits more strongly coupled states at $\lambda \simeq 1$. Conversely, an MD thread in a metastable binding mode at $\lambda \simeq 1$ would have an opportunity to acquire a smaller λ and convert to another binding mode. Of course, the excursions in λ space have to be so that a canonical ensemble of conformations is generated at each alchemical state.

The replica exchange algorithm achieves this by evolving as many MD threads as there are alchemical states. At any one point in time, each MD thread j is assigned the λ value of a unique alchemical state j. The collection of threads, called *replicas*, forms an ensemble of independent canonical systems with the joint canonical statistical weight function

$$\rho_{\mathrm{RE}}(x_1, \ldots, x_n | \lambda_1, \ldots, \lambda_n) = \exp\left[-\sum_{j=1}^{n} \beta \Psi_{\lambda_j}(x_j) \right] \tag{40}$$

where $\Psi_\lambda(x)$ is the alchemical potential energy function (37), x_j denotes the configuration of replica j, and λ_j is the value of λ assigned to it. The joint distribution is sampled by alternating updates of coordinates x_j at a fixed assignment of λ values, which is accomplished independently for each replica by conventional constant temperature MD, with updates of the λ assignments. The latter is performed at fixed by proposing permutations of λ assignments $\{\lambda_1, \ldots, \lambda_n\} \to \{\lambda'_1, \ldots, \lambda'_n\}$ at fixed configurations x_j and accepting and rejecting the move using the Metropolis Monte Carlo algorithm based on the ratio of the values of the proposed and original weight functions

$$\frac{\rho_{\mathrm{RE}}(x_1, \ldots, x_n | \lambda'_1, \ldots, \lambda'_n)}{\rho_{\mathrm{RE}}(x_1, \ldots, x_n | \lambda_1, \ldots, \lambda_n)} \tag{41}$$

There are many variations of replica-exchange differing in the nature of the replicas, the scheme of permutations of state assignments, and the computational implementation.[47] Schemes, such as the one illustrated above, that modify the parameters of the potential energy function are known in the field as Hamiltonian replica exchange algorithms.[48] Kilburg & Gallicchio used the Gibbs Independent Sampling Algorithm[17] for Hamiltonian reassignments and an *asynchronous* implementation[46] of replica-exchange for that allows running the collection of replica simulations on heterogeneous and potentially unreliable computational resources such as on computational grids.[49]

Hamiltonian replica exchange addresses the sampling of intermolecular degrees of freedom. However, because λ couples receptor-peptide interactions, it has only an indirect influence on the rate at which intramolecular degrees of freedom are sampled. Peptides are very flexible and often change conformation upon binding. They often interact with the protein over an extended surface and induce substantial induced-fit reorganization of the receptor. Conformational rearrangements of peptides occur very slowly at room temperature, especially of the cyclic peptides investigated in this study. The temperature replica-exchange algorithm, one of the first versions of replica-exchange proposed,[50] is very useful for accelerating the sampling of the conformational space of peptides and proteins,[51, 52] and is applicable to free energy calculations.[53] Kilburg & Gallicchio adopted a two-dimensional replica-exchange scheme in which both the λ and temperature assignments undergo permutations. The joint canonical weight is generalized as

$$\rho_{\mathrm{RE}}[x_1, \ldots, x_n|(\beta, \lambda)_1, \ldots, (\beta, \lambda)_n)] = \exp\left[-\sum_{j=1}^{n} \beta_j \Psi_{\lambda_j}(x_j)\right] \tag{42}$$

where β_j and λ_j are the inverse temperature and λ assigned to replica j, and (β, λ) is one of the n pair combinations of a set of inverse temperatures and alchemical

states. Kilburg & Gallicchio employed 8 temperatures between 300 to 379 K and 26 alchemical states for a total of 208 replicas for each protein-peptide complex. The multi-dimensional replica-exchange algorithm employed allowed to explore simultaneously multiple conformations of the peptide and multiplied binding modes of each conformation.

3.1.3 Multi-State Free Energy Estimation

While Eq. (19) is formally correct, it is not an optimal free energy estimator. Optimal here refers to a free energy estimator's ability to return a free energy estimate with the smallest bias relative to the true free energy (accuracy) and smallest variance (precision) with a given finite set of samples. Kilburg & Gallicchio employed the Unbinned Weighted Histogram Analysis Method (UWHAM) estimator[44] which is considered an optimal free energy estimator when no information of the system is known other than the samples from the molecular simulations. The statistical and mathematical origins of the method[54, 44] are beyond the scope of this chapter. The main idea is to arrive at an estimate of the free energy $\Delta G(\lambda)$ [Eq. (38)] at λ by using the data collected at all λ-states. UWHAM can be interpreted as an extension of the familiar Weighted Histogram Analysis Method (WHAM),[55] applied to Eq. (20) for the maximum likelihood estimation of the distribution of binding energies in the uncoupled ensemble $p_0(u)$ from the corresponding distributions along the alchemical path $p_\lambda(u)$.

