Supporting Information for:

An ABSINTH-Based Protocol for Predicting Binding Affinities Between Proteins and Small Molecules

Jean-Rémy Marchand[†], Tim Knehans[†‡], Amedeo Caflisch[†], and Andreas Vitalis[†*]

[†] Department of Biochemistry, University of Zürich, CH-8057, Zürich, Switzerland

[‡] Present address: Schrödinger GmbH, Q7, 23, 68161 Mannheim, Germany

Supporting Methods

S.1. Automatic ligand parameters

As mentioned in the main text (see 2.2.3), we used CGenFF[1] to determine the partial charges for ligands. Together with the assignment of solvation groups for ABSINTH,[2] which are described in detail in the main text (2.2.2), this leaves Lennard-Jones parameters as well as bonded parameters to be determined. For the Lennard-Jones parameters, we use a fixed map of Tripos atom types to CAMPARI/ABSINTH atom types. Revisions to the Lennard-Jones parameters of ABSINTH atom types are listed in S6, Tables S4–S7. The map and all other required parameters are either determined by CAMPARI automatically or available from the authors upon request.

Bonded parameters are particular in our approach because the ABSINTH philosophy features two paradigmatic deviations from other force fields. First, sampling is generally assumed to be in an internal rigid-body/torsional space, thus obviating the need for parameters for the constrained bond lengths and angles. The (fixed) values are instead taken directly from crystal structures[3] or from chemical rules represented in public databases such as ZINC.[4] Second, short-range electrostatic interactions are pruned (see S.6 for details), and rotational barriers arising from steric considerations are assumed to be sufficient for the majority of dihedral angles.

However, in organic molecules, there are multifarious bonds with rotational barriers arising from electronic effects, *e.g.*, alkenes, esters, anilines, sulfonamides, *etc.*, which are not found in polypeptides. To prevent arbitrary rotations around these bonds, we use a fourfold heuristic: 1) all ring bonds are constrained and require no parameters; 2) for bonds not expected to isomerize spontaneously, such as double bonds in unconjugated alkenes, a harmonic potential is assigned to limit the rotation around that bond; 3) for relatively clear-cut cases (like aliphatic linear amides), we assign a bimodal potential, which might or might not favor either Z- or E-isomer based on chemical knowledge; 4) in unclear cases, no potential is assigned. The last rule implies that such cases should be inspected in post-processing, but we did not deem this necessary for the results in this manuscript. Note that the input conformer is assumed to be representative, which is primarily relevant for detecting local planarity according to thresholds (15° and 30° for atoms in or not in rings, respectively)

The heuristic just described is meaningful primarily for ligand conformers derived directly or indirectly from experiments or higher-level (quantum) calculations. The major weakness is the inability to sample flexible rings. To overcome this caveat, knowledge-based strategies[5] might be adopted in the future. The strength parameters of the potentials are inferred from values used within CGenFF and CHARMM for similar chemotypes. The code implementing our approach will be included in the upcoming version 4 of CAMPARI[6] (to be released to the public in 2020) and is presently available from the authors upon request.

S.2. General sampling protocol

In this study, our primary data set is a set of structures of complexes of hugely diverse ligands on a number of proteins with known experimental binding affinities. While some proteins, such as HIV protease, are strongly represented in the source data (see 2.3.1 in the main text), our working assumption was that every protein and every ligand is unique. Given that the aim is to predict binding affinities computationally, we need a way to refine the experimental complex structure according to our energy function of choice. Despite the high quality of the original structures, we encountered clashes, missing parts, and geometric inconsistencies. Given that the ABSINTH model is an all-atom, physically motivated force field, we required all atoms, including hydrogens, to be present. In general,

the entire proteins were kept whenever possible. In some cases, such as PDB 2f2h, this implied large structures with higher-order symmetry where only one of several analogous binding sites was considered (the other ones being left empty).

Here, we chose to energetically minimize the binding sites of the completed and relaxed structures (see 2.3.2 and 2.3.3 in the main text) predominantly by molecular dynamics in mixed rigid-body/torsional space[7] at a formal temperature of 250 K. This method has several advantages, which make it well-suited to the task: i) individual, internal coordinate space degrees of freedom can be constrained at will; ii) there are no mass-metric tensor artifacts,[8] *i.e.*, the method is thermodynamically accurate; iii) operating in a pure internal coordinate space allows us to omit most of the parameters for bonded interactions in the ligand; iv) the main drawback of the method, *viz.*, the somewhat artificial dynamics, are not relevant for a thermodynamic prediction. Readers are referred to the original publication[7] for more details. The setup work for the ligands is described briefly in S.1 and in 2.2 in the main text.

The time step was 5 fs, and simulations were run in the NVT ensemble for 10 ps in a droplet boundary condition of virtually infinite size. The Andersen thermostat[9] as adapted for the simple integrator introduced in the original paper[7] was used to maintain constant temperature with a relaxation time of 0.1 ps. The degrees of freedom were specific to each complex. They always included the ligand's rigid-body and torsional coordinates. In addition, certain protein degrees of freedom were allowed to move. These were restricted to those χ-angles, which move atoms that are within a threshold of 6 Å of *all* ligand atoms in the relaxed starting structure. The constraints left between 30 and 132 protein degrees of freedom (quartiles at 80, 88, and 100). All of the sampled dihedral angles were side chain dihedral angles with the backbone serving as the immovable base of motion. It was therefore impossible for loops to relax, for polypeptide chains to move relative to each other, and, in many cases, for the ligand to dissociate even when there were clashes resulting from a poor input structure or a sterically unfavorable protein H-mer.

S.3. Estimation of ligand binding free energies

We estimate binding free energies by directly considering the terms due to complex, protein, and ligand. The most basic version would be as follows:

$$\begin{aligned} &\mathrm{Protein} + \mathrm{Ligand} \rightleftharpoons \mathrm{Complex} \\ &\Delta G_b \approx \langle U(\mathrm{Complex}) \rangle - \langle U(\mathrm{Protein}) \rangle - \langle U(\mathrm{Ligand}) \rangle \end{aligned} \tag{S1}$$

In eq. (S1), the angular brackets denote ensemble averages taken across the latter half of the 10-ps simulations described in S.2. Eq. (S1) represents an approximation: the internal energies are those of the ABSINTH model, which means that they are effective energies. These energies contain solvent entropy terms through the experimental reference free energies of solvation, which are used as parameters of the ABSINTH model (eq. (1) in the main text). In this formalism, conformational entropy terms[10] are neglected. This is expected to give a systematic error that scales with the size of the binding interface and the ligands' intrinsic flexibility.

For a net-neutral ligand, the simple binding reaction described by eq. (S1) is a common approximation: no change to net charge occurs upon binding, and the charge imbalance error in the protein and complex terms are expected to largely cancel. This is true despite the fact that we did not neutralize the various systems, which would theoretically have been possible through changing the protein H-mer in distal parts and/or by adding fixed counterions. The situation is more complicated if both ligand and protein carry net charge groups. In this scenario, we propose here that it is more appropriate to consider charged buffer components explicitly as part of the binding process. The simplest

representation are inorganic ions like K^+ and Cl^-. Let us assume that P denotes a specific protein H-mer, L the ligand, and PL the complex, and that the net charge values of these specific protein and ligand H-mers are *z* and *x*, respectively. Then:

$$\mathrm{P}^z + (n - z - x)\mathrm{K}^+ + n\mathrm{Cl}^- + \mathrm{L}^x \rightleftharpoons (\mathrm{PL})^{z+x} + (n - z - x)\mathrm{K}^+ + n\mathrm{Cl}^- \quad \text{(S2)}$$

Eq. (S2) is the same reaction as that in (S1) except that we add a non-interacting ionic atmosphere ($n > |z|+|x|$). However, writing it in this way highlights that the simple reaction implied by eq. (S1) may correspond to a single microscopic parameter in a more complex thermodynamic network and not to an observable binding affinity. This will be the case if protein species other than P^z are responsible for binding, *e.g.*:

$$\begin{aligned} &\mathrm{P}^z + (n - z - x)\mathrm{K}^+ + n\mathrm{Cl}^- + \mathrm{L}^x \rightleftharpoons (\mathrm{PK})^{z+1} + (n - z - x - 1)\mathrm{K}^+ + n\mathrm{Cl}^- + \mathrm{L}^x \\ &(\mathrm{PK})^{z+1} + (n - z - x - 1)\mathrm{K}^+ + n\mathrm{Cl}^- + \mathrm{L}^x \rightleftharpoons (\mathrm{PL})^{z+x} + (n - z - x)\mathrm{K}^+ + n\mathrm{Cl}^- \end{aligned} \quad \text{(S3)}$$

In the second line of eq. (S3), ligand binding displaces a single potassium ion and releases it into the bulk. This will certainly be relevant if $(PK)^{z+1}$ is more populated in solution than P^z, which is a feasible scenario. We emphasize that the formalism in eq. (S3) can be interpreted as representing all types of ion binding (from specific sites to ionic atmospheres)[11-13] and all types of ions (from single atoms to macromolecules). This poses a complexity that cannot be handled in molecular simulations, let alone in high-throughput applications. In molecular simulations, the typical approach is to neutralize a finite system with simple, monovalent ions, and to add a certain amount of background electrolyte solution, which is the "*n*" in eqs. (S2) and (S3). All species are allowed to move according to their mutual interactions, which means that ions can diffuse to create ionic atmospheres or occupy specific binding sites in competition with the ligand. This eliminates many of the errors incurred by eq. (S1) by performing a Boltzmann integral over the thermodynamically most relevant binding equilibria. However, this method requires a long enough simulation, which renders it infeasible for a screening workflow.

