

Quantifying Protein-Protein Interactions in Molecular Simulations

Alfredo Jost Lopez, Patrick K. Quoika, Max Linke, Gerhard Hummer, [Juergen Koefinger](#)

Submitted date: 20/12/2019 • Posted date: 23/12/2019

Licence: CC BY-NC-ND 4.0

Citation information: Jost Lopez, Alfredo; Quoika, Patrick K.; Linke, Max; Hummer, Gerhard; Koefinger, Juergen (2019): Quantifying Protein-Protein Interactions in Molecular Simulations. ChemRxiv. Preprint.
<https://doi.org/10.26434/chemrxiv.11421201.v1>

We present simple, accurate, and efficient methods to estimate the dissociation constant K_d and the second osmotic virial coefficient B_2 from molecular simulations. We show that for simulations of two proteins in a box, K_d is determined by B_2 and the fraction of bound protein. We present two different methods to calculate B_2 from Monte Carlo and molecular dynamics simulations using implicit or explicit solvent. We derive a surprisingly simple expression for B_2 , adding significantly to the understanding of this important quantity. Non-binding interactions of proteins and other macromolecules shape the physicochemical properties of the crowded environments inside cells and of biomolecular condensates. We show how to extract the contributions of non-binding conformations to B_2 and discuss how these can be determined in analytical ultracentrifugation and SAXS experiments. We expect that our methods will prove to be instrumental in force parameterization efforts and high-throughput studies of large interactomes.

File list (1)

preprint.pdf (1.49 MiB)

[view on ChemRxiv](#) • [download file](#)

Quantifying protein-protein interactions in molecular simulations

Alfredo Jost Lopez,[†] Patrick K. Quoika,^{†,¶} Max Linke,[†] Gerhard Hummer,^{*,†,‡} and
Jürgen Köfinger^{*,†}

[†]*Department of Theoretical Biophysics, Max Planck Institute of Biophysics,
Max-von-Laue-Straße 3, 60438 Frankfurt am Main, Germany*

[‡]*Institute for Biophysics, Goethe University, Max-von-Laue-Straße 9, 60438 Frankfurt am
Main, Germany*

[¶]*Current address: Institute of General, Inorganic and Theoretical Chemistry, University of
Innsbruck, Innrain 80-82, A-6020 Innsbruck, Austria*

E-mail: gerhard.hummer@biophys.mpg.de; juergen.koefinger@biophys.mpg.de

Abstract

The interactions between proteins, nucleic acids, and other macromolecules are essential for their biological functions and shape the physicochemical properties of the crowded environments inside living cells. Binding interactions are commonly quantified by dissociation constants K_d and both binding and non-binding interactions are quantified by second osmotic virial coefficients B_2 . As a measure of non-specific binding and stickiness, B_2 is receiving renewed attention in the context of so-called liquid-liquid phase separation in protein and nucleic acid solutions. We show that K_d is fully determined by B_2 and the fraction of dimer observed in molecular simulations of two proteins in a box. We develop two methods to calculate B_2 . From molecular dynamics or Monte Carlo simulations using implicit solvent, we can determine B_2 from insertion and removal energies by applying Bennett’s acceptance ratio (BAR) method or the (binless) weighted histogram analysis method (WHAM). From simulations using implicit or explicit solvent, one can estimate B_2 from the probability that the two molecules are within a volume large enough to cover their range of interactions. We validate these methods for coarse-grained Monte Carlo

simulations of three weakly binding proteins. The high accuracy of our estimates for K_d and B_2 allows us to separate out the contributions of non-binding interactions to B_2 . Comparison of calculated and measured values of K_d and B_2 can be used to (re-)parameterize and improve molecular force fields by calibrating specific affinities and overall stickiness. The accuracy and efficiency of K_d and B_2 calculations makes them well suited for high-throughput studies of large interactomes.

1 Introduction

In biological cells, most protein, DNA, and RNA molecules have to bind to specific binding partners to perform their biological functions. These specific interactions compete with non-specific interactions and cells have evolved various mechanisms to minimize wasteful non-specific binding.^{1,2} However, non-specific interactions shape the physicochemical properties of the crowded environments inside cells.³ The quantification of binding affinities and interaction strengths of biological macromolecules is thus crucial for the understanding and modeling of cellular processes. In the following, we focus on protein-protein interactions but all our

results are generally applicable to other specific and non-specific binding interactions.

Experimentally, protein interactions are quantified by the dissociation constants K_d and the second osmotic virial coefficient B_{ij} of protein species i and j . We follow the common convention and use B_{22} for self-interactions and B_{23} for cross-interactions. The dissociation constant K_d quantifies the amount of bound proteins and can be measured in isothermal titration calorimetry, surface plasmon resonance, or analytical ultracentrifugation experiments, for example.⁴ The interaction strength of pairs of proteins in binding and non-binding configurations can be quantified by measuring the second osmotic virial coefficient B_{ij} , which relates the microscopic protein interactions to the macroscopic osmotic pressure.^{5–7} Moreover, the second osmotic virial coefficient is related to solubility and used as a predictor for protein crystallization conditions.^{8,9} In experiments, B_{ij} is measured by sedimentation^{10–12} and size-exclusion chromatography.⁸ Scattering experiments, such as static light scattering and small-angle x-ray scattering (SAXS) experiments, can provide approximate estimates for B_{ij} .^{13,14}

K_d and B_{ij} are crucial quantities to relate molecular simulations of interacting proteins to experiment. Such comparisons become increasingly important as molecular simulations of crowded cell-like environments have become computationally feasible, even in full atomic detail.^{15,16} In simulations of strong binders, K_d is usually determined by calculating the binding free energy to specific binding interfaces.¹⁷ If binding interfaces are unknown, K_d values are often calculated from the ratio of bound and unbound populations, as recently applied to RNA-RNA binding.¹⁸ As we will discuss here, this approximation is only accurate for strong binders. B_{ij} can be estimated by integration over configuration space,^{19–21} Mayer-sampling,^{22,23} and from molecular simulations using radial distributions or potentials of mean force.^{24–27}

Here we show that K_d is fully determined by B_{ij} and the fraction $p_b(V)$ of bound proteins estimated from molecular simulations of two pro-

teins in a box with volume V , i.e.,

$$K_d = \frac{1}{N_A p_b(V)(V - 2B_{ij})} \quad (1)$$

N_A is Avogadro’s constant. In the derivation of this equation, we do not make any assumptions about interaction strength or about the degrees of freedom of proteins or solvent. Thus it is generally applicable and valid not only for coarse-grained simulations using implicit solvent but also for fully atomistic molecular dynamics simulations using explicit solvent. We present two different routes to calculate B_{ij} and thus K_d .

For simulations using implicit solvent, we can apply protein insertion and removal moves to estimate the free energy that corresponds to the two-particle partition function determining B_{ij} . The insertion ensemble can be generated with any Monte Carlo or molecular dynamics code to sample from the canonical ensemble without modification. We estimate the partition function by combining the insertion and removal ensemble using either Bennett’s acceptance ratio (BAR) method²⁸ or the binless weighted histogram analysis method (WHAM).^{29–31} In contrast to Mayer sampling,^{22,23} a more general approach using molecular Monte Carlo integration to calculate virial coefficients even of higher-order, we use here exactly the same simulation system for the calculation of B_{ij} as we use to sample from the canonical ensemble.

For simulations using either implicit or explicit solvent, we show that B_{ij} can be calculated accurately by estimating the probability that the two proteins are within a volume that covers the interaction range of the two proteins. We present mathematically simple expressions for B_{ij} and K_d in terms of this probability, which provide insight into their physical interpretations complementary to more common formulations.

We quantify the interactions of the two proteins when they are not bound from our estimates for K_d and B_{ij} . Previously, theoretical models for excluded volumes have been used to extract non-binding interactions from experimentally measured B_{ij} values.³² Here, we use the fact that the contributions of bound config-

urations to B_{ij} are completely determined by K_d and show that the remaining contributions have a simple and clear interpretation. Moreover, we propose that these contributions of non-binding interactions can be estimated in experiments.

The article is organized as follows. In section 2 we derive expressions to calculate the dissociation constant and the second osmotic virial coefficient from simulations. We present the details of our methods in section 3 and present a validation of our methods and results for three weakly binding proteins using coarse-grained simulations in section 4. We end with conclusions in section 5.

2 Theory

For simulations of two proteins in a box, we show that the dissociation constant K_d is determined by the binding probability and the second osmotic virial coefficient B_{ij} of protein species i and j . The latter is determined by the two-particle partition function, which in general can be estimated from the probability that the two proteins are in a volume large enough to cover all inter-protein interactions or, for implicit solvent, by performing a free energy calculation using insertion and removal moves.

2.1 Preliminaries

McMillan and Mayer⁵ have shown how we can apply results of statistical mechanics to describe osmotic properties of solutions. Integrating out solvent degrees of freedom, only solute degrees of freedom remain and solutes interact with each other via effective potentials. For such a system with m solute species, the virial equation of state^{33,34} becomes the osmotic virial equation of state, i.e.,

$$\frac{\Pi V_m}{RT} = \sum_{i=1}^m x_i + \frac{1}{V_m} \sum_{i=1}^m \sum_{j=1}^m x_i x_j B_{ij} + \dots \quad (2)$$

where Π is the osmotic pressure, V_m is the molar volume, R is the gas constant, T the temperature, x_i is the mole fraction of species i , and B_{ij}

is the osmotic second virial coefficient of particles of species i and j .