In this case, Kilburg & Gallicchio collected data as a function of temperature as well as λ on a grid of 208 states. UWHAM provides, in this case, optimal estimates of the dimensionless free energy factor for each state defined as, up to a additive constant,[16]

[16] Note that, because z_{RL} is not dimensionless, the ambiguity of the additive constant is also related to the arbitrariness of the units chosen to evaluate the logarithm.

$$F_r = \ln z_{RL}(\beta_r, \lambda_r) \tag{43}$$

where β_r and λ_r are the values of the inverse temperature and of the alchemical progress parameter of state r and $z_{RL}(\beta, \lambda)$ is defined by Eq. (11). Given the free energy factors, the free energy profile as function of temperature and λ is given by Eq. (38), or[17]

$$\Delta G(\beta_r, \lambda_r) = -k_B T F_r \, . \tag{44}$$

The dimensionless free energy factors minimize the convex objective function[44][18]

$$\frac{1}{N} \sum_{s=1}^{N} \ln \left[\sum_{r=1}^{n} \frac{N_r}{N} e^{-F_r} e^{-v_{rs}} \right] + \sum_{r=1}^{n} \frac{N_r}{N} F_r \tag{45}$$

where N is the total number samples collected at any of the n states, N_r is the number of samples collected at state r, and

$$v_{rs} = \beta_r [\Psi_{0,s} + \lambda_r u_s] \tag{46}$$

is the dimensionless energy of sample s in state r, where $\Psi_{0,s}$ and u_s are, respectively, the values of the decoupled potential energy and of the binding energy of the sample collected during the replica-exchange alchemical simulations. The UWHAM optimizer implemented in the statistical program `R` was used to obtain the dimensionless free energy factors (`cran.r-project.org/web/packages/UWHAM`).[19]

Note that setting to zero the gradient of the UWHAM objective function leads to the self-consistent equations

---

[17] Because the free energy estimates are known up to a temperature-dependent additive factor, differences between free energies at different temperatures are generally meaningless. However differences along λ at different temperature can be compared. For example, the binding free energy at one temperature $\Delta G_b(\beta) = \Delta G(\beta, \lambda = 1) - \Delta G(\beta, \lambda = 0)$ can be compared to the binding free energy estimate at a different temperature to, for example, estimate the binding entropy.

[18] The convexity property guarantees that there is a unique minimum.

[19] Ding, Vilseck, and Brooks[56] developed a GPU implementation of UWHAM called FastMBAR (`github.com/xqding/FastMBAR`).[56]

$$f_r^{-1} = \sum_{s=1}^{N} \frac{e^{-v_{rs}}}{\sum_{r'=1}^{n} N_{r'} f_{r'} e^{-v_{r's}}} \tag{47}$$

where $f_r = e^{-F_r}$. Eq. (47) is the basis of the equivalent Multistate Bennet Acceptance Ratio (MBAR) method to obtain the free energy factors.[57] The UWHAM formulation of multi-state reweighting has been found to be more generalizable than MBAR's.[56] For example it has been recently employed to impose global restraints on the free energy solutions.[58]

3.2 Effect of Mutations on the Binding Affinity of Peptides to PDZ Protein Domains

PDZ protein domains are widespread protein-protein interaction modules. They specifically recognize the 4 to 8 aminoacids at the C-terminus sequence of proteins. Peptides and peptide derivatives that mimic these binding motifs are investigated as potential therapeutics for many diseases.[59] Panel et al.[60] studied the binding free energies between the TIAM1 PDZ domain and a series of peptides derived from its syndecan-1 and caspr4 protein targets (Figure 4) using an alchemical relative binding free energy computational method generally known in the field as Free Energy Perturbation (FEP).[61, 62] The study's goal was to validate the methodology for protein-peptide binding and obtain physical and structural insights into the recognition mechanisms that allow PDZ domain to target specific sequences.

3.2.1 Theory of Relative Binding Free Energy Calculations

The dataset considered by Panel et al.[60] included the TIAM1 PDZ domain bound to the wild-type peptides and a series of single and double mutants. As discussed in Section 2.3.1, peptides are generally too large and complex to be studied by

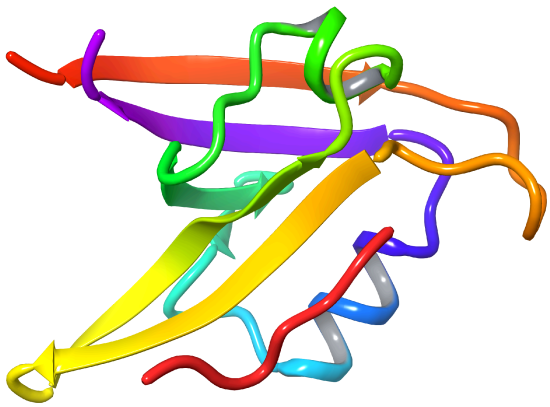

Fig. 4 The 4GVD crystal structure of the complex between the TIAM1 PDZ domain (multi-color ribbons) and the pTKQEEFYA peptide (red tube).[63]

double-decoupling absolute binding free energy calculations with explicit solvation. Instead, the study employed a relative FEP method that yields the difference between a peptide's binding free energies relative to a reference peptide. The approach is based on the thermodynamic cycle illustrated in Figure 5. The reference peptide L_1 is alchemically transformed into a mutant L_2 when bound to the receptor and solvated in water. The difference in the free energies ΔG_{bound} and ΔG_{solv} of these two processes yields the difference in the binding free energy of the two complexes. Therefore, the method allows probing the effect of different mutations on the binding affinity between the peptide and the receptor.

The statistical mechanics formula at the basis of this approach can be derived, for example, from Eq. (21) by considering the expression of the ratio of the binding constants $K_b(2)$ and $K_b(1)$ for the RL_2 and RL_1 complexes, respectively. When taking the ratio, the constant factors and the partition functions of the solvent, of the receptor in the solvent, and of the ligands in vacuum cancel yielding

$$\frac{K_b(2)}{K_b(1)} = \frac{Z_{N,RL_2}}{Z_{N,RL_1}} \frac{Z_{N,L_1}}{Z_{N,L_2}} = e^{-\beta[\Delta G_{\text{bound}} - \Delta G_{\text{solv}}]} \tag{48}$$

where the ratio of partition functions involving the receptor correspond to the free energy difference ΔG_{bound} between the complex with ligand L_2 in the solvent and the same system but with ligand L_2 replaced by L_1. Similarly, the ratio of partition functions of the ligands in solution correspond to the free energy difference ΔG_{solv}. [20] Finally, using Eq. (2), we obtain

$$\Delta\Delta G_b := \Delta G_b(2) - \Delta G_b(1) = \Delta G_{\text{bound}} - \Delta G_{\text{solv}} \tag{49}$$

which is the key formula of the relative binding FEP method.