Here, we pursue an even more simplified approach as follows. We first parsed the partial charges of the ligands into groups such that the net charges per group approximately sum up to an integer value (see 2.2.3 in the main text). The underlying procedure is hierarchical and considers successively larger entities of connected atoms. This is to avoid distant atoms being part of the same charge group. Only groups with net charges of -1, 0, or 1 are searched for. The algorithm yields a partitioning of the atoms of the ligand into sets. Here, the tolerance for matching the target was 0.05 for all molecules except α-amino acids, for which we used 0.15 (for a value of 0.05, the amino and carboxy moieties were not detected separately). Next, given the partitioning, we isolated the positions of the atoms closest to the center of charge for all non-neutral charge groups in the ligand as found at the end of the simulation of the protein-ligand complex. For the computation of protein H-mer energies, a K^+ or Cl^- was introduced at exactly those positions at the beginning of the run. These ions were free to move within a cuboid volume of $(3\ \mathrm{nm})^3$ centered around the ligand and held back by half-harmonic position restraints beyond. Thus, for an example ligand with a single +1 charge group like an ammonium moiety, our approach considers the following reaction:

$$(\mathrm{PK})^{z+1} + \mathrm{L}^+_{(bulk)} \rightleftharpoons (\mathrm{PL})^{z+1} + \mathrm{K}^+_{(bulk)} \quad \text{(S4)}$$

The internal energy of the ion in the bulk in the ABSINTH model is simply given by its reference free energy of solvation: -86.0 kcal/mol for K^+ and -74.5 kcal/mol for Cl^- (compare Table S3) and thus readily included in the difference. Importantly, the ion is free to dissociate from the protein H-mer within that $(3\ \mathrm{nm})^3$ cube. It is, however, critical that ions do not occupy favorable binding sites where they would

not be actually displaced by the ligand. This constraint is difficult to automatize, and an improved heuristic will most likely be implemented in the future. The revised version of eq. (S1) is:

$$\begin{aligned} &\text{Protein/Ions} + \text{Ligand} \rightleftharpoons \text{Complex} + \text{Ions} \\ &\Delta G_b \approx \left\langle U\left(\text{Complex}\right)\right\rangle - \left\langle U\left(\text{Protein/Ions}\right)\right\rangle - \left\langle U\left(\text{Ligand}\right)\right\rangle + U(Ions) \end{aligned} \tag{S5}$$

Eq. (S5) is also given as eq. (3) in 2.3.4 in the main text. "Proteins/Ions" signifies the (possible) presence of explicit counterions in the protein-only simulations. *U(Ions)* is not an ensemble average because we assume a bulk reference state for the unbound ions. The description so far applies to the validation set for ranking known binders (2.3 in the main text). For the enrichment validation (2.4 in the main text), two modifications were incorporated. First, the (only) protein H-mer was considered exactly once for every unique combination of required ions in the screened poses, and the resultant values were used for all matching ligands. To place the ions, the 10 ps of molecular dynamics were preceded by 25000 Monte Carlo steps. The move set for the latter consisted of single ion displacements, either random placements in the entire box (30%) or local moves with a uniform and symmetric sampling interval of 10 Å for every dimension. As described above, ions could move freely inside of a $(3\ \text{nm})^3$ cube but were prevented from leaving it by half-harmonic restraints, which were centered heuristically at the binding site. Second, the tolerance for detecting charge groups was 0.05 without exceptions.

S.4. Estimation of protonation free energies in the ABSINTH model

To estimate the free energy change associated with a change in protonation state of a given protein residue, we use, just like for the ligand binding free energies in S.3, effective energy as a proxy. As explained before, the ABSINTH model expresses solvent entropy terms, which are contained in experimental reference free energies of solvation, as part of an effective, internal energy. Thus, the approximation of the free energy is justifiable only if contributions from conformational entropy are small, which should be the case here due to the very limited degrees of freedom we allow to move (the set is exactly the same as the one used during the corresponding ligand binding free energy calculation). The raw energy differences are affected by differences in the reference state. We remove these by choosing the mean energies of dipeptides as reference state values. These terms form the excess contributions to the estimated free energies. Because we want to treat the buffered proton concentration as an external parameter, the effective protonation free energies are describing the ratio of concentrations at equilibrium of the protonated and deprotonated forms (as in the Henderson-Hasselbalch equation). The final form for, *e.g.*, an aspartate side chain in a protein is:

$$\Delta G_p^{Asp\rightarrow AspH} \approx k_b T \ln\left[f^\circ K_a\right] + \left\langle U\left(\text{Prot-D}^0\right)\right\rangle - \left\langle U\left(\text{Prot-D}^-\right)\right\rangle + \left\langle U\left(\text{Dip-D}^-\right)\right\rangle - \left\langle U\left(\text{Dip-D}^0\right)\right\rangle \tag{S6}$$

In eq. (S6), the angular brackets denote ensemble averages, and "Prot" and "Dip" refer to the protein and dipeptide contexts, respectively. The first term expresses the quantum terms responsible for the intrinsic stability differences between the protonated and deprotonated forms. These differences in electronic energies are not represented in a classical force field and must be added as *ad hoc* parameters. We use here the same reference values as Radak *et al*.[14] for the K_a values. Lastly, f° is a reference state correction for the chosen concentration scale. While it is easy to determine the dipeptide values to sufficient accuracy (see Table S1), the protein values will carry a statistical error. This error is expected to grow with the number of degrees of freedom left flexible in the protein.

Changes in protein H-mers are, just like ligand binding, performed in the presence of possible charge imbalances. However, unlike in the case of the competition between ligand and buffer species, a given protein side chain occupies practically the same volume irrespective of its protonation state. Thus, we cannot devise an automated correction based on inorganic ions in the same way. As an example,

consider a pair of hydrogen-bonded ASP/GLU residues in a binding site, at least one of which is protonated. While the doubly charged conformation may favor the presence of a cation, the ASP/GLU side chains would first have to change conformation to accommodate this ion or ionic group. Thus, for estimating the free energies of conversion between protein H-mers, we do not introduce any additional changes to the ionic environment, which might cause errors.

Table S1. Model compound value for the species given in the top row in kcal/mol. Superscripts indicate charged forms ("+" or "-") and neutral forms ("0", "δ", "ε") where for histidine the index of the protonated nitrogen atom is given. Protonated acids and deprotonated amines are asymmetric but allowed to rearrange through changes in dihedral angles during the simulation runs. The numbers in the first data row are used for all models with two exceptions: for Fig. 2(g) in the main text (based on old parameters), the data in the second row had to be employed while for Fig. S3(f) (reduced rFOS offsets on charged groups) the data in the third row had to be used.

D^-	D^0	E^-	E^0	K^0	K^+	H^δ	H^ε	H^+
-106.35	-19.63	-105.38	-18.24	-19.81	-101.38	-24.23	-26.87	-94.75
-120.61	-19.45	-121.08	-18.07	-19.66	-115.58	-24.16	-26.64	-108.91
-98.13	-19.59	-98.05	-18.25	-19.80	-93.89	-24.22	-26.87	-87.44

The case of multiple ligand H-mers is particular. Here, ligand H-mers with identical net charge groups mean that the unbound protein state is identical, including ions. Because calculations are performed independently for every combination of ligand and protein H-mer, we average the values for identical simulations of unbound proteins in post-processing. Having stated these technical aspects, it is important to stress that the calculation of protein H-mer conversion free energies should follow the same physical model as the one used for ligand binding. Only this setup can ensure that the equilibrium constructed between all bound and unbound species is intrinsically consistent and thus meaningful. This poses a problem for the data shown in Fig. 2(i) in the main text. However, the cost of evaluating electrostatic energies in a Poisson model precluded us from generating mutually consistent data in this approximation, which would have ideally required molecular dynamics of complexes, proteins, and ligands in a Poisson model as described in S.2. As a compromise, complexes were minimized in rigid-body/torsional space in a preliminary ABSINTH model for a maximum of 2000 steps using a Broyden-Fletcher-Goldfarb-Shanno scheme according to Nocedal.[15] These minimized complex structures were then used to derive Poisson estimates of electrostatic binding free energies according to standard methodologies with an assumed low (protein) dielectric of 4.0.[16]

S.5. Combination of binding free energy predictions for multiple H-mers

The selection of protein and ligand H-mers we considered is summarized in the main text (see 2.3.3). Once estimates of ΔG_b have been calculated for a specific combination of ligand and protein H-mer according to eq. (S5), we need to calculate the apparent binding free energy from the set of numbers corresponding to an experimentally unique pair of protein and ligand. To do so, we assume that all species are at chemical equilibrium. As variables we have the concentrations of all species: protein H-mers, ligand H-mers, and complexes of ligand and protein H-mers. The total ligand and total protein concentrations are constants, which we are free to choose (the predicted equilibrium constants do not depend on them beyond numerical precision). We also assume a buffered solution such that the proton concentration can be set and treated as a constant, here pH 7. The predicted free energies for individual reactions (see above) are used as input parameters. Temperature is another external parameter, for which we use a value of 310 K.

