Building intuition for binding free energy calculations: Bound state definition, restraints, and symmetry

Elise Duboué-Dijon, Jérôme Hénin

Building intuition for binding free energy calculations: bound state definition, restraints, and symmetry.

E. Duboué-Dijon[1,2,a] and J. Hénin[1,2,b]

[1)] *CNRS, Université de Paris, UPR 9080, Laboratoire de Biochimie Théorique, 13 rue Pierre et Marie Curie, 75005, Paris, France*

[2)] *Institut de Biologie Physico-Chimique – Fondation Edmond de Rothschild, PSL Research University, Paris, France*

(Dated: 13 May 2021)

The theory behind computation of absolute binding free energies using explicit-solvent molecular simulations is well-established, yet somewhat complex, with counter-intuitive aspects. This leads to frequent frustration, common misconceptions, and sometimes, erroneous numerical treatment. To improve this, we present the main practically relevant segments of the theory with constant reference to physical intuition. We pinpoint the role of the implicit or explicit definition of the bound state (or the binding site), to make a robust link between an experimental measurement and a computational result. We clarify the role of symmetry, and discuss cases where symmetry number corrections have been misinterpreted. In particular, we argue that symmetry corrections as classically presented are a source of confusion, and could be advantageously replaced by restraint free energy contributions. We establish that contrary to a common intuition, partial or missing sampling of some modes of symmetric bound states does not affect the calculated decoupling free energies. Finally, we review these questions and pitfalls in the context of a few common practical situations: binding to a symmetric receptor (equivalent binding sites), binding of a symmetric ligand (equivalent poses), and formation of a symmetric complex, in the case of homodimerization.

I. INTRODUCTION

Binding free energy calculations aim to quantify the binding between two interacting chemical species. Here we focus on explicit, formally exact methods, although many high-throughput approximate methods exist.[1,2] In a biological context, absolute binding free energy calculations are increasingly used in a number of applications,[3,4] for instance the prediction of ligand-enzyme affinities in the search of effective inhibitors,[5,6] or the study of the stability of protein-protein or protein-nucleic acid complexes.[7–10] Such calculations are typically performed to predict the affinity of a ligand to a biomolecule—for instance in the drug discovery community[11,12] to replace long or expensive experiments. They can also be compared with the available experimental data, for instance, to test the quality of a simulation, which can then be used to gain additional atomic-level insight into the binding process or to decompose it into different components. Hence, it is crucial to make sure that the calculated binding free energies or binding constants are comparable to experimentally determined values, which requires special care in the theoretical definitions as well as in the computational protocol.

The computation of absolute binding free energies has been the central topic of a number of works beginning in the 1980s, with seminal theoretical contributions[13–21] that laid the statistical physical foundations of the methods, as well as the practical applications to biological systems.[13,14,22,23] Thanks to the increase in computational power, the democratization of simulation tools[24] and availability of pedagogical tutorials, *e.g.* as listed in Ref. 25, binding free energy calculations are no longer reserved to a small community of experts, but are applied in a wide variety of contexts, outside the groups that specialize in developing the techniques and software.

However, a few key issues are not as consensual as would be expected in a mature field, and the related literature is still scattered, partly contradictory or confusing, and some notions are hardly accessible to non-specialists. This is in particular the case of the use of symmetry corrections and the effects of partial sampling, which are, as we will see, intimately linked to the use of binding restraints and to the question of the binding site definition. It is difficult for a non-specialist to acquire a good physical intuition of these complex matters, which, in our experience, can lead to uncritical application of ill-understood formulae, sometimes erroneously, and is hampering the wider adoption of absolute binding free energy calculations.

Our goal here is to present a clear version of the theory, with constant reference to physical intuition, enabling practitioners to understand in depth each step of a calculation and properly handle possibly tricky steps such as standard state, restraint, and symmetry corrections. Only a rigorous treatment of these points makes it possible to obtain well-defined standard free energies and report meaningful comparisons with experimental data. We present typical cases to illustrate common problems encountered in practice, and discuss how to properly treat them, hoping that these will serve as more general guidelines applicable to a broad range of cases.

In this work, we limit our discussion and practical examples to classical all-atom simulations with explicit solvent. For implicit solvent approaches, one may refer for instance to Ref. 26,27. Two families of methods can be used to quantify binding from simulations: spatial and energy-based. In the first case, simulations are used to compute the statistical distribution of spatial configurations of the receptor and

[a)] Electronic mail: duboue-dijon@ibpc.fr

[b)] Electronic mail: henin@ibpc.fr

ligand. This may be done in unbiased simulations, or, more commonly, using enhanced sampling techniques and estimating a potential of mean force (PMF)—or more generally a free energy surface. This case is treated in detail in Section II A. In the second case, the binding free energy is expressed as function of interaction energies between the receptor, ligand, and solvent. Statistics on these interaction energies are collected in "alchemical" simulations where the potential energy is modified so that the simulations sample often unphysical, but mathematically well-defined states. The most common form of this approach is the double decoupling method, described in Section II B.

In a first part, we present a pedagogical review of the two main classes of binding free energy calculation methods, focusing in more details on the double decoupling method that we will use in our examples. A second theoretical part is devoted to how to properly account for symmetry (both of ligand or binding sites) and partial sampling. Finally, we analyze three typical test cases that are chosen to cover the range of situations involving a symmetric ligand, receptor, or complex.

II. BINDING SITE DEFINITION AND FREE ENERGY ESTIMATORS: A PEDAGOGICAL REVIEW

We consider a binding equilibrium involving three chemical species: a "receptor" R, a "ligand" L, and a complex RL. The receptor could be a macromolecule with a binding pocket that completely surrounds a much smaller ligand, or the structures and roles of both species could be similar: the receptor and ligand can even be the very same species in the case of homodimerization (section IV C, where we show, however, that the macroscopic binding constant is different from that for heteroassociation). For simplicity, we use the conventional names receptor and ligand to cover all cases, but the theory does not depend on each of these having any specific property. Note, however, that differences between them (of size, flexibility, etc.) do have a practical impact on the convergence of numerical quantities from simulations. We will refer loosely to those species as molecules, even though they might be non-molecular species such as monoatomic ions or noble gases.

The binding equilibrium writes $R + L \rightleftarrows RL$. Under constant temperature and pressure conditions, the Gibbs free energy is stationary at equilibrium:

$$\Delta G_{\text{bind}} = \mu_{RL} - (\mu_R + \mu_L) = 0 \tag{1}$$

From now on, we will assume an **ideal solution**, which is relevant to many chemical and biological applications. For a more general treatment including the non-ideal case, see Ref. 28. If the chemical potentials of the solutes exhibit the ideal concentration dependence, the equilibrium condition 1 writes:

$$\Delta G^\circ_{\text{bind}} = -RT \ln\left(\frac{[RL]C^\circ}{[R][L]}\right) \tag{2}$$

$$K^\circ_{\text{bind}} = e^{-\Delta G^\circ_{\text{bind}}/RT} = \frac{[RL]C^\circ}{[R][L]} \tag{3}$$

where C° is the standard concentration, commonly taken to be 1 mol/L. Equation 3 is the law of mass action in terms of volume concentrations. The goal of affinity calculations is to estimate $\Delta G^\circ_{\text{bind}}$, or equivalently, K°_{bind}. Note that the related quantity K_{bind} is often defined without standard state normalization,[29–31] which then appears in the alternate binding free energy relationship $\Delta G^\circ_{\text{bind}} = -k_B T \ln(K_{\text{bind}} C^\circ)$.

In principle, a molecular dynamics (MD) trajectory of a solution containing R and L molecules, run long enough to show many binding and unbinding events, provides enough information to compute the affinity. The probability of association in the microscopic simulation system can be related to the macroscopic binding equilibrium.[32] In practice, the time scales required to sample the binding equilibrium are often out of reach, especially in the case of strong binding. For that reason, direct simulation is rarely used for quantitative affinity estimation. Furthermore, the theoretical treatment necessary to rigorously connect the microscopic and macroscopic statistics is non-trivial,[32] and becomes more complex still in the presence of long-range interactions between the solutes.[33] In the end, this apparently intuitive approach raises considerations that can be counter-intuitive.

To overcome the sampling limitation of the direct approach, binding free energies are thus usually calculated using biased simulations. Two main approaches are most frequently adopted in the literature, namely the Potential of Mean Force (Section II A) and the alchemical double decoupling (Section II B), whose theoretical foundations we will now recall.

A. $\Delta G^\circ_{\text{bind}}$ from the Potential of Mean Force

We consider a case where the association process is described by a one dimensional potential of mean force (PMF) $w(r)$ along a single coordinate r—typically the distance between the ligand and receptor molecules. The PMF $w(r)$ is related to the radial distribution function $g(r)$ through $w(r) = -RT \ln[g(r)]$. This does not assume spherical symmetry of the site or the complex, but rather incorporates all information about the statistics of binding into $w(r)$. We also suppose that the binding site can be delineated by a given range $[r_{\min}, r_{\max}]$ of r. In this case, the binding constant can be obtained by integration of the radial distribution function over the binding site (*i.e.* ensemble of "bound" configurations), with the well-known relationship:

$$K^\circ_{\text{bind}} = \frac{1}{V^\circ}\int_{\text{site}} 4\pi r^2 g(r)\mathrm{d}r = \frac{1}{V^\circ}\int_{\text{site}} e^{-\beta w(r)} 4\pi r^2 \mathrm{d}r \tag{4}$$

This widely valid equation has been derived many times via thermodynamic arguments[15,17,29–31,34]. It is closely related to the dimer-counting approach mentioned above, as explained in Appendix A.

