JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

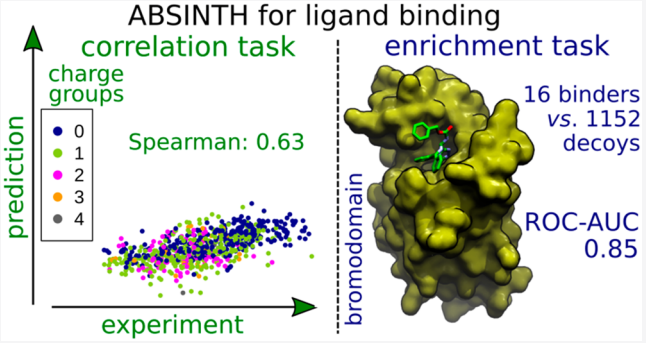
An ABSINTH-Based Protocol for Predicting Binding Affinities between Proteins and Small Molecules

Jean-Rémy Marchand, Tim Knehans, Amedeo Caflisch, and Andreas Vitalis*

Cite This: https://dx.doi.org/10.1021/acs.jcim.0c00558 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The core task in computational drug discovery is to accurately predict binding free energies in receptor−ligand systems for large libraries of putative binders. Here, the ABSINTH implicit solvent model and force field are extended to describe small, organic molecules and their interactions with proteins. We show that an automatic pipeline based on partitioning arbitrary molecules into substructures corresponding to model compounds with known free energies of solvation can be combined with the CHARMM general force field into a method that is successful at the two important challenges a scoring function faces in virtual screening work flows: it ranks known binders with correlation values rivaling that of comparable state-of-the-art methods and it enriches true binders in a set of decoys. Our protocol introduces innovative modifications to common virtual screening workflows, notably the use of explicit ions as competitors and the integration over multiple protein and ligand species differing in their protonation states. We demonstrate the value of modifications to both the protocol and ABSINTH itself. We conclude by discussing the limitations of high-throughput implicit methods such as the one proposed here.



1. INTRODUCTION

The discovery of new molecules that bind to macromolecules of biological and therapeutic interest is a complex task. A great wealth of data from biochemical and biophysical experiments have been accumulated in order to support the development of potent and selective binders of proteins and other macromolecules.[1] The idea of a high-throughput screen plays an important role in the drug discovery pipeline.[2] While the data are often difficult to analyze,[3,4] the generality of the approach retains its fundamental appeal. In cases where experimental assays are unavailable, too expensive, or too unreliable, virtual screening is a viable alternative.[5,6] In particular, the utilization of existing structural information for both the target and ligands appears to be able to elevate enrichment rates; see Marchand et al.[7] for an example. Dozens of software packages for docking have been developed since the 1980s, along with an even larger number of scoring functions.[6,8−11] The computational efficiency of these programs continues to be a point of emphasis because the growth in computational resources is offset by a similar growth in the available chemical space.[12]

The scoring functions implemented in the docking software aim to choose correct poses and to rank compounds according to their binding affinities.[9] Blind prediction challenges have shown that they perform reasonably well at the first task, that is, identify the correct binding pose of an active compound in a set of decoy poses for the same compound, but often perform poorly at ranking compounds according to their measured binding affinities or at distinguishing active compounds from inactive ones.[13,14] To circumvent this limitation, the poses produced by a primary docking campaign are frequently "rescored", that is, a different and putatively more accurate method is used for predicting the affinity of the compounds to the target given these poses. Popular rescoring protocols combine common biomolecular force fields with a continuum treatment of (de)solvation.[15−17] It is of course crucial to accurately describe the transfer of a ligand from an aqueous environment to a protein environment in order to reduce the false positive rate in a virtual screen.[18−21]

The majority of implicit solvent-based scores calculate the electrostatic free energy of solvation by solving the Poisson equation or its Generalized Born (GB) approximation.[15,22−24] A major caveat of these models is that they do not account for the nonpolar contributions to the free energy of solvation. Correction terms exist, but they show poor correlation with experimental results, and it is not clear whether they improve

Received: May 19, 2020
Published: September 8, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the quality of binding affinity calculations.[25−30] ABSINTH is another type of implicit solvent model, introduced in 2008,[31] which addresses this limitation by a different formalism for the solvation energy, inspired by EEF1.[32] In ABSINTH, the solvation component is treated as a direct mean-field interaction (DMFI) between the solute and the solvent and a partial screening of the interactions between atomic partial charges. The DMFI is calculated from the partial desolvation of the reference solvation groups, which carry experimental free energies of solvation (rFOS). The rFOS values originate from experimental data, usually vapor pressure experiments. Because the rFOS are transfer free energies, they include all physical terms of the solvation energy, that is, conformational entropy, polar, and nonpolar contributions. Poisson and GB models, on the other side, calculate solely the polar component of the solvation energy. In addition, the macroscopic treatment of electrostatics in these models can be error-prone because of the requirement to define a sharp boundary between the low-dielectric medium (solute) and the high-dielectric solvent.[33−36] ABSINTH relies on an original description of the solvent exposure of solvation groups, based on the calculation of atomic desolvation via solvent-excluded volumes. This approach removes the requirement to define an explicit boundary of this type.

The ABSINTH model has been applied with particular success to other systems where solvation effects are fundamental. For example, it describes the conformational equilibria of disordered proteins accurately which are often predicted as overly collapsed in simulations with other force fields.[37−39] The paradigm underlying EEF1 and ABSINTH has also been shown to be superior in relative free energy calculations for proteins,[40] and even peptidic systems carrying many charge groups can be modeled in accordance with experimental knowledge.[41−43] It is a particular feature of ABSINTH that explicit ions can be part of the simulations. From these prior applications, it is a straightforward conjecture that the ABSINTH solvation model might be generalizable to describe the physics of drug−macromolecule binding. We thus decided to develop methods to extend the support of ABSINTH and CAMPARI[44] to organic molecules and to design a protocol that would enable predictions of binding affinities from the model.

The rest of the manuscript is structured as follows. First, we briefly summarize the key components of the ABSINTH model (2.1). Next, we describe the aforementioned developments and the test data (2.2−2.5) and provide statistics on the parameterization of small molecules (3.1). Our results demonstrate that the model is able to rank the affinities of known binders (3.2) and to identify such molecules from a pool of decoys (3.3). We finish by discussing further avenues for improvement (4). Notably, it is not useful to evaluate the generalized ABSINTH model for its ability to predict rFOS values of organic molecules, which is a common test for Poisson-like continuum models.[45] This is because ABSINTH uses experimental rFOS values directly and because a validation based on solvation energies is not sufficient to justify the use of a model to rank drug−macromolecule complexes.[30,46]

2. METHODS AND THEORY

2.1. ABSINTH Model and Force Field. The ABSINTH model describes, in an implicit manner, the effects of water on explicit solutes. The original model focused on polypeptides but the paradigm generalizes to all chemotypes for which the following hold. (1) Partial charges are available that can be grouped reasonably into sets of covalently bound atoms; (2) the molecules can be decomposed into substructures corresponding to small molecules with known rFOS values; (3) the conformational preferences for electronically constrained bonds are known or deducible.

To describe the DMFI of water and solute, ABSINTH treats the solvation process holistically, which is an idea pioneered by the EEF1 model[32]

$$\Delta G_{\mathrm{DMFI}} = \sum_{k}^{K} \Delta G_{\mathrm{rFOS}}^{k} \sum_{j}^{J(k)} \lambda_{j}^{k} \upsilon_{f,j}^{k} \tag{1}$$

In eq 1, the ΔG_{rFOS} terms are the rFOS values of the K solvation groups, each containing $J(k)$ atoms, that the (macro)molecule has been decomposed into. The λ are weight factors (usually $J(k)^{-1}$), while the υ_{f} are atomic solvation states. The latter are computed with generalized sigmoidal functions from the per-atom fractions of how much of the surrounding volume is solvent-accessible. We did not change the functional form or parameters for these equations in the present work, and the full descriptions can be found as eqs 2−4 for υ_{f}, and Table 1 for the group assignments as well as the λ in the original work.[31] As the ΔG_{rFOS} are experimental free energies, the DMFI contribution is an effective energy incorporating solvent entropy terms.

While polypeptides are decomposed into building blocks straightforwardly by hand, an automatic and general strategy is needed for the vast chemical space offered by small molecules. The logic behind the group decomposition is that sufficiently independent substructures contribute additively to transfer properties including rFOS values. Thus, the sum of rFOS values across substructures is the theoretically maximal rFOS for the entire molecule if all groups were in a fully solvated state.[47−49] Given the available databases of rFOS values,[50−53] which have been expanded in recent years primarily through the SAMPL challenges,[54−58] we extend ABSINTH here by a cheminformatics-based method to parameterize the DMFI for small molecules (2.2.1 and 2.2.2 below) along with substantial changes to CAMPARI[44] available in version 4 (to be released in 2020).

ABSINTH uses solvent-accessible volumes not only to compute the DMFI but also to implement a GB-like correction to electrostatic interactions[31,59]

$$\begin{aligned} W_{\mathrm{el}} = & \sum_{l}^{L} \sum_{x}^{X(l)} \sum_{m>l}^{L} \sum_{y}^{Y(m)} d_{lm} s_{lm} \frac{q_x^l q_y^m (1 - a\upsilon_{\mathrm{s},x}^l)(1 - a\upsilon_{\mathrm{s},y}^m)}{4\pi\varepsilon_0 |\mathbf{r}_x^l - \mathbf{r}_y^m|} \\ & + \sum_{l}^{L} \sum_{m>l}^{L} (1 - d_{lm}) \frac{Q_l Q_m (1 - a\upsilon_{\mathrm{s},l})(1 - a\upsilon_{\mathrm{s},m})}{4\pi\varepsilon_0 |\mathbf{r}_l - \mathbf{r}_m|} \\ a = & (1 - \varepsilon_{\mathrm{w}}^{-1/2}) \end{aligned} \tag{2}$$

Equation 2 describes the screened electrostatic interactions at the resolution of atoms (quadruple sum) or at the resolution of charge groups (double sum). There are L charge groups in the system. The number of atoms in each group is denoted as either X or Y. Atoms have associated properties of q, their partial charges, of υ_{s}, their solvation states for screening, and of $\mathbf{r}$, their position vectors. The υ_{s} are computed in the same way as the υ_{f} in eq 1 but using different mapping parameters. These

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

parameters were 0.25/0.1 and 0.5/0.9 for steepness/midpoint parameters for the υ_f and υ_s, respectively. The thickness of the solvation shell for calculating solvent-accessibility was 5 Å. These choices are the same as those in the reference publication.[31] Importantly, the υ_s can be calculated per atom or averaged across the atoms in each charge group. The former is the original model and referred to as "atom-based screening" below. The latter has been used for highly charged polypeptides[41] and is referred to as "group-based screening." We evaluate both in this work. The vacuum permittivity and relative dielectric constant in water are denoted as ε_0 and ε_w, respectively. The factor s_{lm}, which takes values of 0 or 1 and is conformation-independent, represents the exclusion rules implemented in ABSINTH. Unlike other force fields, short-range electrostatic interactions are depleted beyond the normal exclusion of pairs of atoms separated by two bonds or less, and this depletion is the origin of a correction we propose here (see Supporting Information, eq S11 and S.6). Specifically, s_{lm} is 1 if and only if every pair of atoms from charge groups l and m is separated by at least three bonds that allow the two atoms to move relative to each other. The last condition is not fulfilled in rigid rings; for example, see Supporting Information, Scheme S1. Equation 2 also represents cutoffs: the factor d_{lm}, which depends on conformation and also takes values of 0 or 1, skips the atom-based sum if the distance between reference atoms is larger than 12 Å. The double sum is a correction for charge groups with net charges Q both different from zero. If and only if they are not accounted for in the first sum, $1\text{-}d_{lm}$ is 1, and a pointwise monopole interaction is calculated. The monopole is represented by the atom in the charge group that is closest to its center of charge (indicated by the subscripts l and m to υ_s and $\mathbf{r}$).