In a first step, all protein H-mers are connected in a tree where each edge corresponds to a single protonation state change of a single side chain. For illustration, we use a hypothetical protein with 2 relevant titration sites, D4 and H15. Histidine has 3 possible states while aspartate has 2, giving rise to 6 different protein states, P1–P6.

Table S2. List of protein H-mers in the example used for illustration. The dots indicate the omitted (identical) sequence context.

Name	Sequence
P1	…D_4^0…H_{15}^{δ}…
P2	…D_4^0…H_{15}^{ε}…
P3	…D_4^0…H_{15}^{+}…
P4	…D_4^{-1}…H_{15}^{δ}…
P5	…D_4^{-1}…H_{15}^{ε}…
P6	…D_4^{-1}…H_{15}^{+}…

A set of 5 H-mer conversion reactions to connect these 6 species would be:

$$\begin{aligned}
&\mathrm{P4}_{(ion)} + \mathrm{H}^+ + \mathrm{L1} \rightleftharpoons \mathrm{P1}_{(ion)} + \mathrm{L1} \\
&\mathrm{P5}_{(ion)} + \mathrm{H}^+ + \mathrm{L1} \rightleftharpoons \mathrm{P2}_{(ion)} + \mathrm{L1} \\
&\mathrm{P6}_{(ion)} + \mathrm{H}^+ + \mathrm{L1} \rightleftharpoons \mathrm{P3}_{(ion)} + \mathrm{L1} \\
&\mathrm{P1}_{(ion)} + \mathrm{L1} \rightleftharpoons \mathrm{P2}_{(ion)} + \mathrm{L1} \\
&\mathrm{P2}_{(ion)} + \mathrm{H}^+ + \mathrm{L1} \rightleftharpoons \mathrm{P3}_{(ion)} + \mathrm{L1}
\end{aligned} \qquad \text{(S7)}$$

In eq. (S7), "L1" denotes the first H-mer of the ligand, which is included for formal clarity. In the actual calculations to derive free energy estimates for these reactions, which are described in S.4, the ligand is not present. The subscript *(ion)* indicates that P1–P6 are all simulated in the presence of m_1 K^+ and m_2 Cl^- ions where m_1 and m_2 are the numbers of identified charge groups with a net charge of +1 and -1, respectively. Unlike the ligand, the ions are present during the protein-only simulations and allowed to move in a $(3\ \mathrm{nm})^3$ cube centered at the binding site (see S.3). As mentioned, the free proton concentration is assumed to be buffered and is only represented as a constant in the equations.

In the next step, the tree is replicated as many times as there are different ligand H-mers, here we assume 2 for our example. Because the ligand is assumed to be in the bulk in the reactions in eq. (S7), the predicted bulk distribution of ligand H-mers can be used to connect the trees to each other. We use an arbitrary protein H-mer for this, e.g.:

$$\mathrm{P1}_{(ion)} + \mathrm{L1} \rightleftharpoons \mathrm{P1}_{(ion)} + \mathrm{L2} \qquad \text{(S8)}$$

There is one such reaction per ligand H-mer after the first one. The predicted equilibrium is precalculated for a specific pH, so these equilibria do not respond explicitly to the pH we choose but must instead be recalculated. Furthermore, the ion sets implied by eq. (S8) for P1 will differ if L1 and L2 differ in their charge groups with a net charge of ± 1. In these cases, there is an additional reaction required to connect the trees (change in ions). For this, we assume that the extra ions unique to only L1 or L2 are released into the bulk, and we apply an appropriate correction based on the same values mentioned in relation to eqs. (S4) and (S5). With this, we obtain a fully connected thermodynamic tree for the 12 combinations of protein and ligand H-mers in the unbound state. To complete the system, we simply connect every vertex of this tree to its corresponding bound form, *i.e.*, we add all 12 binding equilibria as reactions, *e.g.*:

$$\mathrm{P5K}^+_{m_1} Cl^-_{m_2} + \mathrm{L1}_{(bulk)} \rightleftharpoons \mathrm{P5L1} + m_1 \mathrm{K}^+_{(bulk)} + m_2 \mathrm{Cl}^-_{(bulk)} \qquad \text{(S9)}$$

In eq. (S9), the ions are written explicitly to highlight both the differences in their thermodynamic state (bound *vs*. bulk) and the fact that the number is specific to each ligand (m_1 and m_2). We emphasize again that the first species on the left-hand side does not need to be a protein H-mer with ions bound

to specific sites: it simply means that these explicit excess ions were present in a $(3\ \mathrm{nm})^3$ cube centered around the ligand. Thus, we arrive at 26 unknown species concentrations (12 protein H-mers, possibly with ions, 12 complexes, and the 2 ligand forms) with 24 reaction equations for which we have predicted equilibrium constants (12 binding equilibria, 10 protein H-mer conversion equilibria, 1 ligand H-mer equilibrium, and 1 ion exchange equilibrium for the protein H-mer chosen to connect the trees). If, in this example, L1 and L2 were not to differ in both m_1 and m_2, the numbers are reduced accordingly: there would be 12 complexes, 6 protein H-mers with ions, and 2 ligand forms (20 unknowns) along with 18 unique equilibria (12 binding, 5 protein H-mer conversion, and 1 ligand H-mer conversion). The bulk reference state for the ions in eq. (S9) and elsewhere means that their concentrations could be variables although they would have to be coupled on account of electroneutrality. Instead, like for H^+, we choose to assume that they are formally present in a buffered, dilute reference state.

This type of system is straightforwardly solvable in reduced units where binding equilibria, eq. (S9), are in units of [L][P1] and protein H-mer conversions, eq. (S7), are in units of [P1], where P1 is an arbitrarily chosen reference H-mer, and [L] is the total free ligand concentration (including all ligand H-mers). Actual concentrations can be derived in the end by choosing the total ligand and protein concentrations and a subsequent iterative determination of [P1]. The predicted observed (macroscopic) binding equilibrium constant is computable as:

$$K_{obs}^{*} = \frac{\sum_{I}^{NP}\sum_{J}^{NL}\left[\mathrm{P}I\mathrm{L}J_{(aq)}\right]}{\sum_{J}^{NL}\left[\mathrm{L}J_{(aq)}\right]\sum_{J}^{NL}\sum_{I}^{NP}\left[\mathrm{P}I_{(aq,ion(J))}\right]} \qquad \text{(S10)}$$

In eq. (S10), which is also part of eq. (4) in the main text, the numbers of protein and ligand H-mers are abbreviated as *NP* and *NL*, respectively. The asterisk denotes that, due to missing terms and reference state corrections, the numerical values should not be directly compared to experiments whereas the rankings should be. We developed an in-house R-script for solving the equilibria and determining the values of K_{obs}^{*} per complex.

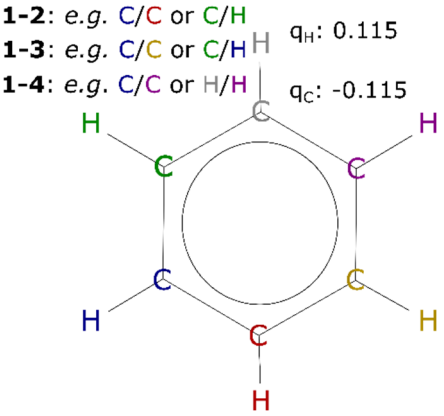
S.6. Changes to the ABSINTH force field

The ABSINTH force field has a few particularities, which make it distinct from other force fields: first, reference free energies of solvation appear as direct parameters (see eq. (1) in the main text), which are values determined experimentally. However, for ionic groups in proteins, to avoid artifacts from excessive salt bridge formation, we added offsets of -30 kcal/mol in the original version.[2] These offsets are retained here and also applied to ionic groups in ligands but with a reduced value of -15 kcal/mol. This change goes along with a homogenization of reference free energies of solvation for ionic compounds according to the work of Kelly and Truhlar.[17] As Table S3 shows, the differences for the polypeptide groups are almost entirely due to the change in offset. The only larger differences due to the homogenization occur for inorganic cations and for arginine.