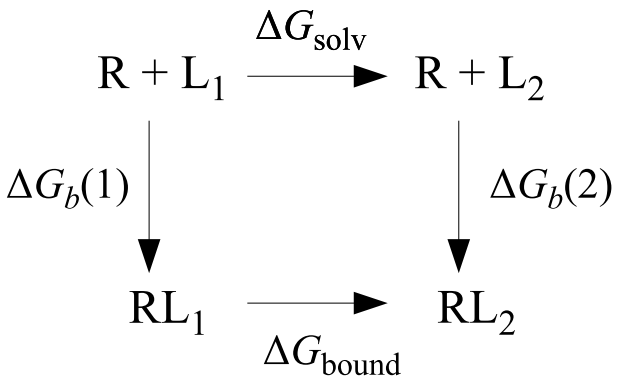

Fig. 5 The thermodynamic cycle used in the relative free energy perturbation method. The vertical transformations correspond to the association equilibrium between the receptor R and one of two ligands L_1 and L_2. The horizontal legs correspond to the alchemical transformation of one ligand into the other alone in solution (top) or in the complex (bottom).

[20] Comparing the free energies of systems with different atomic composition and number of degrees of freedom is arguably physically meaningless at this level of theory. However, note that the overall ratio of partition functions in Eq. (48) if physically well defined. It represents the free energy difference between two systems, the first composed of two solutions one containing the complex with L_2 and the other containing L_1, and the second in which L_2 and L_1 have swapped places. Evidently, the free energy difference $\Delta G_{\text{bound}} - \Delta G_{\text{solv}}$, which is the target of the theory, is physically well defined even though the individual components may not be.

Let's now turn to the evaluation of ΔG_{bound} and ΔG_{solv} by alchemical computer simulations. As usual, the strategy is to compute ratios of partition functions as ensemble averages. However, for example, the expression

$$\frac{Z_{N,L_2}}{Z_{N,L_1}} = \frac{\int d\boldsymbol{x}_{L_2} d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_{L_2}, \boldsymbol{r}_v^N)}}{\int d\boldsymbol{x}_{L_1} d\boldsymbol{r}_v^N \, e^{-\beta U(\boldsymbol{x}_{L_1}, \boldsymbol{r}_v^N)}} \tag{50}$$

cannot be directly turned into the form of an ensemble average because, in general, the number and kind of the internal degrees of freedom of the two ligands differ. Panel et al.[60] adopted the so-called dual topology strategy to address this issue,[21] in which the simulation is conducted with a hybrid peptide in which the wild-type, say, and mutated aminoacid side chains are both represented at the same time (Figure 6). The alchemical potential energy function is constructed so that the environment (the water solution or the solvated receptor) interacts with the atoms of both forms of the sidechain with a strength that depends on the alchemical charging parameter λ. Similarly, the intramolecular potential energy function is designed so that the atoms of the protein backbone interact by bond stretching, bond angle, torsional, and 1,4 non-bonded interactions with both forms of the sidechain. The atoms of the two forms of the sidechain being mutated never interact directly with each other.

Formally, the dual-topology approach is derived from Eq. (48) by multiplying and dividing each term by an appropriate partition function that introduces the additional degrees of freedom to turn each peptide into the hybrid peptide with both forms of the sidechain. For example, if Z_{N,L_1} term represents the peptide with the phenylalinine (PHE) sidechain in solution (Figure 6, red), multiplying and combining it with

$$Z_{\text{ILE}} = \int d\zeta_{\text{ILE}} d\boldsymbol{x}_{\text{ILE}} e^{-\beta U(\zeta_{\text{ILE}})} e^{-\beta U(\boldsymbol{x}_{\text{ILE}})} \tag{51}$$

[21] There is an analogous single-topology strategy[64] which we do not discuss here.

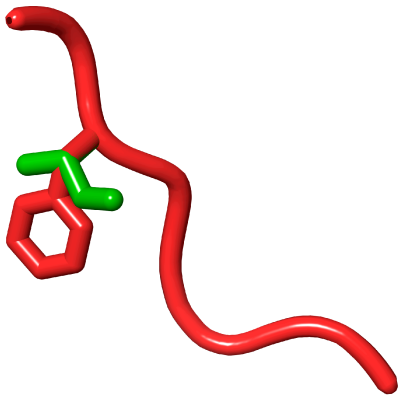

Fig. 6 Representation of the dual-topology alchemical mutation of a phenylalanine (PHE, red) to isoleucine (ILE, green) of the TKQEEFYA peptide considered by Panel et al.[60] The illustration shows the peptide in solution. A similar transformation is applied to the peptide bound to the PDZ domain.

where, as in Eq. (11), ζ_{ILE} represents the six external coordinates that specify the position and orientation of the added isoleucine (ILE) sidechain relative to the peptide backbone, $U(\zeta_{\mathrm{ILE}})$[22] represents the potential energy terms that anchor the ILE side chain to the peptide backbone,[23] $\boldsymbol{x}_{\mathrm{ILE}}$ represents the other internal degrees of freedom of the ILE side chain, and $U(\boldsymbol{x}_{\mathrm{ILE}})$ represents the intramolecular potential energy function that couples atoms of ILE together, [24] transforms it into the partition function, that we will denote by $Z_{N,L_{1(2)}}$, of the hybrid peptide in the PHE state in which the ILE sidechain is "turned off," by which we mean that the ILE sidechain interacts only with the backbone through the $U(\zeta_{\mathrm{ILE}})$ potential and

[22] As explained there, this function acquires in the next section an "SD" superscript.