Two of the most common biomolecular force fields, AMBER and CHARMM, have automatic parameterization pipelines for small molecules, called GAFF[60] and CGenFF.[61] We describe next how we extend ABSINTH to a general force field for small molecules reliant on CGenFF.

2.2. Generation of Parameters for Organic Molecules. *2.2.1. Reference Experimental Data.* In the ABSINTH approach, the solute is decomposed into a set of solvation groups. These groups correspond to model compounds with an experimentally measured rFOS. The backbone and side chain units of a protein are a straightforward decomposition of the polymer. We extended this idea to small molecules by compiling a database of experimentally measured free energies of solvation of model compounds. We extracted from the literature 657 free energies of solvation of neutral organic compounds, 21 molecules with a negative charge, 42 with a positive charge, and 49 corrective values.[50−58,62−67] Corrective values originate from the dissection of the experimental solvation free energies of sets of model compounds into contributions from individual atoms or minimal functional groups[50] and are necessary to extend the coverage of the chemical space beyond direct combinations of the 720 molecules of the database, for example, to assign parameters for small linkers between rings.

2.2.2. Deconstruction of Organic Compounds. The automatic assignment of reference free energies of solvation to the small molecules relies on the RDKit,[68] here release Q1 2018, for chemistry parsing and on NetworkX, here version 1.11. The algorithm is as follows: (1) identification of all substructures from the database of rFOS in the target molecule; (2) creation of a graph, in which each node is a substructure, with edges connecting all nonintersecting nodes, that is, nonoverlapping substructures; (3) selection of cliques that cover the maximum number of atoms in the molecule; (4) selection of the cliques with the fewest numbers of nodes, that is, cliques carrying the maximal description of the small molecule with the minimal count of substructures. If more than one possibility exists at this stage, we calculate the sum of hydration free energies for each combination of substructures and select the combination corresponding to the median sum. Atoms between fused rings are treated separately. Fusion atoms can be accounted for by two different substructures that correspond to the two fused rings, that is, an overlap between substructures is possible for fusion atoms. If the molecule is not complete, the last step is to assign corrective values to the remaining atoms to complete the description of the molecule.

2.2.3. Other Force-Field Parameters. Atomic partial charges are assigned according to the CHARMM generalized force-field (CGenFF) paradigm with the CGenFF software.[61] CGenFF partial charges are convenient because chemical groups in the molecule add up as net neutral groups, a feature that is desirable for the ABSINTH force field. We also tried AM1-BCC partial charges[69] from semi-empirical calculations with the AMSOL package[70] but discontinued their use because of the lack of efficient net neutral grouping possibilities. Van der Waals parameters are refined here while maintaining the spirit of the original ABSINTH publication (see Supporting Information, S.6).[31,43,71] Bonded parameters for simulations in rigid-body/torsional space are automatically generated from atom types, connectivity and input geometry information (see S.1 for details).

2.3. Validation Set for Ranking Experimentally Confirmed Binders. *2.3.1. Compilation of the Data Set.* The PDBbind database contains complexes of organic molecules and proteins with high-quality 3D structures and available binding affinity data. The "refined set" of the PDBbind is filtered to exclude complexes unfit for training a scoring function.[72,73] Greenidge et al. pruned the PDBbind refined set further by enforcing a number of additional rules for testing a GB-based scoring function in 2013.[22] This subset was designed to be as clean as possible for pharmaceutical applications and contains "drug-like" ligands with binding affinities to their targets ranging from the low-nanomolar to high-millimolar ranges. In total, their set contains 855 high-quality structural and affinity data of drug-like molecules in complex with proteins. We could not use the entire set for two main reasons: first, some complexes had to be excluded because of remaining issues with the structural data in the binding site (*ca.* 40), while some had to be excluded because we could not generate the parameters required for the ligand (for example, all phosphorous-containing ligands). In the end, we retained a set of 754 complexes. The data set contains ∼200 different proteins, the rest being proteins that are present multiple times with different ligands. The median number of complexes per protein is 1, the average 3.6, and the maximum 139. Eight proteins are present more than ten times in the data set. The protein with the largest number of ligands is HIV-1 protease, with 139 occurrences, which are not all identical, that is, some are proteins with mutations in the binding site. This is followed by trypsin with 63 ligands, thrombin with 32, and so forth. The complete list of considered PDB codes is provided as Supporting Information, see S.7. The 754 complexes contain 577 unique ligands. Ligand duplications occur for simple endogenous ligands such as glutamate (1II5 and 1XFF) or for

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

complexes of the same inhibitor to different mutants of the same protein, for example, HIV-1 protease versus saquinavir.[74]

2.3.2. Preparation of the Complexes. The complexes were downloaded from the PDBbind database, including ligands with bonding information. Nonprotein atoms were removed. The CHARMM36 force field was used for the parameters of the protein[75] and CGenFF 3.0 for the ligands' partial charges.[76] Missing atoms were added except those in loops, which were generally cut. It was a requirement that such missing loops are far away from the binding site (see 2.3.1). Reconstructed side chains were relaxed through a two-step Monte Carlo procedure in the CAMPARI package.[44] This procedure automatically detects dihedral angles subject to high forces and selectively relaxes them through pivot-style dihedral angle moves on the respective χ-angles. The conformation of the backbone was fixed throughout. We performed two such runs successively with thresholds of 50.0 and 10.0 for keyword FMCSC_TMD_RELAX in CAMPARI and both using 500 elementary steps per side chain.

2.3.3. Generation of Different Protonation States. Many complexes in the data set contain buried side chains of aspartates, glutamates, lysines, and histidines near the ligand, possibly undergoing protonation state variations that could strongly influence force field-estimated binding free energies. All possible protonation states of the aforementioned residues were enumerated if (1) side chain atoms were close to ligand atoms, with distance thresholds of 5.1 Å for Asp_{CG}, Glu_{CD}, Lys_{NZ}, and 6.1 Å for His_{CG} to any ligand atom; and if (2) their burial was over 75%, according to propka 3.1.[77] These criteria led to the generation of 7196 protonation states of the 754 proteins (denoted H-mers below), with a minimum of H-mers per protein of 1, a median of 4, and a maximum of 288. In addition, 16 ligands had clear ambiguities in their protonation states, and the 2−3 possibilities per molecule led to a net total of 7457 protein−ligand H-mer complexes. Equilibrium populations for the H-mers of these 16 ligands at pH 7 were estimated with the help of the calculator plugins of Marvin 15.8.17, 2015, ChemAxon (www.chemaxon.com), while for the rest only the dominant H-mer was retained.

2.3.4. Microscopic Binding Equilibria. For each of the 7457 protein−ligand complexes, we aimed to determine a score representing the microscopic binding equilibrium between bound and unbound forms. For ligands featuring moieties carrying a net charge, we propose here to include a correction based on the idea of displacing inorganic ions. This means that the microscopic equilibrium is modified to include explicit inorganic ions in the unbound state, which are free to occupy the binding site. Conversely, in the bound state, these ions are assumed to be in the bulk.

As shown in the results, this idea of a competitive displacement is required to make the final predicted binding affinities homogeneous across ligands of different charge. The revised equation for an estimated binding free energy ΔG_b contains ensemble averages of potential energies. They can be understood as approximations to the free energy obtained by truncating the cumulant expansion of the Helmholtz free energy after the first term.[78] The approximation implies a neglect of explicit entropy terms and looks as follows

$$\text{protein/ions} + \text{ligand} \rightleftharpoons \text{complex} + \text{ions}$$

$$\Delta G_b \approx \langle U(\text{complex})\rangle - \langle U(\text{protein/ions})\rangle - \langle U(\text{ligand})\rangle + U(\text{ions}) \quad (3)$$

In eq 3, U is the internal energy, which is the sum of DMFI, see eq 1, screened electrostatics, see eq 2, bonded, see S.1, and solute−solute Lennard-Jones terms. U is an effective energy incorporating solvent entropy term according to the ABSINTH implicit water model (see 2.1). "Ions" refers to explicit counterions, for which we use potassium and chloride ions. The number of these ions present is matched to the numbers of detected charge groups in the ligand that carry an integer charge of +1 or −1. Angular brackets indicate ensemble averages (see below). The last term in the second row of eq 3 is not an ensemble average because we assume a dilute, bulk reference state for which the ABSINTH energy of simple ions is known analytically. The procedure is described in detail in the Supporting Information (S.3).

2.3.5. Data Generation. Equation 3 states that binding free energies are estimated from ensemble averages of energies of protein−ligand complexes, proteins (possibly with explicit ions), and ligands. Of course, it is not feasible to consider all degrees of freedom, so we required an algorithm that could sample relevant binding site residues along with the ligand while keeping everything else fixed. Here, all atoms but those of the ligand and those in side chains of residues within 6 Å of ligand atoms were frozen. The ABSINTH paradigm rests on the assumption that sampling is performed in a rigid-body/torsional space (see S.1 and S.2). Thus, we used here the internal coordinate space integrator of Vitalis and Pappu[79] as implemented in CAMPARI. The production data were 10 ps molecular dynamics (MD) simulations at 250 K, with the latter half used for deriving the ensemble averages in eq 3. Note that this is a simulation temperature for an implicit solvent model undergoing no phase transitions. The value was initially chosen as a compromise between stability and sampling efficiency but not optimized thereafter. Prior to the production runs, in very few cases, an additional Monte Carlo relaxation of side chains experiencing large forces was triggered (compare 2.3.2). For complexes and proteins, the probability of simulation crashes was further reduced by short preproduction runs at 2.5 K (effective minimization). The data from these preproduction runs were discarded, and the final structure simply served as the starting point for the production MD. Additional details relevant for this protocol are given in the Supporting Information (see S.2). In few cases, the ligand might move relatively far away from the binding site even in a 10 ps run. However, we maintained all data as relevant up to a threshold of 10 Å (see 3.1).