The second particularity is the short-range interaction model. In ABSINTH, there are no short-range electrostatic interactions unless *all* pairs of atoms between two charge groups are allowed to interact according to the general nonbonded rules. The latter state that the two atoms must be at least three bonds apart and that their distance must not be fixed. Charge groups are defined as minimal sets of connected atoms adding up to a net integer charge (including 0), see S.3 and 2.2.3 in the main text. This depletion is coupled to two other concepts: 1) to perform simulations in an internal coordinate space; 2) to not use dihedral angle potentials on simple rotatable bonds. The former implies keeping most rings frozen, the only handled exception being pucker angles in proline and nucleic acids.[18]

Table S3. Changes to free energies of solvation. These data include the offset contributions. Note that the neutral forms of Asp, Glu, and Lys were not available in the original ABSINTH model. These data are taken from Cabani *et al*.[19]

Compound	*acetate (D^-)*	*propionate (E^-)*	*n-butyl-ammonium (K^+)*	*n-propyl-guanidinium (R^+)*	*4-methyl-imidazolium (H^+)*
New (kcal/mol)	-92.6	-91.2	-86.5	-80.0	-80.0
Old (kcal/mol)	-107.3	-107.3	-100.9	-100.9	-95.0
Compound	*acetic acid (D^0)*	*propionic acid (E^0)*	*n-butylamine (K^0)*	*K^+*	*Cl^-*
New (kcal/mol)	-6.7	-6.5	-4.3	-86.0	-74.5
Old (kcal/mol)	N/A	N/A	N/A	-70.5	-74.6



Scheme S1. Illustration of the short-range electrostatic depletions using benzene. Due to its symmetry, benzene is assigned the same partial charges for all its carbon atoms (-0.115) and all its hydrogen atoms (0.115). Each C–H group forms a charge group with a net charge of zero. In ABSINTH, exclusion rules are such that all atoms in pairs of charge groups must be at least 3 bonds apart (1–4). Thus, in benzene, only opposite C–H groups can theoretically interact via screened electrostatics. In a standard force field, all atoms interact except those separated by 2 bonds or fewer. This means, for example, that adjacent C–H groups interact exclusively via the H—H interaction as if they were fractional-charge monopoles. Additionally, in ABSINTH, a benzene ring is rigid: this implies that the distances between atoms in opposite C–H groups cannot change. Thus, all intramolecular, electrostatic interactions are excluded.

Depleting short-range electrostatic interactions in this way has the advantage that only proper moments (predominantly monopoles and dipoles) interact. This would not be the case if simple 1-2/1-3 exclusion rules are used (see Scheme S1). To determine the charge groups for arbitrary molecules, we use an automatic partition scheme that we refined here with the goal of yielding small, compact groups based on a tolerance threshold. The ABSINTH approach to short-range electrostatic interactions carries with it a twofold caveat: first, important interactions might be missing that would have an influence on local conformational preferences; second, the effective many-body terms used for describing solvation imply that the absolute internal energy of a system will not respond correctly if a fixed conformation is desolvated by the presence of other explicit solutes. This second effect is akin to how in PB/GB models of continuum electrostatics the charging or "self" energy must depend on the size and geometry of the low-dielectric cavity.[20] In ABSINTH, the leading error term would be due to missing monopole-monopole interactions, and we suggest to correct it here. This is a conservative change since in a normal protein with CHARMM-[21] or OPLS-based[22] charges there are no such missing monopole-monopole terms in proteins or nucleic acids (except at uncapped termini). However, in small molecules they can be abundant, *e.g.*, citrate or free amino acids, which occur repeatedly as ligands in our selected PDBbind data set (see 2.3 in the main text).

The corrections use the same monopole approximation we employ for long-range corrections elsewhere,[23] and which is shown as the second sum in eq. (2) in the main text:

$$W_{selfcorr} = W_{self}\left(\{\mathbf{r}\}_{complex}, \{\upsilon\}_{complex}\right) - W_{self}\left(\{\mathbf{r}\}_{free}, \{\upsilon\}_{free}\right)$$

$$W_{self}\left(\{\mathbf{r}\}, \{\upsilon\}\right) = \sum_{i=1}^{ML}\sum_{j=i+1}^{ML} f_{ij}\frac{Q_i Q_j\left(1-a\upsilon_k^i\right)\left(1-a\upsilon_l^j\right)}{4\pi\varepsilon_o\left|\mathbf{r}_l^j - \mathbf{r}_k^i\right|} \qquad \text{(S11)}$$

$$a = \left(1.0 - \varepsilon_w^{-\frac{1}{2}}\right)$$

In eq. (S11), the coordinate vectors and solvation states for individual atoms are written as **r** and υ, respectively. The curly brackets denote sets of these quantities for all atoms in the system. The values of υ for the complex depend on the protein coordinates, *i.e.*, they express the amount of desolvation incurred by the presence of the receptor. They can either be calculated as atom-specific quantities (ABSINTH default) or averaged over charge groups as used in Mao *et al.*[24] The indices *k* and *l* identify the atoms in charge groups *i* and *j*, respectively, which are closest to the corresponding centers of charge. *ML* is the number of charge groups in the ligand, and the integer charges of the two groups are denoted as Q_j and Q_j. The remaining parameters are the relative dielectric of water (ε_w, 78.2) and the vacuum permittivity (ε_0). Finally, the factor f_{ij} is 0.0 whenever this interaction was not excluded at the level of interatomic potentials, and 1.0 otherwise. For the purpose of this work, we implemented the correction as a *post facto* term applied to the final structures only. Since the correction is conservative, a complete addition to the CAMPARI implementation of ABSINTH is envisioned in future work. Note that the correction can only (but need not) play a role for ligands containing at least two charge groups with a net charge of ±1.

The final change to the ABSINTH model is to the Lennard-Jones (LJ) parameters. To support new chemotypes, additional atom types were needed. They are listed in Table S4.

Table S4. Lennard-Jones parameters for new atom types required for parameterizing small molecules. The halogen parameters are those for organohalides and not for ions.

Atom type	σ (Å)	ϵ (kcal/mol)
sulfonyl-like S	3.60	0.40
sulfinyl-like S	3.60	0.40
trivalent P	3.60	0.30
F	2.90	0.05
Cl	3.50	0.30
Br	3.70	0.50
I	4.00	0.60

It has been recognized that the published parameters can be improved in terms of local steric preferences by adding additional energetic restraints[25] and, similarly, that the spatial packing they predict is too tight.[26] Both effects hint at the fact that some of the LJ size parameters are too small. Table S5 summarizes the changes relevant to this work. Most of the interaction strength parameters are unchanged but listed for clarity.

Table S5. **Updated Lennard-Jones parameters along with the original values**.

Atom type	New σ (Å)	Old σ (Å)	New ϵ (kcal/mol)	Old ϵ (kcal/mol)
any N	3.05	3.20 / 2.70	0.15	0.15
sp^3 O	2.90	3.00	0.15	0.15
acid O	3.10	3.00	0.20	0.20
sp^3 C	3.40	3.30	0.10	0.10
sp^2 C (non-ar) or sp C	3.15	3.00	0.10	0.10
sp^2 C (ar)	3.25	3.00	0.10	0.10
polar H	2.35	2.00	0.025	0.025

Cl^-	3.84	4.42	0.15	0.12
K^+	3.20	4.93	0.10	0.0003

Unlike in the original model, where most size parameters were smaller, the updates require that there are exceptions for atoms separated by exactly three bonds. These are needed to avoid rotational barriers becoming too stiff when bond lengths and angles are fixed. A list is provided in Table S6.

Table S6. Lennard-Jones parameters to use as the source for deriving pairwise interaction parameters for atoms separated by exactly three bonds. No such exceptions are present in the original model.

Atom type	σ (Å) (1-4)
any N	2.90
sp^3 O	2.70
acid O	3.00
sp^3 C	3.30
sp^2 C or sp C	3.00
any S/P except phosphate	3.40
polar H	2.20
F	2.70
Cl	3.30
Br	3.50
I	3.80

Lorentz-Berthelot combination rules are applied as in the original model. This holds for both the normal and the 1-4 parameters. As is apparent from Tables S5 and S6, we increased the size of polar hydrogen atoms. However, this is primarily to avoid close contact between such atoms in cases where electrostatic interactions are missing on account of the ABSINTH short-range exclusion model (see above). To compensate for this, nearly all interactions involving polar hydrogen atoms are overridden in a pair-specific manner as shown in Table S7.

Table S7. Pairwise overrides for polar hydrogen atoms. For specific pairs, both for atoms separated by three bonds (4th column) and otherwise (3rd column), the Lennard-Jones parameters are not obtained by combination rules but by the overrides listed.