Application of this PMF approach to complex cases, with the use of restraints and the associated corrections, has been amply discussed from a practical and numerical perspective elsewhere.[35–37]

This well-established relation immediately raises three important points. First, as is now usually known, the binding free

energy is not simply the well depth of the PMF.[29] In the limit of non-interacting species (ideal gas), $w(r) = 0$ and K°_{bind} simply measures the volume of the binding site with respect to the standard volume: $K^\circ_{\text{bind}} = V_{\text{site}}/V^\circ$. Hence, and quite counter-intuitively, $K^\circ_{\text{bind}} \neq 1$ and $\Delta G^\circ_{\text{bind}} \neq 0$. Second, the standard state has to be explicitly accounted for in binding free energy calculations by dividing the integral by the standard volume. Finally, K°_{bind} explicitly depends on a "site" definition, which we comment on in detail in section II C.

B. $\Delta G^\circ_{\text{bind}}$ from alchemical double decoupling

1. *Thermodynamic cycle.*

The PMF method to calculate binding free energies, though arguably the most intuitive, becomes difficult to handle and converge for complex geometries. Thus, many applications favor the so-called "double decoupling" alchemical approach, which is formally equivalent to the PMF calculation, provided all terms are properly estimated and a consistent definition of the binding site is used.[35,38]

The Double Decoupling Method builds upon the fact that, since the Gibbs free energy is a state function, $\Delta G^\circ_{\text{bind}}$ can be calculated using any thermodynamic path between the two states "RL" and "R+L". In particular, the double decoupling method[15,21,30,35] makes use of two alchemical transformations, decoupling the ligand from its environment respectively in the bulk and in the binding site. For these respective transformations, free energy differences ΔG^*_{bulk} and ΔG^*_{site} are estimated using numerical estimators such as the exponential formula, Bennett's Acceptance Ratio (BAR)[39] or its multi-state variant MBAR,[40] the unbinned Weighted Histogram Analysis Method (UWHAM),[41] or Thermodynamic Integration (TI).[42] To improve convergence of these quantities, the transformations go through a number of intermediate states with sufficient overlap.[30,43] If the ligand is charged, the overall charge of the finite simulation box changes during the alchemical transformations, which introduces artifactual free energy contributions that must be accounted for.[44–46]

ΔG^*_{bulk} and ΔG^*_{site} are excess free energies of decoupling, and $\Delta\Delta G^*$ is the *excess free energy of binding*, in the sense that in a system of non-interacting particles, their values would be zero. As a result, the ideal-gas, "cratic" contribution to the binding free energy due to translational entropy[47] can only be accounted for by other terms. Therefore, the standard binding free energy is not simply obtained as the difference of those two terms—*i.e.* $\Delta G^\circ_{\text{bind}} \neq \Delta\Delta G^*$ as was sometimes done in the literature[22,48,49]—because one has to take into account the standard state reference,[47] as proposed in 1986 by Jan Hermans and Shankar Subramaniam.[13]

The complete thermodynamic cycle employed is represented in Fig. 1. Note that modified versions of this cycle can be found in the literature, for instance to facilitate handling ligands with multiple poses.[50] ΔG^*_{site} is the free energy of decoupling the ligand in the "bound" state. In current practice,[24,28,37] the corresponding alchemical transformation is performed using a restraining potential V_{rest} (red triangle in Fig. 1) that ensures that the ligand stays in the active site during the whole transformation. In addition to the two decoupling steps already discussed, one needs to evaluate the free energies associated with restraining the decoupled ligand, $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$—which is often done in the same step as taking into account the standard state reference—and removing the restraints on the coupled ligand in the bound state $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$. The desired standard binding free energy $\Delta G^\circ_{\text{bind}}$ is then obtained by combining all the different steps:[51]

$$\Delta G^\circ_{\text{bind}} = \Delta G^*_{\text{bulk}} + \Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}} - \Delta G^*_{\text{site}} + \Delta G^{\text{rest}\to\text{site}}_{\text{coupled}} \quad (5)$$

Note that in theory, ΔG values include a $P\Delta V$ term, but that term very nearly cancels out when computing the difference $\Delta\Delta G^* = \Delta G^*_{\text{bulk}} - \Delta G^*_{\text{site}}$. As this term does not intervene in any of the specific issues that we discuss, we will omit it henceforth.

2. *Use of restraints and corrections.*

Restraints not only accelerate convergence of the free energy calculation because they limit the amount of phase space to explore in the (partly) decoupled state(s),[13,15,18,24,34,35,37,52,53] but are also, as previously noted,[34] necessary from a theoretical perspective, even in the bound state, contrary to what is sometimes stated in the literature.[49,54] Specifically, restraints are necessary to enforce sampling of (some approximation of) the bound state, because unbinding events, even rare, are incompatible with the estimation of ΔG^*_{site}, which is the free energy of decoupling the ligand **from the binding site** and thus requires the ligand in its coupled state to explore only "bound" configurations. Without restraints, the ligand would eventually leave the active site even in the fully coupled state. This is especially apparent in the case of weak binding,[51] when it happens spontaneously in a relatively short simulation time, but would also be the case even for strong binding in the limit of infinitely long sampling. Historically, it has been overlooked because of the kinetic stability of the bound state on the timescale of relatively short simulation times.

To discuss the practical evaluation and magnitude of the two terms, $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$ and $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$, it is useful to present the commonly used types of restraints. While it is tempting to maintain the ligand in the binding site at all stages of the alchemical transformation with a simple restraint on the distance between the receptor and ligand, the free rotation of the ligand allowed by such simple restraints may make sampling of all different conformations complicated, especially in the intermediate stages of the transformation. Explicit restraints on each translational and rotational degree of freedom have thus been suggested,[21,23] possibly in addition to the internal ligand RMSD[20], or using an external ligand RMSD as sole restraint coordinate.[28] Independently of the chosen degrees of freedom, the restraining potential can be of different forms, typically either harmonic or flat-bottom harmonic, *i.e.* acting only if the chosen degree of freedom is outside a given specified range. Harmonic restraints would restrict the ligand position to remain very close to the reference bound state, while

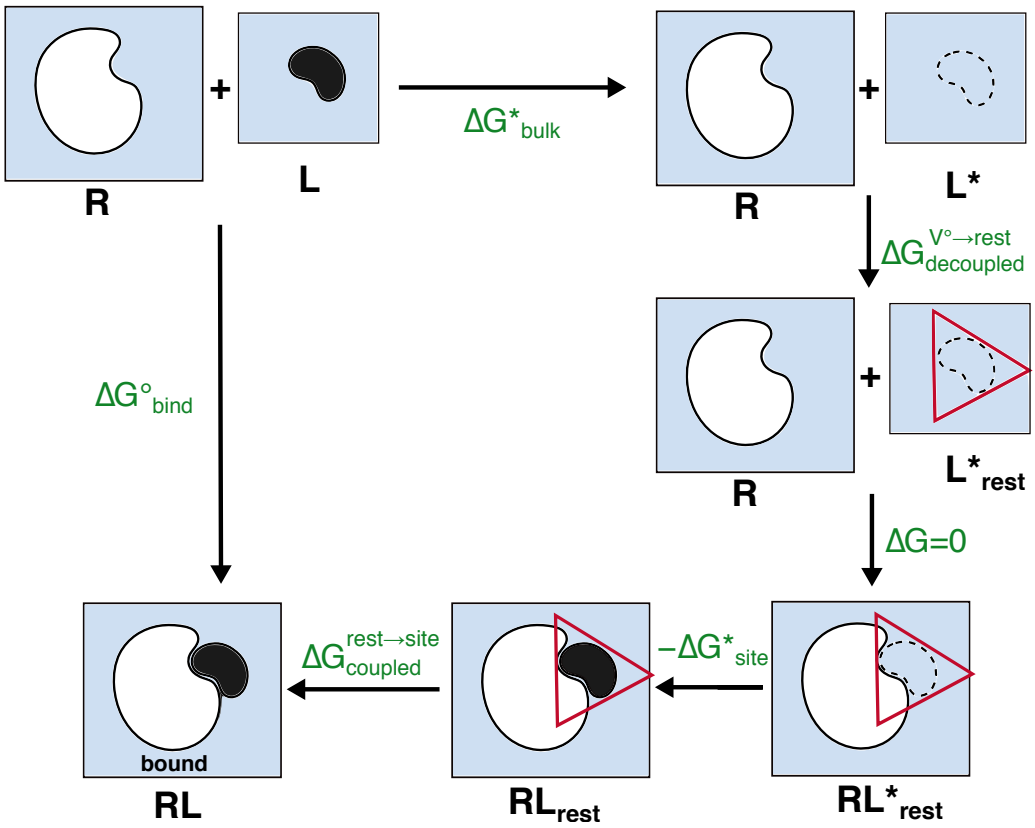

FIG. 1. Thermodynamic cycle for double decoupling with ligand restraints. R: unbound receptor. L: ligand coupled to its environment. L*: decoupled ligand. RL: receptor-ligand complex. The transformation $R + L^*_{\text{rest}} \rightarrow RL^*_{\text{rest}}$ can be seen as a mere translation of the non-interacting ligand, with zero contribution to the free energy.

flat-bottom potentials allow for free movement of the ligand in the "bound" region.