[23] Other attachment modalities, including to the β carbon, are possible.

[24] As further discussed later, here we have explicitly singled-out the ζ degrees of freedom that couple the added sidechain to the backbone to emphasize that they must be appropriately chosen, using, for example, the scheme described by Boresch & Karplus[10], to avoid introducing spurious indirect interactions between backbone atoms that would affect the conformational distribution of the original peptide.[65]

does not otherwise interact with the environment. The same procedure applied to the partition function of the complex of the original peptide bound to the receptor Z_{N,RL_1} in the denominator of Eq. (48) yields the partition function $Z_{N,RL_{1(2)}}$ of the hybrid peptide in the PHE state bound to the receptor. Similarly, multiplying and dividing by the term Z_{PHE} analogous to Eq. (51) to install a PHE sidechain onto the peptide with the ILE sidechain, yields the partition functions $Z_{N,L_{(1)2}}$ and $Z_{N,RL_{(1)2}}$ for the hybrid peptides in solution and bound to the receptor in their ILE states.

With these preparations, finally Eq. (48) is rewritten as

$$\frac{K_b(2)}{K_b(1)} = \frac{Z_{N,RL_{(1)2}}}{Z_{N,RL_{1(2)}}} \frac{Z_{N,L_{1(2)}}}{Z_{N,L_{(1)2}}} = e^{-\beta\Delta G_{\mathrm{bound}}} e^{+\beta\Delta G_{\mathrm{solv}}} \tag{52}$$

where

$$\Delta G_{\mathrm{bound}} = -k_B T \ln \frac{Z_{N,RL_{(1)2}}}{Z_{N,RL_{1(2)}}} = -k_B T \ln\langle e^{-\beta u_2}\rangle_1 \tag{53}$$

where u_2 is the change in potential energy of the system for a given configuration of the solvated complex with the hybrid peptide due to, in this example, turning off PHE sidechain and turning on the ILE sidechain, and $\langle \ldots \rangle_1$ represents the average over the ensemble in which the PHE sidechain is on and the ILE sidechain is off. An analogous ensemble average gives ΔG_{solv} for the transformation of PHE into ILE in solution.

3.2.2 Alchemical Transformations for Relative Binding Free Energies

As discussed in Sections 2.3 and 3.1.1 the free energies ΔG_{solv} and $\Delta G_{\mathrm{bound}}$ for mutating one sidechain into another are calculated in practice using an hybrid alchemical potential energy function $U_\lambda(x)$ parametrized by a progress parameter λ. Panel et al.[60] used the NAMD molecular simulation package[66] which implements the alchemical potential[65]

$$U_\lambda(x) = U_{L_{12}}(x) + (1-\lambda)U_{L_1}(x, 1-\lambda) + \lambda U_{L_2}(x, \lambda) \tag{54}$$

where x is the collection of all of the degrees of freedom of the dual-topology peptide system, $U_{L_{12}}(x)$ contains the potential energy terms that do not depend on λ

$$U_{L_{12}}(x) = U_0(x) + U^{\mathrm{SD}}_{L_1}(\zeta_1) + U^{\mathrm{SD}}_{L_2}(\zeta_2) \tag{55}$$

where $U_0(x)$ is the unperturbed component of the potential energy (including the intramolecular potential energy terms of the dual-topology sidechains not affected by the transformation, but excluding interactions between the two sidechains), and the terms $U^{\mathrm{SD}}_{L_i}(\zeta_i)$ represent the auxiliary restraints used in the dual-topology scheme to anchor each sidechain to the backbone [see Eq. (51)], and

$$U_{L_i}(x, \lambda) = U^{\mathrm{NB}}_{L_i}(x, \lambda) + U^{\mathrm{SS}}_{L_i}(x) + U^{\mathrm{SD}}_{L_i}(x) \tag{56}$$

where $U^{\mathrm{NB}}_{L_i}$ denotes non-bonded interactions between the sidechain atoms and the environment, $U^{\mathrm{SS}}_{L_i}$ denotes the bonded (1-2, 1-3, and 1-4 interactions) among backbone atoms with sidechain i, and $U^{\mathrm{SD}}_{L_i}$ is the corresponding term for bonded interactions between the backbone atoms and the sidechain. [25] As illustrated by Eq. (56), the non-bonded component has an explicit λ dependence due to the use of separation-shifted soft-core pair potentials[67, 65] to describe the non-bonded interactions between the dual-topology sidechains and the rest of the system.

It is straightforward to see that Eq. (54) evaluated at $\lambda = 0$ describes the $L_{1(2)}$ state of the dual-topology peptide with sidechain 2 turned off and, conversely, $\lambda = 1$ describes the $L_{(1)2}$ state. Panel et al.[60] simulated 11 alchemical states from $\lambda = 0$ to $\lambda = 1$. The change in free energy from λ_r to λ_{r+1} was evaluated using the Bennet Acceptance Ratio (BAR) method, which is MBAR [Eq. (47)] for two states and

[25] The S symbol stands for the single-topology region (the backbone in this case), and D stands for dual-topology region (the two sidechains).[67, 65]

where, in this case,

$$v_{rs} = \beta U_{\lambda_r}(x_s) \tag{57}$$

is the alchemical potential energy at λ_r of the conformational sample x_s collected at either λ_r or λ_{r+1}.[26]

3.3 Potential of Mean Force Study of the Binding of the MEEVD peptide to the TPR2A Receptor

The heat shock organizing protein (Hop) binds specifically to the heat shock protein Hsp90 through its tetraticopeptide repeat (TPR) domain TPR2A. TPR modules are widespread protein domains responsible for the specific recognition patterns of many proteins. Due to their molecular recognition characteristics, engineered TPR domains are seen as potential alternatives to antibody-derived biological medicines. Lapelosa[22] studied the binding of the MEEVD peptide from Hsp90 to the TPR2A domain of Hop (Figure 7) using the potential of mean force methodology outlined in Section 2.4. The work yielded an estimate of the standard free energy of binding between TPR2A and MEEVD in good agreement with experimental measurements. It provided structural insights into the entry and exit mechanism of the peptide from the receptor binding site.