Atom type i	Atom type j	σ_{ij} (Å)	σ_{ij} (Å) (1-4)
any N	*polar H*	2.05	1.95
(di)sulfide/thiol S	*polar H*	2.30	2.20
any P or any other S	*polar H*	2.80	2.80
sp^2 O	*polar H*	1.85	1.85
sp^3 O	*polar H*	1.95	1.85
acid O	*polar H*	2.05	2.00
sp^3 C	*polar H*	2.65	2.65
sp^2 C (non-ar) or sp C	*polar H*	2.60	2.50
sp^2 C (ar)	*polar H*	2.575	2.60
nonpolar H	*polar H*	2.10	2.00
F	*polar H*	2.00	2.00
Cl	*polar H*	2.75	2.75
Br	*polar H*	2.85	2.90
I	*polar H*	3.00	3.05

S.7. List of PDB codes retained for 3.2 in the main text

The list is a subset of the 855 complexes identified by Greenidge *et al*.[27] It contains 754 entries as follows:

1a4k 1a4w 1a69 1a8i 1a99 1aaq 1adl 1ai4 1ai5 1ai7 1aid 1ajn 1ajp 1ajq 1ajv
1ajx 1apw 1ax0 1axz 1b05 1b0h 1b1h 1b40 1b46 1b6h 1b6k 1b7h 1b9j 1bcu 1bdq
1bgq 1bhx 1bju 1bjv 1bq4 1br6 1bv7 1bv9 1bwa 1c3x 1c5n 1c5o 1c5p 1c5q 1c5s
1c5t 1c5x 1c5y 1c5z 1c70 1c83 1c84 1c86 1c87 1c88 1ce5 1cea 1ceb 1d4h 1d4i
1d4k 1d4l 1d6v 1d7i 1d7j 1db1 1df8 1dfo 1dhi 1dhj 1dmp 1drk 1dzk 1e1v 1e1x
1e2k 1e2l 1e3v 1e6s 1eb2 1ebw 1eby 1ebz 1ec0 1ec1 1ecv 1efy 1ejn 1ela 1eld
1ele 1enu 1epo 1erb 1ezq 1f0r 1f0u 1f4e 1f4g 1f4x 1fcx 1fcy 1fcz 1fd0 1fh7
1fh8 1fh9 1fhd 1fj4 1fkg 1fkh 1fki 1fl3 1fpc 1g2k 1g2l 1g30 1g32 1g35 1g36
1g74 1g7g 1g85 1gcz 1ghv 1ghw 1ghz 1gi1 1gi7 1gj6 1gja 1gni 1gpn 1gu1 1gyx
1gyy 1gzc 1h1p 1h1s 1h22 1h23 1h9z 1ha2 1hbv 1hmr 1hms 1hmt 1hps 1hpv 1hpx
1hsh 1hvh 1hvi 1hvj 1hvk 1hvl 1hvr 1hvs 1hwr 1hxb 1hxw 1i00 1i5r 1igj 1ii5
1izh 1izi 1j14 1j16 1j17 1j4r 1jak 1jet 1jgl 1jqy 1jsv 1jwt 1jys 1jzs 1k1i
1k1j 1k1l 1k1m 1k1n 1k21 1k22 1k4g 1k4h 1kdk 1kv1 1kyv 1kzk 1kzn 1l2s 1l83
1laf 1lag 1lah 1lbk 1lee 1lf2 1lgw 1li2 1li3 1li6 1lke 1lnm 1lpg 1lpk 1lpz
1lst 1m2q 1m2r 1m48 1mes 1met 1mfi 1mq5 1mq6 1mrw 1mrx 1msm 1msn 1mtr 1mu6
1mu8 1n1m 1n1t 1n2v 1n46 1n4h 1n7m 1n8v 1nc1 1nc3 1nf8 1nfu 1nfw 1nfy 1nhu
1nl9 1nli 1nny 1no6 1nq7 1nt1 1nvq 1nvr 1nvs 1nw7 1nz7 1o2o 1o2q 1o2s 1o2w
1o2x 1o2z 1o30 1o33 1o36 1o3d 1o3i 1o3j 1o3k 1o3p 1o5a 1o5b 1o5c 1o5e 1o5g
1oba 1ocq 1od8 1odi 1odj 1ogd 1ogz 1ohr 1ony 1onz 1os5 1oss 1owe 1owh 1oyq
1oyt 1p1n 1p1o 1p1q 1p57 1pb8 1pb9 1pbq 1pot 1ppc 1pph 1pr1 1pxn 1pxp 1pzi
1q63 1q65 1q66 1q72 1q8t 1qan 1qaw 1qb1 1qbn 1qbo 1qbr 1qbs 1qbu 1qbv 1qi0
1qiw 1qy1 1qy2 1r0p 1r4w 1r5y 1r6n 1r9l 1rd4 1rpj 1s38 1sbg 1sdt 1sdu 1sdv
1sgu 1sh9 1siv 1sqo 1sr7 1srg 1ssq 1stc 1sv3 1sw2 1swg 1swr 1syh 1syi 1t4v
1t7j 1ta2 1ta6 1tcw 1tcx 1td7 1tng 1tnh 1tni 1tog 1toi 1toj 1tok 1tom 1u1w
1ugw 1ugx 1uou 1upf 1urg 1usi 1usk 1utj 1utl 1utm 1utn 1uv6 1uvt 1uw6 1uwf
1uz4 1v0k 1v0l 1v1j 1v2j 1v2k 1v2l 1v2n 1v2o 1v2q 1v2r 1v2s 1v2t 1v2u 1v2w
1vfn 1vj9 1vja 1vyf 1vyg 1vyq 1vzq 1w0z 1w11 1w13 1w3j 1w3k 1w5v 1w5w 1w5x
1w5y 1wcq 1wdn 1we2 1wht 1wm1 1ws4 1ws5 1x8j 1x9d 1xap 1xff 1xgi 1xgj 1xhy
1xk9 1xka 1xkk 1xow 1xt8 1y1z 1y20 1y3n 1y6q 1y6r 1yc1 1yc4 1ydk 1ydr 1yds
1ydt 1yqj 1z1r 1z6e 1zgi 1zhy 1zoe 1zog 1zp8 220l 2a4m 2a8g 2aac 2aj8 2aqu
2avo 2avs 2avv 2ayr 2azr 2b07 2b1v 2b4l 2bak 2bal 2boh 2bok 2bpv 2bpy 2bq7
2bqv 2br1 2brb 2brm 2bt9 2bvd 2bvr 2bvs 2byr 2bys 2bz6 2bza 2c1p 2c3l 2cbu
2cej 2cen 2cgf 2cht 2cji 2csn 2d0k 2d3u 2d3z 2drc 2dri 2e2r 2e7f 2epn 2exm
2f1g 2f2h 2f34 2f35 2f5t 2f80 2f81 2f8g 2fgu 2fgv 2flr 2fpz 2fqw 2fqx 2fr3
2fw6 2fwp 2fx6 2fxu 2g79 2g8r 2g94 2gl0 2glp 2gss 2gst 2gv6 2gv7 2h4g 2h4k
2h6b 2ha2 2ha3 2ha5 2ha6 2hb1 2hb3 2hhn 2hjb 2hs1 2hs2 2hxm 2i0a 2i2b 2i4j
2i4u 2i4v 2i80 2idw 2ien 2ieo 2ihq 2iko 2il2 2iuz 2iwx 2izl 2j2u 2j34 2j47
2j4g 2j4i 2j77 2j7e 2j7g 2j7h 2j9n 2jds 2jfz 2jh0 2jh5 2jh6 2jiw 2nmy 2nmz
2nnk 2nnp 2nt7 2nta 2o0u 2o2u 2o4j 2o4k 2o4r 2o4s 2oag 2ogy 2ojg 2ojj 2ok1
2on6 2oxd 2oxx 2oxy 2oyk 2oym 2p16 2p4j 2p4y 2p7a 2p7z 2p95 2pcp 2pgz 2pk5
2pk6 2pql 2pqz 2psu 2psv 2pu2 2pvh 2pvj 2pvk 2pvl 2pwc 2pwd 2pwg 2pwr 2pyn
2q54 2q55 2q5k 2q63 2q64 2q88 2q89 2qbq 2qbr 2qbs 2qbu 2qci 2qd6 2qd7 2qd8
2qe4 2qfo 2qg0 2qg2 2qhm 2qhy 2qhz 2qi0 2qi1 2qi3 2qi4 2qi5 2qi6 2qi7 2qm9
2qmg 2qnn 2qnq 2qrl 2qt5 2qtg 2qu6 2r2m 2r2w 2r38 2r3t 2r3w 2r43 2r5a 2r5p
2r6w 2r6y 2ra0 2rcb 2ri9 2rkf 2rkg 2sim 2std 2tpi 2uwo 2uxz 2uy0 2v00 2v3u
2v95 2vh6 2vkm 2vnt 2vw5 2vyt 2z4b 2zb0 2zb1 3aid 3b4p 3b50 3b5r 3b65 3b66
3b67 3b68 3b7j 3be9 3bex 3bfu 3bgb 3bgc 3bgq 3bgz 3bra 3brn 3bu1 3buf 3bug
3buh 3bvb 3c2u 3cct 3ccw 3ccz 3cd0 3cd5 3cd7 3cda 3cdb 3cf8 3cj2 3cj4 3cj5
3ckp 3cs7 3cyw 3cyx 3d0b 3d1x 3d1y 3d1z 3d20 3d7z 3d83 3d94 3djk 3e5a 3e5u
3e64 3e92 3e93 3ebl 3eko 3ekr 3eqr 3f8c 3f8f 3gss 3gst 3jdw 3kiv 4ts1 5er1
5std 5yas 6std 7std