We can express the second virial coefficients B_{ij} of an arbitrarily shaped particle of species i and an arbitrarily shaped particle of species j , via one- and two-particle configurational partition functions. To do so, we follow the derivation by Donald A. McQuarrie for spherical particles.⁷ As a starting point, we use the canonical partition function for a mixture of m arbitrarily shaped particles,

$$Q(N_1, \dots, N_m) = \prod_{i=1}^m \frac{1}{N_i!} \left(\frac{Q_i(V)}{\mathcal{Z}_i(V)} \right)^{N_i} \mathcal{Z}(N_1, \dots, N_m) \quad (3)$$

V is the volume and N_i is the number of molecules of species i containing n_i atoms each. $\mathcal{Z}(N_1, \dots, N_m)$ is the corresponding configurational partition function,

$$\mathcal{Z}(N_1, \dots, N_m) = \int_{V^{|X|}} d\mathbf{X} e^{-\beta U(\mathbf{X})} \quad (4)$$

where the potential energy $U(\mathbf{X})$ depends on the set \mathbf{X} of all $|\mathbf{X}| = \sum_i N_i n_i$ atom positions.

We introduce $\mathcal{Z}_i(V)$ for the single-particle canonical partition function, e.g., $\mathcal{Z}_2 = \mathcal{Z}(0, 1, 0, \dots, 0)$. For spherically symmetric particles $\mathcal{Z}_i(V) = V$ and we recover McQuarrie's expression⁷ for $Q(N_1, \dots, N_m)$. For rigid cylindrically symmetric and asymmetric particles $\mathcal{Z}_i(V) = V4\pi$ and $\mathcal{Z}_i(V) = V8\pi^2$, respectively. Note that in the following we use ' \mathcal{Z} ' instead of ' Z ' for these expressions for rigid molecules to distinguish them from the full configurational partition function written as calligraphic ' \mathcal{Z} '. We obtain for the second osmotic virial coefficients

$$B_{ij} = -\frac{V}{2\mathcal{Z}_i\mathcal{Z}_j} [\mathcal{Z}_{ij} - \mathcal{Z}_i\mathcal{Z}_j] \quad (5)$$

where we introduced \mathcal{Z}_{ij} for the two-particle partition function, e.g., $\mathcal{Z}_{12} = \mathcal{Z}(1, 1, 0, \dots, 0)$ for a pair of particles of species 1 and 2 or $\mathcal{Z}_{11} = \mathcal{Z}(2, 0, 0, \dots, 0)$ for a pair of particles of species 1.

2.2 Estimating the dissociation constant

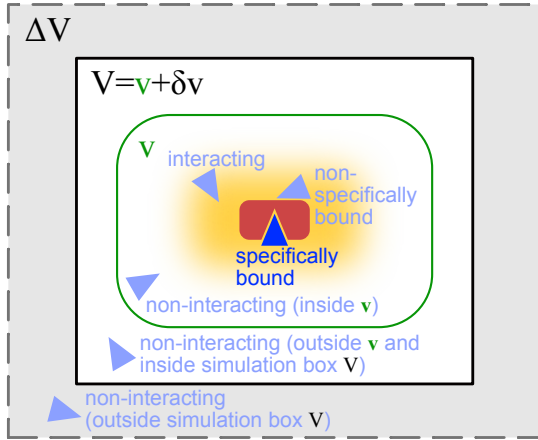


Figure 1: Calculating the dissociation constant K_d and the second osmotic virial coefficient B_{ij} from simulations of two proteins (red and blue) in a box of volume V . The red protein has a single specific wedge-shaped binding site for the triangular blue protein. The light-blue protein configurations illustrate different interaction modes of the two proteins considered in the derivation of eqs 1 and 26. To obtain a K_d estimate independent of box size, we analytically extend the two-particle partition function for the simulation box by the contributions of an extension volume ΔV (gray shaded area) and perform the limit $\Delta V \rightarrow \infty$. We can calculate B_{ij} from the probability $p_v(V)$ that the two proteins are within a sub-volume v (green), which is large enough to cover all protein-protein interactions (yellow shaded area).

To overcome this finite-size effect, we effectively extend the box volume analytically and calculate K_d in the limit of infinite volume (see Figure 1). For asymmetric proteins, we remove the translational and rotational degrees of freedom of the protein of species i , which correspond to a factor $Z_i(V) = 8\pi^2 V$ in the partition function. That is, we fix the position and orientation of the protein of species i . The corresponding two-particle partition function is given by the full two-particle partition function $Z_{ij}(V)$ divided by $Z_i(V)$,

$$\mathcal{Z}(V) = \frac{Z_{ij}(V)}{Z_i(V)} \quad (6)$$

We show how to obtain a box size independent estimate of the dissociation constant K_d from simulations of two proteins in a box. For large enough box-volumes V one would be tempted to estimate the dissociation constant of two proteins A and B directly from the binding probability $p_b(V)$ using the concentrations of free proteins $[A] = [B] = (1 - p_b(V))/(N_A V)$ and the concentration of bound protein $[AB] = p_b(V)/(N_A V)$. The dissociation constant is then given by

$$K_d = \frac{[A][B]}{[AB]} = \frac{(1 - p_b(V))^2}{N_A V p_b(V)} \quad (7)$$

For small box sizes typically used in simulations this estimate suffers from finite-size effects. Accurate estimates using eq 7 would require unusually large boxes, which makes sampling highly inefficient.

We extend this system with fixed position and orientation of particle i by an additional volume ΔV accessible to the second particle. The contribution to the partition function of a protein of species j being in this additional volume ΔV is given by

$$\Delta \mathcal{Z}(\Delta V) = Z_j(\Delta V) \tilde{Z}_i \tilde{Z}_j \quad (8)$$

where $Z_j(\Delta V) = 8\pi^2 \Delta V$ gives the contribution due to the translational and rotational degrees of freedom of an asymmetric protein to the partition function. \tilde{Z}_i and \tilde{Z}_j are the partition functions of individual proteins i and j , whose positions and orientations are fixed in space. That is, \tilde{Z}_i and \tilde{Z}_j contain only contributions due to the respective internal degrees of freedom of free proteins and of solvent degrees in the vicinity of the proteins, which differ from bulk due to the presence of the protein. For rigid protein models in implicit solvent, $\tilde{Z}_i = \tilde{Z}_j = 1$.

The probability $p_b(V_{\text{ex}})$ that the two proteins are bound in the extended volume $V_{\text{ex}} = V + \Delta V$ is now given by the ratio of the partition function $\mathcal{Z}^{(b)}$ of the bound proteins to the partition function of the extended system $\mathcal{Z}(V) + \Delta \mathcal{Z}(\Delta V)$. With the position of particle i fixed, $\mathcal{Z}^{(b)}$ is independent of the size of the volume V and thus the same for the sim-

ulation box and for the extended system, i.e., $\mathcal{Z}^{(b)} = \mathcal{Z}(V)p_b(V)$. Consequently,

$$p_b(V_{\text{ex}}) = \frac{\mathcal{Z}(V)p_b(V)}{\mathcal{Z}(V) + Z_j(\Delta V)\tilde{Z}_i\tilde{Z}_j} \quad (9)$$

To calculate a K_d value unaffected by the finite size of the simulation box, we now substitute eq 9 into eq 7. We then take the limit $\Delta V \rightarrow \infty$ and use that $Z_j(V)/V = 8\pi^2$ to obtain

$$K_d = \frac{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j}{N_A V \mathcal{Z}_{ij}(V)p_b(V)} \quad (10)$$

We can rewrite this equation realizing that the partition function of all bound states of the system, where also particle i can move and rotate, is given by $\mathcal{Z}_{ij}^{(b)}(V) = \mathcal{Z}_{ij}(V)p_b(V)$. Note that $\mathcal{Z}_{ij}^{(b)}(V)$ is proportional to V . Equation 10 becomes

$$K_d = \frac{1}{N_A V} \frac{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j}{\mathcal{Z}_{ij}^{(b)}(V)} \quad (11)$$

Expressing $\mathcal{Z}_{ij}(V)$ by the second osmotic virial coefficient defined in eq 5

$$B_{ij} = -\frac{V}{2} \left[\frac{\mathcal{Z}_{ij}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} - 1 \right] \quad (12)$$

and inserting the resulting expression in eq 10 we obtain the relationship between K_d and B_{ij} given in eq 1,

$$K_d = \frac{1}{N_A p_b(V)(V - 2B_{ij})}$$

As a corollary, the volume dependence of the fraction of bound protein

$$p_b(V) = \frac{1}{N_A K_d (V - 2B_{ij})} \quad (13)$$

is parameterized by K_d and B_{ij} .

For strong binding, K_d and B_{ij} are related via the simple expression³⁵

$$K_d \approx -\frac{1}{2N_A B_{ij}} \quad (14)$$

which we derive in the following. To do so, we write $\mathcal{Z}_{ij}(V)$ as a sum of the partition functions

$\mathcal{Z}_{ij}^{(b)}(V)$ for the bound and $\mathcal{Z}_{ij}^{(u)}(V)$ for the unbound states, i.e., $\mathcal{Z}_{ij}(V) = \mathcal{Z}_{ij}^{(b)}(V) + \mathcal{Z}_{ij}^{(u)}(V)$, and insert this expression in eq 12. We then obtain

$$B_{ij} = -\frac{V}{2} \left[\frac{\mathcal{Z}_{ij}^{(b)}(V) + \mathcal{Z}_{ij}^{(u)}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} - 1 \right] \quad (15)$$

For strongly binding proteins, whose unbound interactions are either attractive, but comparably weak and short-ranged, or repulsive,

$$\frac{\mathcal{Z}_{ij}^{(u)}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} - 1 \ll \frac{\mathcal{Z}_{ij}^{(b)}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} \quad (16)$$

such that

$$B_{ij} \approx -\frac{1}{2} \frac{V \mathcal{Z}_{ij}^{(b)}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} = -\frac{1}{2N_A K_d} \quad (17)$$

where we used eq 11. Rearranging this equation we arrive at eq 14. We now can insert this expression into eq 13 and obtain

$$p_b(V) \approx \frac{1}{1 + N_A K_d V} \quad (18)$$

from which we can express K_d for strong binders as

$$K_d \approx \frac{1}{N_A V} \frac{p_u(V)}{p_b(V)} \quad (19)$$

Here, we introduced the fraction of unbound protein configurations as $p_u(V) = 1 - p_b(V)$.

2.3 Estimating the second osmotic virial coefficient

As we have shown above, we have to estimate B_{ij} to estimate K_d . To do so, we apply the same concepts as we have used for the calculation of K_d . We first remove the translational and rotational freedom of the system by keeping the position and the orientation of protein i fixed (eq 6). Around this protein we define a sub-volume $v < V$ which has to be big enough such that it captures all protein-protein interactions (see Figure 1). Outside this sub-volume, protein-protein interactions can be neglected.

That is, protein j moves freely when it is in volume $\delta v = V - v$.

The probability $p_v(V)$ that protein j is in sub-volume v is given by

$$p_v(V) = \frac{\mathcal{Z}(v)}{\mathcal{Z}(v) + \tilde{\mathcal{Z}}_i \tilde{\mathcal{Z}}_j Z_j(\delta v)} \quad (20)$$

where $\mathcal{Z}(v)$ is given analogous to eq 6 and $Z_j(\delta v) = 8\pi^2 \delta v$ for asymmetric proteins. We can express $\mathcal{Z}(v)$ from eq 20 as

$$\mathcal{Z}(v) = \frac{p_v(V) Z_j(\delta v) \tilde{\mathcal{Z}}_i \tilde{\mathcal{Z}}_j}{1 - p_v(V)} \quad (21)$$

From eq 6 it follows that

$$\mathcal{Z}_{ij}(v) = \frac{p_v(V) Z_j(\delta v) Z_i(v) \tilde{\mathcal{Z}}_i \tilde{\mathcal{Z}}_j}{1 - p_v(V)} \quad (22)$$

$1 - p_v(V)$ is the probability that protein j is in volume δv . Consequently,

$$B_{ij} = -\frac{v}{2} \left[\frac{p_v(V)}{1 - p_v(V)} \frac{Z_j(\delta v)}{Z_j(v)} - 1 \right] \quad (23)$$

Using that $Z_j(v)$ and $Z_j(\delta v)$ are proportional to their argument with the same prefactor (see section 2.1) and that $\delta v = V - v$, where V is the box volume, we obtain

$$B_{ij} = -\frac{v}{2} \left[\frac{p_v(V)}{1 - p_v(V)} \frac{V - v}{v} - 1 \right] \quad (24)$$

Solving for $p_v(V)$, we obtain

$$p_v(V) = \frac{v - 2B_{ij}}{V - 2B_{ij}} \quad (25)$$

which describes the dependence of $p_v(V)$ on the box volume V and the sub-volume v .

We emphasize that eq 24 is generally valid for arbitrary binding partners, without making any assumptions about symmetry or the number of internal degrees of freedom of the binding partners or of the solvent. The only condition is that interactions between binding partners are negligible outside the volume v . We can introduce correction terms based on an effective pair-wise potential acting between the binding partners if this condition is not fulfilled

(see subsection 4.1).

To motivate the interpretation of eq 24, we rewrite it as

$$B_{ij} = -\frac{V}{2} \left[\frac{1 - \frac{v}{V}}{1 - p_v(V)} - 1 \right] \quad (26)$$

Note that the prefactor of eq 26 contains the box volume V whereas the prefactor in eq 24 contains the sub-volume v . This term, determining the two-particle partition function, is the ratio of the probability of finding one particle outside the sub-volume v for the ideal system, $1 - v/V$, to the corresponding probability for the interacting particles, $1 - p_v(V)$. This ratio is independent of the sub-volume v , which has to be just large enough to cover the interaction range. Consequently, the first term in the brackets in eq 26 can be written as $\exp[-\beta F_o^{(\text{ex})}(V)]$ where we introduced the excess free energy of finding the two particles outside their interaction range in the box of volume V as

$$F_o^{(\text{ex})}(V) = -kT \ln \frac{1 - \frac{v}{V}}{1 - p_v(V)} \quad (27)$$

We express K_d as a function of $p_v(V)$ by inserting eq 26 into eq 1 and obtain

$$\begin{aligned} K_d &= \frac{1}{V N_A p_b(V)} \frac{1 - p_v(V)}{1 - \frac{v}{V}} \\ &= \frac{1}{V N_A p_b(V)} e^{-\beta F_o^{(\text{ex})}(V)} \end{aligned} \quad (28)$$

For strong binders, $\exp[-\beta F_o^{(\text{ex})}(V)]$ corresponds to the probability of unbinding as we establish in the following.

We now derive that for strong binders eq 28 can be approximated by eq 19, i.e., that $p_u(V) \approx (1 - p_v(V))/(1 - v/V)$. Proteins are either bound with probability $p_b(V)$ or unbound with probability $p_u(V) = 1 - p_b(V)$. For strong binders, we now assume that the interactions of non-binding states can be neglected. We introduce $v = v' + v_{\text{exc}}$, where v_{exc} is the excluded volume of the proteins for the unbound states. That is, v' is the part of the sub-volume v accessibly to protein j without binding to protein j . Thus, a fraction v'/V' , where

$V' = V - v_{\text{exc}}$, of the unbound states is within volume v and a fraction $1 - v'/V'$ is outside v , i.e., $p_u(V) = p_u(V)v'/V' + p_u(V)(1 - v'/V')$. Within this approximation, $p_v(V)$ is the sum of the fraction of bound states and the fraction of unbound states that are within v , i.e., $p_v(V) = p_b(V) + p_u(V)v'/V'$. Consequently, we obtain

$$p_u(V) = \frac{1 - p_v(V)}{1 - \frac{v}{V}} \frac{1 - \frac{v'}{V'}}{1 - \frac{v'}{V'}} \quad (29)$$

For strong binders, $p_u(V)$ is dominated by the first term such that $p_u(V) \approx (1 - p_v(V))/(1 - v/V)$ and eq 28 can be approximated by eq 19.

We next establish the commonly used relationship of B_{ij} to the partial radial distribution function $g(r)$.⁷ The ratio of $p_v(V)/(1 - p_v(V))$ can be estimated from the probability density of center-of-mass distances $p(r)$ of two particles in a box, for example, which is itself related to the radial distribution function $g(r)$. To do so, we define a spherical volume $v = 4\pi R^3/3$ and a spherical shell around this sphere with volume $\delta v = 4\pi[(\delta R + R)^3 - R^3]/3$. The ratio is then given by

$$\frac{p_v(V)}{1 - p_v(V)} = \frac{\int_0^R p(r)dr}{\int_R^{R+\delta R} p(r)dr} \quad (30)$$

We define a radial distribution function $g(r)$ through

$$p(r) \propto 4\pi r^2 g(r) \quad (31)$$

We choose the proportionality constant such that $g(r) = 1$ for $r > R$ where $p(r) \propto r^2$. Using that $4\pi \int_R^{R+\delta R} g(r)r^2 dr = \delta v$, we may write

$$\frac{p_v(V)}{1 - p_v(V)} = \frac{1}{\delta v} 4\pi \int_0^R g(r)r^2 dr \quad (32)$$

Inserting this expression in eq 24 and using that $Z_j(\delta v)/Z_j(v) = \delta v/v = (V - v)/v$ for proteins i and j of the same symmetry and that $\int_0^R 4\pi r^2 dr = v$, we obtain

$$B_{ij} = -2\pi \int_0^R [g(r) - 1]r^2 dr \quad (33)$$

By introducing an effective interaction potential $\beta w(r) = -\ln g(r)$ we can write eq 33 as it is commonly presented

$$B_{ij} = -2\pi \int_0^R [e^{-\beta w(r)} - 1] r^2 dr \quad (34)$$

where $\beta = 1/(k_B T)$.

Using eq 26 instead of eq 33 or eq 34 we avoid the computation of distance distribution functions and potentials of mean force, respectively, and the subsequent integration. Although this difference might be viewed only as a minor simplification, eq 26 emphasizes that B_{ij} is independent of the detailed shapes of $g(r)$ and $w(r)$ and determined by the excess free energy $F_o^{(\text{ex})}(V)$ of finding the two particles outside of their interaction range. Note that our results also apply to the Kirkwood-Buff integrals $G_{ij} = 4\pi \int_{r=0}^{\infty} [g(r) - 1] r^2 dr = 2B_{ij}$.^{13,36,37}

2.4 The contribution of non-binding interactions to B_{ij}

We can use K_d and B_{ij} to quantify the non-binding interactions of two proteins. Let us first consider two non-binding proteins for which $Z_{ij}^{(b)}(V) = 0$. Consequently, eq 15 becomes

$$B_{ij}^{(u)} = -\frac{V}{2} \left[\frac{Z_{ij}^{(u)}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} - 1 \right] \quad (35)$$

where we use the superscript ‘(u)’ to indicate contributions of the unbound states. For binding proteins $B_{ij}^{(u)}$ is given by difference between $B_{ij} = B_{ij}^{(u)} + B_{ij}^{(b)}$ and the contributions to $B_{ij}^{(b)}$ due to binding,

$$B_{ij}^{(b)} = -\frac{V}{2} \frac{Z_{ij}^{(b)}(V)}{Z_i(V)Z_j(V)\tilde{Z}_i\tilde{Z}_j} = -\frac{1}{2N_A K_d} \quad (36)$$

i.e., we can quantify the non-binding interactions for two binding proteins via

$$B_{ij}^{(u)} = B_{ij} - B_{ij}^{(b)} = B_{ij} + \frac{1}{2N_A K_d} \quad (37)$$

which becomes

$$B_{ij}^{(u)} = -\frac{V}{2} \left[\frac{1 - \frac{v}{V}}{1 - p_v(V)} p_u(V) - 1 \right] \quad (38)$$

As we have shown above, for proteins that do not have non-binding interactions $p_u(V)$ is given by eq 29 such that $B_{ij}^{(u)} = v_{\text{exc}}/2$. For attractive non-binding interactions $B_{ij}^{(u)} < v_{\text{exc}}/2$ and for repulsive non-binding interactions $B_{ij}^{(u)} > v_{\text{exc}}/2$. Note that for asymmetric particles the excluded volume v_{exc} is an excluded region in configuration space, which is spanned by Cartesian coordinates and Euler angles for rigid proteins, for example. Thus, in general v_{exc} should be viewed as an effective volume defining the probability v'/V' , which itself corresponds to a thermodynamic free energy.

2.5 Indistinguishable binding partners (homodimers)

So far, we have assumed that the proteins are distinguishable, i.e., that they form heterodimers, but all expressions derived here are also valid for indistinguishable binding partner forming homodimers. To consider the case of two identical binding partners, we rewrite eq 11 as

$$K_d = \frac{1}{N_A V} \frac{\mathcal{Z}_{ij}^{(\text{free})}(V)}{\mathcal{Z}_{ij}^{(\text{b})}(V)} \quad (39)$$

where we introduced $\mathcal{Z}_{ij}^{(\text{free})}(V)$ for the partition function of two free proteins, which is determined by the product of two single-protein partition function. For indistinguishable binding partners forming homodimers, both $\mathcal{Z}_{ij}^{(\text{free})}(V)$ and $\mathcal{Z}_{ij}^{(\text{b})}(V)$ would have to be multiplied by a factor 1/2 to account for the indistinguishability of the proteins. However these factors then cancel in the ratio in eq 39.

2.6 K_d and B_{ij} from a single simulation

We can estimate K_d and B_{ij} from the fraction of bound protein p_b and the probability $p_v(V)$ of one protein being located in a sub-volume

v around the other. The latter determines B_{ij} according to eq 26, which we then insert into eq 1 to obtain the finite-size corrected estimate of K_d . We call this method the sub-volume method. To calculate B_{ij} , we can also estimate the two-particle partition function $Z_{ij}(V)$, now for simplicity only considering rigid molecules, using free energy methods.³⁸ For implicit solvent, we can use insertion and removal moves of the proteins to efficiently estimate $Z_{ij}(V)$, as explained in the following. We call this method the insertion/removal method.

2.6.1 Estimating two-particle configurational partition functions for implicit solvent

A simulation of a pair of proteins in a box of volume V at reciprocal temperature β gives us immediately the particle-removal energy distribution as the normalized distribution of potential energies. We define $x_i = (\mathbf{r}_i, \Omega_i)$, where \mathbf{r}_i are the Cartesian coordinates of the geometric center of protein i , and Ω_i its Euler angles defining its orientation. We denote the configuration space as $W = V \times \Omega$ to simplify the notation. The particle removal energy distribution is then given by

$$p_{\text{rem}}(E) = \frac{\int_{W^2} dx_2 dx_3 e^{-\beta U(x_2, x_3)} \delta[E - U(x_2, x_3)]}{Z_{23}(\beta)} \quad (40)$$

where $Z_{23}(\beta) = \int_{W^2} dx_2 dx_3 e^{-\beta U(x_2, x_3)}$ and $\delta(\cdot)$ is Dirac's delta function.

The particle-insertion energy distribution $p_{\text{ins}}(E)$ is formally given by

$$p_{\text{ins}}(E) = \frac{\int_{W^2} dx_2 dx_3 \delta[E - U(x_2, x_3)]}{Z_2(\beta=0) Z_3(\beta=0)} \quad (41)$$

where $Z_i(\beta=0) = \int_W dx_i$. Sampling the particle-insertion energy distribution $p_{\text{ins}}(E)$ for a given box-size is straightforward. All one needs is a replica with reciprocal temperature $\beta=0$ exactly. All moves will then be accepted, and the energies saved are those of random insertions. Alternatively, one could make trial moves of the two proteins with MC move widths

$\pm L/2$, where L is the box length, and orientation changes about random axes by $\pm\pi$, and to write out the absolute trial (!) energies (not the energy differences or the accepted energies). With such a move protocol, it would not matter if one or both particles were moved and if moves are accepted or not. It also does not matter what the “acceptance rate” is (i.e., it can be zero!). What is important, though, is that the box volumes in insertion and removal runs are the same.

The normalized removal and insertion energy distributions are related to each other by

$$p_{\text{ins}}(E) = p_{\text{rem}}(E) \frac{e^{\beta E} Z_{23}(\beta)}{Z_2(\beta=0) Z_3(\beta=0)} \quad (42)$$

which follows from

$$\begin{aligned} p_{\text{ins}}(E) e^{-\beta E} &= \frac{\int_{W^2} dx_2 dx_3 e^{-\beta U(x_2, x_3)} \delta[E - U(x_2, x_3)]}{Z_2(\beta=0) Z_3(\beta=0)} \\ &= p_{\text{rem}}(E) \frac{Z_{23}(\beta)}{Z_2(\beta=0) Z_3(\beta=0)} \end{aligned} \quad (43)$$

The ratio of partition functions defines the free energy of going from a system of two non-interacting particles to a system in which they interact,

$$e^{-\beta F} = \frac{Z_{23}(\beta)}{Z_2(\beta=0) Z_3(\beta=0)} \quad (44)$$

An efficient way of determining this free energy is to use the Bennett acceptance ratio (BAR) estimator,²⁸

$$\sum_{i=1}^{N_{\text{ins}}} \frac{1}{1 + \frac{N_{\text{ins}}}{N_{\text{rem}}} e^{\beta(E_i - F)}} = \sum_{i=1}^{N_{\text{rem}}} \frac{1}{1 + \frac{N_{\text{rem}}}{N_{\text{ins}}} e^{\beta(\underline{E}_i + F)}} \quad (45)$$

where E_i are the uncorrelated (by construction) insertion energies, and \underline{E}_i are the uncorrelated removal energies. However, it is clear that this is problematic in cases where the proteins are strongly bound (forming a dimer!) because then one would have very little information about higher energies.

This problem can be remedied by using all the data in a temperature replica exchange simula-

tion. In effect, the high-temperature runs allow us to estimate an accurate density of states to a pretty high energy. The particle-insertion energies complement this density of states on the high-energy side. All the runs at different temperatures can be combined with the list of insertion energies using binless WHAM. As reference, we take the temperature of interest ($\beta = \beta_1$ without loss of generality). The bias energies at replicas with reciprocal temperature β_i are then $\Delta U = (\beta_i/\beta - 1)U$. This formula works also for the insertion energies coming from a run with $\beta = 0$. The insertion energies can be thought of as coming from a run with bias potential $\Delta U = -U$ i.e., on potential zero. A binless WHAM analysis using these bias energies as input will produce the required free energy F as the difference between the reference state and the insertion run.

3 Methods

We chose three weakly binding protein pairs with experimental K_d values covering three orders in magnitude from $\sim \mu\text{M}$ to $\sim \text{mM}$. The lysozyme homodimer has an experimental K_d value of $K_d \approx 2710 \pm 240 \mu\text{M}$ ³⁹ (PDB 6LYZ⁴⁰), the ubiquitin/CUE dimer (PDB 1OTR⁴¹) has a $K_d \approx 155 \pm 9 \mu\text{M}$,⁴¹ and the dimer of the uracil-DNA glycosylase UDG and its uracil-DNA glycosylase inhibitor protein (Ugi) has a $K_d \approx 1.3 \pm 0.3 \mu\text{M}$ ⁴² (PDB 1UUG⁴³).

To simulate these protein pairs, we use the amino-acid level coarse-grained model developed by Kim and Hummer for weakly binding proteins⁴⁴ implemented in the Complexes++ software (<https://www.github.com/bio-phys/complexespp>). We treat all proteins as rigid bodies. In contrast to the original model, which is called KH-model, we shift the original Miyazawa and Jernigan parameters^{45,46} by $e_0 = -1.875$ to account for the solvation energy and we scale the resulting parameters by $\lambda = 0.1243$ to balance them with the electrostatic interactions. In the original model, $e_0 = -2.27$ and $\lambda = 0.159$. The new values have been chosen to better reproduce the B_{22} value of lysozyme and the K_d value of

the ubiquitin/UIM1 complex. We chose residue charges of $-1.0 e$ for Asp and Glu, $+1.0 e$ for Arg and Lys, and $+0.5 e$ for His due to its isoelectric point at pH 7. e is the elementary charge. Consequently, the total charges of the proteins are $+8.5 e$ for lysozyme, $+0.5 e$ for ubiquitin, $-4.5 e$ for CUE, $+7.5 e$ for UDG, and $-11.5 e$ for Ugi. We set the dielectric constant to 80 and the Debye length to 1 nm, corresponding to the conditions in an aqueous solution of 100 mM NaCl.

To generate Boltzmann ensembles of configurations, which also provide the removal energy distributions defined in eq 40, we perform temperature replica exchange Monte Carlo (REMC) simulations using 24 replicas. Temperatures were equally spaced between 300 K and 530 K. In a Monte Carlo sweep, each protein performs one trial move on average, which can be translation or rotation. Replica exchanges are attempted every 10 sweeps. For the rotation move, a rotation axis is randomly generated by drawing a point from a sphere. Then, we rotate around this axis by an angle, which we draw from a box distribution with a width given by twice the maximum angle. This maximum angle is set to 0.1 rad for the coolest replica and to 1.25 rad for the hottest replica, and spaced equidistantly inbetween. Similarly, we set the maximum displacement for the translation move to 0.2 nm in the coolest replica and to 1.35 nm in the hottest replica, with equal spacing inbetween. In our simulations, we use a cut-off radius of 3 nm to truncate our interaction potentials.

To sample the insertion energy distribution defined in eq 41 in simulations, we switch off all interactions by setting all interaction parameters and residue charges to zero. We use a maximum displacement of half the box length and a maximum rotation angle of π . We accept and sample all configurations to generate the insertion ensembles, for which we then recalculate all energies for switched on potentials.

To estimate the two-particle partition function we combine results from REMC simulations (removal ensemble) and the energies calculated for the ensemble of non-interacting proteins (insertion ensemble) using binless

WHAM.^{29–31} To avoid numerical problems, we clip interaction energies at $100 k_b T$. We define two proteins as being bound if their total interaction energy is below $-2 k_b T$ where $T = 300$ K.

For equilibration, we performed 10^6 Monte Carlo sweeps in each replica. For production, we performed 10^7 sweeps and we sampled every 100th sweep, yielding 10^5 structures for each protein pair per replica. We also performed 10^6 insertion moves for each pair, which by design creates uncorrelated configurations.

To study the box volume dependence of the fraction bound $p_b(V)$, we calculated for the coolest replica $p_b = N_{E \leq -2 k_b T} / N$. $N_{E \leq -2 k_b T}$ is the number of structures with energies $E \leq -2 k_b T$ and $N = 10^5$ is the total number of structures. To study the box volume dependence of the sub-volume probability $p_v(V)$, we calculated for the coolest replica $p_v = N_v / N$, where N_v is the number structure within the sub-volume v . We defined this volume as a spherical volume with a radius given by the sum of $(D_i + D_j)/2$, where D_i and D_j are the largest diameters of proteins of species i and j , and our cut-off radius of 3 nm. The resulting radii are between ~ 6.7 nm and ~ 7.4 nm for the three proteins. For each protein pair, we performed simulations for 17 box sizes with volumes ranging from 3375 nm^3 to 10^6 nm^3 . We calculated the standard errors of the mean by block averaging.^{47,48}

We validate the insertion/removal method and the sub-volume method for the smallest boxes used here with volume $\tilde{V} = 15^3 \text{ nm}^3 = 3375 \text{ nm}^3$. With uniform probability, we select 10 000 of the $N = 10^5$ samples and choose at each replica the configurations corresponding to the same 10 000 indices. We also drew 10 000 configurations of the 10^6 configurations in the insertion ensemble with uniform probability. In the insertion/removal method, we then applied WHAM using these 250 000 configurations total to calculate $p_b(\tilde{V})$ and $\mathcal{Z}_{ij}(\tilde{V})$ from which we then estimated K_d and B_{ij} . We repeated this procedure 1000 times and calculated the averages of K_d and B_{ij} and their covariance matrices. We confirmed visually that the distributions of the estimates of K_d and B_{ij} are

distributed according to two-dimensional Gaussians with the estimated covariance matrices. We use the same protocol to obtain estimates and uncertainties from resampling for the sub-volume method, in which we do not use the insertion ensemble.

4 Results

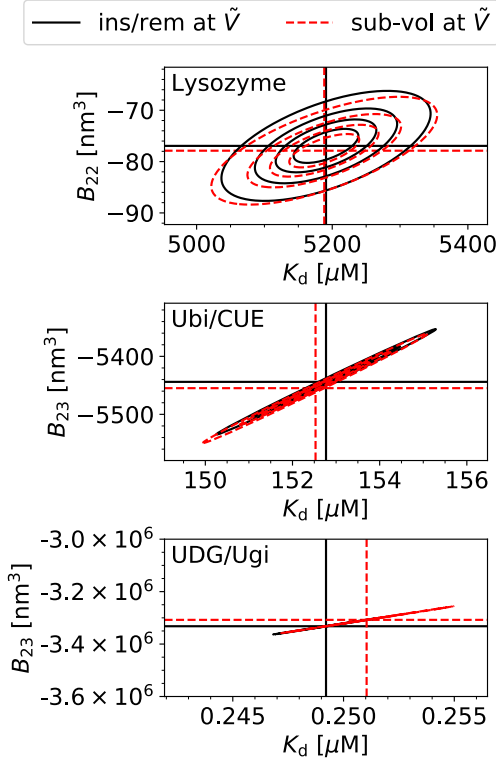


Figure 2: Comparison of the accuracy of the insertion/removal method (‘ins/rem’, black) and the sub-volume method (‘sub-vol’, red) to estimate K_d and B_{ij} for three different protein pairs (top to bottom). The most likely estimates are indicated by horizontal and vertical dashed lines. The contour lines indicate the limits of the 25%, 50%, 75%, and 95% confidence regions. The insertion/removal method (eqs 12 and 1 and the two-particle partition function from WHAM (sec. 2.6.1), black) and the sub-volume method (eqs 24 and 1, red) agree excellently with each other and they have similar uncertainties. For UDG/Ugi, contour lines collapse on to a single line due to the strong correlation between the estimates for K_d and B_{ij} .

We calculated K_d and B_{ij} using the insertion/removal method and the sub-volume method for three protein pairs, i.e., the lysozyme homodimer and the heterodimers ubiquitin/CUE and UDG/Ugi. As we will show, the high accuracy of these estimates allows us to quantify the contributions due to binding and non-binding interactions to B_{ij} .

In the insertion/removal method, we determine K_d and B_{ij} from replica exchange simulations at a box volume \tilde{V} and from insertion ensembles. We first estimated $p_b(\tilde{V})$ and $Z_{ij}(\tilde{V})$ by combining the insertion ensemble and the replicas of our temperature REMC simulations using WHAM. We then evaluated eq 12 to obtain B_{ij} and used this value together with our estimate for $p_b(\tilde{V})$ in eq 1 to obtain K_d . By resampling we estimated the covariance matrix.

In the sub-volume method, we first estimated $p_v(\tilde{V})$ and $p_b(\tilde{V})$ from all replicas using WHAM. We used eq 24 to calculate B_{ij} from $p_v(\tilde{V})$ and used this estimate together with $p_b(\tilde{V})$ to estimate K_d using eq 1.

We find that the estimates for K_d and B_{ij} from the insertion/removal method and the sub-volume method agree excellently with each other (Figure 2 and Table 1). Moreover, the estimates have similar uncertainties. K_d values and B_{ij} values calculated by resampling are correlated for both methods (Fig. 2). A smaller value of B_{ij} , i.e., a more negative value, leads to a smaller value of K_d according to eq 1. For strong binders, the correlation dominates and K_d and B_{ij} fulfill eq 14 to a good approximation.

For additional validation, we use the results for K_d and B_{ij} obtained at the box volume \tilde{V} to predict the box size dependence of the fraction bound $p_b(V)$ and the sub-volume probability $p_v(V)$. We use eq 13 and our estimates for K_d and B_{ij} obtained at a box volume \tilde{V} to calculate $p_b(V)$ (Figure 3). We use eq 25 and our estimates for B_{ij} obtained at a box volume \tilde{V} to calculate $p_v(V)$ (Figure 4). The resulting curves reproduce the box volume dependencies of $p_b(V)$ and $p_v(V)$ observed in the entire range of simulations, covering nearly three orders of magnitude in volume.

For strong binders, the fraction bound $p_b(V)$

Table 1: K_d , B_{ij} , and the contributions of binding interactions, $B_{ij}^{(b)}$, and non-binding interactions, $B_{ij}^{(u)}$, to B_{ij} for three protein complexes (PDB codes 6LYZ, 1OTR, 1UUG) for the insertion/removal method (‘ins/rem’) and the sub-volume method (‘sub-vol’). Errors are standard errors of the mean.

Lysozyme	Method	K_d [μM]	B_{22} [nm^3]	$B_{22}^{(b)}$ [nm^3]	$B_{22}^{(u)}$ [nm^3]
	ins/rem	5191 ± 63	-77 ± 4	-160 ± 2	83 ± 4
	sub-vol	5188 ± 68	-78 ± 4	-160 ± 2	82 ± 3
Ubi/CUE	Method	K_d [μM]	B_{23} [nm^3]	$B_{23}^{(b)}$ [nm^3]	$B_{23}^{(u)}$ [nm^3]
	ins/rem	153 ± 1	-5444 ± 37	-5435 ± 37	-9 ± 3
	sub-vol	153 ± 1	-5455 ± 39	-5444 ± 38	-11 ± 3
UDG/Ugi	Method	K_d [μM]	B_{23} [nm^3]	$B_{23}^{(b)}$ [nm^3]	$B_{23}^{(u)}$ [nm^3]
	ins/rem	0.25 ± 0.002	-3332000 ± 27000	-3332000 ± 27000	-94 ± 5
	sub-vol	0.25 ± 0.002	-3308000 ± 27000	-3308000 ± 27000	-81 ± 7

and the sub-volume probability $p_v(V)$ take on similar values (compare Figures 3 and 4). In these cases $p_v(V)$ is dominated by binding. For small boxes, $p_b(V)$ is close to one and consequently so is $p_v(V)$. For box sizes large enough such that p_b is significantly below one, the contribution of the size of the sub-volume v to p_v is small. For UDG/Ugi, the strongest binding complex considered here, the fraction bound dominates $p_v(V)$ such that the $p_v(V)$ curve in Figure 4 looks nearly identical to the corresponding $p_b(V)$ curve in Figure 3. However, the differences in these curves are significant as they are not only determined by the size of the sub-volume v but also by the non-binding interactions.

We can extract the contributions $B_{ij}^{(u)}$, eq 37, of non-binding interactions to B_{ij} . We can do so even in the case of strong binders for which the K_d value is close to $B_{ij}^{(b)} = -1/(2N_A K_d)$ according to eq 14 (Figure 5, top). Because of the high accuracy of our estimates, either provided by the insertion/removal method or the sub-volume method, we can resolve the small difference $B_{ij}^{(u)} = B_{ij} - B_{ij}^{(b)}$ (Figure 5, center). Focusing on the results from the insertion/removal methods, we find that for lysozyme $B_{ij}^{(u)} \approx 83 \pm 4 \text{ nm}^3 > 0$. This value is close to what one would expect for hard spheres of equal volume, i.e., $B_{ij}^{(u)} = v_{\text{exc}}/2 \approx 70 \text{ nm}^3$. For ubiquitin/CUE, the interactions are clearly attractive but $B_{ij}^{(u)} \approx -9 \pm 3 \text{ nm}^3$ nearly vanishes. For

UDG/Ugi, $B_{ij}^{(u)} \approx -94 \pm 5 \text{ nm}^3$ indicates strong attractive interactions (Figure 5, and Table 1).

Note that for Ubi/CUE and UDG/Ugi the estimates for $B_{23}^{(u)} = B_{23} - B_{23}^{(b)}$ are much smaller than the individual errors of B_{23} and $B_{23}^{(b)}$ ($\sim 27000 \text{ nm}^3$ for UDG/Ugi and $\sim 40 \text{ nm}^3$ for Ubi/CUE, Table 1). Naively, one would think that these large uncertainties preclude reliable estimates for the comparably small difference $B_{23}^{(u)}$ in such a situation. However, the estimates for B_{23} and $B_{23}^{(b)}$ from resampling are highly correlated because of the strong correlation of B_{23} and K_d (Figure 2). That is, the individual errors of B_{23} and $B_{23}^{(b)}$ do not determine the errors of their difference.

Finally, we show that the naïve estimate of K_d from concentrations using eq 7 actually suffers from a finite-size effect and that it converges to the estimates obtained with the insertion/removal and sub-volume methods for large systems sizes (Figure 6). For comparison only, we evaluate eq 7 for our predictions of $p_b(V)$ obtained at a volume \tilde{V} (Figure 3) and extrapolate the naïve estimates for K_d until convergence is reached. For typical box sizes used in simulations, K_d is underestimated by about 10% for the lysozyme homodimer, the weakest binder considered here, and by three orders of magnitude for UDB/Ugi, the strongest binder considered here. To reach convergence when using eq 7, the box volumes have to be increased by a factor ~ 10 for the weakest binder and by a fac-

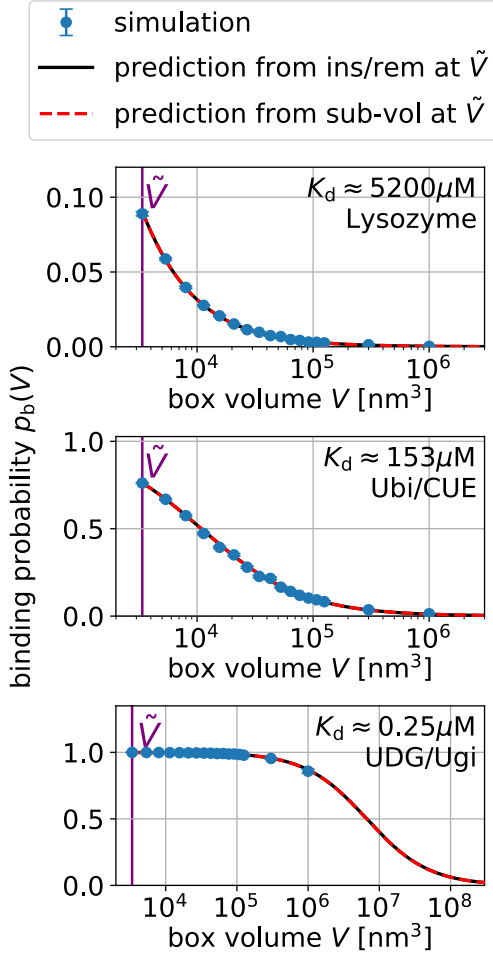


Figure 3: The box size dependence of the binding probability $p_b(V)$ is determined by B_{ij} and K_d via eq 13. We show simulation results (blue) for three protein pairs (top to bottom). Error-bars indicate the blocked standard errors of the mean. The lines are predictions using eq 13 and estimates for K_d and B_{ij} obtained at a box volume $\tilde{V} = 3375 \text{ nm}^3$ (magenta vertical line) using the insertion/removal method (black) and the sub-volume method (red).

tor $\sim 10\,000$ for the strongest binder compared to typical box sizes.

4.1 Practical considerations

In the derivation of K_d and B_{ij} we have assumed that the volume is large enough such that interactions between the protein with fixed position and orientation and the protein in the extended volume can be neglected. If this condition is not fulfilled then we can correct

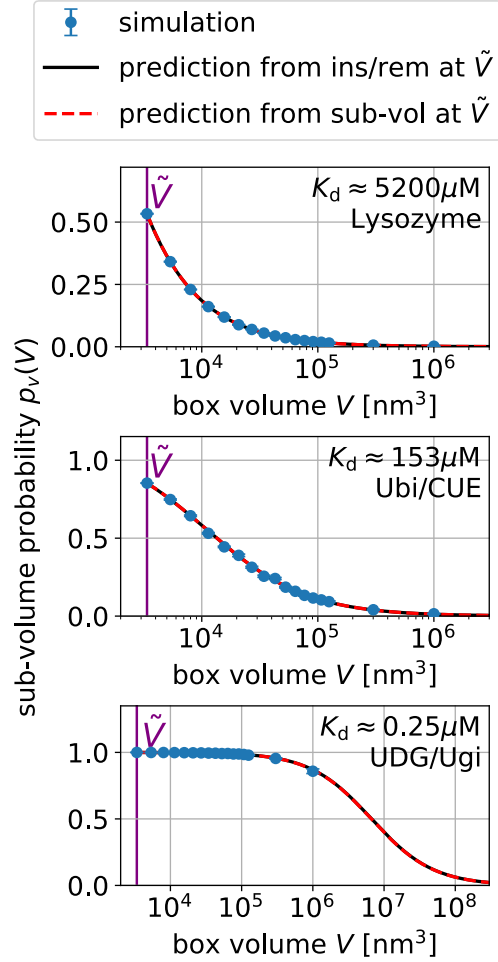


Figure 4: The box size dependence of the sub-volume probability $p_v(V)$ is determined by B_{ij} via eq 25. We show simulation results (blue) for three protein pairs (top to bottom). Error-bars indicate the blocked standard errors of the mean. The lines are predictions using eq 25 and estimates of B_{ij} obtained at a box volume $\tilde{V} = 3375 \text{ nm}^3$ (magenta vertical line) using the insertion/removal method (black) and the sub-volume method (red).

for residual interaction energies using a simple distance dependent interaction potential $\phi(r)$ in the calculation of $Z_j(\Delta V) = 8\pi^2 \times \int_{\mathcal{C}} d\mathbf{r} \exp[-\beta\phi(r)]$, where \mathcal{C} denotes the Cartesian space defining ΔV . For example, at large distances the interaction of charged proteins can be approximated by (screened) Coulomb interactions of the total charges located at the centers of charge. In such a case we would include for the calculation of the fraction bound only configurations of the simulation where the

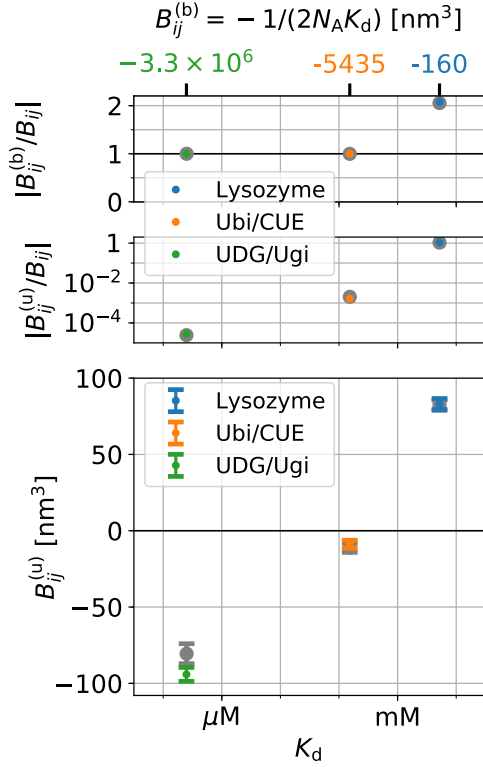


Figure 5: The contributions of binding and non-binding interactions to $B_{ij} = B_{ij}^{(b)} + B_{ij}^{(u)}$ for three protein pairs. We show estimates from the insertion/removal method in color and estimates from the sub-volume method using larger symbols in grey. B_{ij} of the strongest binders is dominated by contributions of binding $B_{ij}^{(b)} = -1/(2N_A K_d)$ such that the ratio of $|B_{ij}^{(b)}|/|B_{ij}|$ is close to one (top). In these cases, non-binding contributions to B_{ij} are relatively small, i.e., $|B_{ij}^{(u)}|/|B_{ij}| \ll 1$ (center).

two proteins are separated less than a cutoff distance, usually given by half the shortest box length. Such a system corresponds to a spherical volume with one particle at its center and the other one moving unrestrained. Doing so, we assume that the residual interaction modeled as a simple pair-potential has a negligible effect on the internal degrees of freedom and the degrees of freedom of the surrounding solvent, i.e., \tilde{Z}_i and \tilde{Z}_j are unchanged.

Suitable definitions of bound states will depend on the molecular model we use for simulations. In our simulations using implicit solvent, we count a state as bound if the interaction energy of the two proteins is smaller than $-2k_b T$.

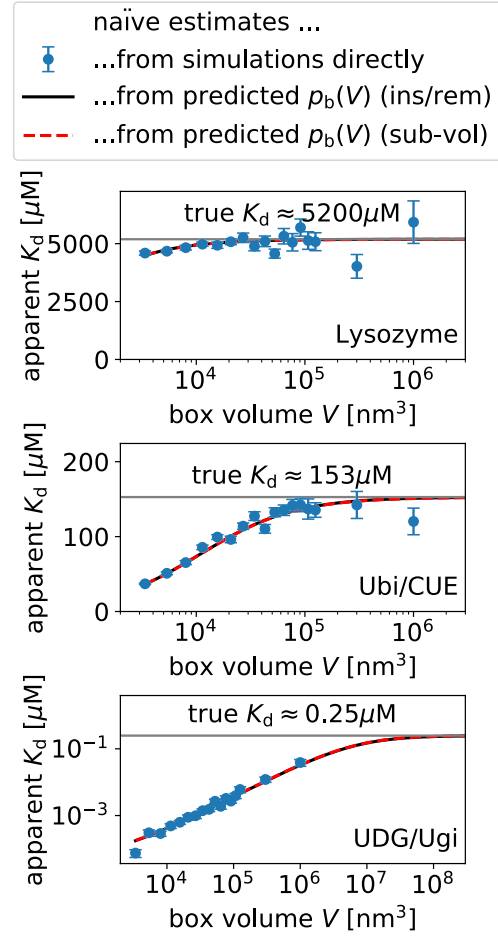


Figure 6: The naïve estimate of K_d from concentrations using eq 7 suffers from finite-size effects and converges to the true values for increasing box size. Blue symbols indicate simulation results with errorbars given by the blocked standard error of the mean. We evaluate eq 7 also for the predictions for $p_b(V)$ from the insertion/removal method (black solid line) and the sub-volume method (red dashed line) to illustrate the convergence to the true values obtained with these methods (gray horizontal line).

Additionally demanding that two proteins have to have a minimum distance smaller than 0.8 nm to be counted as bound does not have a noticeable effect on the binding probability. For molecular dynamics simulations using explicit solvent, a combination of a distance based criterion and a contact time based criterion should be sufficient to distinguish bound states from spurious contacts.

In simulations of two proteins in a box, we

can estimate $p_v(V)$ using a distance-based criterion, for example, as explained in the following. We define a distance between the two proteins, e.g., the center-of-mass distance r . We introduce a maximum distance R such that interactions between proteins are negligible for distances $r > R$. For an ensemble of N structures we count the number of structures N_v for which $r \leq R$. In these structures, the center-of-mass of protein 2 lies within a spherical volume $v = 4\pi R^3/3$ centered at the center-of-mass of protein 1. We then estimate $p_v(V) = N_v/N$.

For strong binders and in boxes of typical size, $p_v(V)$ is close to one. For $p_v(V) = 1$, $(1 - v/V)/(1 - p_v(V))$ diverges. Consequently, $p_v(V)$ has to be determined with sufficient numerical precision to obtain accurate estimates. For example, if we sample 10000 configurations than the numerical precision of $p_v(V)$ is limited to 1/10 000. The precision can be increased by sampling more configurations or, in the case of replica simulation, by including additional replicas using WHAM when calculating $p_v(V)$. For weak binders with $K_d \lesssim 100 \mu\text{M}$, 10000 configurations are sufficient to estimate K_d and B_{ij} without even applying WHAM.

5 Conclusions

We have shown how to calculate the dissociation constant K_d of two proteins in a box from the fraction of bound protein and the second osmotic virial coefficient B_{ij} . We derived and validated two methods to calculate B_{ij} : For implicit solvent, we can use standard Monte Carlo or molecular dynamics simulations of two proteins in a box and determine insertion and removal energy distribution functions. From the latter, we determine the two-particle partition function and thus B_{ij} using BAR/WHAM. For implicit and explicit solvent, we can calculate the probability that the two proteins are within a volume at least covering the interaction range of the two proteins. As we have shown, calculating B_{ij} from the radial distribution function or equally the potential of mean force via an integral is equivalent to this method. For the coarse-grained simulations performed here,

both of our methods provide accurate results with comparable uncertainties.

The relationship between K_d and B_{ij} given by eq 1 is also well suited for the quantification of protein interactions in molecular dynamics simulations using explicit solvent. Fully atomistic simulations of concentrated protein solutions in explicit solvent have become computationally feasible on the microsecond scale.^{15,16} These studies have been facilitated by recent improvements in molecular force fields, which correct, among other things, for an increased stickiness of protein surfaces.^{49–51} These parameterization efforts can benefit from comparisons of K_d and B_{ij} to experiment.

Fully atomistic simulations are coming within reach for the protein pairs considered here. The box volume \tilde{V} corresponds to about 300 000 particles in fully atomistic simulations using explicit solvent. The binding and unbinding of weakly binding proteins like lysozyme can be simulated atomistically without bias.¹⁵ For more strongly binding proteins, enhanced sampling techniques have to be applied. Binding and unbinding events of proteins and other molecules can be simulated efficiently without bias also in molecular dynamics simulations using explicit solvent using the MARTINI model, for example.^{52–54}

The sampling strategy used here for weak binders is different from the sampling strategy commonly used for strong binders. Strong binders usually have specific interfaces and the dissociation constant is determined by the binding free energy to these specific interfaces. If these interfaces are known then we only have to calculate the binding free energy for these specific binding poses dominating K_d .¹⁷ For weak binders, also non-specific binding can contribute significantly to K_d and thus has to be sampled.

B_{ij} plays also an important role in understanding phase separation by which liquid droplets are formed within cells.⁵⁵ Specifically, the Flory-Huggins solution theory is used to model liquid-liquid phase separations.^{56,57} In this framework, the Flory interaction parameter χ is determined by K_d and B_{ij} .⁵⁸ B_{ij} also determines the ‘effective solvation volume’ up

to a proportionality constant, a quantity commonly used in polymer science.⁵⁹

The interactions of proteins in non-binding configurations can be quantified by $B_{ij}^{(u)}$, which is fully determined by K_d and B_{ij} and which is thus a well-defined thermodynamic quantity. These interactions shape the physico-chemical properties of the crowded environments inside cells. For example, non-binding interactions can lead to demixing and therefore to co-localization of binding partners. This co-localization effectively increases the binding probability.

In principle, the contributions $B_{ij}^{(u)}$ of non-binding interactions to B_{ij} can be determined experimentally. SAXS experiments provide information about B_{ij} in the forward scattering and information about dimerization and thus K_d is encoded in the radius of gyration. Varying protein concentrations in equilibrium sedimentation experiments can provide estimates for K_d and B_{ij} .¹⁰ The latter is used to correct for the non-ideality of the protein solution. Equation 1 can be viewed as such a correction for non-ideality. Especially for weak binders we expect that K_d and B_{ij} can be estimated accurately enough such that the contributions $B_{ij}^{(u)}$ of non-binding conformations to B_{ij} can be determined. Similar to the calculations performed here, we expect that in sedimentation experiments the uncertainties in the estimates for $B_{ij}^{(u)}$ will be much smaller than the individual uncertainties in the estimates for K_d and B_{ij} .

The Complexes++ simulation software and the binless-WHAM code can be downloaded free of charge at <https://www.github.com/bio-phys/complexespp> and at <https://github.com/bio-phys/binless-wham>, respectively.

Acknowledgement We thank Drs. Mateusz Sikora, Jakob T. Bullerjahn, Roberto Covino, and Attila Szabo for insightful discussions. We acknowledge financial support by the Max Planck Society.

References

- (1) Johnson, M. E.; Hummer, G. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 603–608.
- (2) Johnson, M. E.; Hummer, G. Evolutionary Pressure on the Topology of Protein Interface Interaction Networks. *J. Phys. Chem. B* **2013**, *117*, 13098–13106, PMID: 23701316.
- (3) Qin, S.; Zhou, H.-X. Protein folding, binding, and droplet formation in cell-like conditions. *Curr. Opin. Struct. Biol.* **2017**, *43*, 28 – 37, Theory and simulation • Macromolecular assemblies.
- (4) Kastiris, P. L.; Bonvin, A. M. J. J. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc., Interface* **2013**, *10*, 20120835.
- (5) McMillan, W. G.; Mayer, J. E. The Statistical Thermodynamics of Multicomponent Systems. *J. Chem. Phys.* **1945**, *13*, 276–305.
- (6) Hill, T. L. Theory of Protein Solutions. I. *J. Chem. Phys.* **1955**, *23*, 623–636.
- (7) McQuarrie, D. A. *Statistical mechanics / Donald A. McQuarrie*; Harper & Row New York, 1975; pp xiv, 641 p. .
- (8) Tessier, P. M.; Vandrey, S. D.; Berger, B. W. et al. Self-interaction chromatography: a novel screening method for rational protein crystallization. *Acta Crystallogr. D* **2002**, *58*, 1531–1535.
- (9) George, A.; Chiang, Y.; Guo, B. et al. *Macromolecular Crystallography Part A; Methods in Enzymology*; Academic Press, 1997; Vol. 276; pp 100–110.

- (10) Harding, S. E.; Rowe, A. J. Insight into protein–protein interactions from analytical ultracentrifugation. *Biochem. Soc. Trans.* **2010**, *38*, 901–907.
- (11) Deszczynski, M.; Harding, S. E.; Winzor, D. J. Negative second virial coefficients as predictors of protein crystal growth: Evidence from sedimentation equilibrium studies that refutes the designation of those light scattering parameters as osmotic virial coefficients. *Biophys. Chem.* **2006**, *120*, 106 – 113.
- (12) Winzor, D. J.; Deszczynski, M.; Harding, S. E. et al. Nonequivalence of second virial coefficients from sedimentation equilibrium and static light scattering studies of protein solutions. *Biophys. Chem.* **2007**, *128*, 46 – 55.
- (13) Blanco, M. A.; Sahin, E.; Li, Y. et al. Reexamining protein–protein and protein–solvent interactions from Kirkwood-Buff analysis of light scattering in multi-component solutions. *J. Chem. Phys.* **2011**, *134*, 225103.
- (14) Wills, P. R.; Winzor, D. J. Rigorous analysis of static light scattering measurements on buffered protein solutions. *Biophys. Chem.* **2017**, *228*, 108 – 113.
- (15) von Bülow, S.; Siggel, M.; Linke, M. et al. Dynamic cluster formation determines viscosity and diffusion in dense protein solutions. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 9843–9852.
- (16) Nawrocki, G.; Karaboga, A.; Sugita, Y. et al. Effect of protein–protein interactions and solvent viscosity on the rotational diffusion of proteins in crowded environments. *Phys. Chem. Chem. Phys.* **2019**, *21*, 876–883.
- (17) Woo, H.-J.; Roux, B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6825–6830.
- (18) Yesselman, J. D.; Denny, S. K.; Bisaria, N. et al. Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 16847–16855.
- (19) Zimm, B. H. Application of the Methods of Molecular Distribution to Solutions of Large Molecules. *J. Chem. Phys.* **1946**, *14*, 164–179.
- (20) Neal, B.; Asthagiri, D.; Lenhoff, A. Molecular Origins of Osmotic Second Virial Coefficients of Proteins. *Biophys. J.* **1998**, *75*, 2469 – 2477.
- (21) Kim, B.; Song, X. Calculations of the second virial coefficients of protein solutions with an extended fast multipole method. *Phys. Rev. E* **2011**, *83*, 011915.
- (22) Singh, J. K.; Kofke, D. A. Mayer Sampling: Calculation of Cluster Integrals using Free-Energy Perturbation Methods. *Phys. Rev. Lett.* **2004**, *92*, 220601.
- (23) Benjamin, K. M.; Singh, J. K.; Schultz, A. J. et al. Higher-Order Virial Coefficients of Water Models. *J. Phys. Chem. B* **2007**, *111*, 11463–11473, PMID: 17850128.
- (24) Grünberger, A.; Lai, P.-K.; Blanco, M. A. et al. Coarse-Grained Modeling of Protein Second Osmotic Virial Coefficients: Sterics and Short-Ranged Attractions. *J. Phys. Chem. B* **2013**, *117*, 763–770, PMID: 23245189.
- (25) Qin, S.; Zhou, H.-X. Calculation of Second Virial Coefficients of Atomistic Proteins Using Fast Fourier Transform. *J. Phys. Chem. B* **2019**, *123*, 8203–8215, PMID: 31490691.
- (26) Mereghetti, P.; Gabdoulline, R. R.; Wade, R. C. Brownian Dynamics Simulation of Protein Solutions: Structural and Dynamical Properties. *Biophys. J.* **2010**, *99*, 3782 – 3791.

- (27) Mereghetti, P.; Martinez, M.; Wade, R. C. Long range Debye-Hückel correction for computation of grid-based electrostatic forces between biomacromolecules. *BMC Biophys.* **2014**, *7*, 4.
- (28) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245 – 268.
- (29) Souaille, M.; Roux, B. Extension to the Weighted Histogram Analysis Method Combining Umbrella Sampling With Free Energy Calculations. *Comp. Phys. Comm.* **2001**, *135*, 40–57.
- (30) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.
- (31) Rosta, E.; Nowotny, M.; Yang, W. et al. Catalytic Mechanism of RNA Backbone Cleavage by Ribonuclease H from Quantum Mechanics/molecular Mechanics Simulations. *J. Am. Chem. Soc.* **2011**, *133*, 8934–8941.
- (32) Harding, S. E.; Horton, J. C.; Jones, S. et al. COVOL: An Interactive Program for Evaluating Second Virial Coefficients from the Triaxial Shape or Dimensions of Rigid Macromolecules. *Biophys. J.* **1999**, *76*, 2432–2438.
- (33) Onnes, H. K. In *Through Measurement to Knowledge: The Selected Papers of Heike Kamerlingh Onnes 1853–1926*; Gavroglu, K., Goudaroulis, Y., Eds.; Springer Netherlands: Dordrecht, 1991; pp 146–163.
- (34) Kamerlingh Onnes, H. Expression of the equation of state of gases and liquids by means of series. *KNAW, Proceedings* **1902**, *4*, 125–147.
- (35) Woolley, H. W. The Representation of Gas Properties in Terms of Molecular Clusters. *J. Chem. Phys.* **1953**, *21*, 236–241.
- (36) Kirkwood, J. G.; Buff, F. P. The Statistical Mechanical Theory of Solutions. I. *J. Chem. Phys.* **1951**, *19*, 774–777.
- (37) Ben-Naim, A.; Navarro, A. M.; Leal, J. M. A Kirkwood–Buff analysis of local properties of solutions. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2451–2460.
- (38) Singh, J. K.; Kofke, D. A. Mayer Sampling: Calculation of Cluster Integrals using Free-Energy Perturbation Methods. *Phys. Rev. Lett.* **2004**, *92*, 220601.
- (39) Sophianopoulos, A. J. Association Sites of Lysozyme in Solution: I. THE ACTIVE SITE. *J. Biol. Chem.* **1969**, *244*, 3188–3193.
- (40) Diamond, R. Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.* **1974**, *82*, 371 – 391.
- (41) Kang, R. S.; Daniels, C. M.; Francis, S. A. et al. Solution Structure of a CUE-Ubiquitin Complex Reveals a Conserved Mode of Ubiquitin Binding. *Cell* **2003**, *113*, 621 – 630.
- (42) Bennett, S. E.; Schimerlik, M. I.; Mosbaugh, D. W. Kinetics of the uracil-DNA glycosylase/inhibitor protein association. Ung interaction with Ugi, nucleic acids, and uracil compounds. *J. Biol. Chem.* **1993**, *268*, 26879–26885.
- (43) Putnam, C. D.; Shroyer, M. J. N.; Lundquist, A. J. et al. Protein mimicry of DNA from crystal structures of the uracil-DNA glycosylase inhibitor protein and its complex with Escherichia coli uracil-DNA glycosylase¹¹Edited by D. C. Rees. *J. Mol. Biol.* **1999**, *287*, 331 – 346.
- (44) Kim, Y. C.; Hummer, G. Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *J. Mol. Biol.* **2008**, *375*, 1416–1433.

- (45) Miyazawa, S.; Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **1985**, *18*, 534–552.
- (46) Miyazawa, S.; Jernigan, R. L. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256*, 623 – 644.
- (47) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
- (48) Straatsma, T.; Berendsen, H.; Stam, A. Estimation of statistical errors in molecular simulation calculations. *Mol. Phys.* **1986**, *57*, 89–95.
- (49) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124, PMID: 25400522.
- (50) Piana, S.; Donchev, A. G.; Robustelli, P. et al. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113–5123, PMID: 25764013.
- (51) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E4758–E4766.
- (52) Marrink, S. J.; Risselada, H. J.; Yefimov, S. et al. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.
- (53) Stark, A. C.; Andrews, C. T.; Elcock, A. H. Toward Optimized Potential Functions for Protein-Protein Interactions in Aqueous Solutions: Osmotic Second Virial Coefficient Calculations Using the MARTINI Coarse-Grained Force Field. *J. Chem. Theory Comput.* **2013**, *9*, 4176–4185.
- (54) Schmalhorst, P. S.; Deluweit, F.; Scherers, R. et al. Overcoming the Limitations of the MARTINI Force Field in Simulations of Polysaccharides. *J. Chem. Theory Comput.* **2017**, *13*, 5039–5053.
- (55) Brangwynne, C.; Tompa, P.; Pappu, R. Polymer physics of intracellular phase transitions. *Nat. Phys.* **2015**, *11*, 899 EP –.
- (56) Huggins, M. L. Solutions of Long Chain Compounds. *J. Chem. Phys.* **1941**, *9*, 440–440.
- (57) Flory, P. J. Thermodynamics of High Polymer Solutions. *J. Chem. Phys.* **1941**, *9*, 660–660.
- (58) Wei, M.-T.; Elbaum-Garfinkle, S.; Holehouse, A. S. et al. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **2017**, *9*, 1118 EP –, Article.
- (59) Harmon, T. S.; Holehouse, A. S.; Rosen, M. K. et al. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife* **2017**, *6*, e30294.

preprint.pdf (1.49 MiB)

[view on ChemRxiv](#) • [download file](#)