2.3.6. Calculation of the Apparent pK_d. There is only a single experimental value for the binding affinity of a given protein−ligand complex. This is in contrast to the individual microscopic binding equilibria for particular combinations of H-mers, for which we estimate binding free energies as described above. We thus need to combine these data. To do so, we consider the total equilibrium across all bound and unbound forms of all considered combinations of ligand and protein H-mers. Once equilibrium concentrations are determined, the apparent constant can be estimated as follows:

$$K^*_{obs} = (K^*_{d,obs})^{-1} = \frac{\sum_I^{NP}\sum_J^{NL}[P_IL_J]}{\sum_J^{NL}[L_J]\sum_J^{NL}\sum_I^{NP}[P_I/ions_J]}$$

$$k^*_{IJ} = \frac{[P_IL_J]}{[L_J][P_I/ions_J]} = \exp(-\Delta G_b/k_bT) \quad (4)$$

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

In eq 4, we represent *NP* protein H-mers (P1, P2, etc.) and *NL* ligand H-mers (L1, L2, etc.), and the square brackets denote concentrations. The asterisks indicate that not all reference state corrections are applied here, which affect numerical values but not correlation coefficients. In total, there are up to *NP*·*NL*·2 + *NL* unique species in the equilibrium for K^*_{obs} (free ligand H-mers, free protein H-mers, and complexes). Protein species may be formally multiplied because of differences in explicit ions, $ions_J$, considered for different ligand H-mers (see 2.3.4 and S.5 in the Supporting Information for details). Consequently, the double sum in the denominator will be simplified if the ion sets used for different ligand H-mers are not all different. Equation 4 implies a choice of total ligand and protein concentration, which are free parameters. The remaining parameters required to calculate the concentrations in eq 4 are individual microscopic equilibria. Here, we choose three types. The most important type are the microscopic binding equilibria for specific H-mer complexes, which connect a unique bound state with its corresponding unbound state, see 2.3.4. For example, for a protein H-mer *I* and a ligand H-mer *J*, the equilibrium constant is calculated as k^*_{IJ} in eq 4 where ΔG_b is calculated from eq 3. Second, we derive equilibria for conversions between free ligands derived from Marvin 15.8.17, 2015, ChemAxon (www.chemaxon.com), see 2.3.3. Third, we use the protein-only simulations (2.3.5 and S.2) to estimate protonation free energies for individual protein side chains within the ABSINTH paradigm. This follows the logic of constant pH simulation methodologies.[80] The method and required reference data are presented in the Supporting Information, S.4 and Table S1, respectively.

We developed and used an in-house R script to combine these microscopic parameters and solve the underlying system of equations for every complex. This provides us with pK_d values according to eq 4, which we compare, in terms of ranking and correlation, to their experimental counterparts. For comparison data obtained with a Poisson model, we could not follow the same approach as in 2.3.5 and directly above. Instead, the most likely H-mer in the unbound state in the ABSINTH model was identified, and a hybrid model was constructed using electrostatic binding free energy predictions from single-point Poisson calculations (see Supporting Information, S.4).

2.4. Validation Set for Predicting Binders. *2.4.1. Compilation and Preparation of the Data Set.* We extracted from the literature 16 ligands of the first bromodomain of bromodomain-containing protein 4 (BRD4(1)), with binding affinities, that is, K_d measured with isothermal calorimetry experiments, ranging from 6 nM to 9 μM, and high-resolution crystal structures (Table 1). Fifty decoys per active compound were generated with the DUD-E webserver,[81,82] each decoy having similar physicochemical properties but dissimilar 2D topology to its corresponding active compound. Decoys were prepared and placed in the binding pocket of 2YEL similarly to previous work with bromodomains.[7,16,83,84] In short, compound H-mers were predicted by ChemAxon, and conformers of the 1152 H-mers with an estimated occupancy of at least 25% were generated using the RDKit ETKDG algorithm (50 runs, 0.5 Å of diversity threshold).[85] The 41,094 conformers were docked in the acetylated lysine pocket with rDock (20 runs, with 6 explicit waters),[86] leaving after clustering 155,076 docked poses to evaluate with either ABSINTH or a Poisson-based scoring function. Active compounds with crystal structures were placed in the binding pocket of 2YEL by superimposition, with no major clashes arising from this procedure. For ABSINTH, all poses underwent the molecular dynamics protocol described in S.2 with the exception that the protein was rigid. The Poisson procedure is described next.

Table 1. True-Positive Ligands Selected for the Enrichment Analysis in the Bromodomain BRD(1)[a]

molecule name	PDB code	K_d in μM	charge	refs
I-BET858	5ACY	0.006	+1	87
I-BET-151	3ZYU	0.009	0	88
RX37	4Z93	0.012	0	88
I-BET726	4BJX	0.023	−1	89
MS417	4F3I	0.036	0	90
GW841819X	2YEL	0.046	0	91
I-BET762	3P5O	0.050	0	92
(+)-JQ1	3MXF	0.050	0	93
BI-2536	4OGI	0.056	+1	94
MS267	4NUE	0.150	0	95
RVX-OH	4MR3	0.153	0	96
BzT-7	3U5L	0.640	0	97
Ms435	4NUC	0.910	0	95
alprazolam	3U5J	2.460	0	97
Olinone	4QB3	3.400	0	98
RVX-208	4J3IS	8.930	0	99

[a]Note that rDock poses were generated only for the neutral forms of I-BET858, I-BET726, and BI-2536. While we considered both neutral and charged forms for scoring those poses (hydrogens generated using OpenBabel 2.4.1), these three ligands are likely disadvantaged relative to the decoys.

2.4.2. Estimation of Poisson-Based Binding Free Energies. The minimization of the binding poses and their scoring was performed similarly to a previously described protocol, applicable in the context of a high-throughput docking campaign.[7,16] Explicit water molecules were removed, and the ligand was minimized for 500 steps of steepest descent and 10,000 steps of conjugate gradient with a convergence criterion of 0.01 kcal mol^{-1} $Å^{-1}$. The protein was fixed throughout. Here, we used CGenFF 4.0 for the ligand parameters rather than 3.0 (compare 2.3.2) because of licensing issues. Binding free energies were evaluated upon minimization as the rigid approximation of the difference of the energy of the complex protein−ligand minus the energies of the isolated protein and isolated ligand. The removal of water molecules was done to allow a straightforward comparison between the two methodologies. The assumed low (protein) dielectric was 4.0.

2.5. Implementation and Availability. The primary implementation platform for the work in this article is CAMPARI.[44] As mentioned above, version 4 will be made available in 2020. A stable development version can be obtained directly from the authors before that. This includes the parameter updates listed in the Supporting Information, Tables S3−S7. The decomposition of the organic molecules (2.4.2) is performed by a Python script and requires a simple text file with the fragment database (2.4.1) as an additional input. The script to integrate results for different H-mers is coded in R and contains the data in Table S1. Along with all of the aforementioned files, required run input files (shell, SLURM, and CAMPARI key-files) can be obtained from the authors upon request.

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

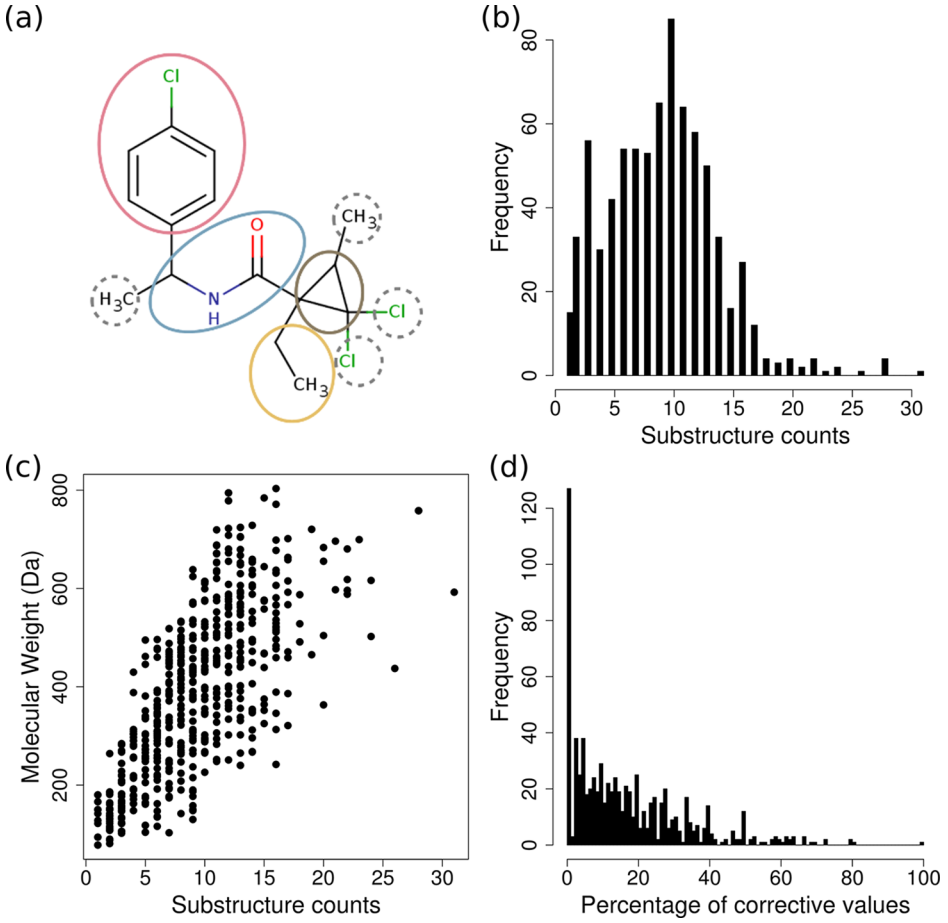


Figure 1. Illustration and analysis of the deconstruction algorithm. (a) An example organic molecule is decomposed into solvation groups. Circles represent different solvation groups as identified from the database of rFOS values. Circles drawn in dashed lines are solvation groups represented by corrective values. The deconstruction of the molecule into four solvation groups and four corrective values is the solution among all cliques that covers the largest number of atoms (excluding corrective values) by the minimum number of substructures (4). (b) Distribution of the number of substructures (corresponding to solvation groups) that were used to parameterize the ligands in the set selected for the ranking test, see 2.3.1. (c) Scatterplot of the count of substructures and molecular weight of the compounds. (d) Distribution of the fraction of heavy atoms in the molecule that are part of corrective values rather than proper fragments.