Supporting Figures

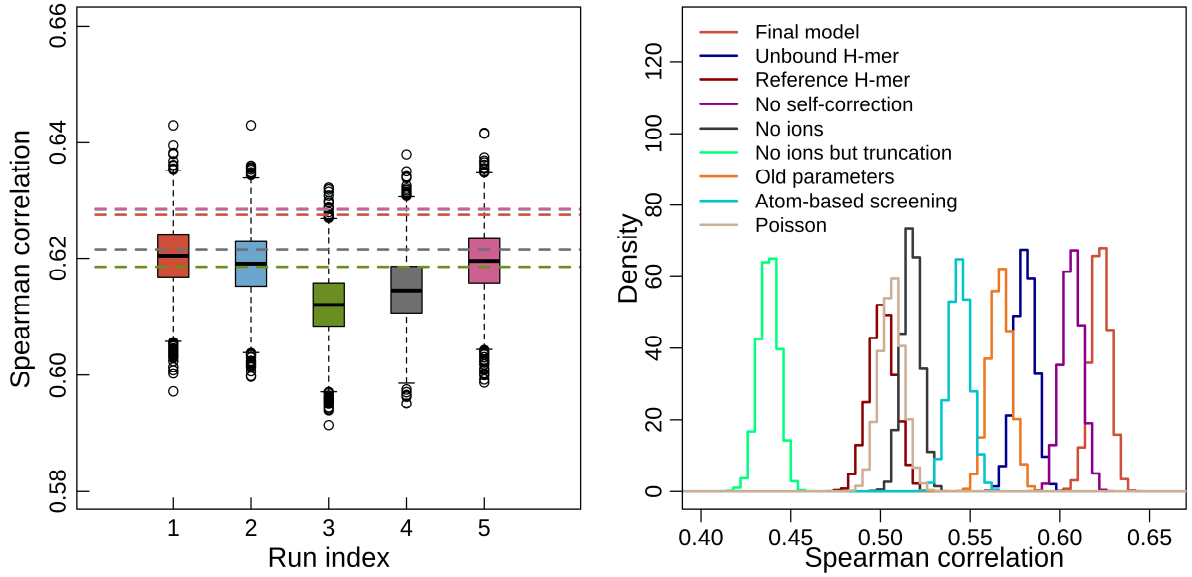


Figure S1. **Statistical analysis of data in Fig. 2 in the main text**. Here we are primarily interested in answering the question whether statistical errors arising from the use of a noisy molecular dynamics protocol can be responsible for the differences in correlation reported in Fig. 2. This is different from, for example, subsampling data to check for robustness with respect to individual data points. Because Spearman rank correlations have a weaker outlier sensitivity than linear correlation coefficients, we use rank correlations in this test. As mentioned in the main text, we reran the entire data set with the final model five times. This is computationally too expensive to do for all cases shown in Fig. 2. Thus, we extracted crude standard deviation estimates per data point from these five runs, restricting ourselves to the 742 complexes where all runs gave an interpretable result. The errors are clearly heteroscedastic, which we highlight in Fig. 3 in the main text. We then resampled the data points assuming zero-mean, normal errors of the inferred widths 5000 times. Each resultant data set was analyzed using the same threshold criteria as used for Fig. 2, and the Spearman correlation was recorded. **(a)** Rank correlation distributions for the five copies of the final model are shown as Tukey-rule boxplots. The horizontal lines are the correlation values for the same data without added noise (the blue one is hidden behind the pink one). The assumed mean per data point is taken as the actual value for each run independently, which means that we add errors to what is itself a noisy estimate. This leads to a slight drop in correlation. More importantly, the distributions resulting from this procedure appear consistent with the spread observed across the five independent runs. **(b)** Using the same errors per data point, other data sets were resampled, and the histograms of correlation values are shown. The data refer, in order, to the nine panels of Fig. 2. The only distribution overlapping with that of the final model is the one for leaving out the self-correction. This is expected since these data are largely identical (see main text). The lack of overlap between the histogram for the final model *vs*. those for the other cases strongly suggests that the differences in rank correlations are significant.

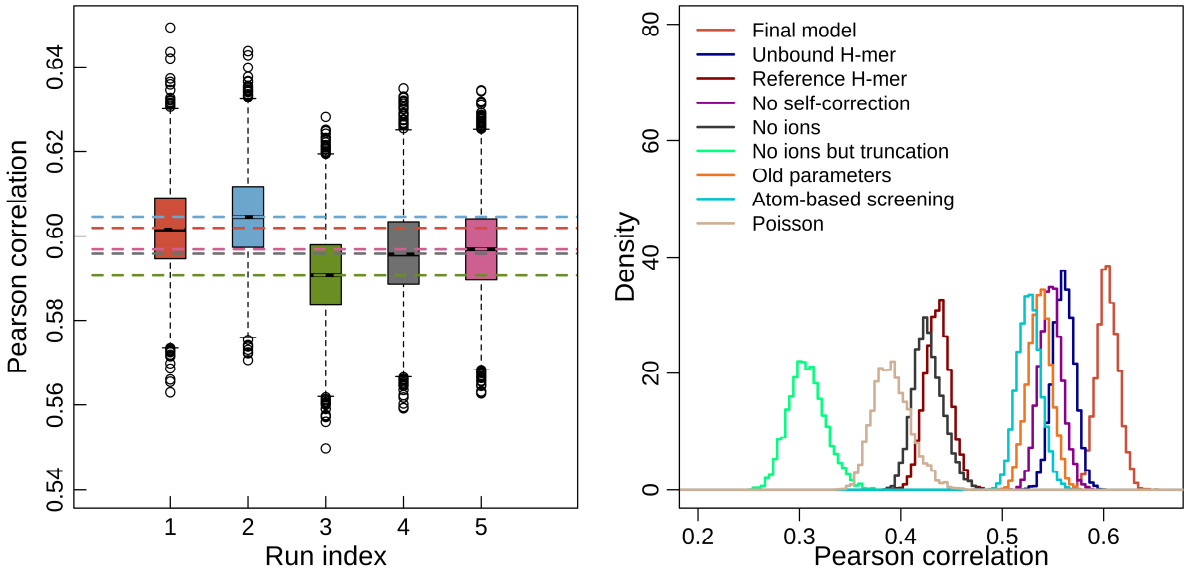


Figure S2. Robustness of Pearson correlation coefficients from bootstrapping. This figure is identical to Fig. S1 except that Pearson correlations are shown and that the underlying distributions were generated from 5000 bootstrap samples. Specifically, for each run, we subsampled the observed data as shown in Fig. 2 in the main text without replacement to ~80% (600 complexes). Under these settings, the conclusions are the same as for Fig. S1 indicating that the improvements highlighted in Fig. 2 are not spurious trends due to few complexes. We chose Pearson correlations for the subsampling to complement the analysis in Fig. S1.

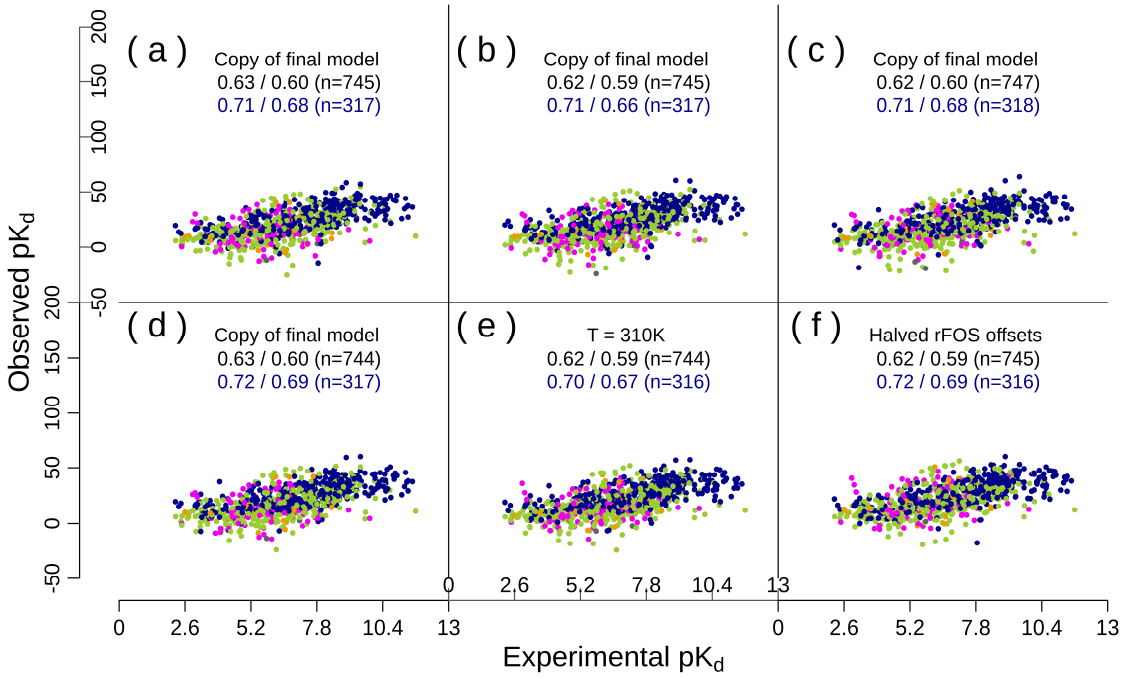


Figure S3. Robustness of data for the final model. This figure is analogous to Fig. 2 in the main text. **(a)–(d)** Data for the 4 copies of the final model mentioned in the main text. Differences across these copies are used to calculate the statistical errors highlighted in Fig. 3 in the main text and used in Fig. S1. **(e)** Data for a rerun of the final model where the simulation temperature was increased to 310 K throughout. The reference dipeptide values listed in Table S1 were not adjusted specifically for this setting as their mutual differences are expected to be largely insensitive to temperature. **(f)** Data for a rerun of the final model where the offsets on the reference free energies of solvation of charge groups in organic molecules and polymers are reduced from -15.0 to -7.5 kcal/mol (see Tables S1 and S3).