The free energy associated with adding restraints on the ligand in the decoupled state (*i.e.* gas phase), initially freely evolving in the standard volume, $\Delta G^{\text{V}^\circ\rightarrow\text{rest}}_{\text{decoupled}}$, does not admit a simple expression in general. With arbitrary complex restraints, one can estimate this term in a separate calculation, for instance gradually switching off the restraints and computing the associated free energy with standard techniques.[23,28,35,51] For specific forms of restraints however, $\Delta G^{\text{V}^\circ\rightarrow\text{rest}}_{\text{decoupled}}$ can be obtained more straightforwardly, for instance with an analytical expression when the six body rigid rotations and translations are restrained using harmonic potentials as suggested by reference 18. Another simple case is when a restraining potential is applied to the distance between the ligand and receptor. $\Delta G^{\text{V}^\circ\rightarrow\text{rest}}_{\text{decoupled}}$ can then be numerically estimated by integration of the restraint potential U_{rest}:[51,55,56]

$$\Delta G^{\text{V}^\circ\rightarrow\text{rest}}_{\text{decoupled}} = -RT \ln\left(\frac{Q}{V^\circ}\right) \text{, where} \tag{6}$$

$$Q = \int_0^\infty 4\pi r^2 e^{-\beta U_{\text{rest}}(r)} \text{dr} \,. \tag{7}$$

Finally, one has to estimate the free energy associated with releasing the restraints on the coupled ligand in the binding site $\Delta G^{\text{rest}\rightarrow\text{site}}_{\text{coupled}}$. It is important here to stress that this is **not** the free energy to remove all restraints on the coupled system and let the ligand freely evolve **in the whole box**. Instead, it corresponds to the free energy difference between the ligand freely evolving **in the binding site** and the ligand evolving under the restraints chosen for the alchemical transformation, which are designed to keep it, more or less tightly, in the binding site. Evaluating this term obviously presupposes prior definition of the so-called bound state, *i.e.* the ensemble of configurations that are considered as bound. We will come back to this point in Section II C. Evaluation of this term has often been overlooked because restraints are usually chosen to weakly perturb ligand fluctuations in the bound state. In the typical case of strong ligand-protein binding, weak restraints give a negligible contribution in the bound state. Flat-bottom restraints can be specifically designed so that the unbiased region covers the bound state ensemble, and the restraints therefore have a negligible free energy contribution.[28] One can even use them directly as the definition of the bound state,[34] so that by definition $\Delta G^{\text{rest}\rightarrow\text{site}}_{\text{coupled}} = 0$. In contrast, harmonic restraints may, depending on their strength, be associated with much more significant corrections. Note that in this framework one can choose as strong restraining potentials as desired for convergence of the alchemical transformation, provided their impact on sampling of the bound state is properly estimated and corrected for.[23,35,57]

In general, if the bound state is defined by an indicator function $I_{\text{site}}(\mathbf{x})$, where $\mathbf{x}$ denotes all the coordinates of the system ($I_{\text{site}} = 1$ in the bound state and $I_{\text{site}} = 0$ ouside), then $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ can be estimated as:

$$\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}} = -RT \ln \left\langle I_{\text{site}}(\mathbf{x}) e^{+\beta U_{\text{rest}}(\mathbf{x})} \right\rangle_{\text{rest}} \tag{8}$$

where $\langle \dots \rangle$ denotes the ensemble average over configurations in presence of the restraining potential. Eq. 8 can be simply numerically estimated on the fully coupled simulation window if the binding site is well sampled in presence of the restraints (as was for instance done by us and others,[55,56] where a simplified version of Eq. 8 was reported, which in those cases is numerically equivalent to Eq. 8). If this is not the case (for instance when using strong harmonic restraints) then Eq. 8 would not numerically converge and $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ has to be calculated through a separate free energy calculation. In particular, if the restraint acts on a single coordinate z and the binding site can also be simply defined by a criterion on that same coordinate, then $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ can be calculated from the free energy profile along z:[23]

$$\mathrm{e}^{-\beta \Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}} = \frac{\int e^{-\beta A(z)} I_{\text{site}}(z) dz}{\int e^{-\beta A(z)} e^{-\beta U_{\text{rest}}(z)} dz} \tag{9}$$

C. $\Delta G^{\circ}_{\text{bind}}$ should be dependent on, but not sensitive to, the site definition.

From the PMF definition of the binding constant (Eq. 4), it is clear that K°_{bind}, and equivalently the binding free energy, formally depends on the binding site definition. The integral in Eq. 4 does not converge at large distances, so a range of integration needs to be defined. As discussed above, in the double decoupling framework, a definition of the binding site is also needed to estimate $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$. This has historically raised many questions[15,58] and is still often a matter of confusion for non-specialists. The origin of the confusion is that it is tempting to argue that since experiments do not need to define the binding site, calculations should not either.[58] This reasoning, however, has repeatedly been demonstrated to be flawed.[15,17,19,34] All theoretical derivations of a binding free energy need to define the configurations that are considered as forming a complex, the bound state, often through a binding site indicator function I, taken as 1 for bound configurations and 0 otherwise.[15,34] Alternative expressions using unrestricted integration[58] do not yield the desired binding free energy but rather relate to the second virial coefficient, probing *non specific interactions* as well as binding.[19] How to define the bound state is then a question that naturally arises. As discussed in several works,[15,17,19] experimental determinations of binding constants implicitly define the bound state as the configurations producing the signal used for detection. Different techniques being sensitive to different types of complexes, the choice of a bound state definition in calculations should in principle depend on the experiment one wants to compare with. For instance, spectroscopic techniques typically detect only closely associated ligand-protein complexes, while calorimetry gives a signal coming from all conformations giving rise to even small heat transfer (even non site-specific association). A definition of the bound state from first principles has been proposed for the case where the observable is a phase transition.[59] Explicit calculation of the experimental signal often being out-of-reach, simulations usually assume a two-state model with "bound" and "unbound" configurations, which is a simplified description of the possibly complex association process. In practice, for strong binding processes, the outcome of the simulations is insensitive to the exact binding site definition, *i.e.* the integration limits in Eq. 4, so long as all the low free energy regions are included in the bound state. This is illustrated by Fig 1 in Ref. 15. A reasonable choice is to put the boundary at free energy maximum so that the calculated binding constant are relatively insensitive to the its exact location.

III. ACCOUNTING FOR SYMMETRY

A. Symmetry number corrections in binding free energies: are they necessary?

Perhaps the most influential expression for a binding free energy is that proposed by the landmark paper of Gilson et al.[15] In addition to the decoupling free energies and pressure-volume term, their main expression for the absolute binding free energy includes a correction

$$\Delta G_{\text{sym}} = RT \ln \left(\frac{\sigma_{RL}}{\sigma_R \sigma_L} \right) , \tag{10}$$

where σ_x denotes the symmetry number of species x. This suggests that the binding free energy should include an explicit term to account for changes in symmetry between the separated components and the complex.

The symmetry term arises from the expressions of the chemical potential of each species based on its classical partition function.[60] When defining a classical partition function, all atoms are labeled as if they were distinguishable. If the potential energy function treats atoms of the same element identically, the resulting atomic partition function must be divided by the number of permutations of atoms of the same element in the system to compensate for the overcounting.

This is not the case in force-field based classical simulations, where molecules are defined by non-dissociative bonds (usually harmonic) between arbitrarily labeled atoms. Most atom permutations break the bonded structure of the molecule and lead to high energies, so that the resulting configuration does not contribute to the computed partition function. However, if a molecule is symmetric, some permutations of atoms—"symmetry permutations", for instance, swapping the two oxygen atoms of a carboxylate group—preserve the molecular structure and the energy. All symmetry-related configurations thus contribute equally to the partition function. Therefore, defining non-dissociative bonds leads to double or multiple counting *only for symmetry-related configurations*. Dividing the partition function by the molecular symmetry number corrects for this overcounting. To be precise,

the relevant molecular symmetry number is the number of permutations of atoms within the molecule that leave the potential energy unchanged, as argued by Gilson and coworkers themselves.[61,62] It can also be seen as the symmetry number of the molecular graph, defined by *non-dissociative bonds* in the force field.

If non-covalent binding is described only by dissociative, "non-bonded" energy terms, as typically done in force-field simulations of biomolecules, the overcounting of bound (RL) and unbound ($R+L$) configurations is precisely the same (the potential energy function has the same symmetry in the bound and unbound states), so that $\sigma_{RL} = \sigma_R \sigma_L$. In this case, the symmetry correction of Eq. 10 becomes:

$$\Delta G_{\text{sym}} = RT \ln \left(\frac{\sigma_{RL}}{\sigma_R \sigma_L} \right) = 0 \,. \tag{11}$$

This runs counter to an intuitive understanding of the symmetry numbers and in particular of the meaning of σ_{RL}. Crucially, it is *not* the symmetry number that a chemist would ascribe to the complex; for example, one would ascribe the monodentate acetate-cation complex (Section IV A 2, Fig. 3) a symmetry number of $\sigma_{RL} = 1$—as opposed to $\sigma_R = 2$ for free acetate—because intuitively, the bound cation breaks the symmetry of the carboxylate group. However, the potential energy remains symmetric with respect to permutation of the two oxygen atoms, regardless of the position of the cation. The correct symmetry number of the complex is thus not 1 but 2, as that of the unbound receptor. (Rigorously speaking, that number should be multiplied by $3! = 6$ to account for permutations of the methyl hydrogen atoms, but that symmetry is of no practical consequence here, and is routinely ignored by practitioners without question.)

The use of often ill-interpreted and counter-intuitive symmetry numbers is a common source of mistakes, including in our own work.[63] The treatment that we propose for double decoupling (Section II B 2) eliminates the need for such a posteriori symmetry corrections. By introducing separately a definition of the bound state and a set of binding restraints, all "symmetry-related" free energy contributions are naturally accounted for by the two restraint-related free energy terms $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ and $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$, in the spirit of early work by Hermans and Wang.[64] This seems to us more intuitive, and less error-prone, than relying on abstract symmetry considerations.[65]

B. Partial sampling: "if you've seen one, you've seen 'em all"

In alchemical perturbation simulations, whether or not there are restraints that restrict the symmetry of the bound complex, the ensemble of configurations that are effectively sampled in finite time can be smaller than the true ensemble, and in particular, some modes of a symmetric distribution can be poorly sampled, or not sampled at all. In this case, it has been argued that unsampled modes lead to missing terms in the calculation of partition functions, which biases the calculated free energies and requires a symmetry correction.[21] We argue here that this statement is incorrect in the case of an alchemical perturbation.

The argument would be valid only if the partition functions for the relevant states were estimated *separately* by integration over configuration space based on independent simulations. Then, any region missed in sampling in the complex would be effectively unaccounted for. However, in alchemical free-energy calculations, this is not the case: partition functions for individual end states are never estimated as *integrals*, but rather, their ratios are computed as ensemble *averages*. Recall that the free energy of alchemical perturbation from state A to state B can be written[66]:

$$e^{-\beta \Delta F_{AB}} = \left\langle e^{-\beta \Delta U_{AB}} \right\rangle_A \tag{12}$$

Numerically this ensemble average is estimated by an average over configurations generated in state A. Suppose that state A exhibits two symmetric modes Γ and Δ, from which n_Γ and n_Δ samples, respectively, are collected. The exponential estimator $\Delta \tilde{F}_{AB}$ then writes:

$$e^{-\beta \Delta \tilde{F}_{AB}} = \frac{1}{n_\Gamma + n_\Delta} \left(\sum_{\mathbf{x}_i \in \Gamma} e^{-\beta \Delta U_{AB}(\mathbf{x}_i)} + \sum_{\mathbf{x}_i \in \Delta} e^{-\beta \Delta U_{AB}(\mathbf{x}_i)} \right) \tag{13}$$

$$= \frac{n_\Gamma}{n_\Gamma + n_\Delta} \left\langle e^{-\beta \Delta U_{AB}} \right\rangle_\Gamma + \frac{n_\Delta}{n_\Gamma + n_\Delta} \left\langle e^{-\beta \Delta U_{AB}} \right\rangle_\Delta \tag{14}$$

where the brackets correspond to empirical averages of samples taken from the respective modes. If the modes are symmetric, then configurations from Γ and Δ yield the same distribution of ΔU_{AB} values, and the computed average does not depend on the number of samples drawn from each mode. Thus, the number of transitions (if any) observed between symmetric modes during the course of the simulation does not impact the convergence nor the accuracy of the free energy estimate.

A significant source of confusion about this question may come from the process of stratification, whereby discrete intermediate states between A and B are simulated to compute the total free energy difference. This suggests that both end points of the transformation are not sampled identically, and *e.g.* only one of the modes is kinetically accessible in the coupled state, but the whole space is readily sampled in the decoupled state. This has been argued to require individual symmetry corrections for each "pair of states" along the transformation.[21] In a stratified setting, the states A and B above would correspond to intermediate states, not the final end-states. Then, as shown above, the free energy difference *for each window* is unbiased by unbalanced sampling of symmetric modes. The fact that this happens to various degrees over different windows is immaterial. This remains true when combining data from all states at once to estimate the free energy using MBAR[40] or UWHAM,[41] as these estimators only depend on the statistical distribution of comparison energies between the states, which is itself insensitive to which of the symmetric modes the samples were collected from.

Shortly put, the first step in all free energy estimators is to take a set of configuration samples $\mathbf{x}$ and map it to a set of comparison energy samples $\Delta U_{AB}(\mathbf{x})$ (or for TI, energy derivatives $\partial U(\lambda,\mathbf{x})/\partial\lambda$). At this step, symmetry-related configurations map to the same energy, and their distribution among symmetric modes becomes irrelevant. This is illustrated numerically below in the case of acetate-cation binding (IV A 2), where we show that partial sampling of symmetric modes does not affect free energy estimates.

In contrast, Equation 14 gives insight on when partial sampling would result in a biased free energy estimate: this happens if and only if *the energy perturbation ΔU_{AB} is asymmetric*. That is the case e.g. for restraint free energy perturbation if the restraint potential breaks the symmetry (states A and B have different symmetry). Then, sampling all modes with their correct statistical weight is critical to obtain the correct free energy contribution.

C. The case of symmetry-breaking restraints

A corollary of the result above, perhaps even more counter-intuitive, is that any restraint that has no other effect than restricting sampling to one of the symmetric modes *does not affect the decoupling free energy*. Suppose that the same binding process is studied with two numerical approaches, either a symmetry-compatible restraint or a symmetry-breaking restraint. Despite the difference in restraints, the two approaches give the same values for both decoupling free energies ΔG^*_{bulk} (of course) and ΔG^*_{site} (because of the arguments above). One may then wonder if the change in restraint contributions to the free energy causes the two approaches to disagree with each other. In fact, differences appear in both restraint free energies $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$ and $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$, and these two contributions cancel out in Equation 5, yielding the same binding free energy for both schemes. This is illustrated by a fully analytical treatment of a toy case in Appendix B.

This is not a purely theoretical consideration, as different applications make one or the other approach more convenient in practice. For example, complexes with spatially distinct binding sites (IV A 1) are more naturally described with a restraint surrounding a single site. Conversely, complexes with fast-exchanging binding modes such as the different orientations of a benzene ligand might be easier to simulate with a simple translational restraint encompassing all binding modes. Therefore it is useful to have a consistent framework to describe these two cases. The thermodynamic cycle and notations of section II B 2 are well-suited for that.

IV. TYPICAL PRACTICAL PROBLEM CASES

Below, we describe three typical situations that cover most of the binding cases where symmetry issues are encountered. They include two ways in which binding can break symmetry (either a symmetric receptor or a symmetric ligand, the symmetry of which is lost in the complex), and a case where binding increases symmetry, when two identical molecules associate to form a symmetric homodimer.

A. Case 1: symmetric receptor – Equivalent sites

1. *Equivalent binding sites on a protein complex*

A common case of equivalent symmetric binding sites is encountered with protein multimers, when symmetric binding sites are found on the supramolecular complex. This is, for instance, the case of phenol binding to the insulin hexamer,[67] which is relevant for pharmaceutical preparation of insulin, and has been recently studied by simulations.[68]

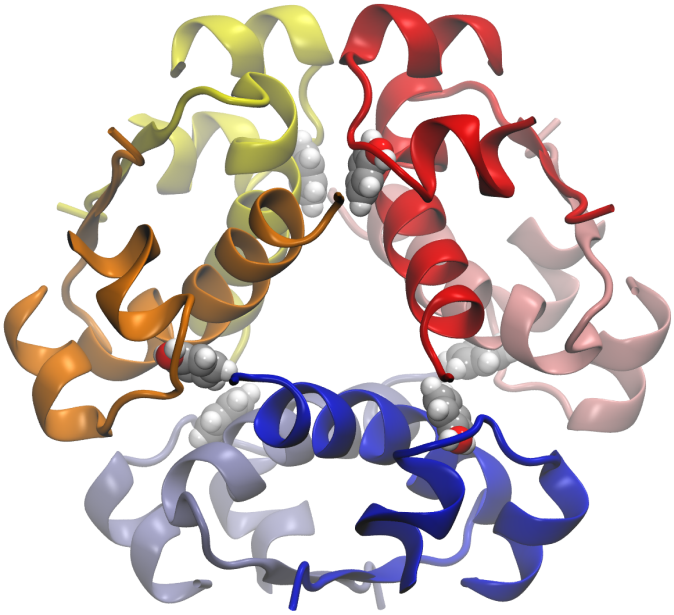

FIG. 2. Structure of the insulin hexamer bound to six phenol molecules, rendered with VMD.[69]

We will first consider the first binding free energy, that is, the binding of one ligand to the protein multimer.

Typically, in such a case, one would model the complex with a single ligand bound, and perform alchemical decoupling using, for instance, a flat-bottom or harmonic restraint on the distance between the ligand center-of-mass and that of selected protein atoms around the binding site. This restraint is asymmetric within the framework developed in section III C.

In practice, in the case of physically well-separated, symmetric binding sites, it is numerically easier to first consider binding to a single site, and evaluate the restraint contribution with respect to this single site definition, $\Delta G^{\text{rest}\to\text{1site}}_{\text{coupled}}$. This procedure yields the binding constant to a single site K^{1site}, which ignores the other, symmetric sites. This is the *microscopic* binding constant of the biophysics literature. The overall first binding constant to any of of the n identical sites is:[70]

$$K_1 = n\,K^{\text{1site}} \tag{15}$$

The a posteriori "symmetry correction" included in Eq. 15 can be viewed in our framework as part of the restraint free energy contribution $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$:

$$\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}} = \Delta G^{\text{rest}\to 1\text{site}}_{\text{coupled}} - RT\ln(n) \tag{16}$$

where the term $-RT\ln(n)$ accounts for changing the site definition.

To predict binding of more ligands to the mono-liganded complex, one must then take into account the fact that this complex is less symmetric than the isolated receptor, so nothing more can be said without further simplifying assumptions. Typically, multimeric proteins like insulin are epected to exhibit cooperative binding. If, however, one assumes independent binding in the n different sites, then the binding constant of i ligands to the receptor can be expressed as a general function of $K^{1\text{site}}$ as:[70]

$$K_i = \frac{n-i+1}{i} K^{1\text{site}} \tag{17}$$

In the more realistic framework of cooperative binding, one would have to simulate explicitly the second binding event, with sufficient timescales to allow relaxation of the interactions that couple the two sites, and taking into account further changes in symmetry, depending on which combination of sites is populated. This is evidently a much more involved endeavor, and beyond our scope here.

2. *Cation binding to carboxylate groups: an example of a symmetric binding site and partial sampling.*

In the case discussed above of symmetric binding sites on large objects such as proteins, sampling is usually attempted—with restraints chosen accordingly— for only one of the binding sites at a time. However, symmetric "receptors" can also present two symmetric binding modes that are spatially close, in which case one would more spontaneously attempt sampling both modes at once. We will discuss in detail such a situation, showing how it relates and differs from the previous one, on the typical case of ion binding to a symmetric carboxylate group—acetate for the sake of simplicity—that we recently encountered, and for which we needed binding predictions to help interpret experiments.[56]

Different modes of cation binding to carboxylate groups have been described.[71,72] A cation can interact directly either with both carboxylate oxygen atoms in the bidentate binding mode, or with only one of the two oxygens in the monodentate mode. Binding can also occur through water molecules without any direct interactions, with the formation of solvent-shared ion pairs. For the purpose of this discussion, we now focus on the monodentate binding mode, our goal being to properly estimate the 1:1 monodentate binding of a cation (*e.g.* Mg^{2+}) to acetate. The binding site corresponding to monodentate binding is symmetric, since the cation can interact symetrically with either of the two oxygen atoms as pictured in Fig. 3.

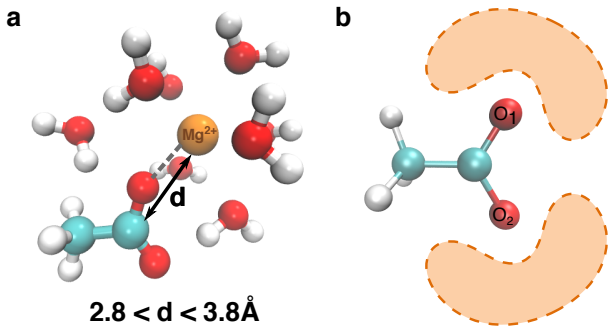

FIG. 3. a) Simulation snapshot of a magnesium cation interacting with an acetate anion in a monodentate mode. b) Qualitative scheme of the two symmetric monodentate binding lobes. The figures were prepared using the VMD visualization software.[69]

The distance d between the carbon atom of the carboxylate group and the cation is the natural coordinate, previously used in the literature,[55,56,73,74] to define the monodentate binding mode and separate it from bidentate and solvent-shared binding modes. For instance, in the case of Mg^{2+} binding to acetate, monodentate bound geometries correspond to $2.8 < d < 3.8$ Å, bidentate complexes typically have $d < 2.8$ Å, and solvent-shared ion pairs correspond to $d > 4$ Å.

The binding free energy corresponding to the monodentate binding mode can then be calculated with the double decoupling method, using the thermodynamic cycle described in Fig. 1. Weak flat-bottom harmonic restraints acting only if $d > 3.8$ Å or $d < 2.8$ Å are imposed during the alchemical decoupling in the bound state to maintain the system in a monodentate geometry. The employed restraint is symmetric and does not restrict the sampling to only one of the two symmetric binding poses, unlike what was discussed previously for symmetric binding sites on protein multimers.

For this practical example, we use standard non-polarizable force fields both for acetate[74] and the magnesium ion.[75] The free energy associated with each alchemical transformation is obtained using the BAR[39] algorithm. Other simulation details are provided in section VI. Table I summarizes the free energies associated with each step of the thermodynamic cycle (Fig. 1), including the appropriate second order correction to $\Delta G^{*}_{\text{bulk}}$ due to the change in overall charge during the alchemical transformation.[44] $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ is estimated on the fully coupled simulation window using Eq. 8, the site being defined strictly as $2.8 < d < 3.8$ Å. Strictly speaking, this region contains the physically accessible site as well as regions that correspond to steric clashes and are effectively forbidden: it is therefore a suitable definition of the site for numerical purposes. As expected with flat-bottom restraints that let the ligand evolve freely in the binding site, $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ is small (in the present case, negligible). $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$ is evaluated with Eqs. 6-7 by numerical integration of the restraining potential. The final standard binding free energy $\Delta G^{\circ}_{\text{bind}} = -37.2$ kJ mol^{-1} is then obtained combining all the different terms according to Eq. 5.

Since our definition of the "bound state" encompasses both

Free energy term	ΔG^*_{bulk}	ΔG^*_{site}	$\Delta G^{\text{V}^\circ\rightarrow\text{rest}}_{\text{decoupled}}$	$\Delta G^{\text{rest}\rightarrow\text{site}}_{\text{coupled}}$	$\Delta G^\circ_{\text{bind}}$
kJ mol^{-1}	1726.7 (0.3)	1769.7 (0.3)	5.91	0.0	−37.1 (0.6)

TABLE I. Decomposition of the free energy terms of Eq. 5, for the acetate-magnesium monodentate binding.

of the two symmetric binding modes (to both oxygen atoms), the obtained binding free energy $\Delta G^\circ_{\text{bind}}$ directly measures the binding to both of the symmetric sites, **without need for any symmetry correction**, despite the symmetry of the binding site.

We did not comment so far on the actual conformations sampled during the alchemical simulation. In practice, in the coupled state and in the first windows of the alchemical transformation, only one of the two binding sites is visited, due to kinetic trapping of the cation. Intuitively, it would then be tempting to try and correct for this partial sampling, since only half of the bound configurations are visited in the fully coupled state. However, we argue that counter-intuitively and despite partial sampling, the decoupling free energy is not affected and no corrections are needed. This derives from the fact that the BAR estimator that we employ depends only the potential energy distributions in each decoupling window. As discussed in III B, as long as the partial sampling does not modify the potential energy distribution, then the same free energy is obtained, irrespective of the sampling. We use the following numerical experiments to illustrate our point.
First, we reanalyze the data from our initial alchemical transformation (Table I) where only binding to O1 was sampled in the first few windows, and keep **in all windows** only the conformations corresponding to this binding mode. We now obtain $\Delta G^*_{\text{site}} = 1769.2 \pm 0.6$ kJ mol^{-1}, which is not different, within the error bars of our calculations, from the value initially obtained using all the samples. To further illustrate our point, we can use a slightly different Mg^{2+} force field,[76] so that both monodentate binding modes are sampled in all the windows of the alchemical transformation. We now evaluate ΔG^*_{site} using, in all windows, only the conformations where the cation is closer to O1 [resp. O2]. The obtained ΔG^*_{site} values do not significantly change—1647.4 ± 0.8 and 1647.5 ± 0.8 kJ mol^{-1} respectively—and are identical within the error bars to that obtained using all the samples, $\Delta G^*_{\text{site}} = 1647.4 \pm 0.6$ kJ mol^{-1}.

In both these examples, we numerically verified that ΔG^*_{site} does not depend at all on the extent of sampling of one or both symmetric binding modes in all or a few windows of the alchemical transformation. As detailed in section III B, this stems from the use of a free energy estimator that depends only on the distribution of comparison potential energies (ΔU_{AB}) in each window. In our case, due to the symmetry of the two binding modes, the distribution is the same irrespective of the (partial) sampling, and thus the same is true for the computed binding free energy.

B. Case 2: symmetric ligand – Phenol binding to lysozyme.

In contrast with large biomolecules, small ligands (*e.g.* phenol, benzene), may exhibit global molecular symmetry. As an example, we consider the classic case of phenol binding to the engineered binding sites of the lysozyme from phage T4.[77,78] Various mutant lysozymes (L99A, L99A/M102Q, L99A/M102H) have been shown to bind a variety of organic molecules, such as phenol, with affinities that depend on the more or less polar nature of the cavity. This system has been extensively characterized experimentally and has become a benchmark for binding free energy (affinity) calculations.[4,23,50,64,79]

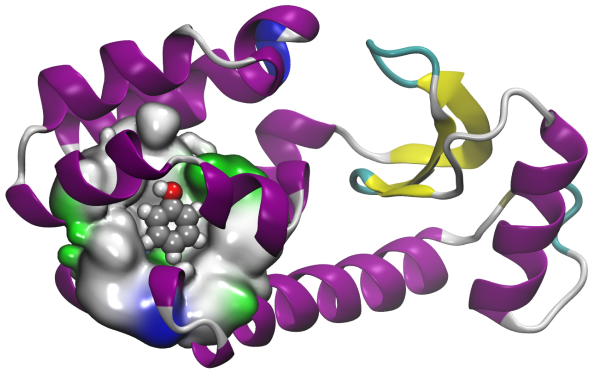

FIG. 4. L99A/M102H mutant of T4 lysozyme complexed with phenol. The protein is rendered as cartoon representation and colored based on secondary structure elements. Residues from the binding site are rendered as a smooth surface and colored based on residue type (white: hydrophobic, green: polar, blue: basic). Phenol is rendered as spacefill and colored by element. Rendered using VMD[69] based on PDB structure 4I7L.[78]

Due to the symmetry of the phenol molecule, 180° rotation of the aromatic ring around the CO bond axis leads macroscopically to the exact same binding pose. However, the two symmetric binding poses can be artificially distinguished in simulations because of the labelling of each atom.

In double decoupling free energy calculations, different kinds of restraints can be employed. If simple harmonic (or flat-bottom) restraints on the distance between the phenol center of mass and the protein binding site are used, then the sampling of both symmetric poses is in principle (if not in practice) allowed. These restraints follow the symmetry of the bound state. In contrast, if orientational restraints (for instance, overall ligand RMSD with respect to the binding site[28]) are employed, then the sampling is restricted to one of the two symmetric modes during the alchemical transformation. Such restraints are thus "asymmetric", in that they break the symmetry of the bound state. As explained in Sec-

tion II B, assuming numerical convergence, the different restraint schemes give the same value of $\Delta G^\circ_{\text{bind}}$, provided the restraint terms $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ and $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$ are properly taken into account.

However, we should underline that all protocols are not equivalent in terms of the rate of convergence of the $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ term. If the restraint is symmetric, then for the same reason as those developed in section III C, the restraint free energy is totally insensitive to partial sampling of one of the two modes, which makes its convergence dependent solely on relaxation within a binding pose, hence generally faster. However, with an asymmetric restraint (allowing sampling of only one of the two modes), the estimation of the restraint free energy $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ crucially depend on proper sampling of the two modes, because the perturbation now involves an asymmetric potential energy term, so comparison energies depend on the sampled mode (if the allowed mode is sampled in the restraint FEP, but not the forbidden mode, a spurious $RT\ln(2)$ appears in the free energy). Hence, for numerical efficiency, *it is generally advisable to use restraints of the same symmetry as the binding site definition*. If, however, asymmetric restraints are preferred for any reason, and assuming that they forbid one of the poses entirely, then we recommend decomposing the evaluation of $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ into two steps: first, estimate the restraint free energy with respect to only one of the binding poses; then add the analytical correction $-RT\ln(2)$ to take into account the existence of the other equivalent pose, a strategy similar to that adopted for symmetric binding sites in Section IV A 1.

C. Case 3: symmetric complex – Homodimerization

Biomolecules such as proteins frequently assemble into multimers, which are homomultimers (homodimers etc.) if they are formed by the assembly of identical molecules. The macroscopic thermodynamic description of homodimerization is slightly different from that of heterodimerization. To pinpoint the effect of symmetry on the definition and calculation of a homodimerization constant, we describe a thought experiment wherein the very same physical process can be described as either homo- or heterodimerization. We reason at the macroscopic level because it allows for a more direct connection with the macroscopic binding free energies that we seek to estimate.

Consider a solution containing solutes that can be arbitrarily considered of the same type M, or as distinct types A and B—a concrete example would be methane molecules containing either ^{12}C or ^{13}C, which could be considered identical or different depending on the method used to detect them.

We first consider the heterodimerization equilibrium:

$$A + B \rightleftharpoons AB$$

Suppose that the system is large enough that it obeys the law of mass action,[32,33] and that the total concentration of both A and B is 1. We also assume that A and B have exactly the same behavior with respect to dimerization. The population of dimers can be thought of as evenly split between AA, BB, AB, and BA, the latter two being of course the same object; thus the heterodimer is twice as concentrated as each homodimer. This can be arrived at by a symmetry argument: this is the distribution that would be obtained by random assignment of the labels A and B to otherwise identical molecules. If we define $x \equiv [AB]$, then $[A_2] = [B_2] = x/2$. The free monomer concentrations are $[A] = [B] = 1 - 2x$.

We can write the law of mass action for heterodimerization:

$$K^\circ_{AB} \equiv \frac{[AB]C^\circ}{[A][B]} = \frac{xC^\circ}{(1-2x)^2} \tag{18}$$

If we compute a PMF $w_{AB}(r)$ between A and B, we can directly apply Eq. (4) to the heterodimerization, and obtain the value of K°_{AB}. Note that this requires no particular assumption on the possible homodimerization of species A and B separately.

Now we re-analyze the very same physical system, ignoring the labels A and B and regarding all solutes as the generic monomer M. The homodimerization equilibrium writes:

$$2M \rightleftharpoons M_2$$

And $[M_2] = [AB] + [A_2] + [B_2] = 2x$, and $[M] = [A] + [B] = 2(1-2x)$.

The law of mass action for homodimerization is:

$$K^\circ_{M_2} \equiv \frac{[M_2]C^\circ}{[M]^2} = \frac{2xC^\circ}{[2(1-2x)]^2} \tag{19}$$

Comparing equations 18 and 19:

$$K^\circ_{M_2} = \frac{1}{2}K^\circ_{AB} \tag{20}$$

That is, keeping the same interactions between the monomers, the homodimerization constant is half the heterodimerization constant. This thought experiment may seem contrived, but any homodimerization equilibrium can be reduced to it by arbitrarily labeling each half of the monomers (isotope labeling is again a concrete analogy). Thus, this result is valid for homodimerization in general.

Since A and B are identical to M with respect to intermolecular interactions, the self-association PMF of M is that of A and B: $w_{MM}(r) = w_{AB}(r)$. Therefore the homodimerization constant $K^\circ_{M_2}$ can be expressed by modifying Eq. (4):

$$K^\circ_{M_2} = \frac{1}{2V^\circ}\int_{\text{site}} 4\pi r^2 e^{-\beta w_{MM}(r)}\text{d}\mathbf{r} \tag{21}$$

This expression has been arrived at by another route.[80] However, on many occasions, the factor of 2 has been omitted, including by us.[81,82] The resulting discrepancy of $RT\ln(2)$ is small enough to be usually inconspicuous, given the error margins of computational—and experimental—estimates of binding free energies.

V. CONCLUSION

We have recalled the main exact theoretical results allowing for the practical determination of macroscopic, absolute binding free energies based on explicit-solvent molecular simulations. Adding to the existing practical advice on

how to perform absolute binding free calculations using alchemical transformations,[24] we point to several controversial steps in the calculations that can lead to erroneous numerical treatment–even though the resulting error is often hidden in the large error bars of the alchemical transformation itself. We argue that a key step to obtain well-defined binding constants (or binding free energies) that can be compared to experimental data is to precisely define the binding site (or mode), the thermodynamics of which you wish to characterize. This should be done as consistently as possible with the experimental measurements that can be predicted by or compared with simulations. This often requires simplifications (such as assuming a two-state model), unless direct computation of the experimental signal is possible.

Another key point is to carefully correct for the restraints both in the decoupled and coupled state, in a way that is consistent with the site definition. We show that this clarifies several practical situations and, in the case of symmetric receptor or ligand, eliminates the need for a specific and often ill-interpreted symmetry correction term. Instead, symmetry effects are accounted for when evaluating the restraint contributions to the free energy in both the coupled and decoupled states, and we discuss the potential pitfalls of different protocols. We have also shown that counter to a very common intuition, in symmetric cases, estimators of the excess free energies of decoupling are not affected by total, partial, or non-existent sampling of some symmetry-equivalent modes.

Finally, we have shown that the computation of macroscopic constants for homodimerization equilibria requires a special correction factor, and explained its origin.

VI. COMPUTATIONAL DETAILS

Computation of the binding free energy of Mg^{2+} to acetate (IV A 2) was performed using the Gromacs 5.1.1 software.[83] The simulation box contained one acetate anion, one cation and 1723 water molecules. The simulations were performed in the constant temperature/constant pressure (NpT) ensemble, using the same setup as previously reported.[56] Water molecules were described by the SPC/E force field,[84] and two different force fields[75,76] for Mg^{2+}, giving rise to two different behaviors with respect to sampling of one or both symmetric sites, were used. Note that these force fields are known to strongly overestimate ion-acetate binding, as we discussed in a previous work,[56] and their use here is only meant to illustrate typical issues encountered in free energy calculations. Free energy calculations were performed using the double decoupling procedure described in II B. The free energy difference associated with each alchemical transformation was reconstructed using the Bennett Acceptance Ratio[39] (BAR) method as implemented in Gromacs. The restraint used during the alchemical transformation in the bound state is a flat bottom well potential, flat for $2.8 < d < 3.8$, and harmonic on each side with a $k = 100000$ kJ mol^{-1} nm^{-2} force constant.

Appendix A: Theoretical connection between the dimer counting and PMF formalisms

Here we derive the connection between Eq. 4, which expresses the binding constant in terms of an integral over the PMF, and dimerization statistics in a microscopic simulation system.

Most binding free energy calculations rely on simulations of a unitary system of volume V containing 1 R and 1 L. Calling x the probability of the bound state (or bound fraction) in that system, and naively applying the law of mass action (Eq. 3) would lead to:

$$K^\circ_{\text{bind}} = \frac{x}{(1-x)^2}\frac{V}{V^\circ}\,, \tag{A1}$$

where $V^\circ = 1/C^\circ$ is the standard volume. However, it has been shown[32] that this expression is not applicable to a microscopic simulation box containing a single R and L: the law of mass action results from statistics over many copies of R and L. In a unitary system where the solutes do not interact outside of the bound state,[33] K°_{bind} may instead be expressed as:

$$K^\circ_{\text{bind}} = \frac{x}{(1-x)}\frac{V}{V^\circ} \tag{A2}$$

We now show that this expression is closely related to the PMF-based expression of the binding constant (Eq. 4). Consider the probability distribution $p(\mathbf{x})$ of the 3D ligand position with respect to the receptor, normalized to be 1 in the bulk. The framework of Eq. 4 requires that the bound state be defined solely by the presence of $\mathbf{x}$ in a region of space corresponding to the binding site, without regard to non-translational degrees of freedom. We call W the weight of the bound state in this distribution:

$$W = \int_{\text{site}} p(\mathbf{x})\mathrm{d}\mathbf{x} \tag{A3}$$

$$= \int_{\text{site}} g(r)4\pi r^2 \mathrm{d}r \tag{A4}$$

which we note is the integral in Eq. 4. Under the assumptions of Eq. A2 that the partners are either bound or non-interacting ($w(r) = 0$ outside the site),[33] the weight of the unbound state is:

$$\int_{\overline{\text{site}}} p(\mathbf{x})d\mathbf{x} = V - V_{\text{site}} \approx V\,, \tag{A5}$$

with the common assumption that $V \gg V_{\text{site}}$, so that

$$\int_V p(\mathbf{x})\mathrm{d}\mathbf{x} = W + V \tag{A6}$$

Therefore

$$x = \frac{W}{W+V} \text{ and } 1-x = \frac{V}{W+V} \tag{A7}$$

and

$$\frac{x}{1-x} = \frac{W}{V}\,, \tag{A8}$$

which, substituted into Eq. A2, gives Eq. 4.

Site def.	Restraint	$\Delta G^*_{\text{bulk}} - \Delta G^*_{\text{site}}$	$\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$	$\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$	$\Delta G^\circ_{\text{bind}}$
asym	asym	$\Delta\Delta G^*$	$-RT\ln(V^1_{\text{rest}}/V^\circ)$	0	$\Delta\Delta G^* - RT\ln(V^1_{\text{rest}}/V^\circ)$
sym	asym	$\Delta\Delta G^*$	$-RT\ln(V^1_{\text{rest}}/V^\circ)$	$-RT\ln(n)$	$\Delta\Delta G^* - RT\ln(V^1_{\text{rest}}/V^\circ) - RT\ln(n)$
sym	sym	$\Delta\Delta G^*$	$-RT\ln(n\times V^1_{\text{rest}}/V^\circ)$	0	$\Delta\Delta G^* - RT\ln(V^1_{\text{rest}}/V^\circ) - RT\ln(n)$

TABLE II. **Decomposition of the absolute binding free energy (Eq. 5), in the case of binding modes with or without order-n symmetry, with symmetric or asymmetric flat-bottom restraints.** Simple analytical expressions apply because of assumptions listed in the text. In a more general case, restraint free energies may be estimated numerically, and the $RT\ln(n)$ terms listed here are, at convergence, accounted for by the numerical estimators, and need not be explicitly added. See section IV B for a discussion of cases where these might not converge in practice, and strategies to handle it.

Appendix B: Analytical treatment of symmetry contributions to restraint free energies under simplifying assumptions

To make the symmetry contributions to the restraint free energies apparent, we describe a case where $\Delta G^{\text{V}^\circ\to\text{rest}}_{\text{decoupled}}$ and $\Delta G^{\text{rest}\to\text{site}}_{\text{coupled}}$ admit simple analytical expressions. In most practical applications, at least one of these terms has to be computed numerically. Suppose that we want to estimate the binding affinity of a bound state that comprises either n symmetry-equivalent binding poses (symmetric ligand) or n symmetry-equivalent binding sites (symmetric receptor): these two cases are largely equivalent from a formal perspective. Crucially, we apply flat-bottom restraints that do not perturb binding to an individual mode. Different definitions of both the site and the restraints are possible. Table II summarizes the value of the different free energy terms depending on these choices. As stated above, the values of the excess free energies of decoupling, ΔG^*_{bulk} and ΔG^*_{site}, are independent from the symmetry of the restraint scheme or that of the site definition. We call V^1_{rest} the restraint volume for one mode. If we use an asymmetric site definition, *i.e.* limited to a single binding site or binding pose, then we naturally employ asymmetric restraints (first line of Table II) and the obtained binding free energy is then:

$$\Delta G^\circ_{\text{bind, asym}} = \Delta\Delta G^* - RT\ln(V^1_{\text{rest}}/V^\circ)\,. \tag{B1}$$

In contrast, if we define the binding site as symmetric, then different restraints can be used (last two lines of Table II), either obeying or breaking the symmetry of the binding site. Both setups eventually yield the same standard free energy of forming the n-fold-symmetric complex:

$$\Delta G^\circ_{\text{bind}} = \Delta\Delta G^* - RT\ln(V^1_{\text{rest}}/V^\circ) - RT\ln(n)\,. \tag{B2}$$

ACKNOWLEDGMENTS

This work was supported by the "Initiative d'Excellence" program from the French State (Grants "DYNAMO", ANR-11-LABX-0011, and "CACSICE", ANR-11-EQPX-0008). Computational work was performed using HPC resources from LBT/HPC. We thank our colleagues H. Martinez-Seara, P. Jungwirth and V. Palivec (UOCHB, Prague, CZ), C. H. Robert (LBT, IBPC, Paris), and G. Brannigan (Rutgers Camden, NJ, USA) for stimulating discussions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at http://doi.org/10.5281/zenodo.4519498.

[1] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham, "Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models," Acc. Chem. Res. **33**, 889–897 (2000).

[2] S. Holderbach, L. Adam, B. Jayaram, R. C. Wade, and G. Mukherjee, "RASPD+: Fast protein-ligand binding free energy prediction using simplified physicochemical features," Front. Mol. Biosci. **7** (2020), 10.3389/fmolb.2020.601065.

[3] C. Chipot, A. E. Mark, V. S. Pande, and T. Simonson, "Applications of free energy calculations to chemistry and biology," in *Free Energy Calculations Theory and Applications in Chemistry and Biology*, edited by C. Chipot and A. Pohorille (Springer, 2007) pp. 463–492.

[4] D. L. Mobley and M. K. Gilson, "Predicting binding free energies: Frontiers and benchmarks," Annu. Rev. Biophys. **46**, 531–558 (2017).

[5] D. L. Mobley and P. V. Klimovich, "Perspective: Alchemical free energy calculations for drug discovery," J. Chem. Phys. **137**, 230901 (2012).

[6] M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp, and P. C. Biggin, "Predictions of ligand selectivity from absolute binding free energy calculations," J. Am. Chem. Soc. **139**, 946–957 (2017).

[7] L. Jiang, Y. Gao, F. Mao, Z. Liu, and L. Lai, "Potential of mean force for protein-protein interaction studies," Proteins: Structure, Function and Genetics **46**, 190–196 (2002).

[8] D. Suh, S. Jo, W. Jiang, C. Chipot, and B. Roux, "String method for protein–protein binding free-energy calculations," J. Chem. Theory Comput. **15**, 5829–5844 (2019).

[9] T. Siebenmorgen and M. Zacharias, "Computational prediction of protein–protein binding affinities," Wiley Interdiscip. Rev.: Comp. Mol. Sci. **10**, 1–18 (2020).

[10] D. Jakubec and J. Vondrášek, "Efficient Estimation of Absolute Binding Free Energy for a Homeodomain-DNA Complex from Nonequilibrium Pulling Simulations," J. Chem. Theory Comput. **16**, 2034–2041 (2020).

[11] L. F. Song and K. M. Merz, "Evolution of alchemical free energy methods in drug discovery," J. Chem. Inf. Model. **60**, 5308–5318 (2020).

[12] Z. Cournia, B. K. Allen, T. Beuming, D. A. Pearlman, B. K. Radak, and W. Sherman, "Rigorous free energy simulations in virtual screening," J. Chem. Inf. Model. **60**, 4153–4169 (2020).

[13] J. Hermans and S. Subramaniam, "The Free Energy of Xenon Binding to Myoglobin from Molecular Dynamics Simulation," Israel J. Chem. **27**, 225–227 (1986).

[14] B. Roux, M. Nina, R. Pomès, and J. Smith, "Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study," Biophys. J. **71**, 670–681 (1996).

[15] 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).

[16] V. Helms and R. C. Wade, "Hydration energy landscape of the active site cavity in cytochrome p450cam," Proteins: Structure, Function, and Genetics **32**, 381–396 (1998).

[17] H. Luo and K. Sharp, "On the calculation of absolute macromolecular binding free energies," Proc. Nat. Acad. Sci. **99**, 10399–10404 (2002).

[18] 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).

[19] M. Mihailescu and M. K. Gilson, "On the theory of noncovalent binding," Biophys. J. **87**, 23–36 (2004).

[20] H.-J. Woo and B. Roux, "Calculation of absolute protein-ligand binding free energy from computer simulations." Proc. Nat. Acad. Sci. **102**, 6825–30 (2005).

[21] D. L. Mobley, J. D. Chodera, and K. A. Dill, "On the use of orientational restraints and symmetry corrections in alchemical free energy calculations," J. Chem. Phys. **125**, 084902 (2006).

[22] S. Miyamoto and P. A. Kollman, "Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches," Proteins: Struct. Func. Bioinf. **16**, 226–245 (1993).

[23] Y. Deng and B. Roux, "Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant," J. Chem. Theory Comput. **2**, 1255–1273 (2006).

[24] A. S. J. S. Mey, B. Allen, H. E. B. Macdonald, J. D. Chodera, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern, and H. Xu, "Best Practices for Alchemical Free Energy Calculations," Living J. Comp. Mol. Sci. , 1–51 (2020).

[25] "http://www.alchemistry.org/wiki/Tutorials," .

[26] J. M. Swanson, R. H. Henchman, and J. A. McCammon, "Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy," Biophys. J. **86**, 67–74 (2004).

[27] S. Genheden and U. Ryde, "The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities Samuel," Expert Opin. Drug Discov. **10**, 449–461 (2015).

[28] R. Salari, T. Joseph, R. Lohia, J. Hénin, and G. Brannigan, "A Streamlined, General Approach for Computing Ligand Binding Free Energies and Its Application to GPCR-Bound Cholesterol," J. Chem. Theory Comput. **14**, 6560–6573 (2018).

[29] W. L. Jorgensen, "Interactions between Amides in Solution and the Thermodynamics of Weak Binding," J. Am. Chem. Soc. **111**, 3770–3771 (1989).

[30] J. Pranata and W. L. Jorgensen, "Monte Carlo Simulations Yield Absolute Free Energies of Binding theory and the OPLS potential functions lar dynamics," Tetrahedron **41**, 2491–2501 (1991).

[31] D. Shoup and A. Szabo, "Role of diffusion in ligand binding to macromolecules and cell-bound receptors," Biophys. J. **40**, 33–39 (1982).

[32] D. H. D. Jong, L. V. Schafer, A. H. D. Vries, S. J. Marrink, H. J. C. Berendsen, and H. Grubmüller, "Determining Equilibrium Constants for Dimerization Reactions from Molecular Dynamics Simulations," J. Comput. Chem. **32**, 1919–1928 (2011).

[33] A. Jost Lopez, P. K. Quoika, M. Linke, G. Hummer, G. Hummer, and J. Köfinger, "Quantifying Protein-Protein Interactions in Molecular Simulations," J. Phys. Chem. B **124**, 4673–4685 (2020).

[34] E. Gallicchio and R. M. Levy, *Advances in Protein Chemistry and Structural Biology*, 1st ed., Vol. 85 (Elsevier Inc., 2011) pp. 27–80.

[35] Y. Deng and B. Roux, "Computations of standard binding free energies with molecular dynamics simulations," J. Phys. Chem. B **113**, 2234–2246 (2009).