3. RESULTS AND DISCUSSION

3.1. Free Energy of Solvation Parameters for Small Molecules. The deconstruction of small organic compounds as "polymers" of building blocks is nontrivial. Drug-like molecules are diverse, originating both from natural compounds, with complex chemistry, and, on the contrary, from simple scaffolds and simple chemistry. We aimed to maximize the use of direct experimental measurements in the assignment of solvation groups. Our database of rFOS contains 720 values for molecular fragments and 49 corrective values. The algorithm is designed to assign the minimum amount of corrective values and automatically find the largest substructures with rFOS data, Figure 1a. We investigated the quality of the deconstruction algorithm on the ligands of the PDBbind validation set (see 2.3.1). The minimum number of substructures per ligand, including correction factors, is 1, that is, complete description of the (fragment) molecule by a single solvation group, and the maximum is 31. The median is 9, as well as the mean, Figure 1b. As expected, there is a natural bias toward more substructures for larger compounds, Figure 1c. Corrective values are used sparsely, Figure 1d, with minimum, median, and mean uses of 0, 12.5, and 17%, respectively, for the fraction of heavy atoms in the molecule that are not included in proper molecular fragments but instead described by corrective values. This level of the use of corrective values is expected to impact the resultant rFOS values only slightly. For 791 ligands from the Greenidge et al.[22] set, the mean, relative unsigned difference is ∼15% if we *only* use corrective values instead of relying on the complete database, while the correlation coefficients (Spearman/Pearson) are both >0.99. The deviation is almost exclusively toward more negative rFOS values (the signed difference is *ca.* −15%).

3.2. ABSINTH Ranks Known Actives as Good as State-Of-The-Art Techniques. We investigated the ability of a scoring function based on the ABSINTH solvation model to rank known ligands according to their experimental binding affinities to their protein targets. There are several deviations from common scoring protocols and from the original ABSINTH model, which we included and which are evaluated below. A short description is found in 2.3. Because the methodological details are quite expansive, most of them are found as Supporting Information (sections S.1−S.6).

The salient aspects are summarized as follows. For each selected complex (2.3.1), we analyzed the protein's ligand binding site to derive a list of residues that could possibly exist in multiple protonation states. From this list, we prepared,

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

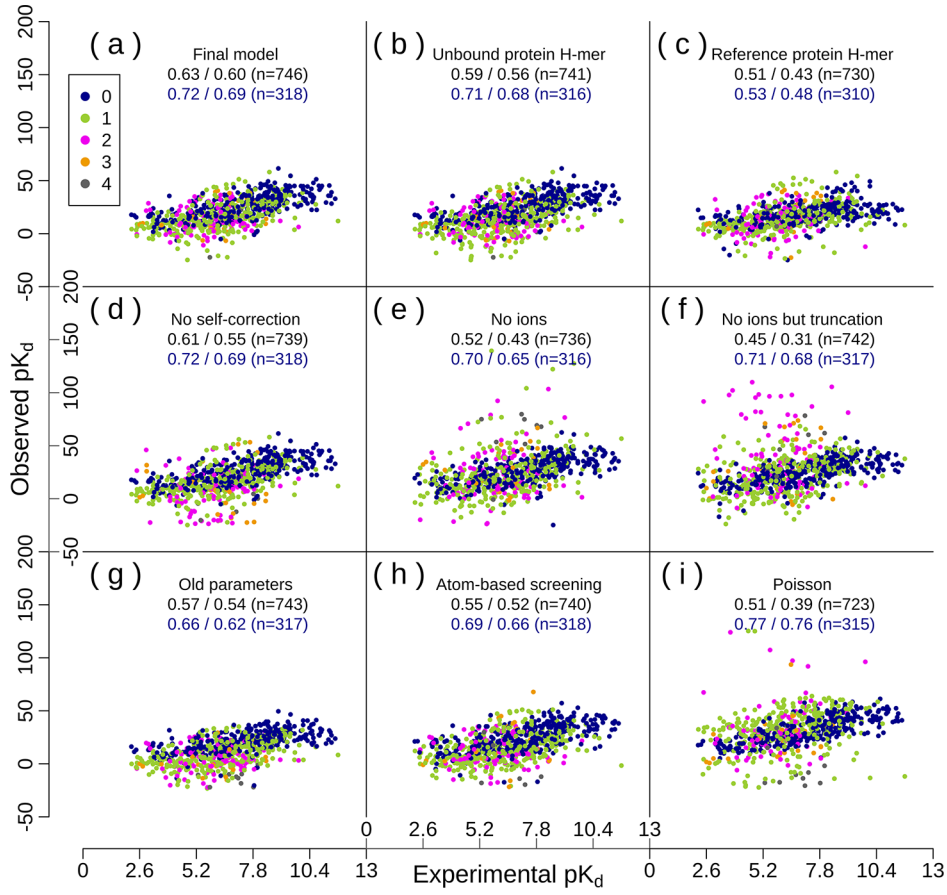


Figure 2. Scatterplots of experimental and predicted pK_d values for different models. In all cases, the color code differentiates ligands with different numbers of charge groups carrying a nonzero net charge (equilibrium-weighted in cases of multiple ligand H-mers). Correlation values are given (Spearman/Pearson) along with the data set size both overall and separately for compounds devoid of formal charges (blue font). For both plotting and correlation, values were retained for an interval from −25 to 150 in predicted pK_d, and the corresponding numbers of points (n, out of 754) are provided in the figure. (a) One replicate data set from the final model (for others, see Figure S3a−d). (b) Same as (a) but choosing only the microscopic binding equilibrium for the most populated H-mer(s) in the unbound state as score. (c) Same as (b) for the reference protein H-mer. (d) Same as (a) but ignoring the *post facto* self-correction (see Supporting Information, eq S11). (e) Data set collected without ions in the protein simulations. (f) Data set collected without ions in the protein simulations and truncating all electrostatic interactions, including monopole−monopole terms, to residue-based cutoffs at 12 Å. (g) Data set collected using old parameters (Lennard-Jones and free energy of solvation offsets, see Supporting Information, S.6, also compare Figure S3f). (h) Data set collected using atom-based rather than group-based screening (see 2.1). (i) Data set where all electrostatic and solvation terms (eqs 1 and 2) are replaced with results from Poisson calculations (see Supporting Information, S.4, for details). These values were obtained using published protocols[7] without further minimization within a Poisson or related model.

combinatorially, all possible complexes of protein H-mers with putative relevant forms of the PDB ligand (2.3.3). These 7457 structures were relaxed (2.3.2) and subsequently underwent an MD simulation protocol in rigid-body/dihedral angle space (S.2) involving a binding site-centric subset of degrees of freedom (2.3.5) at 250 K. For each complex, simulations were also performed for ligand and protein separately with identical settings, thus accounting for both ligand and protein stress. For the protein-only simulations, K^+ and Cl^- were added according to the ligands' charge groups carrying a nonzero net charge (2.3.4 and S.3).

For an individual complex, the binding energy was estimated from the mean force-field energy difference including reference state corrections for ions (if any, eq 3). The total ABSINTH energy, which is a sum of contributions from bonded potentials, van der Waals terms, screened electrostatic interactions, eq 2, and the DMFI, eq 1, was augmented by a self-correction for ligands carrying multiple charged functionalities, eq S11 (see S.6). The screened electrostatics term used group-consistent screening following a prior study.[41] We added and updated force−field parameters (Tables S3−S7) inspired in part by independent results.[43,100]

The experimentally observable binding free energy of a complex was approximated from the binding energies of all its considered H-mers and the conversion energies between these H-mers (S.4). The latter were calculated from the same MD data collected as described above, specifically the protein-only part. This required reference energies for protein residues, which were taken from dipeptide simulations (Table S1). The final prediction from this model is a single computational estimate of the experimentally observed pK_d or pK_i values, eq 4.

Of course, a few of the 7457 systems are unstable because of steric or electrostatic conflicts, which might cause the ligand to move away from the binding site. We decided to be lenient in keeping points where the root mean square deviation (RMSD) between simulated and experimental complex was large (threshold of 10 Å): discarding points based on the RMSD might mask predictions of favorable alternative binding poses and thus obfuscate a relevant source of error. On average,

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

~100 cases did not produce interpretable data, some also because of MD simulations of complex or protein becoming unstable. The latter can happen for poor combinations of the protonation states of adjacent side chains, for example, a protonated histidine next to a protonated acid. In the equilibrium calculation, such missing data were treated as highly unfavorable, both for microscopic binding and protein H-mer reactions, which effectively discards the problematic species from the equilibrium.

The results of several calculations are plotted in Figure 2. The data in (a), which correspond to the final model, represent one example taken from 5 replicates. For these replicates, the entire calculation pipeline, which is stochastic because of velocity initialization and the thermostat (see S.2), was repeated in independent sets. The Spearman correlations were 0.63, 0.63, 0.62, 0.62, and 0.63 for these (Pearson: 0.60, 0.60, 0.59, 0.60, and 0.60) indicating a very small statistical error. Thus, almost all of the differences in Figure 2 appear to be significant, which we confirmed by a resampling strategy (Figures S1 and S2). The comparatively large size of the data set (>700 complexes) and its inherent diversity make it unlikely that the observed improvements are because of spurious trends resulting from peculiarities of individual or few complexes.

Figure 2b,c highlight the importance of using the correct protein H-mer. In particular, predictions by assuming the H-mer derived using the most likely protonation states of isolated amino acid side chains at neutral pH, (c), are clearly worse. The most likely protein H-mer in the unbound state is estimable independent of any ligand (and vice versa), so this information is also, unlike the estimation of the most likely H-mer(s) in the bound state, useable in screening campaigns. As observed consistently in force-field scoring functions with implicit treatment of desolvation costs,[23] the estimation of the binding affinity of neutral compounds is an easier task than that of charged compounds. This is confirmed by (e) and (f): we find that the ion corrections, which resemble a competitive titration experiment, are needed to produce quantitatively comparable values for compounds carrying charge groups. From these data, it is not advisable to discard protein context to limit the impact of net charges, which is what is done effectively in (f). Presumably, this is because a net-charge complementarity between the whole protein and ligand is a contributor to binding affinity. With all corrections in place, there is still a difference between strictly neutral compounds and the rest (Spearman 0.72 versus 0.52), but there is appreciable correlation also for the latter.

Figure 2d shows that the self-correction we introduce here, eq S11, is needed for ligands with multiple charged functionalities in close proximity, for example, free amino acids or citrate. It, by construction, cannot affect any other cases, which is why the remaining data points are identical to (a). Finally, Figure 2g,h demonstrate the usefulness of improvements to the ABSINTH model, in particular to use group-consistent charge screening. These results are also relevant for the further development of the ABSINTH model in other contexts.