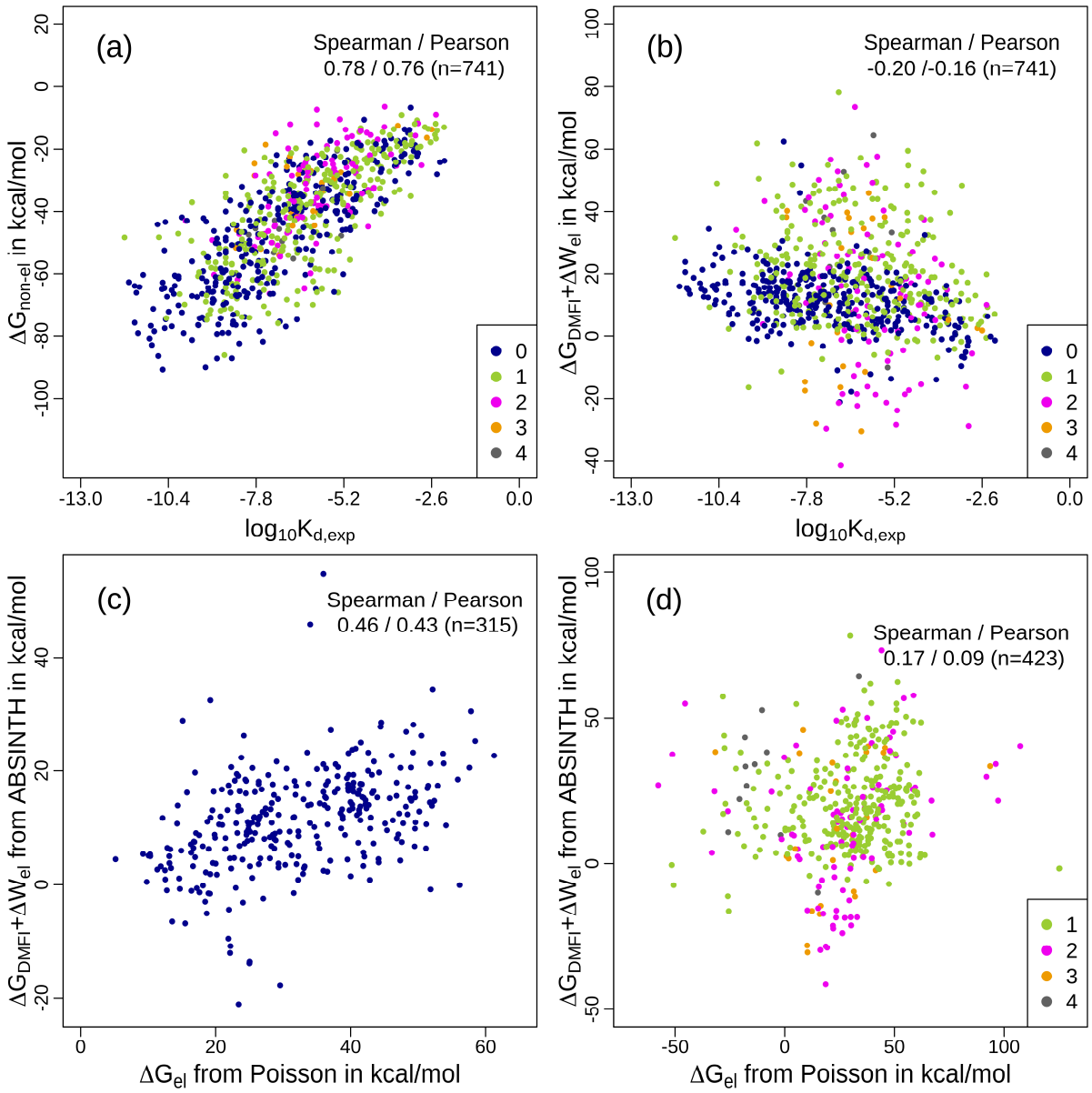


Figure S4. **Analysis of individual contributions to estimated binding free energies**. These plots are created from the same source data as Figs. 2(b) and 2(i) in the main text, *i.e.*, they are for the single microscopic equilibrium per complex involving the most likely H-mer(s) in the unbound state. The estimated binding free energies have been split into non-electrostatic and electrostatic/solvation contributions. For ABSINTH, the latter include nonpolar solvation, which is not a separable contribution in the ABSINTH/EEF1 paradigm. For Poisson, electrostatic contributions are the only available data, and they do not include contributions from nonpolar solvation. The color code is the same as in Fig. 2, *i.e.*, color signifies the number of charge groups carrying a nonzero net charge per ligand. **(a)** Correlation of experimental data with non-electrostatic contributions. **(b)** Correlation of experimental data with electrostatic/solvation contributions from ABSINTH. Note the consistent and expected anticorrelation for molecules with zero charge groups that is largely lost for the remaining ligands. **(c)** Correlation of electrostatic/solvation contributions between Poisson and ABSINTH data (based on Figs. 2(i) and 2(b), respectively) for molecules free of any charge groups with nonzero net charge. There is clear correlation, and the ranges of values are comparable. **(d)** The same as (c) for all other ligands yielding interpretable data. The dramatic deterioration of correlation is clear evidence that large differences between ABSINTH and Poisson arise due to contributions from charge interactions and not due to nonpolar solvation. Note that in (a) and (b) the directions of both axes are flipped relative to Fig. 2.

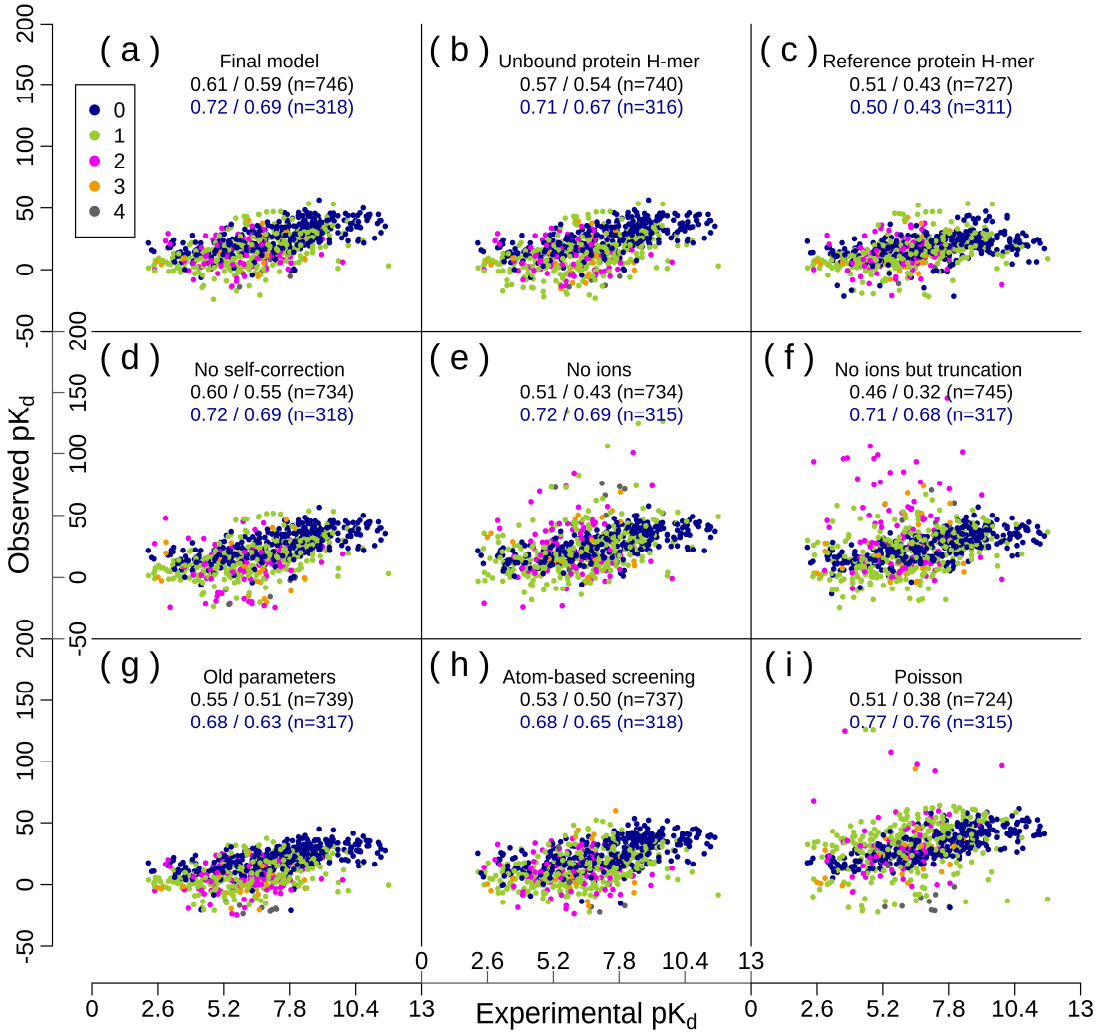


Figure S5. Analysis of the impact of using single-point energies calculated from minimized poses instead of trajectory averages. This figure is analogous to Fig. 2 in the main text. For every panel except (i), the change is that the predicted binding affinity for an individual combination of protein and ligand H-mers is calculated by eq. (S5) but using single-point energy values instead of trajectory averages. Single conformations for complex, protein, and ligand were obtained separately by minimization for a maximum of 2000 steps using a Broyden-Fletcher-Goldfarb-Shanno scheme according to Nocedal.[15] In (i), only the Lennard-Jones contributions are replaced in this way, and the Poisson-derived electrostatic and solvation components, which are obtained similarly (see S.4 above), are the same as in Fig. 2. Note that the differences to Fig. 2 are generally within the expected level of statistical errors (Fig. S1).

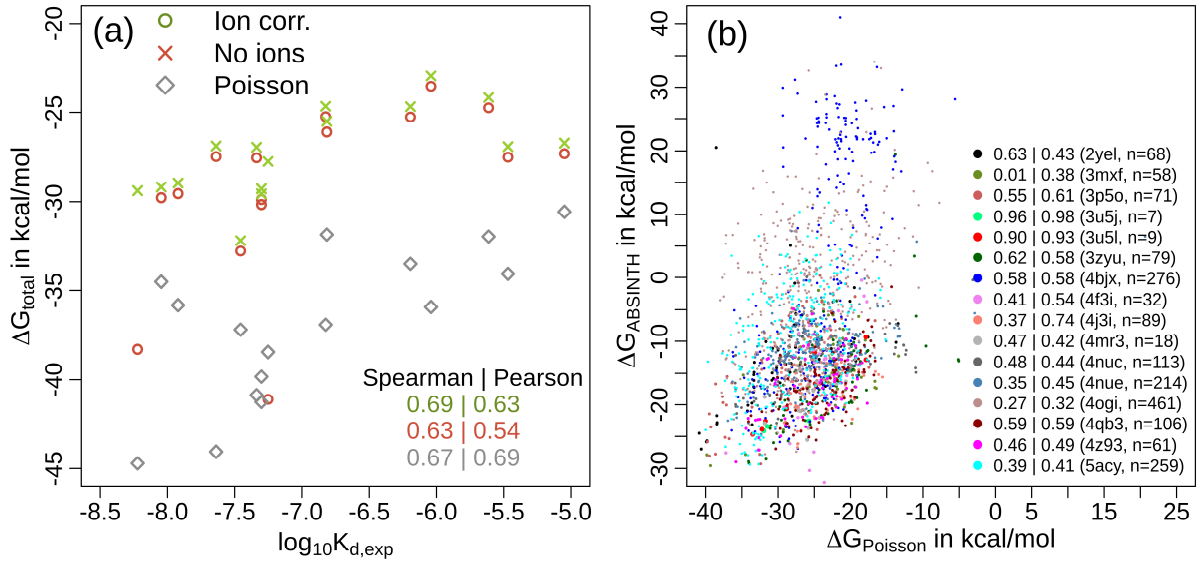


Figure S6. Further analyses for the enrichment data set. **(a)** The decadic logarithms of the experimental affinities (Table I in the main text) are plotted against the best energies across poses for the 16 positive controls. The color code is the same as that in Fig. 4 in the main text. Spearman and Pearson correlation coefficients are listed. **(b)** The total predicted binding free energies from the MM/P protocol are plotted against the corresponding ABSINTH values with ion corrections for individual poses obtained for the positive controls. Three points are cut from the plot for ease of visualization. Note that the number of available poses per positive control differs considerably. These numbers of valid poses per compound are listed along with the corresponding correlation coefficients (Spearman|Pearson) in the figure (color code in the legend). Here, Pearson correlation coefficients are sometimes significantly larger than Spearman ones (*e.g.*, for 4j3i or 3mxf), which is due to individual outlier poses having particularly favorable or unfavorable energies in both models. The overall correlation coefficients across all of these poses are 0.32 and 0.37 (Spearman and Pearson, respectively).

Supporting References

1. Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Automation of the CHARMM general force field (CGenFF) II: Assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **2012,** *52* (12), 3155-3168.
2. Vitalis, A.; Pappu, R. V., ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2009,** *30* (5), 673-699.
3. Engh, R. A.; Huber, R., Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr. A* **1991,** *47* (4), 392-400.
4. Sterling, T.; Irwin, J. J., ZINC 15 – Ligand discovery for everyone. *J. Chem. Inf. Model.* **2015,** *55* (11), 2324-2337.
5. Schärfer, C.; Schulz-Gasch, T.; Hert, J.; Heinzerling, L.; Schulz, B.; Inhester, T.; Stahl, M.; Rarey, M., CONFECT: Conformations from an expert collection of torsion patterns. *ChemMedChem* **2013,** *8* (10), 1690-1700.
6. Vitalis, A. CAMPARI website. http://campari.sourceforge.net/ (accessed July 31, 2020).
7. Vitalis, A.; Pappu, R. V., A simple molecular mechanics integrator in mixed rigid body and dihedral angle space. *J. Chem. Phys.* **2014,** *141* (3), 034105.
8. Fixman, M., Classical statistical mechanics of constraints: A theorem and application to polymers. *Proc. Natl. Acad. Sci. USA* **1974,** *71* (8), 3050-3053.
9. Andersen, H. C., Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980,** *72* (4), 2384-2393.

10. Menzer, W. M.; Li, C.; Sun, W.; Xie, B.; Minh, D. D. L., Simple entropy terms for end-point binding free energy calculations. *J. Chem. Theory Comput.* **2018,** *14* (11), 6035-6049.
11. Manning, G. S.; Ray, J., Counterion condensation revisited. *J. Biomol. Struct. Dyn.* **1998,** *16* (2), 461-476.
12. Vitalis, A.; Baker, N. A.; McCammon, J. A., ISIM: A program for grand canonical Monte Carlo simulations of the ionic environment of biomolecules. *Mol. Simul.* **2004,** *30* (1), 45-61.
13. Grochowski, P.; Trylska, J., Continuum molecular electrostatics, salt effects, and counterion binding—A review of the Poisson–Boltzmann theory and its modifications. *Biopolymers* **2008,** *89* (2), 93-113.
14. Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B., Constant-pH molecular dynamics simulations for large biomolecular systems. *J. Chem. Theory Comput.* **2017,** *13* (12), 5933-5944.
15. Nocedal, J., Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980,** *35* (151), 773-782.
16. Marchand, J.-R.; Dalle Vedove, A.; Lolli, G.; Caflisch, A., Discovery of inhibitors of four bromodomains by fragment-anchored ligand docking. *J. Chem. Inf. Model.* **2017,** *57* (10), 2584-2597.
17. Kelly, C. P.; Cramer, C. J.; Truhlar, D. G., Aqueous solvation free energies of ions and ion−water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton. *J. Phys. Chem. B* **2006,** *110* (32), 16066-16081.
18. Radhakrishnan, A.; Vitalis, A.; Mao, A. H.; Steffen, A. T.; Pappu, R. V., Improved atomistic Monte Carlo simulations demonstrate that poly-L-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks. *J. Phys. Chem. B* **2012,** *116* (23), 6862-6871.
19. Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L., Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solution Chem.* **1981,** *10* (8), 563-595.
20. Bashford, D.; Case, D. A., Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000,** *51* (1), 129-152.
21. Huang, J.; MacKerell, A. D., CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **2013,** *34* (25), 2135-2145.
22. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L., Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001,** *105* (28), 6474-6487.
23. Vitalis, A.; Caflisch, A., 50 Years of Lifson–Roig models: Application to molecular simulation data. *J. Chem. Theory Comput.* **2012,** *8* (1), 363-373.
24. Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V., Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. USA* **2010,** *107* (18), 8183-8188.
25. Choi, J.-M.; Pappu, R. V., Improvements to the ABSINTH force field for proteins based on experimentally derived amino acid specific backbone conformational statistics. *J. Chem. Theory Comput.* **2019,** *15* (2), 1367-1382.
26. Arnon, Z. A.; Vitalis, A.; Levin, A.; Michaels, T. C. T.; Caflisch, A.; Knowles, T. P. J.; Adler-Abramovich, L.; Gazit, E., Dynamic microfluidic control of supramolecular peptide self-assembly. *Nat. Commun.* **2016,** *7* (1), 13190.
27. Greenidge, P. A.; Kramer, C.; Mozziconacci, J.-C.; Wolf, R. M., MM/GBSA binding energy prediction on the PDBbind data set: Successes, failures, and directions for further improvement. *J. Chem. Inf. Model.* **2013,** *53* (1), 201-209.