3.3.1 Calculation of the Standard Binding Free Energy

Lapelosa[22] computed a 1-dimensional radial potential of mean force (PMF), $\Delta F(r)$, along the center of mass separation r between the receptor and the peptide (Figure 7) using the Adaptive Biasing Force (ABF) method described in the

[26] The numerator and the denominator of Eq. (47) are often combined to cast the formula in terms of energy differences $v_{r's} - v_{rs}$.

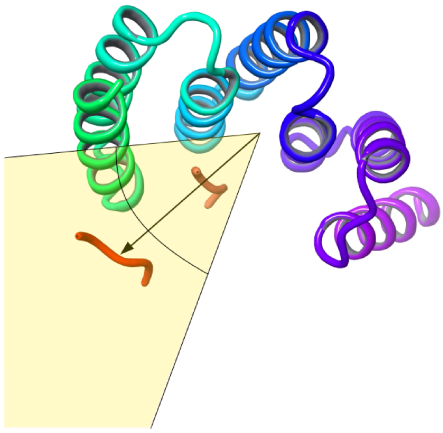

Fig. 7 Illustration of the calculation of the binding free energy of the complex between the Hop TPR2A domain (multi-colored ribbon) with the MEEVD peptide (red and pink tubes). The MEEVD peptide is shown in its position in the crystal structure (PDB id: 1ELR[68], pink tube) and in a representative position and orientation (red tube) within the simulation cone (the yellow shaded region). The potential of mean force is collected along the distance (black arrow) between the center of mass of the receptor and the center of mass of the peptide while the peptide is kept within the cone. The arc across the cone delineates the binding site region $r < r_b$.

next section. The PMF was then employed to compute the free energy of binding. The expression of the binding constant in terms of the radial PMF is derived from Eqs. (34) and (35) by expressing the integral in terms of spherical polar coordinates (r, θ, ϕ), where r is the distance between the centers of mass of the receptor and the peptide, θ is the angle between the line connecting the centers of masses and the axis connecting the C-α atoms of two chosen residues of the receptor, and ϕ is an azimuthal angle (which can be considered arbitrary because neither the conical sampling region nor the binding site region depends on it). Following a procedure similar to the one that yielded Eq. (34) from Eq. (27), we carry out the integration in Eq. (34) over the θ and ϕ coordinates to obtain

$$K_b = C^\circ \int_0^{r_b} dr r^2 \frac{p(r)}{p(r^*, \theta^*, \phi^*)}$$

where $p(r^*, \theta^*, \phi^*) = p(\boldsymbol{r}_L^*)$, and

$$p(r) = \int_{\cos\theta_0}^{1} d(\cos\theta) \int_0^{2\pi} d\phi p(r, \theta, \phi) \tag{58}$$

is the polar angle-averaged probability density of the ligand position in the conical region. Considering the value of the radial probability density at distance r^* in the canonical region far away from the receptor and integrated over the polar angles,

$$p(r^*) = 2\pi(1 - \cos\theta_0) p(r^*, \theta^*, \phi^*) \tag{59}$$

we finally obtain [27]

$$K_b = 2\pi(1 - \cos\theta_0) C^\circ \int_0^{r_b} dr r^2 e^{-\beta \Delta F(r)} \tag{60}$$

where $r_b = 20$ Å is the limiting radial distance of the binding site region and $\theta_0 = 60°$ is the angle of aperture of the cone, and

$$\Delta F(r) = -k_B T \ln \frac{p(r)}{p(r^*)} \tag{61}$$

is the radial PMF relative to the bulk distance $r^* = 30$ Å.

3.3.2 Calculation of the Potential of Mean Force Using the Adaptive Biasing Force Method

The peptide's radial PMF, $\Delta F(r)$, was evaluated using the Adaptive Biasing Force (ABF) method.[69] ABF serves the dual purpose of accelerating the sampling of the peptide positions relative to the receptor and providing an estimate of the PMF.

[27] Probably because of a typo, the 2π factor is missing in the corresponding expression (equation 2) of the paper by Lapelosa[22].

ABF introduces a fictitious biasing force $f_b(r)$ along the radial direction such that the observed distribution of distances with the addition of the biasing force, $p_{\text{obs}}(r)$, is flat within the sampling region (in this case the region within the cone illustrated in Figure 7 with $\theta_0 = 60^\circ$ angle of aperture and up to $r < r^* = 30$ Å).