We emphasize that the PDBbind data set includes only known binders. It has been established repeatedly that, for this or similar data sets, very simple scores measuring effectively the interaction interface produce high levels of correlation.[8,101,102] We confirm this here by analyzing the correlation of experimental log K_d values with two contributions to the estimated binding free energies: the nonelectrostatic (van der Waals and covalent stress terms) versus electrostatic/solvation contributions for the most likely H-mer(s) in the unbound state (Figure S4a–b). The resultant rank correlation of 0.78 in (a) is close to a realistic ceiling when considering errors in experimental measurements.[103] This means that the remaining terms are essentially a source of noise. Given this interpretation, we replaced the electrostatic/solvation contributions in Figure S4b with values derived from a continuum electrostatics treatment using the Poisson equation. We could not produce a completely analogous data set with a Poisson model because of computational feasibility and because aspects of our protocol (like the use of explicit ions or the MD integrator) are not supported in this paradigm and its implementations.

The hybrid Poisson data set is shown in Figure 2i. While the overall results are most similar to (e), a focus on compounds not carrying net-charge groups reveals this model to approach the aforementioned ceiling (Spearman 0.77) and to outperform the model in Figure 2a for this subset (Spearman: 0.72). This allows four inferences: first, the Lennard-Jones contributions can be treated independently. Second, the electrostatic treatments, despite their different paradigms, produce correlated results for neutral compounds (Figure S4c). Third, a particular treatment of nonpolar solvation is not necessary to achieve good *correlation* with this experimental data set (absolute errors are a different issue).[104] Fourth, problems with charged compounds need to be addressed at a more fundamental level than at the accuracy of the continuum model (compare Figure S4d), and the inclusion of ions is one strategy for this. Of course, the Poisson data in Figure 2i are somewhat artificial, but the choice of using the values for the most likely H-mer in the unbound state is not at fault: selecting instead the value for the reference H-mer, or the median, minimum, or maximum value across H-mers leave the results largely unchanged or make them worse. Conversely, the correlation for charged compounds is likely to improve upon following a true MM/PBSA approach where snapshots are taken from fully flexible MD simulations, and the results are averaged. Table 1 of Sun et al.[105] shows that correlations differ strongly for a minimization-based protocol from an MD average for MM/PBSA but much less so for the chosen MM/GBSA model.[106] This sensitivity might be exacerbated for charged compounds, and we hypothesize this to be the most likely reason for the comparatively poor performance for charged compounds in Figure 2i. Similar to the cited MM/GBSA results, the ABSINTH results do not change significantly if the protocol in 2.3.5 is changed to instead perform a single-point calculation after minimization (see Supporting Information, Figure S5). The chosen low dielectric, which was 4.0 here, is known to have an even larger impact on Poisson models, in particular for charged compounds, as it controls the amplitude of electrostatic terms. We point out that we did not modify the ABSINTH model's use of a true vacuum reference (a relative dielectric of 1.0) in this work. We instead conjecture that appropriate corrections can be introduced that recognize and partially address weaknesses in common scoring protocols. The use of inorganic ions is one such correction (see Supporting Information, S.3).

As a test of robustness, we also checked the impact of two comparatively artificial choices in our model. First, we chose a simulation temperature of 250 K, which, while feasible in a continuum model, is much lower than both ambient and

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

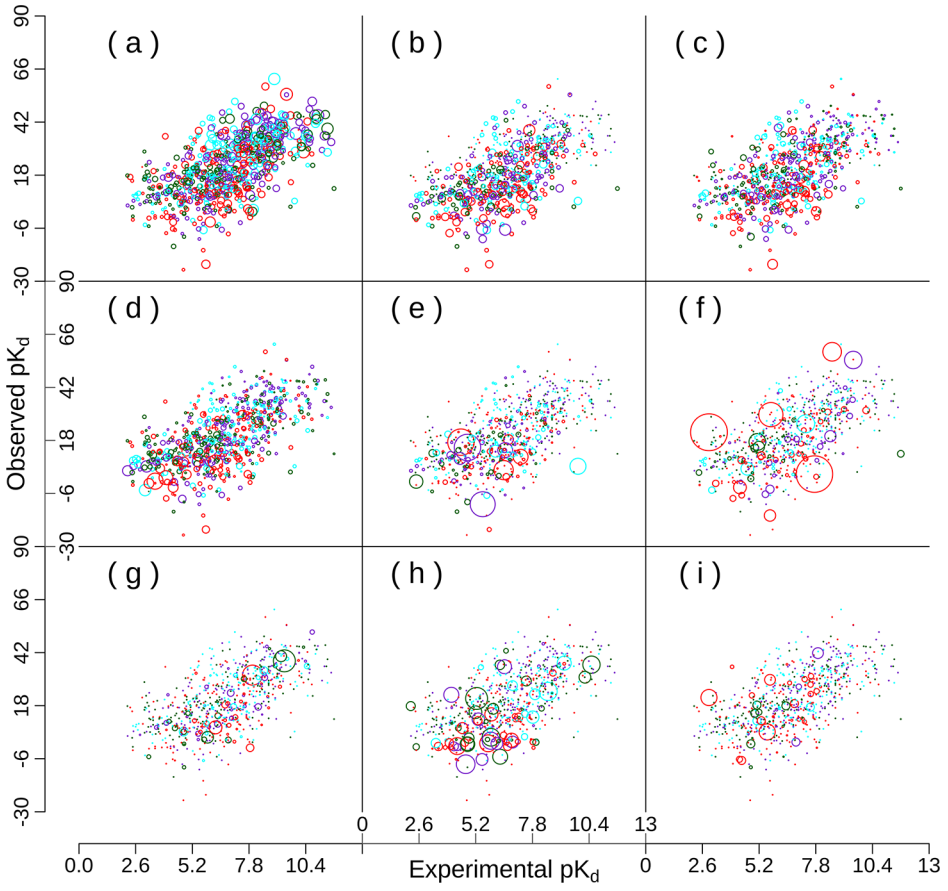


Figure 3. Scatterplots of experimental versus predicted pK_d values for the final model. In each panel, the data and color code are identical: colors correspond, from green to red, to the quartiles of statistical errors in increasing order. Errors are calculated as the min/max ranges across five identical replicates. The quartile boundaries were 2.80, 4.02, and 6.09, respectively. Points are restricted to the 742 complexes yielding interpretable data in all five replicates. Above a minimum, the sizes of the circles' scale linearly with selected properties of the complex or ligand. Where necessary, the properties were obtained as weighted averages following the predicted equilibrium distribution across the complex H-mers. (a) The number of dihedral degrees of freedom in the ligand. (b) The absolute value of the ligand's net charge. (c) The absolute value of the ligand's reference free energy of solvation; note that there are only two ligands with a positive value, both very close to 0.0. (d) The RMSD of the final ligand conformation from the crystal pose. (e) The product of net ligand and protein charges if positive. (f) The absolute value of the product of net ligand and protein charges if negative. (g) The difference in the predicted pK_d relative to the pK_d predicted from the microscopic equilibrium involving just the most likely H-mer(s) in the unbound state. (h) The value of 1.0 minus the fractional occupation of the most populated species. These data depend on total ligand and protein concentrations which were both set to 1.0 μM. (i) The number of protein H-mers in the equilibrium.

physiological temperatures. However, choosing 310 K has virtually no impact on the quality of the results (Figure S3e), and in practice any value in this range is likely to be acceptable as long as the receptor stays largely rigid. Second, we were interested in how much the free energy of solvation offsets for net-charge groups on organic molecules and polymers (compare Supporting Information, Table S3) might affect our results. These offsets are correction parameters introduced in the original model,[31] which are halved here. Halving them again from −15 to −7.5 kcal/mol has no discernible impact either (Figure S3f). This suggests a low sensitivity and high robustness to these values.

We next were interested in elucidating whether the observation that a simple Lennard-Jones model is superior on this data set is because of any remaining systematic errors in the complete model. To answer this question, we show in Figure 3 the same scatterplot as in Figure 2a, but highlighting a number of properties of the underlying complexes. Figure 3b,c confirm that neutral and hydrophobic compounds are most suitable for binding very tightly to their protein target.

Figure 3d is a reassuring result in that it demonstrates that larger deviations from the crystal pose occur predominantly for experimentally and computationally weak binders, as expected. The same trend is seen in Figure 3e, again in line with expectation: complexes where the charges of ligand and protein are of the same sign and their product is large are excluded from high affinities. Figure 3f shows the opposite case of favorable charge complementarity: here, complexes with large values tend to be both outliers and associated with large statistical errors. However, these are rare cases, and the complexes with moderate charge complementarity exhibit no particular trend.

In the last row of Figure 3, we analyze the impact of the H-mer equilibrium. However, the error in pK_d that would have been made by considering the most likely protein H-mer in the unbound state alone, (g), the relevance of multiple species at equilibrium, (h), or the number of protein H-mers, (i), all fail to reveal obvious trends. Based on Figure 3, we thus conclude that there are no clear avenues to pursue for eliminating further

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

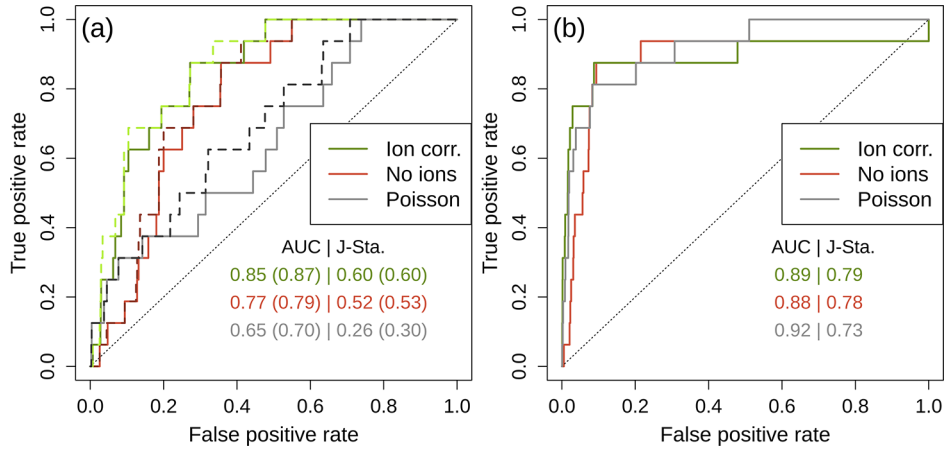


Figure 4. ROC plots for the emulation of a prospective screening campaign. Values are shown for three different models, ABSINTH with ion corrections, without these corrections, and MM/P. In all cases, the total predicted binding energy per individual complex is the underlying score. The area-under-the-curve (AUC) and Youden's J-statistic are reported (J-Sta.). (a) ROC plots for the enrichment of molecules. For each unique molecule, the predicted pose with the best score was identified. The resulting plot reflects the ranks of the 16 known binders in the set of ∼1000 molecules (solid lines). To see the impact of possible deficiencies in the docking algorithm, the dashed lines and the numbers in parentheses show the same upon including the crystallographic poses. Relative to a numerical null model (random classifier, based on picking 10^6 sets of 16 positive controls at random), the estimated p-values for AUC/J-Sta. are $<1 \times 10^{-6}/2 \times 10^{-6}$, $4 \times 10^{-5}/1 \times 10^{-4}$, and 0.019/0.096 for ABSINTH with ions, ABSINTH without ions, and MM/P, respectively. (b) ROC plots for the enrichment of poses. Rather than using only a single pose per molecule, it is assumed that the known crystallographic poses are the only true positives in the entire set of $\sim 1.57 \times 10^5$ poses. The assumed negatives include all predicted poses of known binders.

systematic errors from the predictions. Of course, this conclusion is restricted to the properties we chose to analyze.