[36] S. Doudou, N. A. Burton, and R. H. Henchman, "Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force," J. Chem. Theory Comput. , 909–918 (2009).

[37] J. C. Gumbart, B. Roux, and C. Chipot, "Standard binding free energies from computer simulations: What is the best strategy?" J. Chem. Theory Comput. **9**, 794–802 (2013).

[38] R. A. Corey, O. N. Vickery, M. S. P. Sansom, and P. J. Stansfeld, "Insights into membrane protein–lipid interactions from free energy calculations," Journal of Chemical Theory and Computation **15**, 5727–5736 (2019).

[39] C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," J. Comput. Phys. **22**, 245–268 (1976).

[40] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states." J. Chem. Phys. **129**, 124105 (2008).

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

[42] J. G. Kirkwood, "Statistical mechanics of fluid mixtures," J. Chem. Phys. **3**, 300–313 (1935).

[43] N. Lu, D. A. Kofke, and T. B. Woolf, "Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods," J. Comput. Chem. **25**, 28–40 (2003).

[44] T. Simonson and B. Roux, "Concepts and protocols for electrostatic free energies," Mol. Sim. **42**, 1090–1101 (2016).

[45] G. J. Rocklin, D. L. Mobley, K. A. Dill, and P. H. Hünenberger, "Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects," J. Chem. Phys. **139**, 4–8 (2013).

[46] M. M. Reif and C. Oostenbrink, "Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation," J. Comput. Chem. **35**, 227–243 (2014).

[47] J. Janin, "For Guldberg and Waage, With Love and Cratic Entropy," Proteins Struct. Funct. Gen. , 0–1 (1996).

[48] W. L. Jorgensen, J. K. Buckner, S. Boudon, and J. Tirado-Rives, "Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water," J. Chem. Phys. **89**, 3742–3746 (1988).

[49] H. Fujitani, Y. Tanida, and A. Matsuura, "Massively parallel computation of absolute binding free energy with well-equilibrated states," Phys. Rev. E **79**, 1–12 (2009).

[50] Y. Sakae, B. W. Zhang, R. M. Levy, and N. Deng, "Absolute Protein Binding Free Energy Simulations for Ligands with Multiple Poses, a Thermodynamic Path That Avoids Exhaustive Enumeration of the Poses," J. Comput. Chem. **41**, 56–68 (2020).

[51] N. Deng, D. Cui, B. W. Zhang, J. Xia, J. Cruz, and R. Levy, "Comparing alchemical and physical pathway methods for computing the absolute binding free energy of charged ligands," Phys. Chem. Chem. Phys. **20**, 17081–17092 (2018).

[52] V. Helms and R. Wade, "Thermodynamics of water mediating protein-ligand interactions in cytochrome p450cam: a molecular dynamics study," Biophys. J. **69**, 810–824 (1995).

[53] P. Procacci, "Alchemical determination of drug-receptor binding free energy: Where we stand and where we could move to," J. Mol. Graph. Model. **71**, 233–241 (2017).

[54] H. Fujitani, Y. Tanida, M. Ito, G. Jayachandran, C. D. Snow, M. R. Shirts, E. J. Sorin, and V. S. Pande, "Direct calculation of the binding free energies of FKBP ligands," J. Chem. Phys. **123**, 1–5 (2005).

[55] M. Kumar, T. Simonson, G. Ohanessian, and C. Clavaguéra, "Structure and thermodynamics of Mg:phosphate interactions in water: A simulation study," ChemPhysChem **16**, 658–665 (2015).

[56] D. M. De Oliveira, S. R. Zukowski, V. Palivec, J. Hénin, H. Martinez-Seara, D. Ben-Amotz, P. Jungwirth, and E. Duboué-Dijon, "Binding of divalent cations to acetate: Molecular simulations guided by Raman spectroscopy," Phys. Chem. Chem. Phys. **22**, 24014–24027 (2020).

[57] J. Wang, Y. Deng, and B. Roux, "Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials," Biophys. J. **91**, 2798–2814 (2006).

[58] R. D. Groot, "The association constant of a flexible molecule and a single atom: Theory and simulation," J. Chem. Phys. **97**, 3537–3549 (1992).

[59] R. M. Neumann, "Molecular association at the microscopic level," Canadian J. Phys. **89**, 793–797 (2011).

[60] D. Chandler and L. R. Pratt, "Statistical mechanics of chemical equilibria and intramolecular structures of nonrigid molecules in condensed phases," J. Chem. Phys. **65**, 2925–2940 (1976).

[61] M. K. Gilson and K. K. Irikura, "Symmetry numbers for rigid, flexible, and fluxional molecules: Theory and applications†," The Journal of Physical Chemistry B **114**, 16304–16317 (2010).

[62] M. K. Gilson and K. K. Irikura, "Correction to "symmetry numbers for rigid, flexible, and fluxional molecules: Theory and applications"," The Journal of Physical Chemistry B **117**, 3061–3061 (2013).

[63] D. N. LeBard, J. Hénin, R. G. Eckenhoff, M. L. Klein, and G. Brannigan, "General anesthetics predicted to block the GLIC pore with micromolar affinity," PLoS Comput. Biol. **8**, e1002532 (2012).

[64] J. Hermans and L. Wang, "Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. application to a complex of benzene and mutant t4 lysozyme," J. Am. Chem. Soc. **119**,

2707–2714 (1997).

[65] Our formalism differs from Gilson et al.'s, because their definition of the complex includes an indicator function of bound configurations, which is also used as a restraint in simulations. Such restraint terms can behave as non-dissociative bonds. If they are integrated into the definition of the complex and reduce its symmetry, then they give rise to a non-zero symmetry correction. Thus, the symmetry correction of Gilson et al. plays the role of the symmetry contributions to $\Delta G^{\mathrm{rest}\to\mathrm{site}}_{\mathrm{coupled}}$ in our formalism.

[66] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. i. nonpolar gases," J. Chem. Phys. **22**, 1420–1426 (1954).

[67] U. Derewenda, Z. Derewenda, E. J. Dodson, G. G. Dodson, C. D. Reynolds, G. D. Smith, C. Sparks, and D. Swenson, "Phenol stabilizes more helix in a new symmetrical zinc insulin hexamer," Nature **338**, 594–596 (1989).

[68] V. Palivec, C. M. Viola, M. Kozak, T. R. Ganderton, K. Křížková, J. P. Turkenburg, P. Halušková, L. Žáková, J. Jiráček, P. Jungwirth, and A. M. Brzozowski, "Computational and structural evidence for neurotransmitter-mediated modulation of the oligomeric states of human insulin in storage granules," J. Biol. Chem. **292**, 8342–8355 (2017).

[69] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: visual molecular dynamics," J. Mol. Graph. **14**, 33–8, 27–8 (1996).

[70] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry. Part III: The Behavior of Biological Macromolecules* (WH Freeman, 1980).

[71] T. Dudev and C. Lim, "Effect of Carboxylate-Binding Mode on Metal Binding / Selectivity," Acc. Chem. Res. **40**, 53–56 (2007).

[72] H. Einspahr and C. E. Bugg, "The geometry of calcium carboxylate interactions in crystalline complexes," Acta Cryst. B **37**, 1044–1052 (1981).

[73] M. D. Daily, M. D. Baer, and C. J. Mundy, "Divalent Ion Parameterization Strongly Affects Conformation and Interactions of an Anionic Biomimetic Polymer," J. Phys. Chem. B **120**, 2198–2208 (2016).

[74] T. Martinek, E. Duboué-Dijon, Š. Timr, P. E. Mason, K. Baxová, H. E. Fischer, B. Schmidt, E. Pluharová, and P. Jungwirth, "Calcium ions in aqueous solutions: Accurate force field description aided by ab initio molecular dynamics and neutron scattering," J. Chem. Phys. **148**, 222813 (2018).

[75] P. Li, B. P. Roberts, D. K. Chakravorty, and K. M. Merz, "Rational design of particle mesh ewald compatible lennard-jones parameters for +2 metal cations in explicit solvent," J. Chem. Theory Comput. **9**, 2733–2748 (2013).

[76] K. M. Callahan, N. N. Casillas-Ituarte, M. Roeselová, H. C. Allen, and D. J. Tobias, "Solvation of magnesium dication: molecular dynamics simulation and vibrational spectroscopic study of magnesium chloride in aqueous solutions." J. Phys. Chem. A **114**, 5141–5148 (2010).

[77] A. E. Eriksson, W. A. Baase, J. A. Wozniak, and B. W. Matthews, "A cavity-containing mutant of t4 lysozyme is stabilized by buried benzene," Nature **355**, 371–373 (1992).

[78] M. Merski and B. K. Shoichet, "Engineering a model protein cavity to catalyze the kemp elimination," Proc. Nat. Acad. Sci. **109**, 16179–16183 (2012).

[79] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill, and B. K. Shoichet, "Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site," J. Mol. Biol. **394**, 747–763 (2009).

[80] J. Domański, G. Hedger, R. B. Best, P. J. Stansfeld, and M. S. P. Sansom, "Convergence and sampling in determining free energy landscapes for membrane protein association," J. Phys. Chem. B **121**, 3364–3375 (2016).

[81] J. Hénin, A. Pohorille, and C. Chipot, "Insights into the recognition and association of transmembrane α-helices. The free energy of α-helix dimerization in glycophorin A," J. Am. Chem. Soc. **127**, 8478–8484 (2005).

[82] J. Hénin, A. Pohorille, and C. Chipot, "Erratum: Insights into the recognition and association of transmembrane α-helices. the free energy of α-helix dimerization in glycophorin A," J. Am. Chem. Soc. **132**, 9510–9510 (2010).

[83] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," J. Comput. Chem. **26**, 1701–1718 (2005).

[84] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, "The missing term in effective pair potentials," J. Phys. Chem. **91**, 6269–6271 (1987).