Analytic Model of the Free Energy of Alchemical Molecular Binding

Denise Kilburg[†,‡] and Emilio Gallicchio[*,†,‡,¶]

†*Department of Chemistry, Brooklyn College of the City University of New York, Brooklyn, NY 11210, USA*

‡*Ph.D. Program in Chemistry, the Graduate Center of the City University of New York, New York, NY 10016, USA.*

¶*Ph.D. Program in Biochemistry, the Graduate Center of the City University of New York, New York, NY 10016, USA*

E-mail: egallicchio@brooklyn.cuny.edu

Abstract

We present a parameterized analytic statistical model of the thermodynamics of alchemical molecular binding within the solvent potential of mean force formalism. The model describes the free energy profiles of linear single-decoupling alchemical binding free energy calculations accurately. The parameters of the model, which are physically motivated, are derived by maximum likelihood inference from data obtained from alchemical molecular simulations. The validity of the model has been assessed on a set of host-guest complexes. The model faithfully reproduces both the binding free energy profiles and the probability densities of the perturbation energy as a function of the alchemical progress parameter. The model offers a rationalization for the characteristic shape of binding free energy profiles. The parameters obtained from the model are potentially useful descriptors of the association equilibrium of molecular complexes.

Potential applications of the model for the classification of molecular complexes and the design of alchemical molecular simulations are envisioned.

Introduction

The primary goal of a quantitative model of molecular binding is to provide an estimate of the standard free energy of binding, ΔG_b°, or, equivalently, of the equilibrium constant, K_b, for the association equilibrium $R + L \leftrightharpoons RL$, between two molecules R and L. For example, the binding of a drug molecule to a receptor. A brute-force molecular simulation approach to the calculation of the binding constant, based on following the motion of the ligand in and out of the receptor, is generally not feasible due to the long times between binding and unbinding events.[1] Biased methods have been developed to accelerate the dynamics of association and obtain the free energy profile of ligand binding along pathways in and out of the receptor.[2–10]

Alchemical descriptions of the binding equilibrium provide an alternative to the study of physical binding/unbinding paths.[11–15] The idea is that, because a free energy change depends only on the end states, one can connect the bound and unbound states of the molecular system by any thermodynamic path, whether physical or unphysical. In alchemical methods, the potential energy function is modified parametrically in a series of steps traced by a progress parameter λ to go from a description of the unbound state to that of the bound state. These methods effectively "grow" the ligand in place within the binding site. The field has a long history,[16–19] but only relatively recently it has converged into a unified statistical thermodynamics theory of biomolecular binding.[12,20–22] The double-decoupling method,[11,20,23] which is used to compute absolute binding free energies, is so called because it involves free energy calculations to decouple the ligand to an intermediate gas phase from the bound and solution states of the ligand. Free energy perturbation methods,[24–27] are suitable for the analysis of relative binding, such as in drug optimization.

We have developed an alchemical single-decoupling methodology, based on an implicit description of the solvent,[28] that enables the transfer of the ligand directly into the binding site rather than through multiple thermodynamic pathways.[29–31] Among other advantages, the single-decoupling approach leads naturally to a statistical representation of the equilibrium in terms of probability distributions of the binding energy. For example, it is possible to relate the binding free energy to the probability distribution, $p_0(u)$, of the binding energy in the absence of receptor-ligand interactions.[12]

Analogously to approaches based on physical binding pathways, alchemical binding free energy calculations yield free energy profiles along the thermodynamic transformations. Alchemical free energy profiles are functions of the alchemical progress parameter λ, rather than, for instance, the ligand-receptor distance. A typical alchemical calculation involves collecting distributions of perturbation energies as a function of the alchemical progress parameter λ. These are merged using thermodynamic reweighting algorithms[32,33] to yield the free energy profile along λ. Typically, only the difference between the endpoints of the free energy profile, which is the binding free energy, is considered. However, the shape of the free energy profile can also yield useful information regarding the physical characteristics of the molecular complex. For example, a quadratic dependence on λ, typical of linear response, is often observed during the alchemical transformation.

In this work, we present a method to relate the shape of the free energy profile to physical observables of the complex. Working within the single-decoupling framework, we develop a statistical analytic model of binding and we construct a procedure to estimate the parameters of the model from data generated by alchemical molecular simulations. The model is based on the statistics of ligand-receptor interaction energies when the ligand uniformly explores the binding site volume as if the receptor atoms were not present. This general strategy has a long history in the treatment of solvation (examples are scaled particle theory, particle insertion, and information/fluctuation theories[34–38]) but it has not been fully explored to study molecular recognition. The main distinction is that a receptor, unlike a homogeneous

solvent, has a specific shape and distribution of interaction sites. We show that the single decoupling theory offers a useful starting point to think about this problem.

Theory and Methods

Statistical mechanics theory of non-covalent molecular association

The standard free energy of binding, ΔG_b°, between a receptor R and a ligand L is given by

$$\beta \Delta G_b^\circ = -\ln K_b, \tag{1}$$

where $\beta = 1/(k_B T)$, T is the absolute temperature, k_B is Boltzmann's constant and K_b is the dimensionless binding constant that, assuming ideal solutions, is expressed as

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

where $[\ldots]$ are equilibrium concentrations and C° is the standard state concentration (conventionally set as 1M or 1 molecule/1668 Å^3).

In a widely employed classical statistical mechanics theory of non-covalent association,[12,20] the binding constant is expressed as

$$K_b = C^\circ V_{\text{site}} \langle e^{-\beta \Delta U} \rangle_0, \tag{3}$$

where $U(x,\zeta) = V(x,\zeta) + W(x,\zeta)$ is the effective potential energy function of the receptor-ligand complex, expressed in terms of the internal degrees of freedom, x, of receptor and ligand, and the external degrees of freedom (i.e. overall translation and rotations),[21] ζ, of the ligand with respect to the receptor. The function $\Delta U(x,\zeta) = U(x,\zeta) - U_0(x)$ is the binding energy of the complex in conformation (x,ζ), where $U_0(x)$ is the effective potential energy of the system when receptor and ligand are at infinite separation. V_{site} is the chosen

volume of the binding site, that is the volume of the region of positions and orientations ζ of the ligand relative to the receptor which are considered to correspond to the bound state of the complex.[1] The average $\langle \ldots \rangle_0$ in Eq. (3) is conducted over the decoupled equilibrium ensemble corresponding to $U_0(x)$, in which receptor and ligand do not interact, while the ligand samples uniformly the binding site volume. Finally, $V(x,\zeta)$ is the potential energy of the system and $W(x,\zeta)$ is the solvent potential of mean force, which represents the solvation free energy of the complex in conformation (x,ζ). The solvent potential of mean force, which is based on the partial averaging over the solvent degrees of freedom,[12,39] is a quantity of general applicability and, in principle, does not introduce any new approximations into the theory proposed here.

Inserting Eq. (3) into Eq. (1) yields

$$\beta\Delta G_b^\circ = -\ln C^\circ V_{\text{site}} + \beta\Delta G_{\text{exc.}}, \tag{4}$$

where $-k_BT \ln C^\circ V_{\text{site}}$ is the concentration-dependent component of the standard free energy of binding independent of the specific form of the potential energy, and

$$\beta\Delta G_{\text{exc.}} = -\ln\langle e^{-\beta\Delta U}\rangle_0 \tag{5}$$

is the excess free energy of the complex.

In the following, we focus on the excess component of the standard free energy of binding. To simplify the notation, we henceforth denote the excess free energy as ΔG, and we measure all energies and free energies in units k_BT thereby omitting factors of β throughout.

[1] Eq. (3) refers to the case in which only overall translations are used to define the binding site volume. In general, a term corresponding to the integration over orientational degrees of freedom is also present.[12,21]

Alchemical binding free energy models

Molecular simulations aimed at computing the excess free energy of binding based on Eqs. (4) and (5) are referred to as "alchemical" in that they sample the unphysical uncoupled state in which receptor and ligand, while being close to each other, behave as if the other were not present. In practice, Eq. (5) converges very slowly because, due to atomic overlaps, in the uncoupled state large and positive values ofΔU (and, consequently, negligibly small values of $\exp(-\Delta U)$) are much more likely to be sampled than favorable ones, causing the average to be dominated by the infrequent occurrences of overlap-free configurations. To overcome this obstacle, it is common to adopt a stratification scheme based on an alchemical hybrid potential $U(x, \zeta; \lambda)$, dependent on an alchemical progress parameter λ, conventionally ranging from 0 and 1. This strategy implies a λ-dependent excess free energy defined as

$$\Delta G(\lambda) = -\ln K(\lambda), \tag{6}$$

where

$$K(\lambda) = \langle e^{-\Delta U(\lambda)} \rangle_0 \ , \tag{7}$$

is the λ-dependent binding constant, and where, using the notation introduced above,

$$\Delta U(\lambda) = U(x, \zeta; \lambda) - U_0(x) \tag{8}$$

is the perturbation energy at λ for the complex in conformation (x, ζ). In the following we will refer to $\Delta G(\lambda)$ as the *alchemical free energy profile* and $K(\lambda)$ as the *binding constant profile*.

The stratification approach above leads to the familiar computational algorithms for the calculation of free energy differences based on the accumulation of the effects of small progressive increments of λ. For instance, Eq. (7) is easily generalized to yield an expression

of the ratio of equilibrium constants at nearby values of λ:

$$\frac{K(\lambda')}{K(\lambda)} = \langle e^{-[\Delta U(\lambda') - \Delta U(\lambda)]} \rangle_\lambda \ , \tag{9}$$

which is the basis of the Free Energy Perturbation (FEP) method. [2] Similarly, inserting Eq. (7) into Eq. (6) and differentiating with respect to λ, leads to the well-known Thermodynamic Integration (TI) formula:[40]

$$\frac{d\Delta G(\lambda)}{d\lambda} = \langle \frac{\partial U(\lambda)}{\partial \lambda} \rangle_\lambda \tag{10}$$

which, when integrated, yields the free energy profile.

Being related to ensemble averages, it is helpful for the current purpose to note that both the FEP and TI formulas can be expressed in terms of probability density functions. For instance, Eq. (7) can be rewritten as

$$K(\lambda) = \int_{-\infty}^{+\infty} d(\Delta U_\lambda) e^{-\Delta U_\lambda} p_0(\Delta U_\lambda) \ , \tag{11}$$

where $p_0(\Delta U_\lambda)$ is the probability density of the perturbation energy, $\Delta U(\lambda)$, at λ in the $\lambda = 0$ ensemble. Analogously, denoting $u(\lambda) = \partial U(\lambda)/\partial\lambda$, Eq. (10) is rewritten as

$$\frac{d\Delta G(\lambda)}{d\lambda} = \int_{-\infty}^{+\infty} du \, u p_\lambda(u) \tag{12}$$

where $p_\lambda(u)$ is the probability density of the $\partial U/\partial\lambda$ function in the ensemble at λ.