3.3. ABSINTH Enriches Actives in a Set of Decoys. The most valuable property of a scoring function is to be able to identify true binders in a set of non-binders, which is the purpose of a virtual screen. We compared the performances of a reference model (molecular mechanics/Poisson, MM/P) to two ABSINTH scores to identify known binders in a set of decoys on the bromodomain of BRD4(1). The test set comprised 16 known binders with affinities in the nM−μM range. We use two levels of decoys: first, ∼1000 molecules were selected with similar properties to the known binders; second, for every molecule, including known binders, diverse poses were generated using rDock (see 2.4.1 for details). To make this test similar to actual applications and to the MM/P data set, in particular under resource constraints, the protocol for ABSINTH was simplified relative to that described in 2.3.2−2.3.6 and the Supporting Information: there is only a single protein H-mer, the few ligand H-mers were treated independently, and no protein residues were allowed to move. Moreover, the ion corrections were not performed for each ligand separately as this would not be meaningful for diverse and electrostatically unfavorable poses (see Supporting Information, S.3, for details).

Figure 4 shows receiver operating characteristic (ROC) curves for two different splits of the data into (true) positives and negatives. This evaluates the models as binary classifiers, meant to distinguish binders from nonbinders. The solid lines in Figure 4a show the ROC results as one would have obtained in a prospective screening campaign. Clearly, even without knowing crystallographic poses, both ABSINTH scores perform much better than a random classifier based on both AUC (random: 0.5) and Youden's J-statistic (random: ∼0.14 for this number of true positives and negatives). The ion corrections produce a superior model because without them, for this target, positively charged compounds are ranked too favorably and push down the ranks of the (mostly neutral) true binders. The performance of the MM/P scores is significantly worse in this test. It is interesting to note that inclusion of crystal poses (dashed lines) makes a larger difference for MM/P than for ABSINTH. This is unexpected because the primary effect should be to eliminate limitations of rDock in identifying a representative pose. As Figure S6a demonstrates, the MM/P performance is worse in the enrichment task, even though its performance in the correlation task for this single-receptor data set of 16 compounds is comparable if not slightly superior (Pearson/Spearman 0.69/0.67 versus 0.63/0.69 for ABSINTH with ion corrections). This highlights the point made above: a model that exhibits good correlation with experimental data for binders can be a surprisingly poor choice in a virtual screen, the obvious caveat being that, generally speaking, model performance is highly variable across different systems.

Figure 4b demonstrates that all models excel at distinguishing experimental from predicted poses. The performance of ABSINTH is weakened slightly by the fact that one of the crystal poses led to an unstable simulation and is ranked last (technically, its rank is a tie with all other poses for which this happened, about 10% of all poses for ABSINTH). The importance of the ion corrections is visible again. The MM/P score performs similarly in this test, which highlights that the weakness of this method observed in (a) is not because of poor scores assigned to good poses or vice versa. Instead, it appears that MM/P offers too little contrast between good poses for nonbinders and good poses for binders. As expected based on this result, the total scores from both methods are correlated at the level of individual poses (Figure S6b) albeit not as tightly as one might expect given the similarity in Figure 4b. Repetitions of the ABSINTH calculations yielded AUC/J-statistic values of 0.87/0.62 and 0.83/0.60 (ions) as well as 0.79/0.55 and 0.76/0.52 (no ions) for the data in Figure 4a highlighting the statistical robustness of these results. As in the ranking test (3.2), the stochasticity of the results stems from the molecular dynamics sampler (see S.2).

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

4. CONCLUSIONS

In this work, we propose and subsequently demonstrate that the ABSINTH implicit solvent model and force field can be generalized to describe the interactions of organic molecules with proteins. Our approach involves a number of modifications to typical virtual screening workflows, most notably the inclusion of inorganic ions as competitors for charged ligands, and the explicit enumeration of protein and ligand H-mers. The model performs well in comparison to reference models based on the Poisson equation to describe electrostatic contributions (Figures 2, S4c, and 4). Literature results suggest that a correlation coefficient of around 0.8 is a ceiling for scoring functions in predicting experimental affinities of known binders across a range of targets.[8,22,103,105] As we show in Figure S4a, and as is found equivalently from machine learning approaches,[8,101] simple geometric descriptors, which encompasses the Lennard-Jones functional form, are enough to provide this correlation of ∼0.8. Evidently, the electrostatic contributions will be un- or anti-correlated with the experiment as seen in Figure S4b. This raises the question whether implicit solvent models, where the focus is on polar interactions, for example, GB/PB models, should be evaluated or optimized at all for such a test. A critical reading of this work, Sun et al.,[105] Huang et al.,[107] or Greenidge et al.[22] might suggest that the proposed improvements and parameters are primarily "silencing" the noise from electrostatic terms. For example, Sun et al. found that a dielectric constant of 4.0 is clearly superior to smaller values, that is, it appears best to reduce the magnitude of descreening and desolvation effects. This is why it is critically important that the ABSINTH model also does well in the second task (Figure 4), where a significant enrichment of true actives is observed.

It is important to emphasize that the ABSINTH model as used here has not been modified substantially since its inception and the only conceptual amendment we propose here is the self-correction for molecules containing charged moieties in close (topological) proximity (eq S11, Figure 2d). The remaining changes are to parameters: partial charges are modular entities in ABSINTH, and we have replaced the original OPLS[108] values with those from CHARMM36,[75] primarily to be consistent with the use of CGenFF.[61] Along with minor updates to rFOS parameters (Table S3), revised Lennard-Jones parameters (Tables S4−S7) are suggested here. Similar parameter updates have already been proposed and validated for simulations of nanotube formation by zwitterionic molecules.[43] A systematic evaluation of the manual adjustments of Lennard-Jones parameters in the broader context of the ABSINTH model is part of ongoing work. In addition, the results in 3.2 rely on a tunable radius, which determines the extent of receptor flexibility in the vicinity of the ligand. However, for a campaign on an individual receptor, the set of mobile residues can and will usually be hand-picked. Larger values increase the amount of sampling required to keep statistical errors at bay.

As a result of the conservative nature of extending ABSINTH to small molecules, we are able to retain the model's appealing properties: (1) the use of experimental rFOS values means that the DMFI will frequently encapsulate realistic desolvation costs for both ligand and protein; (2) the parameterization is comparatively simple because most bonded interactions can be omitted in rigid-body/torsional space (see Supporting Information, S.1); (3) by employing solvent-accessible volumes, concerns related to the treatment of boundaries of low-dielectric cavities are avoided;[35] (4) single-point energy calculations are fast and suitable for deployment in the rescoring of virtual screening campaigns. The last point holds despite the fact that there is potential for further optimization given the peculiarities of this type of application, for example, in the treatment of constrained, intrareceptor interactions. A caveat of the current protocol is that ring flexibility, as for macrocycles, is not handled by the MD integrator.[79] The simplest workaround is to treat one of the ring bonds as a set of harmonic restraints, but we did not use this or other solutions for the results presented here. When working with large data sets, it is always a challenge to ensure that setup and parameters are appropriate.[109] The largest source of errors in assigning rFOS parameters for ABSINTH are difficult-to-detect moieties carrying a net charge such as those occurring in nitrogen-containing heteroaromatic systems. We are currently not resorting to atomic partial charges in the solvation group assignment, but this will likely have to change in the future. At the same time, it is an ongoing process to extend the chemical space to which rFOS parameters can be assigned by adding values to the reference database (2.2.1).

Aside from such technical points, the primary limitations in improving the results further are difficult to pin down. As Figure 3 shows, the final ABSINTH model no longer suffers from systematic errors that can be mapped cleanly to simple properties of proteins or ligands. Of course, our data do support the conclusion (Figures 2 and S4b) that correlations are weaker for ligands carrying net-charge groups, but this is not manifested as systematic deviations. The ion corrections require care as follows. In a screening campaign, the receptor is usually a single structure, and this structure might offer putative binding sites for ions. We are currently developing strategies to make sure that ions can be placed in such a way that they are sufficiently relaxed yet do not explore sites inaccessible to ligands. More importantly, successful applications of free energy methodologies to the prediction of binding affinities,[110,111] in particular also for bromodomains,[112] suggest that the virtual screening protocol itself, in particular the approximation in eq 3, is to blame. Specifically, Aldeghi et al. found that Poisson models were outperformed by absolute free energy methods even after incorporation of entropy corrections or inclusions of variable amounts of explicit water molecules.[113] Thus, we will, in future work, evaluate an ABSINTH-based framework for absolute free-energy calculations. In this context, it is useful to remind the reader that ABSINTH is an efficient implicit solvent model developed for molecular simulations. For the system and data in Figure 4, the computational cost for a single-point energy calculation was between two and three orders of magnitude below that of the corresponding single-point calculation using finite-difference Poisson. This is what enables us to even pursue the strategies explored in the manuscript such as the integration of H-mers, the use of explicit ions, and the reliance on trajectory averages.

We have presented here the use of an ABSINTH force field for small molecules in the context of (re)scoring poses that were generated by other means, either experimentally or by a docking program. While it might seem challenging to rely on energy values averaged over MD runs to derive scores, our results suggest that this is not a major issue even if the receptor is partially flexible. The low cost of single-point energy evaluations in ABSINTH means that the pose search can itself

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

employ the ABSINTH model or a simplified version thereof, but we have not tested this in the present work. Finally, with respect to the parameter refinements for ABSINTH, it will be interesting to see how valuable the modifications proposed here prove in the context of simulations of the folding, binding, and assembly of biomacromolecules.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00558.