A derivation of ABF is beyond the scope of this chapter, however to motivate it, first note that differentiation of Eq. (31) leads to the conclusion that the gradient of the PMF with respect of ζ_L is the average gradient of the system potential energy function

$$\frac{\partial \Delta F(\zeta_L)}{\partial \zeta_L} = \langle \frac{\partial U}{\partial \zeta_L} \rangle_{\zeta_L} \tag{62}$$

where U is the potential energy function of the solvated system and $\langle \ldots \rangle_{\zeta_L}$ represents an ensemble average at fixed ζ_L. In other words, the negative of the gradient of the PMF is the system force averaged over the degrees of freedom of the system other than those along which the PMF is defined, thereby justifying the name *potential of mean force* for $\Delta F(\zeta_L)$. The same conclusion applies to forms of the PMF averaged over some coordinates such as ligand orientations [Eq. (30)], including the 1-dimensional radial PMF, $\Delta F(r)$, considered in the work of Lapelosa.[28]

Also, note that the PMF along a coordinate is proportional to the logarithm of the probability distribution for that coordinate [Eq. (30)]. Thus, a flat distribution indicates that the overall force, the mean force, plus the biasing force along the coordinate is zero or, equivalently, that the added biasing force is equal and opposite to the mean force. This implies that the potential of mean force can be obtained by integrating the biasing force that flattens the radial distribution. The additional benefit of having a flat distribution is that the dynamics along the chosen coordinate are more likely to be diffusive and not impeded by free energy barriers. Indeed,

[28] In this case the radial force is interpreted in terms of the force of a central potential, and Eq. (62) has additional terms due to the Jacobian of the radial coordinate.[69]

several independent binding/unbinding events have been reported in the study by Lapelosa.[22]

4 Conclusion

This chapter has shown how a statistical mechanics formulation of the non-covalent molecular association from first principles gives rise to different computational methods to estimate the binding free energies of protein-peptide complexes. The three case studies illustrate the application of each method to particular molecular complexes and how they are tailored to achieve specific goals. It is much more challenging to apply rigorous binding free energy estimation methods to protein-peptide complexes relative to small-molecule binding. We hope that this chapter illustrates how a good appreciation of the underlying theories and their computational implementations helps understand the practices connected with each approach and its strengths and limitations.

5 Acknowledgements

E.G. acknowledges support from the National Science Foundation (NSF CAREER 1750511).

References

1. Panagiotis L Kastritis and Alexandre MJJ Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, 10(79):20120835, 2013.

2. Denise Kilburg and Emilio Gallicchio. Recent advances in computational models for the study protein-peptide interactions. *Adv. Prot. Chem. Struct. Biol.*, 105:27–57, 2016.
3. Ilda D'Annessa, Francesco Saverio Di Leva, Anna La Teana, Ettore Novellino, Vittorio Limongelli, and Daniele Di Marino. Bioinformatics and biosimulations as toolbox for peptides and peptidomimetics design: where are we? *Frontiers in Molecular Biosciences*, 7, 2020.
4. Mihail Mihailescu and Michael K Gilson. On the theory of noncovalent binding. *Biophys. J.*, 87:23–36, 2004.
5. Corinne LD Gibb and Bruce C Gibb. Binding of cyclic carboxylates to octa-acid deep-cavity cavitand. *J. Comp. Aided Mol. Des.*, pages 1–7, 2013.
6. Eva Judy and Nand Kishore. Discrepancies in thermodynamic information obtained from calorimetry and spectroscopy in ligand binding reactions: Implications on correct analysis in systems of biological importance. *Bulletin of the Chemical Society of Japan*, 2020.
7. Thomas Simonson. The physical basis of ligand binding. *In Silico Drug Discovery and Design*, pages 3–43, 2016.
8. Manuel Tsiang, Gregg S. Jones, Magdeleine Hung, Susmith Mukund, Bin Han, Xiaohong Liu, Kerim Babaoglu, Eric Lansdon, Xiaowu Chen, Jacob Todd, Terrence Cai, Nikos Pagratis, Roman Sakowicz, and Romas Geleziunas. Affinities between the binding partners of the hiv-1 integrase dimer-lens epithelium-derived growth factor (in dimer-ledgf) complex. *Journal of Biological Chemistry*, 284(48):33580–33599, 2009.
9. Anirudh Ranganathan, Philipp Heine, Axel Rudling, Andreas Pluckthun, Lutz Kummer, and Jens Carlsson. Ligand discovery for a peptide-binding gpcr by structure-based screening of fragment-and lead-like chemical libraries. *ACS chemical biology*, 12(3):735–745, 2017.
10. S Boresch, F Tettinger, M Leitgeb, and M Karplus. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B*, 107:9535–9551, 2003.
11. Joseph Marcotrigiano, Anne-Claude Gingras, Nahum Sonenberg, and Stephen K Burley. Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G. *Mol Cell*, 3(6):707–716, 1999.
12. Joanna Wysocka. Identifying novel proteins recognizing histone modifications using peptide pull-down assay. *Methods*, 40(4):339–343, 2006.
13. M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.*, 72:1047–1069, 1997.
14. Emilio Gallicchio and Ronald M Levy. Recent theoretical and computational advances for modeling protein-ligand binding affinities. *Adv. Prot. Chem. Struct. Biol.*, 85:27–80, 2011.

15. Terrell L. Hill. *An Introduction to Statistical Thermodynamics*. Dover, New York, 1986.
16. B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.
17. Rajat K Pal and Emilio Gallicchio. Perturbation potentials to overcome order/disorder transitions in alchemical binding free energy calculations. *J. Chem. Phys.*, 151(12):124116, 2019.
18. Emilio Gallicchio, Mauro Lapelosa, and Ronald M. Levy. Binding energy distribution analysis method (BEDAM) for estimation of protein-ligand binding affinities. *J. Chem. Theory Comput.*, 6:2961–2977, 2010.
19. A. Ben Naim. *Water and Aqueous Solutions*. Plenum, New York, 1974.
20. E. Gallicchio, M. M. Kubo, and R. M. Levy. Entropy-enthalpy compensation in solvation and ligand binding revisited. *J. Am. Chem. Soc.*, 120:4526–27, 1998.
21. Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci.*, 110(16):6358–6363, 2013.
22. Mauro Lapelosa. Free energy of binding and mechanism of interaction for the meevd-tpr2a peptide–protein complex. *J. Chem. Theory Comput.*, 13(9):4514–4523, 2017.
23. Jeffrey Cruz, Lauren Wickstrom, Danzhou Yang, Emilio Gallicchio, and Nanjie Deng. Combining alchemical transformation with a physical pathway to accelerate absolute binding free energy calculations of charged ligands to enclosed binding sites. *J. Chem. Theory Comput.*, 16(4):2803–2813, 2020.
24. P. Cherepanov, G. Maertens, P. Proost, B. Devreese, J. Van Beeumen, Y. Engelborghs, E. De Clercq, and Z. Debyser. Hiv-1 integrase forms stable tetramers and associates with ledgf/p75 protein in human cells. *Journal of Biological Chemistry*, 278(1):372–381, 2003.
25. Thomas S Peat, David I Rhodes, Nick Vandegraaff, Giang Le, Jessica A Smith, Lisa J Clark, Eric D Jones, Jonathan AV Coates, Neeranat Thienthong, Janet Newman, et al. Small molecule inhibitors of the ledgf site of human immunodeficiency virus integrase identified by fragment screening and structure based design. *PloS one*, 7:e40147, 2012.
26. Lee D Fader, Eric Malenfant, Mathieu Parisien, Rebekah Carson, François Bilodeau, Serge Landry, Marc Pesant, Christian Brochu, Sébastien Morin, Catherine Chabot, et al. Discovery of BI 224436, a noncatalytic site integrase inhibitor (NCINI) of HIV-1. *ACS Med. Chem. Lett.*, 5(4):422–427, 2014.
27. Feng-Hua Zhang, Bikash Debnath, Zhong-Liang Xu, Liu-Meng Yang, Li-Rui Song, Yong-Tang Zheng, Nouri Neamati, and Ya-Qiu Long. Discovery of novel 3-hydroxypicolinamides as selective inhibitors of HIV-1 integrase-LEDGF/p75 interaction. *Eur. J. Med. Chem.*, 125:1051–1063, 2017.

28. Peter Cherepanov, Andre LB Ambrosio, Shaila Rahman, Tom Ellenberger, and Alan Engelman. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc. Natl. Acad. Sci.*, 102(48):17308–17313, 2005.
29. David I. Rhodes, Thomas S. Peat, Nick Vandegraaff, Dharshini Jeevarajah, Janet Newman, John Martyn, Jonathan A. V. Coates, Nicholas J. Ede, Philip Rea, and John J. Deadman. Crystal structures of novel allosteric peptide inhibitors of hiv integrase identify new interactions at the ledgf binding site. *ChemBioChem*, 12(15):2311–2315, 2011.
30. Emilio Gallicchio, Nanjie Deng, Peng He, Alexander L. Perryman, Daniel N. Santiago, Stefano Forli, Arthur J. Olson, and Ronald M. Levy. Virtual screening of integrase inhibitors by large scale binding free energy calculations: the SAMPL4 challenge. *J. Comp. Aided Mol. Des.*, 28:475–490, 2014.
31. Denise Kilburg and Emilio Gallicchio. Assessment of a single decoupling alchemical approach for the calculation of the absolute binding free energies of protein-peptide complexes. *Frontiers in Molecular Biosciences*, 5:22, 2018.
32. Mauro Lapelosa, Emilio Gallicchio, and Ronald M. Levy. Conformational transitions and convergence of absolute binding free energy calculations. *J. Chem. Theory Comput.*, 8:47–60, 2012.
33. J. L. Banks, J. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, A. K. Felts, T. A. Halgren, D. T. Mainz, J. R. Maple, R. Murphy, D. M. Philipp, M. P. Repasky, L. Y. Zhang, B. J. Berne, R. A. Friesner, E. Gallicchio, and R. M. Levy. Integrated modeling program, applied chemical theory (IMPACT). *J. Comp. Chem.*, 26:1752–1780, 2005.
34. Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comp. Bio.*, 13(7):e1005659, 2017.
35. Daniele Di Marino, Ilda D'Annessa, Holly Tancredi, Claudia Bagni, and Emilio Gallicchio. A unique binding mode of the eukaryotic translation initiation factor 4E for guiding the design of novel peptide inhibitors. *Prot. Sci.*, 24:1370–1382, 2015.
36. G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105:6474–6487, 2001.
37. E. Gallicchio and R. M. Levy. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.*, 25:479–499, 2004.

38. Emilio Gallicchio, Kristina Paris, and Ronald M. Levy. The AGBNP2 implicit solvation model. *J. Chem. Theory Comput.*, 5:2544–2564, 2009.

39. Baofeng Zhang, Denise Kilburg, Peter Eastman, Vijay S. Pande, and Emilio Gallicchio. Efficient gaussian density formulation of volume and surface areas of macromolecules on graphical processing units. *J. Comp. Chem.*, 38:740–752, 2017.

40. Chipot and Pohorille (Eds.). *Free Energy Calculations. Theory and Applications in Chemistry and Biology*. Springer Series in Chemical Physics. Springer, Berlin Heidelberg, Berlin Heidelberg, 2007.

41. Robert W Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

42. William L. Jorgensen and Laura L. Thomas. Perspective on free-energy perturbation calculations for chemical equilibria. *J. Chem. Theory Comput.*, 4:869–876, 2008.

43. Sheenam Khuttan, Solmaz Azimi, Joe Wu, and Emilio. Gallicchio. Alchemical transformations for concerted hydration free energy estimation with explicit solvation. *J. Chem. Phys.*, page In press, 2021.

44. Zhiqiang Tan, Emilio Gallicchio, Mauro Lapelosa, and Ronald M. Levy. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.*, 136:144102, 2012.

45. Emilio Gallicchio and Ronald M Levy. Advances in all atom sampling methods for modeling protein-ligand binding affinities. *Curr. Opin. Struct. Biol.*, 21:161–166, 2011.

46. Emilio Gallicchio, Junchao Xia, William F Flynn, Baofeng Zhang, Sade Samlalsingh, Ahmet Mentes, and Ronald M Levy. Asynchronous replica exchange software for grid and heterogeneous computing. *Computer Physics Communications*, 196:236–246, 2015.

47. J.D. Chodera and M.R. Shirts. Replica exchange and expanded ensemble simulations as gibbs sampling: Simple improvements for enhanced mixing. *J. Chem. Phys.*, 135:194110, 2011.

48. Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113:6042–6051, 2000.

49. Junchao Xia, William Flynn, Emilio Gallicchio, Keith Uplinger, Jonathan D Armstrong, Stefano Forli, Arthur J Olson, and Ronald M Levy. Massive-scale binding free energy simulations of hiv integrase complexes using asynchronous replica exchange framework implemented on the ibm wcg distributed network. *J. Chem. Inf. Model.*, 59(4):1382–1397, 2019.

50. Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

51. A. K. Felts, Y. Harano, E. Gallicchio, and R. M. Levy. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins: Struct. Funct. Bioinf.*, 56:310–321, 2004.

52. Michael Andrec, Anthony K Felts, Emilio Gallicchio, and Ronald M Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc Natl Acad Sci U S A*, 102:6801–6806, 2005.

53. Steven W. Rick. Increasing the efficiency of free energy calculations using parallel tempering and histogram reweighting. *J. Chem. Theory Comput.*, 2:939–946, 2006.

54. Zhiqiang Tan. On a likelihood approach for monte carlo integration. *J. Am. Stat. Assoc.*, 99:1027–1036, 2004.

55. E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy. Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B*, 109:6722–6731, 2005.

56. Xinqiang Ding, Jonah Z Vilseck, and Charles L Brooks III. Fast solver for large scale multistate bennett acceptance ratio equations. *Journal of chemical theory and computation*, 15(2):799–802, 2019.

57. Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.*, 129:124105, 2008.

58. Timothy J Giese and Darrin M York. Variational method for networkwide analysis of relative ligand binding free energies with loop closure and experimental constraints. *J. Chem. Theory Comput.*, 2021.

59. Vanitha Krishna Subbaiah, Christian Kranjec, Miranda Thomas, and Lawrence Banks. Pdz domains: the building blocks regulating tumorigenesis. *Biochemical Journal*, 439(2):195–205, 2011.

60. Nicolas Panel, Francesco Villa, Ernesto J Fuentes, and Thomas Simonson. Accurate pdz/peptide binding specificity with additive and polarizable free energy simulations. *Biophys. J.*, 114(5):1091–1102, 2018.

61. Anthony J Clark, Tatyana Gindin, Baoshan Zhang, Lingle Wang, Robert Abel, Colleen S Murret, Fang Xu, Amy Bao, Nina J Lu, Tongqing Zhou, Peter D. Kwong, Lawrence Shapiro, Barry Honig, and Richard A. Friesner. Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of hiv-1. *J. Mol. Biol.*, 429(7):930–947, 2017.

62. Anthony J Clark, Christopher Negron, Kevin Hauser, Mengzhen Sun, Lingle Wang, Robert Abel, and Richard A Friesner. Relative binding affinity prediction of charge-changing sequence mutations with fep in protein–protein interfaces. *J. Mol. Biol.*, 431(7):1481–1493, 2019.

63. Xu Liu, Tyson R Shepherd, Ann M Murray, Zhen Xu, and Ernesto J Fuentes. The structure of the tiam1 pdz domain/phospho-syndecan1 complex reveals a ligand conformation that modulates protein dynamics. *Structure*, 21(3):342–354, 2013.

64. Antonia S. J. S. Mey, Bryce K. Allen, Hannah E. Bruce Macdonald, John D. Chodera, David F. Hahn, Maximilian Kuhn, Julien Michel, David L. Mobley, Levi N. Naden, Samarjeet Prasad, Andrea Rizzi, Jenke Scheen, Michael R. Shirts, Gary Tresadern, and Huafeng Xu. Best practices for alchemical free energy calculations [article v1.0]. *Living Journal of Computational Molecular Science*, 2(1):18378, 12 2020.

65. Wei Jiang, Christophe Chipot, and Benoît Roux. Computing relative binding affinity of ligands to receptor: An effective hybrid single-dual-topology free-energy perturbation approach in namd. *J. Chem. Inf. Model.*, 59(9):3794–3802, 2019.

66. James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with namd. *J. Comp. Chem.*, 26(16):1781–1802, 2005.

67. Thomas Steinbrecher, David L Mobley, and David A Case. Nonlinear scaling schemes for lennard-jones interactions in free energy calculations. *J Chem Phys*, 127:214108, 2007.

68. Clemens Scheufler, Achim Brinker, Gleb Bourenkov, Stefano Pegoraro, Luis Moroder, Hans Bartunik, F Ulrich Hartl, and Ismail Moarefi. Structure of tpr domain–peptide complexes: critical elements in the assembly of the hsp70–hsp90 multichaperone machine. *Cell*, 101(2):199–210, 2000.

69. Jeffrey Comer, James C Gumbart, Jérôme Hénin, Tony Lelièvre, Andrew Pohorille, and Christophe Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *The Journal of Physical Chemistry B*, 119(3):1129–1151, 2015.