[2] It should be noted that, while Eq. (9) is mathematically exact, modern numerical implementations of FEP employ more efficient BAR and MBAR free energy estimators.[32]

Linear alchemical transformations

Eqs. (11) and (12) take a particular convenient form when the alchemical potential energy function $U(x,\zeta;\lambda)$ varies linearly with respect to λ:

$$U(x,\zeta;\lambda) = U_0(x) + \lambda u(x,\zeta) \tag{13}$$

where $U_0(x,\zeta)$ is the potential energy of the decoupled state and $u(x,\zeta)$ is the so-called binding energy function of the complex, which is assumed as independent of λ. By comparing Eqs. (13) and (8), it is straightforward to show that for an alchemical potential of the form (13) the perturbation potential is proportional to the binding energy function

$$\Delta U(x,\zeta;\lambda) = \lambda u(x,\zeta) \tag{14}$$

and that the λ-derivative employed in the TI formula is independent of λ and equal to the binding energy function:

$$\frac{\partial U(x,\zeta;\lambda)}{\partial \lambda} = u(x,\zeta)\,. \tag{15}$$

Inserting Eq. (14) into Eq. (11) we obtain

$$K(\lambda) = \int_{-\infty}^{+\infty} du\, e^{-\lambda u} p_0(u)\ , \tag{16}$$

where $p_0(u)$, which plays a central role in this work, is the probability density of the binding energy function in the uncoupled state, that is in the state in which the ligand is uniformly distributed in the binding site region and receptor and ligand do not interact with each other. Mathematically, Eq. (16) expresses the fact that the binding constant profile $K(\lambda)$ is given by the two-sided Laplace transform of $p_0(u)$. In turn, the binding free energy profile $\Delta G(\lambda)$ is related to $K(\lambda)$ by Eq. (6), and the excess binding free energy is $\Delta G(\lambda = 1)$. Finally, the Potential Distribution Theorem[41] provides a relationship between $p_0(u)$ and the binding

energy distributions at any other value of λ:

$$p_\lambda(u) = e^{\Delta G(\lambda)} e^{-\lambda u} p_0(u). \tag{17}$$

It is therefore apparent that knowledge of $p_0(u)$ determines all of the other quantities that characterize the alchemical transformation, including the binding free energy profile and the binding free energy. In this respect, the function $p_0(u)$ serves the same role in the alchemical theory of binding that the density of states $\Omega(E)$ plays in classical statistical mechanics. For instance, note the parallel between Eq. (17) and the well know Boltzmann's relationship $p_\beta(E) \propto \exp[-\beta E]\Omega(E)$, which gives the energy distribution of a system at any temperature given the density of states.

The main aim of the work presented here is to develop an analytic model for $p_0(u)$ from which to derive all of the other quantities discussed above and, conversely, to estimate the parameters of the model against the results of alchemical molecular simulations.

Statistical model for $p_0(u)$

In this section, we turn to the derivation of a model for the probability distribution, $p_0(u)$, of the binding energy in the uncoupled ensemble at $\lambda = 0$, that is in the state when the ligand and the receptor are not interacting. Note the critical distinction between the state from which samples are collected (the uncoupled ensemble), and the quantity being sampled (the binding energy function): we are interested in the distribution of binding energies, which are in general not zero, when receptor and ligand configurations are sampled in the absence of receptor-ligand interactions. As illustrated in Fig. 1, since in the absence of interactions clashes between ligand and receptor atoms are likely, a long tail at large and positive values of the binding energy characterizes $p_0(u)$. $p_0(u)$ also has a much smaller, but finite, tail at favorable binding energies. The low energy tail of $p_0(u)$ is amplified by the $\exp(-u)$ exponential term, to yield, through Eq. (17), the expected distribution of binding energies

in the bound state narrowly centered around a favorable mean binding energy (see Fig. 1).

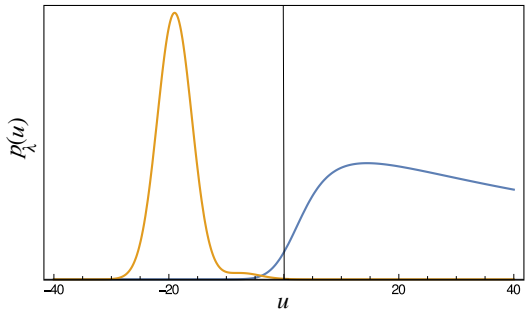

Figure 1: $p_0(u)$ (blue, curve on the right) and $p_1(u)$ (yellow, curve on the left) from Eq. (26) for $\bar{u}_B = -10$, $\sigma_B = 3$, $\epsilon_{LJ} = 1$, $\tilde{u}_c = 10$, $n_l = 2$, $p_b = 10^{-6}$, and $p_m = 0$. The scale of the y-axis is arbitrary and probability densities are not normalized. Energy values are expressed in units of $k_B T$.

To start thinking about a functional form for $p_0(u)$, consider the model illustrated in Fig. 2, which depicts the binding site volume containing receptor atoms (large circles) arranged in some configuration, and a ligand represented by small blue circles. Because at $\lambda = 0$ ligand-receptor interactions are turned off, the ligand atom occupies the binding site with uniform probability. The binding site volume is divided into two main regions. In the region outside any of the receptor atom circles, as location "B" in the white region of Fig. 2, the interaction energy between the ligand atom and the receptor is the result of many, relatively weak electrostatic and dispersion interactions of similar magnitude. This mode of interaction describes the behavior of $p_0(u)$ at favorable values of the binding energy. When, instead, a ligand atom is found within the inner core of a receptor atom (shaded in light and dark gray in Fig. 2), such as at locations "C" and "M," the repulsion energy of that individual interaction dominates all of the others. This interaction mode is expected to be essential to describe the high energy tail of $p_0(u)$. The atomic core of an atom is considered here as its most immediate region where the interaction potential dominates over all other interactions. Because receptor atoms cannot overlap to more than a certain degree, strong repulsive interactions can be understood as the result of a single pair interaction rather than of cooperative contributions of many interactions. (The distinction between the light and dark gray regions within the repulsive interaction core region is explained below).

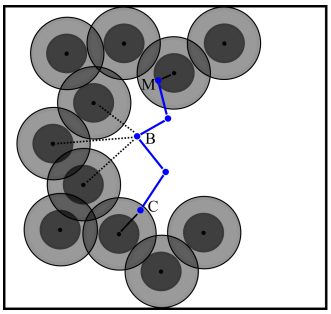

Figure 2: Illustration of the model of the uncoupled state of the receptor-ligand complex. The box represents the receptor site volume. Ligand atoms are blue connected by bonds. Receptor atoms are black surrounded by circles representing the extent of the ligand-receptor repulsive interaction potential. A ligand atom in the white region (such as at location labeled "B") interacts with many receptor atoms by means of soft long-ranged electrostatic and dispersion interactions represented by dashed lines. A ligand atom in the light gray region (such as at location "C") interacts mainly with the closest receptor atom by means of repulsive 12-6 potential (represented by a continuous line). The dark gray region (such as the ligand atom location labeled "M") represents the region where the repulsive interaction energy is constant and capped at the maximum value $u_{\max}$.

To model the two distinct properties of repulsive and attractive interactions, it is useful to think of the ligand-receptor binding energy as the results of two contributions

$$u = u_C + u_B, \tag{18}$$

where u_C represent the *collisional* component, which corresponds to short-ranged repulsive interactions predominant within the atomic cores and well represented by the a single pairwise interaction, and u_B is the *background* component given by the sum of contributions of many weak and favorable long-ranged pairwise interactions.

Motivated by the central limit theorem, we model the probability distribution of the background component by a Gaussian distribution:

$$p_B(u_B) = g(u_B; \bar{u}_B, \sigma_B) = \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left[-\frac{(u_B - \bar{u}_B)^2}{2\sigma_B^2}\right], \tag{19}$$

where $\bar{u}_B$ is the mean and σ_B is the standard deviation of the distribution of the distribution of the total background ligand-receptor interaction energies, u_B, obtained by summing over

all ligand-receptor atom pairs

The collisional energy u_C is assumed to be zero in the region outside the atomic cores. Inside one of the atomic cores, u_C is assumed to be represented by the repulsion energy between the pair of atoms with the most severe clash. Here we represent the repulsive pairwise interactions by the Weeks-Chandler-Andersen (WCA)[42] form of the Lennard-Jones (LJ) potential

$$u_{WCA}(r) = \begin{cases} 4\epsilon_{LJ}\left[\left(\frac{\sigma_{LJ}}{r}\right)^{12} - \left(\frac{\sigma_{LJ}}{r}\right)^{6}\right] + \epsilon, & r < 2^{1/6}\sigma_{LJ} \\ 0 & r > 2^{1/6}\sigma_{LJ}\ , \end{cases} \tag{20}$$

which, as shown in the Appendix, for a single ligand atom leads to the collisional binding energy distribution

$$p_{WCA}(u_C) = \frac{H(u_C - \tilde{u}_C)(1+\tilde{x}_C)^{1/2}}{4\epsilon_{LJ}x(1+x)^{3/2}} \tag{21}$$

where $H(\cdot)$ is Heaviside's step function, $x = \sqrt{u_C/\epsilon_{LJ}}$, $\tilde{x}_C = \sqrt{\tilde{u}_C/\epsilon_{LJ}}$ and $\tilde{u}_C > 0$ is an adjustable energy parameter that defines the level set of the boundary of the core of receptor atoms. The parameter $\tilde{u}_C$ is implicitly defined as the repulsive energy above which the energy of the collision follows the probability density (21).

Eq. (21), derived for a monoatomic ligand, can be generalized to a polyatomic ligand. In doing so, it is critical to note that, even though the total collisional energy can be expressed as the sum of the collisional energies for each ligand atom, the central limit theorem is not applicable because the mean and variance of each contribution, described by probability density (21), are undefined. We can assume however that the collisional energy is dominated by the largest repulsive interaction among all of the ligand atoms: $u_C \simeq \max_{i=1,N}[u_C(i)]$, where $u_C(i)$ is the collisional energy of ligand atom i. The probability density of the maximum, $x_{\max}$, of a set of N independent random variables, x_i, distributed according to the probability density $f(x)$ is given by the expression[43]

$$p(x_{\max}) = N\left[F(x_{\max})\right]^{N-1} f(x_{\max}) \tag{22}$$

where $F(x)$ is the integrated form of $f(x)$, that is the cumulative distribution corresponding to $f(x)$. In general, the positions of the N atoms of the ligand are not statistically independent so Eq. (22) is an approximation. It is expected however that this form, with an effective number of statistically independent number of atoms groups, n_l, is of general applicability. If the ligand is small and rigid it will behave as a single atom. On the other extreme, a large and flexible ligand can be thought of being composed of groups of atoms with nearly uncorrelated position.

Combining Eqs. (35), (40), and (22) finally yields

$$p_{WCA}(u_C) = n_l \left[1 - \frac{(1+x_C)^{1/2}}{(1+x)^{1/2}}\right]^{n_l-1} \frac{H(u_C - \tilde{u}_C)}{4\epsilon_{LJ}} \frac{(1+x_C)^{1/2}}{x(1+x)^{3/2}} \tag{23}$$

for the probability distribution of the collisional energy related to the repulsive WCA potential for a polyatomic ligand. The factor of n_l in front of the expression is the normalization constant and the other symbols have the same meanings as in Eq. (21).

In alchemical molecular simulations, it is customary to adopt soft-core interaction potentials to smoothly cap the maximum value of pair-wise interactions and avoid discontinuities near the uncoupled state.[33,44,45] In this work we model this feature by capping to a maximum value $u_{\max}$ the repulsive WCA potential at short interatomic distances (see Fig. 10). We thus consider the inner core region of the receptor region, denoted by dark gray shading in Fig. 2, where the repulsive interaction energy is constant and equal to $u_{\max}$. In this work, we set $u_{\max} = 1 \times 10^6$ kcal/mol.

Combined collisional and background interaction energy model

We now turn the derivation of the probability distribution $p_0(u) = p(u_B + u_C)$ of the ligand-receptor energy in the uncoupled ensemble. While the background component u_B is assumed

to occur for any configuration of the complex, the probability density of the collisional component is conditional on there being at least one atomic clash defined as $u_C > \tilde{u}_C$. We denote by $p_{\rm b}$ the probability that no such collision occurs in the uncoupled ensemble and when the ligand is within the binding site volume, by p_c the probability that a collision occurs in the region corresponding to the continuous repulsive part of the WCA potential (the light gray region in Fig. 2), and by p_m the probability that the clash occurs within the inner core region (the dark gray region in Fig. 2) where the repulsive energy is $u_{\rm max}$. The probabilities p_b, p_c, and p_m are not all independent parameters of the model since it is required that they sum to 1: $p_b + p_c + p_m = 1$.

Under these assumptions, the probability distribution of the collisional component of the interaction energy is written as

$$p_C(u_C) = p_b\delta(u_C) + p_m\delta(u_C - u_{\rm max}) + p_c p_{WCA}(u_C) \tag{24}$$

where $p_{WCA}(u_C)$ is given by Eq. (23) and the δ-functions express the fact that outside the core region the collisional energy is zero and that inside the inner core region it is equal to $u_{\rm umax}$. Finally, assuming that the background and collisional contributions are statistically independent, the probability density of the total binding energy $u = u_B + u_C$ is given by the convolution of the respective probability densities:

$$p_0(u) = p_0(u_C + u_B) = \int_{-\infty}^{+\infty} p_C(u')p_B(u - u')du'. \tag{25}$$

Substituting in Eq. (25) the definitions given in Eqs. (19) and (24) we obtain:

$$p_0(u) = p_b g(u; \bar{u}_B, \sigma_B) + p_m g(u; \bar{u}_B + u_{\rm max}, \sigma_B) + p_{\rm c} \int_{\tilde{u}_C}^{+\infty} p_{WCA}(u')g(u - u'; \bar{u}_B, \sigma_B)du', \tag{26}$$

where $g(u; \bar{u}, \sigma)$ is the normalized Gaussian distribution of mean $\bar{u}$ and standard deviation σ [see Eq. (19)].

While the integral in Eq. (26) is not available in analytical form, it is amenable to numerical computation by for example Gauss-Hermite quadrature (see Methods). Fig. 1 shows $p_0(u)$ for a particular choice of the parameters $\bar{u}_B$, σ_B, ϵ_{LJ}, $\tilde{u}_c$, and p_b and p_c. Also shown in this figure is $p_1(u) \propto e^{-u} p_0(u)$ [see Eq. (17)]. These distributions indeed reflect the behavior of binding energy distributions obtained from actual molecular simulations (see Results).

Model for the free energy profile

Since the Laplace transform of a convolution of two functions is the product of their Laplace transforms, from Eq. (16) and Eqs. (24) and (19), for the binding constant profile we have

$$K(\lambda) = K_C(\lambda) K_B(\lambda), \tag{27}$$

where

$$K_C(\lambda) = \int_{-\infty}^{+\infty} p_C(u) e^{-\lambda u} du = p_b + p_m e^{-\lambda u_{\max}} + p_c K_{WCA}(\lambda) \tag{28}$$

where $K_{WCA}(\lambda)$ is the two-sided Laplace transform of $p_{WCA}(u)$. From Eq. (23):

$$K_{WCA}(\lambda) = \int_{\tilde{u}_C}^{u_{\max}} p_{WCA}(u) e^{-\lambda u} du \tag{29}$$

Finally, the two-sided Laplace transform of $p_B(u) = g(u; \bar{u}_B, \sigma_B)$ is:

$$K_B(\lambda) = \int_{-\infty}^{+\infty} p_B(u) e^{-\lambda u} = e^{\sigma_B^2 \lambda (\lambda/2 - \bar{u}_B/\sigma_B^2)} \tag{30}$$

An illustrative binding free energy profile, $\Delta G(\lambda) = -\ln K(\lambda)$, obtained from Eqs. (27), (28), (29) and (30) for some choice of parameter values is shown in Fig. 3. Free energy profiles from simulations indeed follow have the shape illustrated in Fig. 3 (see Results). Note that in this model $\Delta G(\lambda)$ is given by the sum of the free energies corresponding to the

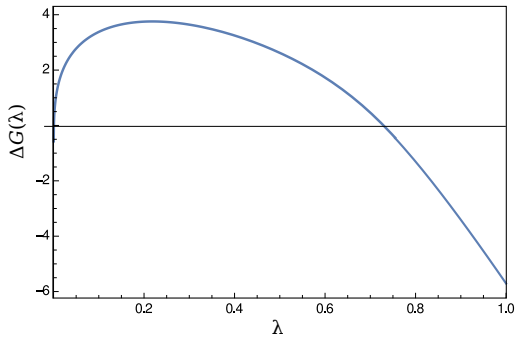

Figure 3: The binding free energy $\Delta G(\lambda) = -\ln K(\lambda)$ from Eqs. (27)–(30) as a function of λ for $\bar{u}_B = -10$, $\sigma_B = 3$ $\epsilon_{LJ} = 1$, $\tilde{u}_c = 10$, $n_l = 2$, $p_b = 10^{-6}$, and $p_m = 0$. Energy is expressed in units of $k_B T$.

collisional and background processes:

$$\Delta G(\lambda) = -\ln K_C(\lambda) - \ln K_B(\lambda) = \Delta G_C(\lambda) + \Delta G_B(\lambda). \tag{31}$$

Mixture model of background component

The analytic model described so far predicts Gaussian-distributed binding energies at $\lambda \simeq 1$, where the collisional contribution is negligible. In practice, however, we encounter systems displaying bimodal binding energy distributions in this regime (see for example Fig. 9). These occurrences are interpreted as the system undergoes a conformational transition from a high-entropy/high-energy state to a low-entropy/low-energy state as λ is increased. We found that these systems can be described well by a mixture model of the background binding energy component described by the weighted sum of two Gaussian distributions:

$$p_B(u_B) = P_a g(u_B; \bar{u}_a, \sigma_a) + P_b g(u_B; \bar{u}_b, \sigma_b)\,, \tag{32}$$

where P_a and P_b ($P_a + P_b = 1$) are the probabilities of occurrence of conformational states a and b at $\lambda = 0$, respectively, and $(\bar{u}_a, \sigma_a)$ and $(\bar{u}_b, \sigma_b)$ are the corresponding average and standard deviation parameters. (In general, any number of conformational states can be considered by introducing the average binding energy and standard deviation parameters for

each.)

To formulate the full model of $p_0(u)$ for this case, Eq. (32) replaces the single Gaussian functions $g(u; \bar{u}_B, \sigma_B)$ in Eq. (26). In the case of the mixture model Eq. (30) becomes

$$K_B(\lambda) = \int_{-\infty}^{+\infty} p_B(u) e^{-\lambda u} = P_a e^{\sigma_a^2 \lambda (\lambda/2 - \bar{u}_a/\sigma_a^2)} + P_b e^{\sigma_b^2 \lambda (\lambda/2 - \bar{u}_b/\sigma_b^2)} \tag{33}$$

Otherwise the remainder of the analytical theory is unchanged. Note that this model can be expanded to an arbitrary number of states and that it reduces to the single-state model [Eq. (19)] when only one state is present (that is $P_a = 1$, for example).

Model parameterization

The analytical model of binding defined by Eq. (26) with Eqs. (23) and (19) depends on seven independent parameters: $\bar{u}_B$, the average background binding energy in the coupled state, σ_B, the standard deviation of the background binding energy in the decoupled state, ϵ_{LJ}, the effective Lennard-Jones ϵ parameter of the repulsive potential within the atomic core, $\tilde{u}_c$, the closest contact dominates the collisional binding energy contribution, n_l, the effective number of statistically independent atom groups of the ligand, p_b, the probability that in the uncoupled state the system is free of atomic clashes, and p_c, the probability of occurrence of one or more atomic clashes described by the repulsive component of the WCA potential. The parameter p_m, the probability of occurrence of an atomic clash of interaction energy $u_{\max}$, is derived from p_b and p_c so that they collectively sum to one.

The mixture model (Section) introduces three additional parameters of the background energy model (the relative occupancy of the two states, and one additional set of average and standard deviation parameters of the background component). In this work, it has been relatively straightforward to identify by manual inspection the cases displaying bimodal binding energy distributions which required the mixture model. Future work will explore unsupervised model selection approaches[46] to automate the search for the most suitable

parameterization for each complex.

The parameters of the selected model are obtained by Maximum Likelihood (ML) inference[47] using as input the binding energy values collected from alchemical molecular simulations at a series of values of λ. ML seeks the parameters that maximize the likelihood function $\mathcal{L}$, or, equivalently, minimize the negative of its logarithm:

$$-\ln \mathcal{L}(\theta) = -\sum_i \ln p_{\lambda_i}(u_i|\theta) = -\sum_i \ln \frac{e^{-\lambda_i u_i} p_0(u_i|\theta)}{K(\lambda_i|\theta)} \tag{34}$$

where the summation runs on the samples of binding energies u_i collected in the ensemble at $\lambda = \lambda_i$, θ represents the set of model parameters above, and we have used Eq. (17). Computational details of the ML parameter estimation procedure are given in the appendix.

Computational details

The host-guest complexes were prepared as described.[48–50] Single-decoupling[29] Hamiltonian Replica-exchange Molecular dynamics simulations[51] employed 22 intermediate λ steps as follows: $\lambda = 0$, 1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.002, 0.004, 0.008, 0.01, 0.02, 0.04, 0.07, 0.1, 0.17, 0.25, 0.35, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. The calculation employed the OPLS-AA force field[52,53] and the AGBNP2 implicit solvent model.[28] We employed a soft-core binding energy function[33] with $u_{\max} = 1 \times 10^6$ kcal/mol. The replica-exchange simulations were started from energy-minimized and thermalized structures from manually docked models. A flat-bottom harmonic restraint with a tolerance of 5 Å between the centers of mass of the host and the guest was applied to define the binding site volume. Each cycle of a replica lasted for 100 picoseconds with 1 fs time-step. The average sampling time for a replica was approximately 10 ns. Calculations were performed on the campus computational grid at Brooklyn College. The binding energies obtained from all replicas were analyzed using UWHAM[33] method and the R-statistical package to compute the binding free energy profile $\Delta G_{\rm b}(\lambda)$.

Results

We tested the analytical model of binding presented above on four host-guest complexes: cyclohexanol, nabumetone, and N-tBOC-L-alanine binding to β-cyclodextrin[48] and trans-4-methylcyclohexanoate binding to the octa-acid cavitand host[50] (Figs. 4, 6 and 8). The results for the complexes with cyclohexanol and nabumetone, are shown in Fig. 5 and Table 1. The results for the complexes with trans-4-methylcyclohexanoate and N-tBOC-L-alanine, which undergo λ-dependent conformational transitions, are presented in Figs. 7 and 9, and Table 2. The analytic model fits very well the binding energy distributions and free energy profiles from the numerical simulations for all of the complexes we studied.

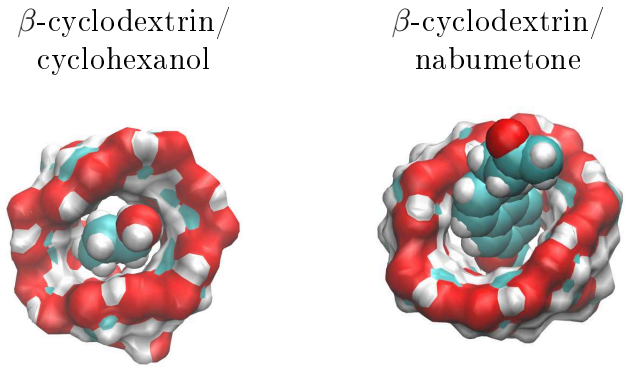

Figure 4: Molecular representations of two of the four host-guest complexes studied in this work. The host is shown in surface representation and the guest is shown using van der Waals atomic spheres.

Table 1: Model parameters for the complexes of cyclohexanol and nabumetone with β-cyclodextrin.

	$\bar{u}_B{}^a$	$\sigma_B{}^a$	p_b	p_c	$\tilde{u}_c{}^a$	$\epsilon_{LJ}{}^a$	n_l
cyclohexanol	1.00	2.95	1.0×10^{-2}	2.0×10^{-1}	0.5	20	2.1
nabumetone	−2.23	2.91	3.9×10^{-4}	6.9×10^{-2}	0.5	20	3.2

^{a} In kcal/mol

In the case of cyclohexanol and nabumetone, for example, the model correctly interpolates the Gaussian behavior of the binding energy distributions at $\lambda \simeq 1$ and the diffuse and asymmetric aspects of the distributions at $\lambda \simeq 0$ (Fig. 5). The binding energy distributions at intermediate λ values present characteristics of both limits and are also correctly described

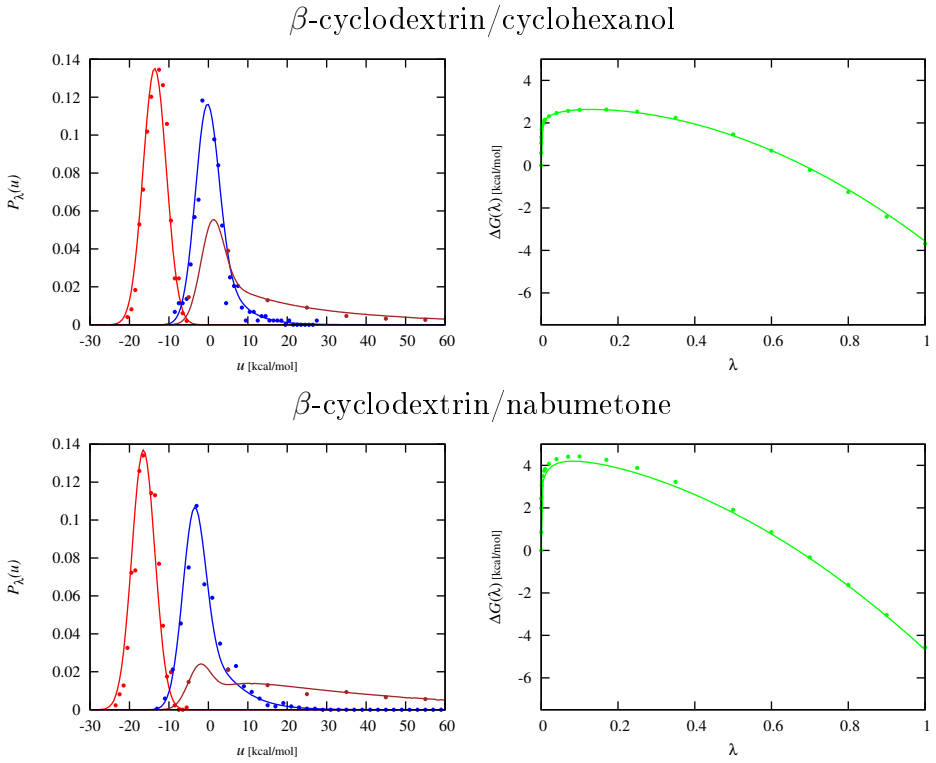

Figure 5: Binding energy probability densities, $p_\lambda(u)$, and binding free energy profiles for the complexes of cyclohexanol and nabumetone with β-cyclodextrin. Binding energy probability densities are shown for (from left to right) for $\lambda = 1$ (red), $\lambda = 0.1$ (blue), and $\lambda = 0.01$ (brown) with corresponding histogram estimates from alchemical molecular calculations (filled circles). Analytical binding free energy profiles (right, green) are compared to UWHAM numerical estimates.

by the model. Free energy profiles (Fig. 5, right panels) are also closely described by the analytic model. For large values of λ ($\lambda > 0.3$, approximately), the free energy profiles vary quadratically with λ, consistent with linear response behavior. The quadratic regime is preceded by a highly non-linear variation of the free energy near $\lambda = 0$. The analytic model correctly captures the singularity of the first derivative of the free energy profile at $\lambda = 0^+$.[54] The maximum of the free energy corresponds to the value of λ at which the average binding energy is zero. In general, as it can be shown from Eqs. (10) and (15), the first derivative of the free energy profile is proportional to the average binding energy. The singularity of the first derivative at $\lambda = 0^+$ is, thus, consistent with the undefined first moment of the $p_0(u)$ probability density. As the data in Fig. 5 illustrates, the analytic model successfully interpolates between the linear response regime at $\lambda \simeq 1$ and the collisional regime at $\lambda \simeq 0$.

The values of the free energy profile at $\lambda = 1$ are the excess binding free energies, which match the numerical estimates (Fig. 5).

The model parameters obtained by fitting the analytic predictions to the numerical results for the complexes with cyclohexanol and nabumetone are listed in Table 1. The stronger binding affinity of nabumetone (-3.9 kcal/mol) relative to cyclohexanol (-3.0 kcal/mol) is driven by stronger interaction energies as reflected by the $\bar{u}_B$ parameter. The average binding energies at the bound state $\lambda = 1$ match closely the linear response predictions from Eq. (47): $\langle u \rangle_1 = -13.7$ and -16.6 kcal/mol, from Eq. (47) and fitted $\bar{u}_B$, σ_B parameters (Table 1), for cyclohexanol and nabumetone, respectively, compared to the direct numerical estimates $\langle u \rangle_1 = -13.2$ and -15.7 kcal/mol, from direct numerical averaging of the binding energies from the $\lambda = 1$ simulation replicas.

The most stable bound states of trans-4-methylcyclohexanoate and N-tBOC-L-alanine have significantly more favorable interaction energies than those of cyclohexanol and nabumetone (-14.0 and -11.06 kcal/mol, respectively, Table 2). However, the trend toward stronger interaction energies is partially offset by the progressively smaller probabilities of fitting the guest into the host without causing atomic clashes, as illustrated by the p_b parameter (Table

2, 5th column). For example, the estimates indicate that it is almost 3 orders of magnitude more difficult to fit N-tBOC-L-alanine into the β-cyclodextrin cavity than cyclohexanol. This feature presumably reflects the larger size and more complex structure of N-tBOC-L-alanine. The variations of p_b could also represent the probabilities of occurrence of binding-competent conformations of the host.

As expected, a common set of values of the $\tilde{u}_c$ and ϵ_{LJ} parameters, corresponding loosely to the magnitude and softness of the core inter-atomic repulsion potential, describes all of the complexes investigated. The magnitude of the fitted ϵ_{LJ} parameter ($\epsilon_{LJ} = 20$ kcal/mol) is significantly larger than typical Lennard-Jones ϵ force field parameters. This confirms the expectation that these parameters should be interpreted to represent the shape and intensity of the repulsive potential exercised by groups of atoms, rather than by individual atoms.

Finally, in Tables 1 and 2 we report the fitted values of the the n_l parameter (8th and 9th columns, respectively) which represents the number of statistically independent number of atom groups of the guests. Indeed, n_l values roughly scale as the size of the guest. For example for nabumetone binding to β-cyclodextrin we find $n_l = 3.2$ compared to $n_l = 2.1$ for cyclohexanol. Despite the smaller size, the n_l value for trans-4-methylcyclohexanoate binding to the octa-acid cavitand is similar to that of nabumetone and N-tBOC-L-alanine, possibly reflecting the fact that this parameter is influenced by the shape of the receptor cavity as well.

Table 2: Model parameters for the complexes which display multiple binding modes: trans-4-methylcyclohexanoate with the octa-acid cavitand and of N-tBOC-L-alanine with β-cyclodextrin.

	${P_{\text{state}}}^b$	${\bar{u}_B}^a$	${\sigma_B}^a$	p_b	p_c	${\tilde{u}_c}^a$	${\epsilon_{LJ}}^a$	n_l
t-4-m-cyclohexanoate								
state a	~ 1.0	-2.0	2.95	2.2×10^{-4}	3×10^{-2}	0.5	20	3.0
state b	3×10^{-4}	-14.0	1.8					
N-tBOC-L-alanine								
state a	~ 1.0	-0.01	2.56	4.68×10^{-5}	8×10^{-2}	0.5	20	3.185
state b	5.5×10^{-8}	-11.06	2.56					

[a] In kcal/mol.[b] Population of the indicated conformational state in the uncoupled ensemble.

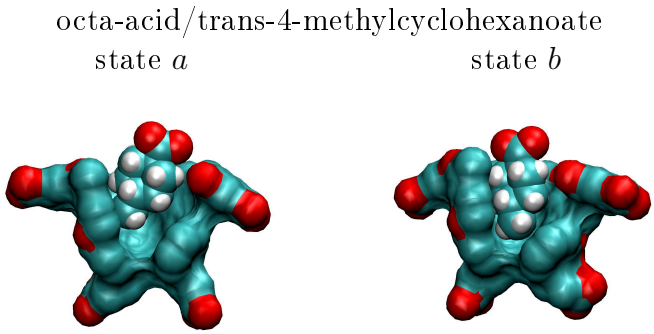

Figure 6: Molecular representations of two conformations of complex between trans-4-methylcyclohexanoate with the octa-acid cavitand representative of the conformational states a and b discussed in the text. State b (right), in which the methyl substituent is inserted deeply in the lower cavity of the host, is characterized by a more favorable binding energy than state a. However, the conformational state a is many times more likely than state b in absence of guest/host interactions. The complex undergoes a transition from state a to state b as λ increases. The cavitand is shown in surface representation with the atoms occluding the view of the guest removed. Trans-4-methylcyclohexanoate is shown in Van der Waals representation.

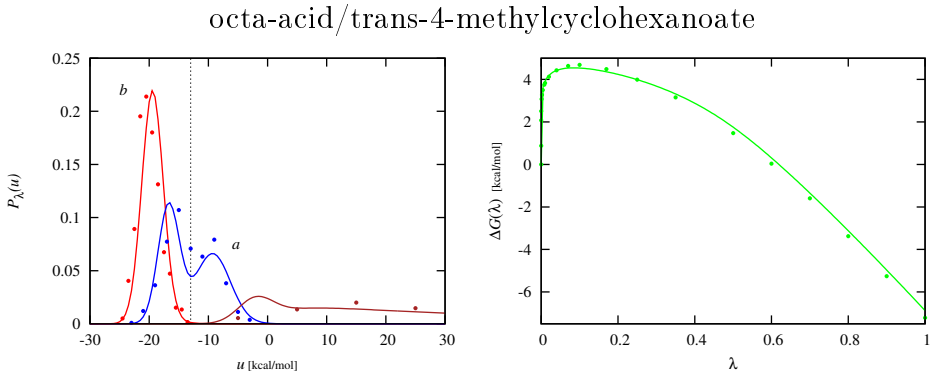

Figure 7: Binding energy probability densities, $p_\lambda(u)$, and binding free energy profiles for the complex of trans-4-methylcyclohexanoate with the octa-acid cavitand. Binding energy probability densities are shown for (from left to right) for $\lambda = 1$ (red), $\lambda = 0.5$ (blue), $\lambda = 0.01$ (brown) with corresponding histogram estimates from alchemical molecular calculations (filled circles). A transition from a high binding energy state a to a low binding energy state b occurs at $\lambda \simeq 0.5$. The vertical dotted line separates the probability density peaks characteristic of the two states. Analytical binding free energy profiles (right, green) are compared to UWHAM numerical estimates.

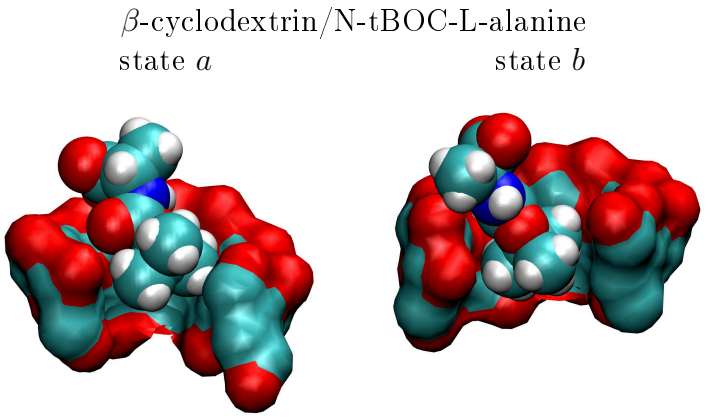

Figure 8: Molecular representations of two conformations of theβ-cyclodextrin/N-tBOC-L-alanine complex representative of the conformational states a and b discussed in the text. State b (right), in which the carboxylate group is oriented toward the solvent and the tert-butyl group is deeper within the host cavity, is characterized by a more favorable binding energy than state a. However, the conformational state a is many times more likely than state b in absence of guest/host interactions. The complex undergoes a transition from state a to state b as λ increases. The β-cyclodextrin host is shown in surface representation with the atoms occluding the view of the guest removed. N-tBOC-L-alanine is shown in Van der Waals representation.

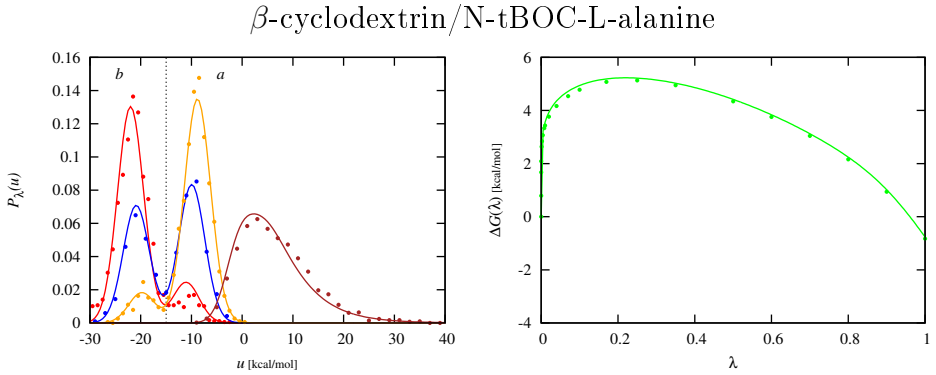

Figure 9: Binding energy probability densities, $p_\lambda(u)$, and binding free energy profiles for the complex of N-tBOC-L-alanine with β-cyclodextrin. Binding energy probability densities are shown for (from left to right) for $\lambda = 1$ (red), $\lambda = 0.9$ (blue), $\lambda = 0.8$ (orange), and $\lambda = 0.1$ (brown) with corresponding histogram estimates from alchemical molecular calculations (filled circles). A transition from a high binding energy state a to a low binding energy state b (see Fig. 8) occurs at $\lambda \simeq 0.9$. The vertical dotted line separates the probability density peaks characteristic of the two states. Analytical binding free energy profiles (right, green) are compared to UWHAM numerical estimates.

The complexes of trans-4-methylcyclohexanoate with the octa-acid cavitand (Fig. 6) and that of N-tBOC-L-alanine with β-cyclodextrin (Fig. 8) undergo λ-induced transitions along the alchemical path from a more probable but more weakly interacting conformational state (state a in Figs. 6 and 8) to a more stable bound state (state b).

The more stable bound pose of trans-4-methylcyclohexanoate corresponds to the state in which the methyl substituent occupies the deep and narrow pocket of the octa-acid cavitand (Fig. 6, state b), as opposed to being loosely bound as in state a. State b is favored by stronger intermolecular interactions (-19 kcal/mol at $\lambda = 1$ compared to -16.6 for state a, from Eq. 46 and the parameters in Table 2) but its probability of occurrence at $\lambda = 0$ is predicted to be 4 orders of magnitude smaller than the loosely bound state. As λ is increased, the influence of the host-guest interactions grows, and state b becomes predominant despite being less likely. The complex with N-tBOC-L-alanine undergoes a similar transition (Fig. 8) from a loosely bound state (state a) to a more stable state (state b in Fig. 8), in which the carboxylate group is rotated toward the solvent, and the body of the aminoacid, including the tert-butyl moiety, is more buried in the host interior.

The optimized parameters (Table 2) indicate that the stable bound state of N-tBOC-L-alanine is extremely unlikely relative to the loosely bound state ($P_b = 5.5 \times 10^{-8}$, 2nd column in Table 2) as compared to trans-4-methylcyclohexanoate ($P_b = 3 \times 10^{-4}$). The small probability of occurrence of the stable bound state causes the binding affinity of N-tBOC-L-alanine to be rather weak ($\Delta G_b = -0.5$ kcal/mol) compared to that of trans-4-methylcyclohexanoate ($\Delta G_b = -6.5$ kcal/mol). Another interesting difference between these two complexes is that, as evidenced by the optimized standard deviation parameters σ_B in Table 2, the stable bound state b of trans-4-methylcyclohexanoate is significantly more energetically restrained than the loosely bound state a. In contrast, the fluctuations of the binding energy, measured by σ_B, is unchanged in going from state a to state b of N-tBOC-L-alanine. This feature is significant because the standard deviation parameter σ_B, which controls the curvature of the binding energy profile, has a strong influence on the binding

free energy. A transition to a state with smaller binding energy fluctuations, such as in the care of trans-4-methylcyclohexanoate, disfavors binding.

The $\lambda-$induced conformational transitions are particular evident in the distributions of binding energy values as a function of λ (Figs. 7 and 9). For the other complexes studied (Fig. 5) the peaks of the binding energy distributions linearly shift toward more negative values as λ is increased. In contrast, the binding energy distributions for trans-4-methylcyclohexanoate and N-tBOC-L-alanine become bimodal starting at some critical λ, and develop by growing the low energy peak (corresponding to the stable bound state b) at the expense of the high energy one. In the case of N-tBOC-L-alanine, for example, the binding energy distribution at $\lambda \simeq 0.8$ is clearly bimodal (Fig. 9) with a predominant high energy mode (corresponding to state a) centered near $u = -9$ kcal/mol and a low energy mode (corresponding to state b) near $u = -22$ kcal/mol. As λ is increased, population shifts to state b, which becomes the predominant state at $\lambda = 1$. At $\lambda = 0.9$ the two states have almost the same population. This behavior is the hallmark of a pseudo first-order phase equilibrium,[55] in which two phases, characterized by compensating differences in average energy and entropy, coexist within the same free energy range. The conformational transition is also apparent in the abrupt change of slope of the binding free energy profile near $\lambda = 0.9$ (Fig. 9). As mentioned, the slope of the binding free energy profile corresponds to the average binding energy as a function of λ. Correspondingly, at $\lambda \simeq 0.9$, the system transitions to a state of lower binding energy thereby causing the change in slope. Note that, while the transition appears slight in the binding free energy profile, the shift in the slope causes a significant decrease (by about 1 kcal/mol) of the binding free energy. The shift in slope of the binding free energy profile and the bimodal character of the binding energy distributions cannot be described without invoking the mixture model.

Discussion

The results obtained as part of this work indicate that it is feasible to represent alchemical binding free energy profiles and binding energy distributions by parameterized analytic functions. The model we proposed offers a rationalization for the shape of the free energy profile and the binding energy distributions. The critical feature of the model is the ability to bridge the two limiting behaviors of the free energy profile, the region near $\lambda \simeq 0$ determined by atomic clashes and the region near $\lambda \simeq 1$ characterized by linear response. The main conceptual advance that enabled this versatility of the model is the description of the binding energy in the uncoupled state of the complex as the sum of two interaction energy components with radically distinct statistical signatures. The first, termed "collisional" interaction energy, describes atomic clashes dominated by nearest neighbor pairs and follows the statistics of the maximum of a set of random variables. The second, that we termed "background" interaction energy, describes the sum of many weak and favorable interatomic interactions and follows the central limit theorem. The two statistical components, assumed statistically independent, are then combined using standard convolution to obtain the distribution of the total binding energy and, through of a Laplace transformation, the binding free energy profile.

The general strategy of describing free energy changes along a thermodynamic path by means of probability models applied to the "decoupled" end point has a long history in the treatment of solvation phenomena in condensed phases. Examples are scaled particle theory, particle insertion models, and information/fluctuation theories.[34–36,38,56] Early work in this area by Pratt & Chandler,[57] introduced the connection between the solubility of hard sphere particles[58] and the probability of formation of suitable cavities in the neat solvent, a prediction that was confirmed by Pangali, Rao, and Berne[59] and subsequent computer simulation work.[60–62] Both Pohorille and Pratt[63] and Hummer et al.,[36] elaborated on the concept of, $p_0(r)$, the probability that a cavity of size r occurs in a neat liquid, which was first introduced in scaled particle theory[56,64] to model the probability of occurrences of cavities

based on the moments of the number of solvent molecules that occupy the solute volume in neat water.

The same essential concepts have been used here to formulate a model connecting the free energy of inserting a ligand molecule into a receptor binding site to probability distributions collected in the decoupled state. The main difference between the solvation process, seen as solute insertion, and binding, seen as ligand insertion, is that, unlike a homogeneous solution, the distribution of receptor atoms is not homogeneous. In particular, there are regions in the receptor binding site where a ligand can fit without requiring conformational reorganization. Conversely, there are interior regions of the receptor from where the ligand is effectively excluded. The model we formulated takes into account these complex geometric and energetic effects in terms of effective physical parameters which are optimized by maximum likelihood inference to reproduce the results of alchemical molecular simulations. The close agreement obtained here between model predictions and molecular simulations of a set of relatively simple but yet chemically-relevant host-guest complexes is evidence that the model is sound and deserving of further investigation and development.

The primary advantage of the theory developed here is that, unlike numerical reweighting methods such as MBAR and UWHAM,[32,33] it yields physical parameters characterizing the thermodynamics of binding of each complex. These parameters can be useful in the classification of molecular complexes. For instance, the $\bar{u}_B$ and σ_B parameters measure the strength of favorable electrostatic and dispersion receptor-ligand interactions as a function of λ [Eq. (47)]. In particular, the σ_B parameter measures the linear response of the complex to the establishment of favorable interactions. A larger σ_B can be an indication, for example, of larger polarizability of the receptor and can be interpreted in terms of local dielectric constant.[65–68] On the other hand, the p_b parameter, which is the probability that ligand and receptor do not overlap while uncoupled, is a measure of the entropic and reorganization costs that oppose the formation of the complex. The model relates these thermodynamic driving forces to interpretable physical parameters such as the size of the binding cavity, if

present, relative to the size of the ligand, or, alternatively, the likelihood of the formation of a suitable binding cavity that can fit the ligand. Similarly, the n_l parameter is interpreted as a measure of ligand size and ligand flexibility.

As discussed, the mixture model parameters indicate the presence of multiple conformational states of the complex and their average interaction energies and relative probabilities. The model parameters attempt to capture the trade-off between energetic gains (the $\bar{u}_B$ parameter) and entropic costs (the P_{state} parameter). In the systems examined we observed transitions from loosely bound states with a high probability of occurrence to strongly bound but entropically disfavored conformational states. For alchemical states near the uncoupled state, the weight of the binding energy component of the alchemical potential energy function [Eq. (13)] is small and the complex tends to visit exclusively the loosely coupled state given its overwhelmingly large probability. However, as the coupled state is approached the strongly bound state becomes competitive with the loosely bound state due to the increase the weight of the interaction energy.

Taken together, the parameters of the model, offer useful insights into the binding equilibrium. When tabulated over a series of systems, they can potentially be employed to characterize and categorize receptor-ligand complexes and, when correlated with binding affinities, can inform receptor and ligand design.

Future work will assess the potential usefulness of the analytic model toward the improvement of alchemical simulation protocols. Because, it builds upon a physically-motivated ansatz dependent on a relatively small number of parameters, the model could be the basis of a free energy estimator with a smaller variance than general-purpose approaches.[32,33] For example, a potential application of the model is as a framework to analyze and measure free energy changes near the decoupled state without the need for extrapolation[23] or soft-core alchemical potentials.[33,45] As analyzed by Simonson[54] and reproduced by our model, the first derivative of the free energy profile has a singularity at $\lambda = 0$. This singularity causes problems for numerical free energy estimators,[32,69] which are usually addressed by the adoption

of non-linear soft-core alchemical potentials.[70,71] These difficulties can also be addressed by replacing the numerical estimation of free energies near the singularity with the estimation of the parameters (which are free of singularities) of the analytic free energy function (31). The analytic model can also be potentially useful to evaluate alchemical thermodynamic lengths to optimize the λ schedule[72,73] of alchemical transformations.

The model, as currently expressed, is limited to single-decoupling linear alchemical transformations.[12] Single-decoupling requires pre-averaging to the solvent degrees of freedom by means of a solvent potential of mean force treatment[39] implemented here using the AGBNP2[28] implicit solvent model. The requirement of linearity of the alchemical transformation with respect to the charging parameter λ is introduced to deploy potential distribution theorem identities[41] relating binding energy distributions at different values of λ. Future work will attempt to extend the model to non-linear coupling schemes and explicit solvation models. Binding free energy calculations with explicit solvation are typically conducted according to the double-decoupling scheme,[20] which is based on the difference of the free energies of coupling the ligand to the hydrated receptor and the free energy of solvation. Hence, it is conceivable that analogous analytic models can be developed for double-decoupling alchemical calculations by considering each free energy leg separately.

Conclusion

We have presented a parameterized analytical model describing the free energy profile of linear single-decoupling alchemical binding free energy calculations. The parameters of the model, which are physically motivated, are obtained by fitting model predictions to numerical simulations. The validity of the model has been assessed on a set of host-guest complexes. The model faithfully reproduces the binding free energy profiles and the probability densities of the perturbation energy as a function of the alchemical progress parameter λ. The model offers a rationalization for the characteristic shape of the free energy profiles. The parameters

obtained from the model are potentially useful descriptors of the association equilibrium of molecular complexes.

Acknowledgements

We acknowledge support from the National Science Foundation (CAREER 1750511) and the Levy-Kosminsky Professorship in Physical Chemistry at Brooklyn College. Molecular simulations were conducted on the WEB computational grid at Brooklyn College of the City University of New York. We thank Ronald Levy, Tai-Sung Lee, and Zhiqiang Tan for guidance and suggestions on an early formulation of the linear response model.

Appendix

Derivation of Eq. (21)

Consider two particles interacting by the pair potential (20) in which one particle (representing the receptor) is fixed at the origin and the other (representing the ligand) is uniformly distributed in a sphere of radius r_C centered at the origin (Fig. 10). Here we assume that $r_c < r_0 = 2^{1/6}\sigma_{LJ}$ (the distance beyond which the Lennard-Jones WCA potential is zero). We will derive the probability density $p_C(u_C)$ of the interaction energy $u_{\mathrm{WCA}}(r)$, where r is the distance between the two particles, by differentiating the cumulative probability function $P_C(u_C)$ defined as the probability that, given that the ligand particle is uniformly distributed in the sphere, the interaction energy $u_{\mathrm{WCA}}(r)$ is greater than the given value u_C. The value of the WCA potential at r_c is denoted by $\tilde{u}_c$; $\tilde{u}_c$ is therefore the smallest allowed interaction energy.

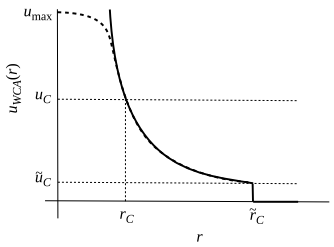

Figure 10: Representation of the repulsive WCA component of the Lennard-Jones potential [Eq. (20)] used in the derivation of Eq. (21). Here $\tilde{r}_C$ represents the radius of the spherical core region around a receptor atom and $\tilde{u}_C$ the corresponding repulsion potential energy. Similarly, r_C is a generic distance between the ligand atom and the receptor atom within the core and u_C is the corresponding potential energy. The dashed curve represents the soft-core interaction potential capped at $u = u_{\max}$.

The probability that the pair interaction energy is smaller than u_C is given by:

$$P_{WCA}(u_C) = H(u_C - \tilde{u}_c)\frac{V_C - V(u_C)}{V_C} \tag{35}$$

where the Heaviside function imposes the requirement that u_C be larger than the minimum

values, V_C is the volume of the sphere of radius r_c and $V(u_C)$ is the volume of the sphere of radius $r(u_C)$, where $r(u_C)$ is inter-particle distance at which the LJ WCA potential has value u_C. From Eq. (20) we have

$$r(u_C) = \frac{r_0}{(1+x_C)^{1/6}}; \quad u_C \geq 0 \tag{36}$$

where $r_0 = 2^{1/6}\sigma_{LJ}$ is the minimum of the Lennard-Jones pair potential and

$$x_C = \sqrt{u_C/\epsilon_{LJ}} \tag{37}$$

Inserting Eq. (36) into Eq. (35) and differentiating with respect to u_C yields Eq. (21), which expresses a normalized distribution as it can be verified by direct integration using the fact that

$$\int \frac{dx}{(1+x)^{3/2}} = \frac{2}{(1+x)^{1/2}} \tag{38}$$

Now consider a receptor composed of M atoms interacting with a monoatomic ligand with the WCA repulsive potential (20). The cumulative probability is given by the expression $P_{WCA}(u_C) = H(u_C - \tilde{u}_c)(1 - V(u_C)/V_C)$, as in Eq. (35), where now $V(u_C)$ is the volume of the region of the receptor where the WCA potential is larger than u_C and, similarly, $V_C \geq V(u_C)$ is the volume where the WCA potential is larger than $\tilde{u}_C$. We can approximate $V(u_C)$ by the van der Waals volume $V[r(u_C)]$ of a molecule with M atoms with van der Waals radii $r(u_C)$, given by Eq. (36) below, corresponding to distance at which the value of WCA repulsive pair potential is equal to u_C. Differentiating the cumulative distribution with respect to u_C, yields:

$$p_{WCA}(u_C) = -\frac{1}{V_C}\frac{dV(r)}{dr}\frac{dr(u_C)}{du_C} = \frac{H(u_C - \tilde{u}_C)}{12\epsilon_{LJ}}\frac{A(r)r}{V_C}\frac{1}{x(1+x)^{3/2}} \tag{39}$$

where $A(r)$ is the van der Waals surface of the receptor when the atomic radii are set to r, and r and x are both functions of u_C [see Eqs. (36) and (37)].

Eq. (39) is interesting because it links the probability density of the collisional interaction energy to the shape of the receptor. There are numerical algorithms (some analytical) to obtain the van der Waals surface area of a molecule.[74] For large u, $r(u)$ is small and atomic overlaps between receptor atoms can be ignored. In this limit $A[r(u)] \simeq M4\pi r(u)^2$, and assuming that $V_C \simeq M4\pi r(\tilde{u}_C)^3/3$, we finally obtain

$$p_{WCA}(u_C) = \frac{H(u_C - \tilde{u}_C)}{4\epsilon_{LJ}} \frac{(1+x_C)^{1/2}}{x(1+x)^{3/2}} \tag{40}$$

which has the same form as the probability density of the collisional energy for one receptor atom.

Maximum Likelihood parameter estimation

The maximum likelihood optimization function in Eq. (34), net of additive terms independent of the parameters, and using Eqs. (27) is

$$-\ln \mathcal{L}(\theta) = -\sum_i \ln p_0(u_i|\theta) + \sum_k N_k \ln K_B(\lambda_k) + \sum_k N_k \ln K_C(\lambda_k) \tag{41}$$

where N_k is the number of binding energy samples collected at $\lambda = \lambda_k$, $p_0(u_i|\theta)$ is given by Eq. (26), $K_B(\lambda_k)$ is given by Eq. (33), and $K_C(\lambda_k)$ by Eqs. (28) and (29). Given an initial set of parameters, the $p_0(u_i|\theta)$ terms are evaluated using Eq. (26). The convolution integral, or, in the case of the mixture model, the sum of two integrals in Eq. (26) is each of the form:

$$\int_{-\infty}^{+\infty} p_{WCA}(u') e^{-(u-u'-\bar{u})^2/2\sigma^2} du' \tag{42}$$

which, upon the variable transformation $y = (u' - u + \bar{u})/\sqrt{2}\sigma$ becomes of the standard form:

$$\sqrt{2}\sigma \int_{-\infty}^{+\infty} f(y) e^{-y^2} dy \tag{43}$$

where $f(y) = p_{\mathrm{WCA}}(\sqrt{2}\sigma y + u - \bar{u})$, which is amenable to evaluation by Gauss-Hermite quadrature:

$$\sqrt{2}\sigma \int_{-\infty}^{+\infty} f(y) e^{-y^2} dy \simeq \sqrt{2}\sigma \sum_n w_n f(y_n) \tag{44}$$

where $w_n$and y_n are the Gauss-Hermite weights and nodes, respectively. In this work we used 15 Gauss-Hermite nodes. The calculation of $K_C(\lambda_k)$ in Eq. (41) requires the numerical evaluation of the integral in Eq. (29) which is accomplished using the variable transformation $u = \exp[y/2] - 1$ and a linear interaction grid of 100 y-values distributed from $y = 0$ to $y = 2 \ln u_{\max}$.

Minimization of the function (41) was implemented in a Python application (`github.com/egallicc/femodel-tf-optimizer`) using TensorFlow[75], which derives function gradients on the fly. Because of the presence of the step function, TensorFlow was unable to compute the first derivative of Eq. (23), even though it is continuous at $u = \tilde{u}_C$. To mimic the behavior of Eq. (23) without a step function we used the following expression

$$p_{WCA}(u_C) \simeq n_l \left[1 - \tanh(z^{12})^{1/12}\right]^{n_l - 1} \frac{s(u_C - \tilde{u}_C)}{4\epsilon_{LJ}} \frac{(1 + x_C)^{1/2}}{x(1 + x)^{3/2}} \tag{45}$$

where $z = (1 + x_C)^{1/2}/(1 + x)^{1/2}$ and $s(u) = [1 + \exp(-20u/\tilde{u}_C)]^{-1}$ is a sigmoid function. The advantage of Eq. (45) is that it allows optimization of $\tilde{u}_C$ in TensorFlow.

We observed that successful optimization progress in TensorFlow was achieved only when starting with good initial guesses of the parameters. When ligand-receptor interactions are established, atomic collisions are unlikely and the binding energy is mainly determined by the background component. Thus, histograms obtained from molecular dynamics trajectories near $\lambda = 1$ are most useful in the estimation of the background binding energy parameters $\bar{u}_B$ and σ_B. An initial first guess for these parameters can be extracted from the average $\langle u_B \rangle_{\lambda=1}$ and standard deviation $\sqrt{\langle \delta u_B^2 \rangle_{\lambda=1}}$ of the binding energies at $\lambda = 1$, observing that, because the background energy is assumed to be Gaussian-distributed, its parameters follow

linear response behavior upon variation of λ:

$$\langle \delta u_B^2 \rangle_\lambda = \langle \delta u_B^2 \rangle_0 = \sigma_B^2 \tag{46}$$

$$\langle u_B \rangle_\lambda = \langle u_B \rangle_0 - \lambda \sigma_B = \bar{u}_B - \lambda \sigma_B^2 \, , \tag{47}$$

which can be easily derived by applying the potential distribution theorem [Eq. (17)] to the Gaussian distribution of u_B at λ: $g[u_B; \bar{u}_B(\lambda), \sigma_B(\lambda)] \propto \exp[-\lambda u] g(u_B; \bar{u}_B, \sigma_B)$.

Conversely, the histograms at small λ values are most useful to estimate the collisional energy parameters ϵ_{LJ}, $\tilde{u}_c$, n_l, p_b, and p_c once a first guess for the values of $\bar{u}_B$ and σ_B is available. We observed (see Results), as it would be expected, a high degree of universality of the parameters ϵ_{LJ} and $\tilde{u}_c$, which describe the extent and softness of the repulsive potential within the atomic cores common to all complexes investigated. We varied the p_b parameter, which regulates the relative magnitude of the two components in Eq. (26), to match the shape of histograms at intermediate values of λ. Finally, we employed the n_l parameter to reproduce the shape of the high energy tail of histograms at small λ values (with larger n_l values describing slower decaying tails). Given the difficulty of binning the unbound high energy portion of binding energies, this last step was performed by also matching at the shape of the free energy profile $\Delta G(\lambda)$ at small λ. Final refinement of the parameters was performed numerically with TensorFlow starting with these initial guesses.

The mixture model (Section) introduces three additional parameters related to the background energy model including the relative occupancy of the two states at $\lambda = 0$, and one additional set of average and standard deviation parameters of the background component. In the case of N-tBOC-L-alanine, for which the binding energy distribution at large values of λ were clearly bimodal, we obtained initial guesses of the parameters by exploiting the linear response behavior of each of the average and standard deviation parameters [Eqs. (46)

and (47)], and those of the state probabilities:

$$P_a(\lambda) = \frac{P_a e^{-(\bar{u}_a^2 - \langle u_a \rangle_\lambda^2)/2\sigma_a^2}}{M(\lambda)} \tag{48}$$

$$P_b(\lambda) = 1 - P_a(\lambda) \tag{49}$$

where

$$M(\lambda) = P_a e^{-(\bar{u}_a^2 - \langle u_a \rangle_\lambda^2)/2\sigma_a^2} + P_b e^{-(\bar{u}_b^2 - \langle u_b \rangle_\lambda^2)/2\sigma_b^2} \; . \tag{50}$$

which can be derived by application of the potential distribution theorem to the Gaussian mixture distribution (32). In the case of trans-4-methylcyclohexanoate binding to the octa-acid cavitand the relative populations of the two components were not clearly resolved. However we were able to distinguish the value of the standard deviation of the two components and obtain their corresponding average binding energies using iterative ML optimization using TensorFlow.

References

(1) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today* **2013**, *18*, 667–673.

(2) Gumbart, J. C.; Roux, B.; Chipot, C. Efficient determination of protein–protein standard binding free energies from first principles. *J. Chem. Theory Comput.* **2013**, *9*, 3789–3798.

(3) Limongelli, V.; Bonomi, M.; Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci.* **2013**, *110*, 6358–6363.

(4) Cavalli, A.; Spitaleri, A.; Saladino, G.; Gervasio, F. L. Investigating Drug–Target Association and Dissociation Mechanisms Using Metadynamics-Based Algorithms. *Accounts of chemical research* **2014**, *48*, 277–285.

(5) Di Leva, F. S.; Novellino, E.; Cavalli, A.; Parrinello, M.; Limongelli, V. Mechanistic insight into ligand binding to G-quadruplex DNA. *Nucl. Acids Res.* **2014**, 5447–5455.

(6) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The adaptive biasing force method: everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* **2014**, *119*, 1129–1151.

(7) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci.* **2015**, *112*, E386–E391.

(8) Sandberg, R. B.; Banchelli, M.; Guardiani, C.; Menichetti, S.; Caminati, G.; Procacci, P. Efficient nonequilibrium method for binding free energy calculations in molecular dynamics simulations. *J. Chem. Theory Comput.* **2015**, *11*, 423–435.

(9) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian accelerated molecular dynam-

ics: Unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.* **2015**, *11*, 3584–3595.

(10) Saglam, A. S.; Chong, L. T. Highly Efficient Computation of the Basal kon using Direct Simulation of Protein-Protein Association with Flexible Molecular Models. *The Journal of Physical Chemistry B* **2015**, *120*, 117–122.

(11) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: Progress and challenges. *Curr. Opin. Struct. Biol.* **2011**, *21*, 150–160.

(12) Gallicchio, E.; Levy, R. M. Recent Theoretical and Computational Advances for Modeling Protein-Ligand Binding Affinities. *Adv. Prot. Chem. Struct. Biol.* **2011**, *85*, 27–80.

(13) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **2012**, *137*, 230901.

(14) Cuendet, M. A.; Tuckerman, M. E. Alchemical free energy differences in flexible molecules from thermodynamic integration or free energy perturbation combined with driven adiabatic dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 3504–3512.

(15) Tony, L.; Mathias, R.; Gabriel, S. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press, 2014.

(16) Tembe, B. L.; McCammon, J. A. Ligand-Receptor Interactions. *Comput. Chem.* **1984**, *8*, 281.

(17) Jorgensen, W. L. Interactions between amides in solution and the thermodynamics of weak binding. *J. Am. Chem. Soc.* **1989**, *111*, 3770–3771.

(18) Payne, V. A.; Matubayasi, N.; Reed Murphy, L.; Levy, R. M. Monte Carlo Study of the Effect of Pressure on Hydrophobic Association. *J. Phys. Chem. B* **1997**, *101*, 2054–2060.

(19) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Entropy-Enthalpy Compensation in Solvation and Ligand Binding Revisited. *J. Am. Chem. Soc.* **1998**, *120*, 4526–27.

(20) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.

(21) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B* **2003**, *107*, 9535–9551.

(22) 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.

(23) Deng, N.-j.; Zhang, P.; Cieplak, P.; Lai, L. Elucidating the energetics of entropically driven protein–ligand association: calculations of absolute binding free energy and entropy. *J. Phys. Chem. B* **2011**, *115*, 11902–11910.

(24) Lybrand, T. P.; McCammon, J. A.; Wipff, G. Theoretical calculation of relative binding affinity in host-guest systems. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 833–835.

(25) Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-Ligand Complexes: Computation of the Relative Free Energy of Different Scaffolds and Binding Modes. *Journal of Chemical Theory and Computation* **2007**, *3*, 1645–1655.

(26) de Ruiter, A.; Oostenbrink, C. Free energy calculations of protein–ligand interactions. *Curr. Op. Chem. Biol.* **2011**, *15*, 547–552.

(27) Wang, L. et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(28) Gallicchio, E.; Paris, K.; Levy, R. M. The AGBNP2 Implicit Solvation Model. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.

(29) Gallicchio, E.; Lapelosa, M.; Levy, R. M. Binding Energy Distribution Analysis Method (BEDAM) for Estimation of Protein-Ligand Binding Affinities. *J. Chem. Theory Comput.* **2010**, *6*, 2961–2977.

(30) Gallicchio, E.; Levy, R. M. Prediction of SAMPL3 Host-Guest Affinities with the Binding Energy Distribution Analysis Method (BEDAM). *J. Comp. Aided Mol. Design.* **2012**, *25*, 505–516.

(31) Gallicchio, E. Role of Ligand Reorganization and Conformational Restraints on the Binding Free Energies of DAPY Non-Nucleoside Inhibitors to HIV Reverse Transcriptase. *Mol. Biosc.* **2012**, *2*, 7–22.

(32) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(33) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.* **2012**, *136*, 144102.

(34) Widom, B. Potential-distribution theory and the statistical mechanics of fluids. *J. Phys. Chem.* **1982**, *86*, 869–872.

(35) Pratt, L. R.; Pohorille, A. Theory of hydrophobicity: transient cavities in molecular liquids. *Proc Natl Acad Sci U S A* **1992**, *89*, 2995–2999.

(36) Hummer, G.; Garde, S.; García, A. E.; Pohorille, A.; Pratt, L. R. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 8951–8955.

(37) Simonson, T. Dielectric Relaxation in Proteins: Microscopic and macroscopic Models. *International Jrnl. of Quantum Chemistry* **1999**, *73*, 45–57.

(38) Huang, D. M.; Chandler, D. The Hydrophobic Effect and the Influence of Solute-Solvent attractions. *J. Phys. Chem. B* **2002**, *106*, 2047–2053.

(39) Roux, B.; Simonson, T. Implicit Solvent Models. *Biophys. Chem.* **1999**, *78*, 1–20.

(40) Straatsma, T.; Berendsen, H. Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J. Chem. Phys.* **1988**, *89*, 5876–5886.

(41) Beck, T. L.; Paulaitis, M. E.; Pratt, L. R. *The Potential Distribution Theorem and Models of Molecular Solutions*; Cambridge University Press, New York, 2006.

(42) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. *J. Chem. Phys.* **1971**, *54*, 5237–47.

(43) Gumbel, E. J. *Statistics of Extremes*; Dover Publications: New York, 2012.

(44) Steinbrecher, T.; Joung, I.; Case, D. A. Soft-core potentials in thermodynamic integration: Comparing one- and two-step transformations. *J. Comput. Chem.* **2011**, *32*, 3253–3263.

(45) Buelens, F. P.; Grubmüller, H. Linear-scaling soft-core scheme for alchemical free energy calculations. *J. Comput. Chem.* **2012**, *33*, 25–33.

(46) Burnham, K. P.; Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach*; Springer Science & Business Media, 2003.

(47) Lee, T.-S.; Radak, B. K.; Pabis, A.; York, D. M. A new maximum likelihood approach for free energy profile construction from molecular simulations. *J. Chem. Theory Comput.* **2012**, *9*, 153–164.

(48) Wickstrom, L.; He, P.; Gallicchio, E.; Levy, R. M. Large Scale Affinity Calculations of Cyclodextrin Host-Guest Complexes: Understanding the Role of Reorganization in the Molecular Recognition Process. *J. Chem. Theory Comput.* **2013**, *9*, 3136–3150.

(49) Gallicchio, E.; Chen, H.; Chen, H.; Fitzgerald, M.; Gao, Y.; He, P.; Kalyanikar, M.; Kao, C.; Lu, B.; Niu, Y.; Pethe, M.; Zhu, J.; Levy, R. M. BEDAM Binding Free Energy Predictions for the SAMPL4 Octa-Acid Host Challenge. *J. Comp. Aided Mol. Des.* **2015**, *29*, 315–325.

(50) Pal, R. K.; Haider, K.; Kaur, D.; Flynn, W.; Xia, J.; Levy, R. M.; Taran, T.; Wickstrom, L.; Kurtzman, T.; Gallicchio, E. A Combined Treatment of Hydration and Dynamical Effects for the Modeling of Host-Guest Binding Thermodynamics: The SAMPL5 Blinded Challenge. *J. Comp. Aided Mol. Des.* **2016**, *31*, 29–44.

(51) Gallicchio, E.; Xia, J.; Flynn, W. F.; Zhang, B.; Samlalsingh, S.; Mentes, A.; Levy, R. M. Asynchronous replica exchange software for grid and heterogeneous computing. *Computer Physics Communications* **2015**, *196*, 236–246.

(52) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Developement and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(53) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and reparameterization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(54) Simonson, T. Free energy of particle insertion: an exact analysis of the origin singularity for simple liquids. *Molecular Physics* **1993**, *80*, 441–447.

(55) Kim, J.; Straub, J. E. Generalized simulated tempering for exploring strong phase transitions. *J. Chem. Phys.* **2010**, *133*, 154101.

(56) Reiss, H.; Frisch, H.; Lebowitz, J. Statistical mechanics of rigid spheres. *J. Chem. Phys.* **1959**, *31*, 369–380.

(57) Pratt, L. R.; Chandler, D. Theory of the Hydrophobic Effect. *J. Chem. Phys.* **1977**, *67*, 3683–3704.

(58) Stamatopoulou, A.; Ben-Amotz, D. Cavity formation free energies for rigid chains in hard sphere fluids. *J. Chem. Phys.* **1998**, *108*, 7294–7300.

(59) Pangali, C.; Rao, M.; Berne, B. J. Hydrophobic Hydration around a pair of Apolar Species in Water. *J. Chem. Phys.* **1979**, *71*, 2975–2981.

(60) Wallqvist, A.; Berne, B. J. Molecular Dynamics Study of the Dependence of Water Solvation Free Energy on Solute Curvature and Surface Area. *J. Phys. Chem.* **1994**, *99*, 2885–2892.

(61) Berne, B. J. Inferring the hydrophobic interaction from the properties of neat water. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 8800–8803.

(62) Gallicchio, E.; Kubo, M. M.; Levy, R. M. Enthalpy-Entropy and Cavity Decomposition of Alkane Hydration Free Energies: Numerical Results and Implications for Theories of Hydrophobic Solvation. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.

(63) Pohorille, A.; Pratt, L. R. Cavities in molecular liquids and the theory of hydrophobic solubilities. *J. Am. Chem. Soc.* **1990**, *112*, 5066–5074.

(64) Pierotti, R. A. A Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chemical Reviews* **1976**, *76*, 717–26.

(65) Archontis, G.; Simonson, T. Dielectric relaxation in an enzyme active site: molecular dynamics simulations interpreted with a macroscopic continuum model. *J. Am. Chem. Soc.* **2001**, *123*, 11047–11056.

(66) Simonson, T. Gaussian fluctuations and linear response in an electron transfer protein. *Proc. Natl. Acad. Sci.* **2002**, *99*, 6544–6549.

(67) Simonson, T. Dielectric relaxation in proteins: the computational perspective. *Photosynthesis research* **2008**, *97*, 21–32.

(68) Nymeyer, H.; Zhou, H.-X. A method to determine dielectric constants in nonhomogeneous systems: application to biological membranes. *Biophys. J.* **2008**, *94*, 1185–1193.

(69) Shirts, M. R.; Pande, V. S. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J Chem Phys* **2005**, *122*, 144107.

(70) Pohorille, A.; Jarzynski, C.; Chipot, C. Good practices in free-energy calculations. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

(71) Shirts, M. R.; Mobley, D. L. An introduction to best practices in free energy calculations. *Biomolecular Simulations: Methods and Protocols* **2013**, 271–311.

(72) Schön, J. A thermodynamic distance criterion of optimality for the calculation of free energy changes from computer simulations. *J. Chem. Phys.* **1996**, *105*, 10072–10083.

(73) Shenfeld, D. K.; Xu, H.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Minimizing thermodynamic length to select intermediate states for free-energy calculations and replica-exchange simulations. *Phys. Rev. E* **2009**, *80*, 046705.

(74) Zhang, B.; Kilburg, D.; Eastman, P.; Pande, V. S.; Gallicchio, E. Efficient Gaussian Density Formulation of Volume and Surface Areas of Macromolecules on Graphical Processing Units. *J. Comp. Chem.* **2017**, *38*, 740–752.

(75) Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; `https://www.tensorflow.org/`, Software available from tensorflow.org.

Table of Contents Abstract & Graphics

The binding equilibrium between a ligand molecule and a receptor is simulated alchemically by "turning on" the ligand within the binding site. Alchemical calculations are nowadays widely employed in the computational design of drugs, catalysts and advanced materials. Most often these studies entail the numerical analysis of molecular dynamics simulations. In this work we develop a fully analytic statistical model of the thermodynamics of alchemical binding. The parameters of the model, which have intuitive physical interpretation, are obtained by maximum likelihood inference from data extracted from alchemical calculations. Applications of the model for the classification of molecular complexes and the design of alchemical molecular simulations are envisioned.