Parameterization, sampling, estimation of binding free energies, estimation of protonation free energies for protein H-mers, calculation of the equilibrium, and ABSINTH improvements, list of the PDB codes used in 3.2, statistical and robustness analyses of data in Figure 2 (Figures S1−S3), and further analyses for data in Figures 2 and 4 (Figures S4−S6) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Andreas Vitalis − *Department of Biochemistry, University of Zürich, CH 8057 Zürich, Switzerland;* orcid.org/0000-0002-5422-5278; Email: a.vitalis@bioc.uzh.ch

Authors

Jean-Rémy Marchand − *Department of Biochemistry, University of Zürich, CH 8057 Zürich, Switzerland;* orcid.org/0000-0002-8002-9457

Tim Knehans − *Department of Biochemistry, University of Zürich, CH 8057 Zürich, Switzerland*

Amedeo Caflisch − *Department of Biochemistry, University of Zürich, CH 8057 Zürich, Switzerland;* orcid.org/0000-0002-2317-6792

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00558

Author Contributions

J.R.M. and T.K. wrote the deconstruction algorithm to assign rFOS data. A.V. developed and implemented all other novel methods. J.R.M. compiled the test sets. J.R.M. and A.V. ran the calculations, analyzed the data, and wrote the manuscript. A.V. and A.C. designed the study. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by grants 149897 and 169007 from the Swiss National Science Foundation to AC.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2013**, *42*, D1083−D1090.

(2) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of early drug discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239−1249.

(3) Pu, M.; Hayashi, T.; Cottam, H.; Mulvaney, J.; Arkin, M.; Corr, M.; Carson, D.; Messer, K. Analysis of high-throughput screening assays using cluster enrichment. *State Med.* **2012**, *31*, 4175−4189.

(4) List, M.; Schmidt, S.; Christiansen, H.; Rehmsmeier, M.; Tan, Q.; Mollenhauer, J.; Baumbach, J. Comprehensive analysis of high-throughput screens with HiTSeekR. *Nucleic Acids Res.* **2016**, *44*, 6639−6648.

(5) Westermaier, Y.; Barril, X.; Scapozza, L. Virtual screening: An in silico tool for interlacing the chemical universe with the proteome. *Methods* **2015**, *71*, 44−57.

(6) Torres, P. H. M.; Sodero, A. C. R.; Jofily, P.; Silva, F. P., Jr. Key topics in molecular docking for drug design. *Int. J. Mol. Sci.* **2019**, *20*, 4574.

(7) 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*, 2584−2597.

(8) Khamis, M. A.; Gomaa, W. Comparative assessment of machine-learning scoring functions on PDBbind 2013. *Eng. Appl. Artif. Intell.* **2015**, *45*, 136−151.

(9) Liu, J.; Wang, R. Classification of current scoring functions. *J. Chem. Inf. Model.* **2015**, *55*, 475−482.

(10) Glaab, E. Building a virtual ligand screening pipeline using free software: a survey. *Briefings Bioinf.* **2016**, *17*, 352−366.

(11) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **2017**, *9*, 91−102.

(12) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566*, 224−229.

(13) Gaieb, Z.; Liu, S.; Gathiaka, S.; Chiu, M.; Yang, H.; Shao, C.; Feher, V. A.; Walters, W. P.; Kuhn, B.; Rudolph, M. G.; Burley, S. K.; Gilson, M. K.; Amaro, R. E. D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 1−20.

(14) Gathiaka, S.; Liu, S.; Chiu, M.; Yang, H.; Stuckey, J. A.; Kang, Y. N.; Delproposto, J.; Kubish, G.; Dunbar, J. B.; Carlson, H. A.; Burley, S. K.; Walters, W. P.; Amaro, R. E.; Feher, V. A.; Gilson, M. K. D3R grand challenge 2015: Evaluation of protein−ligand pose and affinity predictions. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 651−668.

(15) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods? *Curr. Med. Chem.* **2006**, *13*, 3583−3608.

(16) Marchand, J.-R.; Lolli, G.; Caflisch, A. Derivatives of 3-amino-2-methylpyridine as BAZ2B bromodomain ligands: in silico discovery and in crystallo validation. *J. Med. Chem.* **2016**, *59*, 9919−9927.

(17) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery* **2015**, *10*, 449−461.

(18) Kolb, P.; Huang, D.; Dey, F.; Caflisch, A. Discovery of kinase inhibitors by high-throughput docking and scoring based on a transferable linear interaction energy model. *J. Med. Chem.* **2008**, *51*, 1179−1188.

(19) Davis, C. M.; Gruebele, M.; Sukenik, S. How does solvation in the cell affect protein folding and binding? *Curr. Opin. Struct. Biol.* **2018**, *48*, 23−29.

(20) Yoshida, N. Role of solvation in drug design as revealed by the statistical mechanics integral equation theory of liquids. *J. Chem. Inf. Model.* **2017**, *57*, 2646−2656.

(21) Willow, S. Y.; Xie, B.; Lawrence, J.; Eisenberg, R. S.; Minh, D. D. L. On the polarization of ligands by proteins. *Phys. Chem. Chem. Phys.* **2020**, *22*, 12044−12057.

(22) 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*, 201−209.

(23) Marchand, J.-R.; Caflisch, A. In silico fragment-based drug design with SEED. *Eur. J. Med. Chem.* **2018**, *156*, 907−917.

(24) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, *48*, 4040−4048.

(25) Czaplewski, C.; Ripoll, D. R.; Liwo, A.; Rodziewicz-Motowid?o, S.; Wawak, R. J.; Scheraga, H. A. Can cooperativity in hydrophobic association be reproduced correctly by implicit solvation models? *Int. J. Quantum Chem.* **2002**, *88*, 41−55.

(26) Feig, M.; Brooks, C. L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.
(27) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute−solvent interaction energy. *J. Am. Chem. Soc.* **2003**, *125*, 9523−9530.
(28) Pitera, J. W.; van Gunsteren, W. F. The importance of solute−solvent van der Waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.* **2001**, *123*, 3163−3164.
(29) Shimizu, S.; Chan, H. S. Anti-cooperativity and cooperativity in hydrophobic interactions: Three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. *Proteins: Struct., Funct., Bioinf.* **2002**, *48*, 15−30.
(30) Harris, R. C.; Pettitt, B. M. Examining the assumptions underlying continuum-solvent models. *J. Chem. Theory Comput.* **2015**, *11*, 4593−4600.
(31) Vitalis, A.; Pappu, R. V. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2009**, *30*, 673−699.
(32) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins: Struct., Funct., Bioinf.* **1999**, *35*, 133−152.
(33) Nina, M.; Beglov, D.; Roux, B. Atomic radii for continuum electrostatics calculations based on molecular dynamics free energy simulations. *J. Phys. Chem. B* **1997**, *101*, 5239−5248.
(34) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. Limitations of atom-centered dielectric functions in implicit solvent models. *J. Phys. Chem. B* **2005**, *109*, 14769−14772.
(35) Swanson, J. M. J.; Wagoner, J. A.; Baker, N. A.; McCammon, J. A. Optimizing the Poisson dielectric boundary with explicit solvent forces and energies: Lessons learned with atom-centered dielectric functions. *J. Chem. Theory Comput.* **2007**, *3*, 170−183.
(36) Tjong, H.; Zhou, H.-X. On the dielectric boundary in Poisson−Boltzmann calculations. *J. Chem. Theory Comput.* **2008**, *4*, 507−514.
(37) Mao, A. H.; Lyle, N.; Pappu, R. V. Describing sequence−ensemble relationships for intrinsically disordered proteins. *Biochem. J.* **2013**, *449*, 307−318.
(38) Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J. Am. Chem. Soc.* **2016**, *138*, 15323−15335.
(39) Sherry, K. P.; Das, R. K.; Pappu, R. V.; Barrick, D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E9243−E9252.
(40) Cumberworth, A.; Bui, J. M.; Gsponer, J. Free energies of solvation in the context of protein folding: Implications for implicit and explicit solvent models. *J. Comput. Chem.* **2016**, *37*, 629−640.
(41) 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. U.S.A.* **2010**, *107*, 8183−8188.
(42) Das, R. K.; Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392−13397.
(43) 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*, 13190.
(44) Vitalis, A. CAMPARI Website. http://campari.sourceforge.net/ (accessed July 31, 2020).
(45) Mohan, V.; Davis, M. E.; McCammon, J. A.; Pettitt, B. M. Continuum model calculations of solvation free energies: accurate evaluation of electrostatic contributions. *J. Phys. Chem.* **1992**, *96*, 6428−6431.
(46) Scarsi, M.; Caflisch, A. Comment on the validation of continuum electrostatics models. *J. Comput. Chem.* **1999**, *20*, 1533−1536.
(47) Murphy, K. P.; Gill, S. J. Group additivity thermodynamics for dissolution of solid cyclic dipeptides into water. *Thermochim. Acta* **1990**, *172*, 11−20.
(48) Makhatadze, G. I.; Privalov, P. L. Contribution of hydration to protein folding thermodynamics: I. The enthalpy of hydration. *J. Mol. Biol.* **1993**, *232*, 639−659.
(49) Privalov, P. L.; Makhatadze, G. I. Contribution of hydration to protein folding thermodynamics: II. The entropy and Gibbs energy of hydration. *J. Mol. Biol.* **1993**, *232*, 660−679.
(50) 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*, 563−595.
(51) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. Affinities of amino acid side chains for solvent water. *Biochemistry* **1981**, *20*, 849−855.
(52) 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*, 16066−16081.
(53) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 711−720.
(54) Guthrie, J. P. A blind challenge for computational solvation free energies: Introduction and overview. *J. Phys. Chem. B* **2009**, *113*, 4501−4507.
(55) Sulea, T.; Wanapun, D.; Dennis, S.; Purisima, E. O. Prediction of SAMPL-1 hydration free energies using a continuum electrostatics-dispersion model. *J. Phys. Chem. B* **2009**, *113*, 4511−4520.
(56) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 259−279.
(57) Geballe, M. T.; Guthrie, J. P. The SAMPL3 blind prediction challenge: transfer energy overview. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 489−496.
(58) Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. Blind prediction of solvation free energies from the SAMPL4 challenge. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 135−150.
(59) Vitalis, A.; Caflisch, A. 50 Years of Lifson−Roig models: Application to molecular simulation data. *J. Chem. Theory Comput.* **2012**, *8*, 363−373.
(60) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general AMBER force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(61) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671−690.
(62) Parsons, G. H.; Rochester, C. H.; Wood, C. E. C. Effect of 4-substitution on the thermodynamics of hydration of phenol and the phenoxide anion. *J. Chem. Soc. B* **1971**, *0*, 533−536.
(63) Ooi, T.; Oobatake, M.; Némethy, G.; Scheraga, H. A. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086−3090.
(64) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. *J. Phys. Chem.* **1996**, *100*, 16385−16398.
(65) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. New model for calculation of solvation free energies: Correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects. *J. Phys. Chem.* **1996**, *100*, 11775−11788.
(66) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. Estimation of absolute free energies of hydration using continuum methods:

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Accuracy of partial charge models and optimization of nonpolar contributions. *J. Chem. Theory Comput.* **2006**, *2*, 128−139.
(67) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769−779.
(68) Landrum, G. *RDKit: Open-Source Cheminformatics*, 2018.
(69) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(70) Hawkins, G. D.; Giesen, D. J.; Lynch, G. C.; Chambers, C. C.; Rossi, I.; Storer, J. W.; Li, J.; Thompson, J. D.; Winget, P.; Lynch, B. J.; Rinaldi, D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G., *AMSOL*-version 7.0, Based in Part on AMPAC-version 2.1; University of Minnesota: Minneapolis, MN 2003.
(71) Wyczalkowski, M. A.; Vitalis, A.; Pappu, R. V. New estimators for calculating solvation entropy and enthalpy and comparative assessments of their accuracy and precision. *J. Phys. Chem. B* **2010**, *114*, 8166−8180.
(72) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(73) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405−412.
(74) Tie, Y.; Kovalevsky, A. Y.; Boross, P.; Wang, Y.-F.; Ghosh, A. K.; Tozser, J.; Harrison, R. W.; Weber, I. T. Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 232−242.
(75) Huang, J.; MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **2013**, *34*, 2135−2145.
(76) 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*, 3155−3168.
(77) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.
(78) 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*, 6035−6049.
(79) Vitalis, A.; Pappu, R. V. A simple molecular mechanics integrator in mixed rigid body and dihedral angle space. *J. Chem. Phys.* **2014**, *141*, 034105.
(80) 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*, 5933−5944.
(81) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(82) Réau, M.; Langenfeld, F.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys selection in benchmarking datasets: Overview and perspectives. *Front. Pharmacol.* **2018**, *9*, 11.
(83) Marchand, J.-R.; Caflisch, A. Binding mode of acetylated histones to bromodomains: Variations on a common motif. *ChemMedChem* **2015**, *10*, 1327−1333.
(84) Spiliotopoulos, D.; Zhu, J.; Wamhoff, E.-C.; Deerain, N.; Marchand, J.-R.; Aretz, J.; Rademacher, C.; Caflisch, A. Virtual screen to NMR (VS2NMR): Discovery of fragment hits for the CBP bromodomain. *Bioorg. Med. Chem. Lett.* **2017**, *27*, 2472−2478.
(85) Riniker, S.; Landrum, G. A. Better informed distance geometry: Using what we know to improve conformation generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562−2574.
(86) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.* **2014**, *10*, e1003571.
(87) Sullivan, J. M.; Badimon, A.; Schaefer, U.; Ayata, P.; Gray, J.; Chung, C.-w.; von Schimmelmann, M.; Zhang, F.; Garton, N.; Smithers, N.; Lewis, H.; Tarakhovsky, A.; Prinjha, R. K.; Schaefer, A. Autism-like syndrome is induced by pharmacological suppression of BET proteins in young mice. *J. Exp. Med.* **2015**, *212*, 1771−1781.
(88) Ran, X.; Zhao, Y.; Liu, L.; Bai, L.; Yang, C.-Y.; Zhou, B.; Meagher, J. L.; Chinnaswamy, K.; Stuckey, J. A.; Wang, S. Structure-based design of γ-carboline analogues as potent and specific BET bromodomain inhibitors. *J. Med. Chem.* **2015**, *58*, 4927−4939.
(89) Gosmini, R.; Nguyen, V. L.; Toum, J.; Simon, C.; Brusq, J.-M. G.; Krysa, G.; Mirguet, O.; Riou-Eymard, A. M.; Boursier, E. V.; Trottet, L.; Bamborough, P.; Clark, H.; Chung, C.-w.; Cutler, L.; Demont, E. H.; Kaur, R.; Lewis, A. J.; Schilling, M. B.; Soden, P. E.; Taylor, S.; Walker, A. L.; Walker, M. D.; Prinjha, R. K.; Nicodème, E. The discovery of I-BET726 (GSK1324726A), a potent tetrahydroquinoline ApoA1 up-regulator and selective BET bromodomain inhibitor. *J. Med. Chem.* **2014**, *57*, 8111−8131.
(90) Zhang, G.; Liu, R.; Zhong, Y.; Plotnikov, A. N.; Zhang, W.; Zeng, L.; Rusinova, E.; Gerona-Nevarro, G.; Moshkina, N.; Joshua, J.; Chuang, P. Y.; Ohlmeyer, M.; He, J. C.; Zhou, M.-M. Down-regulation of NF-κB transcriptional activity in HIV-associated kidney disease by BRD4 inhibition. *J. Biol. Chem.* **2012**, *287*, 28840−28851.
(91) Chung, C.-w.; Coste, H.; White, J. H.; Mirguet, O.; Wilde, J.; Gosmini, R. L.; Delves, C.; Magny, S. M.; Woodward, R.; Hughes, S. A.; Boursier, E. V.; Flynn, H.; Bouillot, A. M.; Bamborough, P.; Brusq, J.-M. G.; Gellibert, F. J.; Jones, E. J.; Riou, A. M.; Homes, P.; Martin, S. L.; Uings, I. J.; Toum, J.; Clément, C. A.; Boullay, A.-B.; Grimley, R. L.; Blandel, F. M.; Prinjha, R. K.; Lee, K.; Kirilovsky, J.; Nicodeme, E. Discovery and characterization of small molecule inhibitors of the BET family bromodomains. *J. Med. Chem.* **2011**, *54*, 3827−3838.
(92) Nicodeme, E.; Jeffrey, K. L.; Schaefer, U.; Beinke, S.; Dewell, S.; Chung, C.-w.; Chandwani, R.; Marazzi, I.; Wilson, P.; Coste, H.; White, J.; Kirilovsky, J.; Rice, C. M.; Lora, J. M.; Prinjha, R. K.; Lee, K.; Tarakhovsky, A. Suppression of inflammation by a synthetic histone mimic. *Nature* **2010**, *468*, 1119−1123.
(93) Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I.; Philpott, M.; Munro, S.; McKeown, M. R.; Wang, Y.; Christie, A. L.; West, N.; Cameron, M. J.; Schwartz, B.; Heightman, T. D.; La Thangue, N.; French, C. A.; Wiest, O.; Kung, A. L.; Knapp, S.; Bradner, J. E. Selective inhibition of BET bromodomains. *Nature* **2010**, *468*, 1067−1073.
(94) Chen, L.; Yap, J. L.; Yoshioka, M.; Lanning, M. E.; Fountain, R. N.; Raje, M.; Scheenstra, J. A.; Strovel, J. W.; Fletcher, S. BRD4 structure−activity relationships of dual PLK1 kinase/BRD4 bromodomain inhibitor BI-2536. *ACS Med. Chem. Lett.* **2015**, *6*, 764−769.
(95) Zhang, G.; Plotnikov, A. N.; Rusinova, E.; Shen, T.; Morohashi, K.; Joshua, J.; Zeng, L.; Mujtaba, S.; Ohlmeyer, M.; Zhou, M.-M. Structure-guided design of potent diazobenzene inhibitors for the BET bromodomains. *J. Med. Chem.* **2013**, *56*, 9251−9264.
(96) Picaud, S.; Wells, C.; Felletar, I.; Brotherton, D.; Martin, S.; Savitsky, P.; Diez-Dacal, B.; Philpott, M.; Bountra, C.; Lingard, H.; Fedorov, O.; Müller, S.; Brennan, P. E.; Knapp, S.; Filippakopoulos, P. RVX-208, an inhibitor of BET transcriptional regulators with selectivity for the second bromodomain. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 19754−19759.
(97) Filippakopoulos, P.; Picaud, S.; Fedorov, O.; Keller, M.; Wrobel, M.; Morgenstern, O.; Bracher, F.; Knapp, S. Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family. *Bioorg. Med. Chem.* **2012**, *20*, 1878−1886.
(98) Gacias, M.; Gerona-Navarro, G.; Plotnikov, A. N.; Zhang, G.; Zeng, L.; Kaur, J.; Moy, G.; Rusinova, E.; Rodriguez, Y.; Matikainen, B.; Vincek, A.; Joshua, J.; Casaccia, P.; Zhou, M.-M. Selective chemical modulation of gene transcription favors oligodendrocyte lineage progression. *Chem. Biol.* **2014**, *21*, 841−854.

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(99) McLure, K. G.; Gesner, E. M.; Tsujikawa, L.; Kharenko, O. A.; Attwell, S.; Campeau, E.; Wasiak, S.; Stein, A.; White, A.; Fontano, E.; Suto, R. K.; Wong, N. C. W.; Wagner, G. S.; Hansen, H. C.; Young, P. R. RVX-208, an inducer of ApoA-I in humans, is a BET bromodomain antagonist. *PLoS One* **2013**, *8*, e81390.
(100) 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*, 1367−1382.
(101) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein−ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169−1175.
(102) Gabel, J.; Desaphy, J.; Rognan, D. Beware of machine learning-based scoring functions-on the danger of developing black boxes. *J. Chem. Inf. Model.* **2014**, *54*, 2807−2815.
(103) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The experimental uncertainty of heterogeneous public K_i data. *J. Med. Chem.* **2012**, *55*, 5165−5173.
(104) Wang, C.; Nguyen, P. H.; Pham, K.; Huynh, D.; Le, T.-B. N.; Wang, H.; Ren, P.; Luo, R. Calculating protein−ligand binding affinities with MMPBSA: Method and error analysis. *J. Comput. Chem.* **2016**, *37*, 2436−2446.
(105) Sun, H.; Li, Y.; Tian, S.; Xu, L.; Hou, T. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Phys. Chem. Chem. Phys.* **2014**, *16*, 16719−16729.
(106) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383−394.
(107) Huang, K.; Luo, S.; Cong, Y.; Zhong, S.; Zhang, J. Z. H.; Duan, L. An accurate free energy estimator: Based on MM/PBSA combined with interaction entropy for protein−ligand binding affinity. *Nanoscale* **2020**, *12*, 10737−10750.
(108) 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*, 6474−6487.
(109) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 775−786.
(110) Bhati, A. P.; Wan, S.; Wright, D. W.; Coveney, P. V. Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Comput.* **2017**, *13*, 210−222.
(111) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 host−guest binding affinity prediction challenge. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 937−963.
(112) Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **2016**, *7*, 207−218.
(113) Aldeghi, M.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Statistical analysis on the performance of molecular mechanics Poisson−Boltzmann surface area versus absolute binding free energy calculations: Bromodomains as a case study. *J. Chem. Inf. Model.* **2017**, *57*, 2203−2221.

https://dx.doi.org/10.1021/acs.jcim.0c00558
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX