JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Editors' Choice
Open Access

pubs.acs.org/jcim

Article

Current State of Open Source Force Fields in Protein–Ligand Binding Affinity Predictions

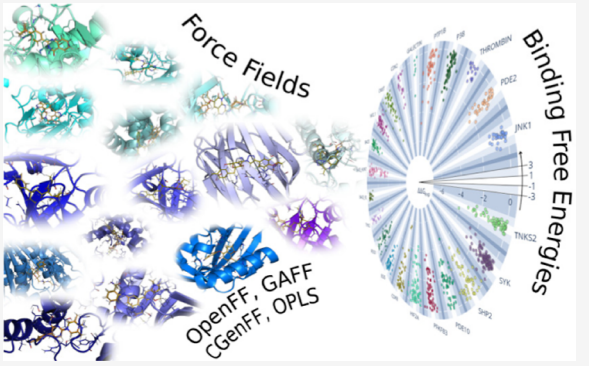
David F. Hahn,* Vytautas Gapsys, Bert L. de Groot, David L. Mobley, and Gary Tresadern

Cite This: J. Chem. Inf. Model. 2024, 64, 5063−5076

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In drug discovery, the in silico prediction of binding affinity is one of the major means to prioritize compounds for synthesis. Alchemical relative binding free energy (RBFE) calculations based on molecular dynamics (MD) simulations are nowadays a popular approach for the accurate affinity ranking of compounds. MD simulations rely on empirical force field parameters, which strongly influence the accuracy of the predicted affinities. Here, we evaluate the ability of six different small-molecule force fields to predict experimental protein−ligand binding affinities in RBFE calculations on a set of 598 ligands and 22 protein targets. The public force fields OpenFF Parsley and Sage, GAFF, and CGenFF show comparable accuracy, while OPLS3e is significantly more accurate. However, a consensus approach using Sage, GAFF, and CGenFF leads to accuracy comparable to OPLS3e. While Parsley and Sage are performing comparably based on aggregated statistics across the whole dataset, there are differences in terms of outliers. Analysis of the force field reveals that improved parameters lead to significant improvement in the accuracy of affinity predictions on subsets of the dataset involving those parameters. Lower accuracy can not only be attributed to the force field parameters but is also dependent on input preparation and sampling convergence of the calculations. Especially large perturbations and nonconverged simulations lead to less accurate predictions. The input structures, Gromacs force field files, as well as the analysis Python notebooks are available on GitHub.



INTRODUCTION

Prioritizing the synthesis of compounds by means of computationally predicted binding affinities among equally important absorption, distribution, metabolism, excretion, and toxicity properties has become one of the central strategies in small-molecule drug discovery.[1] There are different methods, ranging from data-driven artificial intelligence to more rigorous physics-based models. Among the latter, the calculation of relative binding free energies (RBFE) from alchemical molecular dynamics (MD) simulations is probably the most frequently used and accurate method, given the accessible time scales for the size of the ligand−protein complexes. RBFE calculations involve alchemical perturbations, where a ligand is changed into another via a chemically unrealistic pathway. This can only be achieved in silico, such as by changing the atoms of one element into those of another. Following the alchemical pathways across the thermodynamic cycle will result in the same double free energy difference for the perturbation in solvent and protein as when traversing the physical pathways, i.e., monitoring the unbinding of one ligand and the binding of another. However, the alchemical transitions offer a clear sampling advantage over the physical ligand binding/unbinding pathway, thus reducing the computational cost of free energy calculations. In addition, RBFE calculations benefit from the cancellation of errors arising from calculating the separate solvation and protein legs for similar ligands.[2] The final result of the calculation is the relative affinity of the ligand to a protein with respect to the other ligand. The reader is referred to a recent review of alchemical methods and recommendations for their use.[3]

Due to tremendous algorithmic advances, the development of user-friendly software, and the continuous increase in accuracy and computational power in the last decades, these calculations are nowadays frequently utilized. However, the calculations are still costly (compute costs of approximately 10 US$ per relative free energy difference[4] and, in addition, potential software licensing costs). The accuracy with respect to experimental affinities is typically in the range of 1−2 kcal mol^{-1} with the best performing cases arguably capable of approaching experimental accuracy.[5−11] When comparing to experiments, there are mainly four sources of error encountered in binding free energy calculations: system

Received: March 10, 2024
Revised: April 23, 2024
Accepted: April 25, 2024
Published: June 19, 2024

setup, force field (FF) parameters, sampling time, and experimental uncertainty. First, the setup of the system has a significant impact on the prediction accuracy. This includes the exact chemical composition of the system, consisting of proteins, ligands, solvents, potential ions, and cofactors.[12] All the molecules need to be in their relevant tautomeric and charge states. Also, the initial coordinates of all atoms will strongly affect the results, as well as the simulation parameters mimicking the experimental conditions.[13,14] Here, careful preparation and well-considered parameters keep this error contribution low, but this typically involves extensive manual work. The potential pitfalls and best practices to circumventing errors in system preparation were recently summarized.[15] Furthermore, there are many approximations required to model such systems, which include the number of degrees of freedom treated, the treatment of finite-size effects, and especially the FF parameters used in classical mechanic simulations.

Another source of error in free energy estimates comes from finite sampling. Current computational power allows reaching microsecond simulation time scales, yet in large scale free energy scans, shorter sampling (up to tens of nanoseconds) is often employed. Depending on the system, such short sampling times may not be sufficient to converge the populations along the relevant degrees of freedom, e.g., ligand pose changes, amino acid rotamer motions, and water positions in the binding site. Therefore, the limited sampling does not always ensure a proper representation of the thermodynamic ensemble underlying the modeled system. This issue may be minimized by employing different or enhanced sampling protocols[16] such as replica exchange[17,18] or related replica methods,[19,20] metadynamics/local elevation,[21,22] and umbrella sampling or well designed sampling (MC) moves.[23] Performing multiple-independent simulation repeats allows for more reliable phase space exploration and uncertainty estimation.[24,25] Sampling improvements in relative free energy calculations may also arise by optimally planning the perturbations to be calculated,[26−28] altering the alchemical pathway,[29−32] using different atom mapping as in the separated topology approach,[33] or using no atom mapping at all as in enveloping distribution sampling.[34−37] To sample the water position sufficiently, enhanced water sampling protocols[38,39] can be employed. Multiple options exist, such as explicit water perturbations,[40] Monte Carlo moves,[23,41,42] or grand canonical ensemble simulations.[43−45]

Finally, uncertainty in the experimental measurements for the reference data limits the achievable prediction accuracy.[11] Typically, one compares the result of calculations to the experimentally measured bioactivity data, which itself has errors and is only an approximation or model to the ideal or true affinity. Additionally, the experimental data might be unsuitable for comparison because the experimental conditions differ from the simulation conditions (e.g., the temperature) or because the experiment did not measure the same observables (e.g., phenotypic vs functional assays). To keep this error low, one should use high-quality and well curated data for the comparison and above all appreciate the maximum expected performance given the underlying experimental error.[46]

While some analyses suggest that the sampling, FF, and experimental errors might contribute in a quantitatively similar manner,[47] generally, the magnitude of each source of error is unknown and will likely be case-dependent. In the current work, we concentrate on quantifying FF-related errors by comparing six small molecule mechanic FFs in a benchmark of relative protein−ligand binding free energy calculations. For each FF, we obtained up to 1116 $\Delta\Delta G$ estimates across 22 protein targets. The large and diverse set of systems allows a statistically meaningful comparison of not only distinct FF families—GAFF, CGenFF, OPLS, and OpenFF—but also different versions of OpenFF: v1.0, v1.2, and v2.0. With OpenFF presenting a novel direction in FF development,[48−50] here, we demonstrate the ability of this FF to deliver high accuracy binding free energy predictions.

METHODS

Dataset. The employed benchmark dataset is listed in the Supporting Information, Table S.1. A total of 22 protein targets, 598 ligands, and 1116 alchemical perturbations were considered.

In order to compare them to other calculations, we selected benchmark sets from previously published literature. Eight datasets originate from Wang et al.[5] and contain the targets JNK1, TYK2, BACE, MCL1, CDK2, THROMBIN, PTP1B, and P38. Another eight datasets were assembled in the benchmark study of Schindler et al.[7] Furthermore, we included protein−ligand systems that have appeared in various other free energy perturbation (FEP) studies: GALECTIN-3,[51] PDE2,[52] PDE10,[53] ROS1,[54] and two additional BACE datasets.[55−58] To keep our results as comparable as possible to prior calculations, we used the same input coordinates of the prepared systems as were previously used in the studies of Gapsys et al.,[8] Schindler et al.,[7] and Pérez-Benito et al.[54] The input structures are provided in the protein−ligand-benchmark repository, release 0.2.1.[59]

Calculation Details. *pmx/GROMACS Nonequilibrium Switching Approach.* The prepared protein and ligand structures were parameterized using the corresponding FF parameters (see below). The remainder of the preparation and the simulation protocol followed the nonequilibrium thermodynamic integration protocol from the study of Gapsys et al.[8] and is summarized as follows. For each perturbation, hybrid coordinates and topologies were generated from the physical end state ligand coordinates and topologies using pmx.[60] A mapping between the atoms of two molecules was established following a predefined set of rules to ensure minimal perturbation and system stability during the simulations. The pmx method follows a sequential, dual mapping approach. In the first step, pmx identifies the maximum common substructure between the two molecules and proposes this as a basis for mapping. In the second step, pmx superimposes the molecules and suggests a mapping based on the interatomic distances. Finally, the mapping with more atoms identified for direct morphing between the ligands is selected. Additionally, pmx incorporates a number of empirical rules to ensure simulation stability, e.g., avoiding ring and bond breaking, preventing mappings that result in disconnected fragments, and disallowing mapping heavy atoms to hydrogens. The obtained mapping is used to create hybrid structures and topologies following a combination of single and dual topology approach.

The two branches of the thermodynamic cycle were prepared for simulation: ligand in water and ligand bound to the protein. The systems were placed in a dodecahedral box with a minimal distance of 1.5 nm to the box wall. The solutes were solvated with the TIP3P[61] water, and sodium and

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

chloride ions were added to neutralize the system and reach a 150 mM salt concentration.

The Amber99sb*ILDN[62−64] FF was used to parameterize the proteins for the simulations with OpenFF and GAFF2.1x ligand FFs. The ion parameters for these simulations were taken from Joung and Cheatham.[61] The Charmm36m[65] protein FF was used in combination with the MATCH/CGenFF ligand parameters.

To calculate relative free energy differences, first, every system was simulated at equilibrium in its physical state, e.g., ligand X representing state A and ligand Y representing state B. The simulation protocol involved energy minimization, followed by a brief 10 ps *NVT* equilibration and finally a production run for 6 ns in the *NPT* ensemble, where frames were written to file every 47 ps. From the generated trajectories, the first 48 frames (2.256 ns simulation time) were discarded, and from the rest, 80 snapshots were extracted. These configurations were used to perform rapid (50 ps) alchemical transitions between the physical states: from state A to state B when starting from the equilibrium ensemble generated at the state A and vice versa. The whole procedure, starting with energy minimization and ending with the fast alchemical transitions, was repeated 3 times. Each repeat used different random initial ion coordinates and initial velocities for the *NVT* equilibration. All in all, the simulation time for one leg of the thermodynamic cycle of 3 replicas adds up to 60 ns for each double free energy difference. This is an equivalent simulation time to a classical equilibrium FEP approach using twelve 5 ns lambda windows, which happens to be the default in the commercial FEP+ software and is used in many published studies.[5]

The simulation temperature was kept at 298 K by means of the stochastic dynamics integrator with a friction of 0.5 ps^{-1}.[66] This protocol is in line with that previously described in ref 8 except that ref 8 used MD integrator in combination with the velocity rescaling thermostat[67] with a time constant of 0.1 ps. The pressure was controlled by means of the Parrinello−Rahman barostat[68] with a time constant of 5 ps, keeping pressure at 1 bar. Electrostatic interactions were treated by means of the particle mesh Ewald (PME) method[69,70] with a direct space cutoff of 1.1 nm, a relative strength of interactions at a cutoff of 10^{-5}, and a Fourier grid spacing of 0.12 nm. Van der Waals interactions were switched starting at 1.0 nm distance, and were completely turned off for the distances reaching 1.1 nm. Dispersion correction was used to adjust energy and pressure. Nonbonded interactions during the alchemical transitions were softened. The functional form of the softcore potential described in ref 29 (with the default set of parameters) was used for the transitions in PDE2, GALECTIN, BACE (Hunt), BACE, BACE (P2), CMET, JNK1, TYK2, MCL1, CDK2, THROMBIN, PTP1B, and P38 systems. For the alchemical transitions in the other systems, the softcore potential described in ref 71 was used with the parameters α = 0.3 and σ = 0.25 nm. The bonds were constrained by means of the LINCS algorithm.[72]

From the alchemical transitions, work values were collected, and free energy differences were calculated based on the Crooks fluctuation theorem[73] using a maximum likelihood estimator.[74]

Free Energy Perturbation Using FEP+. The free energy calculations using Schrodinger's FEP+[5] were retrieved from published results, and the calculation details can be found therein.[7,8,54] The calculations use the same input structures as those available in the reference dataset as well as the same alchemical perturbations.[59] The previously published FEP+ results were generated by the automated Schrodinger protocol with default settings, i.e., 5 ns simulation time, 12−24 λ points per perturbation, Hamiltonian replica exchange, and the replica exchange solute tempering protocol. The proteins and ligands were parameterized using the OPLS3e FF with custom parameters,[75] as described in the respective publications.[7,8,54] The results for targets BACE, BACE (HUNT), BACE (P2), CDK2, GALECTIN, JNK1, MCL1, P38, PDE2, PTP1B, THROMBIN, and TYK2 are retrieved from ref 8. Reference 7 is the source of the results for targets CDK8, CMET, EG5, HIF2A, PFKFB3, SHP2, SYK, and TNKS2. Finally, the results of targets PDE10 and ROS1 are taken from ref 54.

Small Molecule Force Field Parameterizations. Below, we provide small molecule parameterization details. As the simulation data was collected from multiple literature sources, we summarize the particular FF version used for each system in the Supporting Information, Table S.1.

Open Force Field. Open Force Field (OpenFF) parameters were used in 3 different versions (Parsley v1.0.0[49] and v1.2.1 and Sage v.2.0.0[50]). The OpenFF toolkit 0.8.4[48,76] was used to parameterize the ligands with Austin Model 1-bond charge correction (AM1-BCC) charges.[77,78] In the following, the three FFs are named OpenFF-1.0, OpenFF-1.2, and OpenFF-2.0, without the last patch number of the release.

GAFF2.1x. GAFF parameters were assigned by means of Antechamber[79] and ACPYPE.[80] The AM1-BCC partial charge model was used.[77,78] Off-site charges on chlorine and bromine were added according to the rules, as described in ref 81. The effect of the off-site charges in perturbations concerning chlorine and bromine atoms is analyzed in the Supporting Information, Figure S.15. We specify the FF as "GAFF2.1x" as results across the dataset are pulled from two different studies, with some systems using GAFF2.1[8] and a later study using GAFF2.11.[82] Table S.1 lists the exact FF used for each target.

*CGenFF/MATCH**. Small molecule parameterization with the CGenFF[83] was performed by assigning atom types with the MATCH[84] tool and subsequently replacing the bonded parameters with those in CGenFF v3.0.1. For the BACE inhibitor sets, the MATCH algorithm was unable to identify the appropriate atom types; therefore, in these cases, a web-based atom-typing and parameter assignment server[85,86] was used in combination with the CGenFF v4.1 parameters. As for GAFF2.1x above, virtual charged sites were added to chlorine and bromine containing ligands (Supporting Information, Figure S.15).[87] Throughout the article, we refer to this parameterization as CGenFF/MATCH* to mark that several different tools were employed in the parameterization procedure, which may lead to differences in assigned parameters depending on the atom-typing, generalized FF version, and even structure converter used.[88]

OPLS3e. The Schrodinger FF OPLS3[89] and OPLS3e[75] were used in the FEP+ results presented, which were taken from published sources.[7,8,54] Table S.1 lists the source of the results for each target. For simplicity, we labeled all the FEP+ results in the plots and tables as "OPLS3e". Note that differences in results between OPLS3e and the other FFs are not only due to the FF parameters, but may additionally originate from the different MD engine and sampling protocol.

Consensus Approach. For the consensus approach "Consensus", the results were averaged over the first repeat of the simulations using OpenFF-2.0, GAFF2.1x, and

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

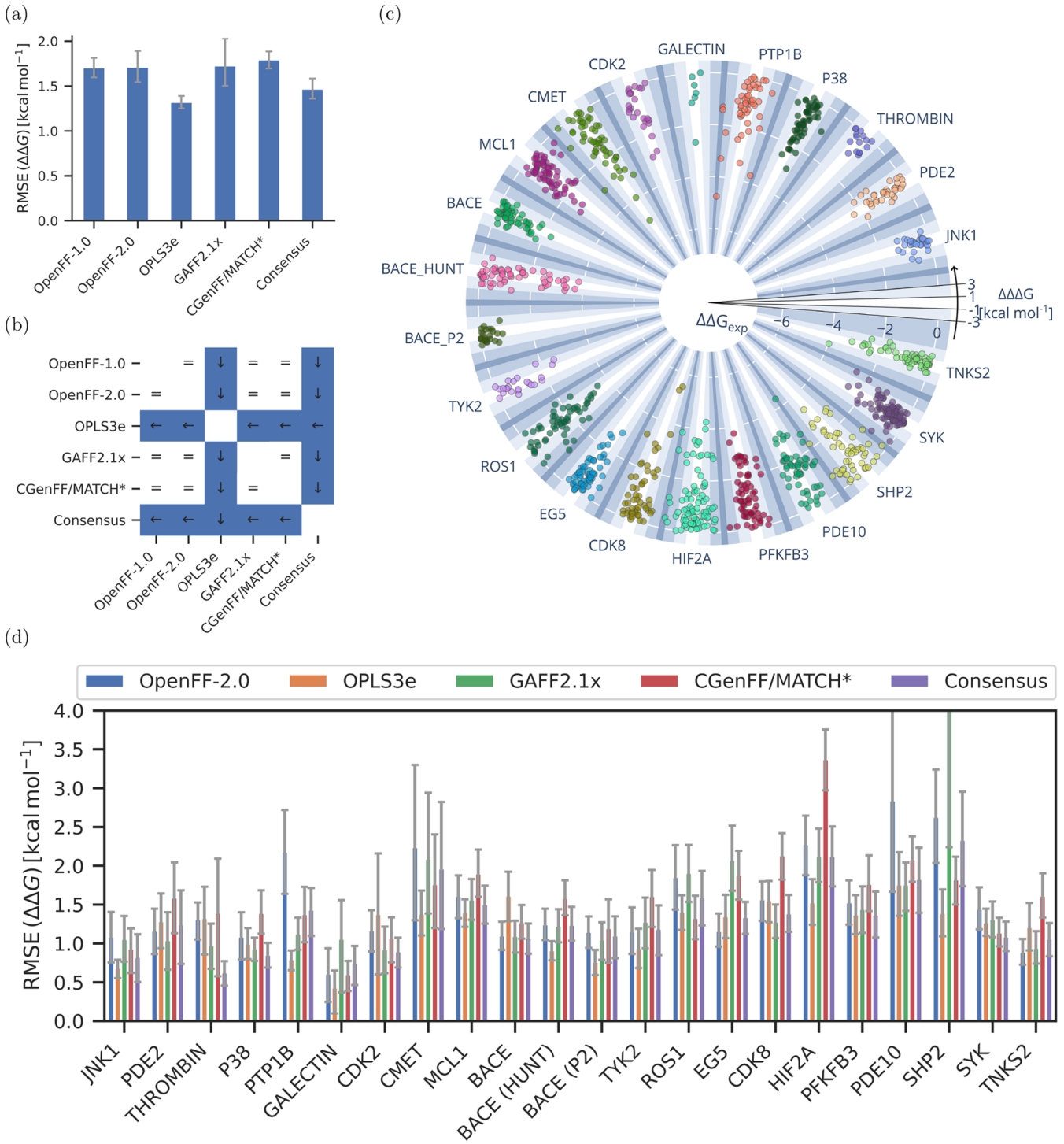


Figure 1. Comparison of ΔΔ*G* values of the perturbations obtained from calculations using the five force fields OpenFF-1.0, OpenFF-2.0, GAFF2.1x, CGenFF/MATCH*, and OPLS3e and the consensus approach. (a) Overall RMSE comparison across all targets and all 1116 perturbations. (b) Illustration of significant differences between pairs of force fields. White matrix element with an equal sign ("=") means that the differences between the two force fields are statistically insignificant. Colored matrix element denotes a significant difference considering a 95% confidence interval. Arrow in a blue matrix element points at the force field, which has the lower error (either left or down). (c) Comparison of all experimental and calculated binding free energy differences for the OpenFF-2.0 Sage force field. All edges belonging to one target are shown in one color in a segment of the circle. Radial distance denotes the experimental $\Delta\Delta G_{exp}$. Deviation of the calculation from experiment is shown on the angular axis as deviation from the segment center (white background). Scale of this deviation is illustrated in the right segment and also coded in background color. (d) RMSE values for each target separately. Each group represents a target set with the RMSE values between experimental and calculated value for the respective force fields in different colors. Lower and upper bound of the 95% confidence interval are given as error bars. Corresponding graph with MUE instead of RMSE can be found in the Supporting Information, Figure S.2.

CGenFF/MATCH*. This sums up to the same sampling time as the results from the single FFs.

Two alternative consensus approaches were calculated, which are presented in the Supporting Information. The first

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

Table 1. Comparison of the Five Force Fields OpenFF-1.0, OpenFF-2.0, *GAFF2.1x*, *CGenFF/MATCH**, OPLS3e, and the Consensus Approach Based on the RMSE of the ΔΔ*G* Values of the Perturbations[a]

	N	RMSE [kcal mol^{-1}]					
		OpenFF 1.0	OpenFF 2.0	CGenFF/MATCH*	GAFF 2.1x	OPLS 3e	Consensus
ALL	1116	$1.7^{1.8}_{1.6}$	$1.7^{1.9}_{1.6}$	$1.8^{1.9}_{1.7}$	$1.7^{2.0}_{1.5}$	$1.3^{1.4}_{1.3}$	$1.5^{1.6}_{1.4}$
BACE	58	$1.0^{1.2}_{0.8}$	$1.1^{1.3}_{0.9}$	$1.3^{1.5}_{1.0}$	$1.1^{1.3}_{0.9}$	$1.6^{1.9}_{1.3}$	$1.1^{1.3}_{0.9}$
BACE (HUNT)	60	$1.1^{1.3}_{0.9}$	$1.3^{1.4}_{1.0}$	$1.5^{1.8}_{1.4}$	$1.2^{1.4}_{1.0}$	$0.9^{1.0}_{0.8}$	$1.2^{1.5}_{1.0}$
BACE (P2)	26	$1.1^{1.3}_{0.9}$	$1.2^{1.3}_{1.0}$	$1.2^{1.6}_{0.8}$	$1.1^{1.3}_{0.8}$	$0.8^{0.9}_{0.6}$	$1.1^{1.3}_{0.8}$
CDK2	25	$1.0^{1.2}_{0.8}$	$1.2^{1.4}_{0.9}$	$1.0^{1.4}_{0.8}$	$0.9^{1.2}_{0.6}$	$1.4^{2.1}_{0.6}$	$0.9^{1.1}_{0.7}$
CDK8	54	$1.7^{2.0}_{1.4}$	$1.6^{1.8}_{1.3}$	$2.1^{2.4}_{1.8}$	$1.2^{1.5}_{1.1}$	$1.5^{1.8}_{1.3}$	$1.4^{1.6}_{1.2}$
CMET	57	$1.9^{2.6}_{1.4}$	$2.2^{3.3}_{1.3}$	$1.7^{2.4}_{1.2}$	$2.1^{2.9}_{1.4}$	$1.3^{1.7}_{1.1}$	$2.0^{2.9}_{1.2}$
EG5	65	$1.7^{2.2}_{1.4}$	$1.1^{1.4}_{1.0}$	$1.8^{2.2}_{1.6}$	$2.1^{2.5}_{1.6}$	$1.3^{1.6}_{1.1}$	$1.4^{1.5}_{1.1}$
GALECTIN	7	$1.0^{1.4}_{0.5}$	$0.6^{0.9}_{0.3}$	$0.6^{0.8}_{0.4}$	$1.0^{1.6}_{0.4}$	$0.4^{0.6}_{0.1}$	$0.7^{1.0}_{0.5}$
HIF2A	80	$2.2^{2.7}_{1.8}$	$2.3^{2.7}_{1.9}$	$3.5^{3.8}_{3.0}$	$2.1^{2.5}_{1.8}$	$1.4^{1.8}_{1.2}$	$2.1^{2.5}_{1.8}$
JNK1	31	$0.9^{1.2}_{0.7}$	$1.1^{1.4}_{0.8}$	$0.9^{1.2}_{0.6}$	$1.0^{1.4}_{0.8}$	$0.7^{0.8}_{0.6}$	$0.8^{1.1}_{0.5}$
MCL1	71	$1.5^{1.8}_{1.3}$	$1.6^{1.9}_{1.3}$	$1.8^{2.2}_{1.6}$	$1.6^{1.8}_{1.3}$	$1.4^{1.6}_{1.2}$	$1.5^{1.7}_{1.3}$
P38	56	$1.3^{1.6}_{1.1}$	$1.0^{1.4}_{0.8}$	$1.3^{1.7}_{1.1}$	$0.9^{1.1}_{0.8}$	$1.0^{1.2}_{0.8}$	$0.9^{1.0}_{0.7}$
PDE10	59	$1.9^{2.3}_{1.5}$	$2.9^{4.2}_{1.6}$	$2.1^{2.4}_{1.8}$	$1.7^{2.1}_{1.4}$	$1.7^{2.1}_{1.4}$	$1.7^{2.3}_{1.4}$
PDE2	34	$1.3^{1.7}_{0.9}$	$1.1^{1.4}_{0.8}$	$1.5^{2.0}_{1.2}$	$1.0^{1.4}_{0.7}$	$1.2^{1.6}_{0.9}$	$1.2^{1.7}_{0.7}$
PFKFB3	66	$1.8^{2.1}_{1.6}$	$1.5^{1.8}_{1.2}$	$1.6^{2.1}_{1.4}$	$1.4^{1.7}_{1.1}$	$1.4^{1.6}_{1.1}$	$1.4^{1.6}_{1.1}$
PTP1B	49	$1.6^{2.1}_{1.1}$	$2.3^{2.7}_{1.6}$	$1.4^{1.8}_{1.0}$	$1.1^{1.3}_{0.9}$	$0.8^{0.9}_{0.7}$	$1.5^{1.7}_{1.1}$
ROS1	61	$2.3^{3.3}_{1.8}$	$1.8^{2.2}_{1.4}$	$1.3^{1.6}_{1.1}$	$1.9^{2.3}_{1.5}$	$1.5^{1.6}_{1.2}$	$1.6^{1.9}_{1.2}$
SHP2	56	$2.6^{3.1}_{2.3}$	$2.6^{3.2}_{2.0}$	$1.8^{2.1}_{1.5}$	$4.3^{6.1}_{2.3}$	$1.3^{1.7}_{1.1}$	$2.3^{3.0}_{1.7}$
SYK	101	$1.3^{1.5}_{1.2}$	$1.4^{1.7}_{1.1}$	$1.1^{1.3}_{1.0}$	$1.4^{1.5}_{1.1}$	$1.2^{1.4}_{1.1}$	$1.1^{1.3}_{0.9}$
THROMBIN	16	$1.3^{1.6}_{1.0}$	$1.3^{1.5}_{1.1}$	$1.5^{2.1}_{0.5}$	$1.0^{1.2}_{0.6}$	$1.2^{1.7}_{0.9}$	$0.6^{0.8}_{0.4}$
TNKS2	60	$0.9^{1.1}_{0.7}$	$0.9^{1.1}_{0.7}$	$1.6^{1.9}_{1.3}$	$0.9^{1.2}_{0.7}$	$1.2^{1.5}_{0.9}$	$1.0^{1.3}_{0.8}$
TYK2	24	$1.1^{1.5}_{0.8}$	$1.1^{1.5}_{0.9}$	$1.6^{2.0}_{1.2}$	$1.3^{1.6}_{0.9}$	$1.0^{1.2}_{0.7}$	$1.1^{1.5}_{0.8}$

[a]Each row represents a target set (or "ALL" for all target sets combined) with a specified number *N* of perturbations followed by the RMSE between experimental and calculated values for the respective FF. The upper and lower bounds of the 95% confidence interval are given as sub- and superscript. All values are in kcal mol^{-1}. The corresponding table with MUE instead of RMSE can be found in the Supporting Information, Figure S.2.

one was obtained from an average over GAFF2.1x and OpenFF2.0 (referred to as "Consensus (OFF, GAFF)"), while the second one was obtained as an average over GAFF2.1x, OpenFF2.0, CgenFF/MATCH*, and OPLS3e (referred to as "Consensus (all)").

Analysis. All the graphs and analyses presented in this article can be followed and reproduced with the Python notebooks available at https://github.com/dfhahn/protein-ligand-benchmark-analysis.[90]

Calculation of ΔΔG and ΔG Values. For the RBFE (ΔΔ*G*) values, we used the raw values without any cycle closure correction as they reflect better potential shortcomings of FFs. For the pmx results, we calculated the ΔΔ*G* values as averages over three repeats, and the standard deviation across the repeats was used as an error estimate.

For the binding free energy estimates (Δ*G*), we calculated the maximum likelihood estimate with the package arsenic[91] for ΔΔ*G* values coming both from FEP+ and pmx.

Metrics. The performance of the calculations employing different FFs is evaluated based on various error and ranking metrics. The aggregated statistics are calculated as the pairwise root mean squared error (RMSE) and mean unsigned error (MUE) of the calculated relative binding free energies (ΔΔ*G*) compared to the experimental values. These were calculated for the individual target sets and the whole set of 1116 edges.

For the final binding free energies of ligands (Δ*G*), the node-based RMSE and MUE were calculated, as well as the ranking coefficients Kendall's τ_K and Spearman's ρ. Again, we calculated the statistics for various subsets of the full dataset as well as for the whole set of 598 ligands. For the calculation of Kendall's $\tau_{K,overall}$ considering the whole dataset, we calculated the weighted average of the Kendall's τ_K of all individual targets

$$\tau_{K,overall} = \frac{1}{N}\sum_{targets} N_{target}\tau_{K,target} \quad (1)$$

where *N* is the sum of all considered ligands across targets, N_{target} is the number of ligands of a target, and $\tau_{K,target}$ is the corresponding Kendall's τ_K of the target. Note that only resulting RMSE values and Kendall's τ_K are discussed in the main text, but values for MUE and Spearman's ρ can be found in the Supporting Information.

Error Calculation. If not stated otherwise, all results are given with a 95% confidence interval, obtained from bootstrapping using 1000 bootstrap samples. The lower and upper bounds of the interval are given as sub- and superscripts behind the actual value.

Significance Test. To evaluate if there is a significant difference between two calculated sets compared to the experiment, we calculated the significance by bootstrapping using a confidence interval of 95%.

Convergence Criteria for Perturbations. To discriminate the error of FF parameters from sampling errors, the set of all edges was filtered according to two convergence criteria indicating issues with sampling. The first criterion is the convergence criterion α based on the overlap of the work distributions from the nonequilibrium sampling. α is defined in the range $-1 \leq \alpha \leq 1$ and is described in more detail in ref 92, eq 5. The second criterion is the standard deviation of the ΔΔ*G* values $\sigma(\Delta\Delta G)$ over the three repeats. For a

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

perturbation to be considered converged, both requirements $\alpha < 0.8$ and $\sigma(\Delta\Delta G) < 1.5$ kcal mol^{-1} must be true.

Parameter Analysis. We performed a parameter analysis to investigate the influence of certain OpenFF parameters on the errors. For each perturbation, the FF parameters involved in the perturbations were identified, i.e., only the parameters that were either changed or annihilated during the perturbation. For each parameter, the RMSE across all perturbations involving this parameter was calculated. As parameters are often used in the same combination (e.g., the bond, angle, and torsion parameters describing an ester group), the correlation between parameters used in the same edges was calculated using the Matthew's correlation coefficient,[93] as it is suited to correlate binary vectors (parameters either used or not used in edges). The obtained correlation matrix between parameters was then clustered with spectral clustering[94] to identify groups of parameters, which are used simultaneously in perturbations. To analyze the influence of a parameter change from OpenFF-1.0 to OpenFF-2.0 on the prediction error, the ΔRMSE of parameter p was calculated as

$$\Delta\mathrm{RMSE}(p) = \mathrm{RMSE}_{\mathrm{OpenFF-2.0}}(p) - \mathrm{RMSE}_{\mathrm{OpenFF-1.0}}(p) \quad (2)$$

where $\mathrm{RMSE}_{\mathrm{FF}}(p)$ is the RMSE between predicted $\Delta\Delta G$ with FF and experimental $\Delta\Delta G$ of all perturbations involving a perturbation of parameter p.

RESULTS AND DISCUSSION

Prediction Accuracy. *Overall Performance of Various Force Fields Analyzed Based on* $\Delta\Delta G$. The general summary of the benchmark study is provided in Figure 1 illustrating all performed RBFE calculations (1116 edges) for 22 targets. In Figure 1c, we used the recent OpenFF, OpenFF-2.0 (Sage), to exemplify the accuracy achievable with the open source FF. The results for each target are shown in different colors in separate segments of the circle. The radial distance denotes experimental $\Delta\Delta G_{\mathrm{exp}}$, showing that there are varying dynamic ranges among the targets. The deviation of the calculation from experiment $\Delta\Delta\Delta G = \Delta\Delta G_{\mathrm{calc}} - \Delta\Delta G_{\mathrm{exp}}$ is shown on the angular axis as a deviation from the segment center (white background). Based on the $\Delta\Delta G$ values of the edges, a RMSE of $1.7^{1.9}_{1.6}$ kcal mol^{-1} (MUE = $1.2^{1.3}_{1.1}$ kcal mol^{-1}) was obtained. This is in line with current industry standards.[7]

Overall, the open source FFs performed comparably to one another and did not show significant differences in terms of $\Delta\Delta G$ prediction for the results averaged over the whole set of targets and chemical series (Figure 1a,b). The obtained RMSE values from the experiment are: GAFF2.1x $1.7^{2.0}_{1.5}$, OpenFF-1.0 $1.7^{1.8}_{1.6}$, OpenFF-2.0 $1.7^{1.9}_{1.6}$, and CGenFF/MATCH* $1.8^{1.9}_{1.7}$ kcal mol^{-1}. It is interesting to note that a consensus variant constructed as a linear combination over three open source FFs significantly outperformed each of the open source FFs considered separately (RMSE of $1.5^{1.6}_{1.4}$). The OPLS3e FF shows a significantly lower RMSE of $1.3^{1.4}_{1.3}$ kcal mol^{-1} when averaged over all $\Delta\Delta G$ values calculated in this work. Note that more recent versions of FEP+ using the OPLS4[95] FF should lead to more accurate results.[10] However, we refrain from comparing to OPLS4 results as there are no results available using the same input structures.

Table 1 and Figure 1d list the per-target accuracy reached by each FF in terms of $\Delta\Delta G$ RMSE from experimental measurement. The corresponding $\Delta\Delta G$ MUE values can be found in Table S.3 and Figure S.2. This illustrates well that the prediction accuracy is case-dependent. For example, the predicted $\Delta\Delta G$ for GALECTIN in Figure 1 all fall close to the experimentally measured values. Whereas, several other cases, e.g., HIF2A and SHP2, have a widespread distribution of calculated relative free energy differences when compared to the experimental measurement.

Although the aggregated RMSE statistics overall (Figure 1a) or per-target (Figure 1d) do not show a significant difference between the public FFs, the differences become more apparent by looking at the number of outliers. Figure 2 shows the ratio

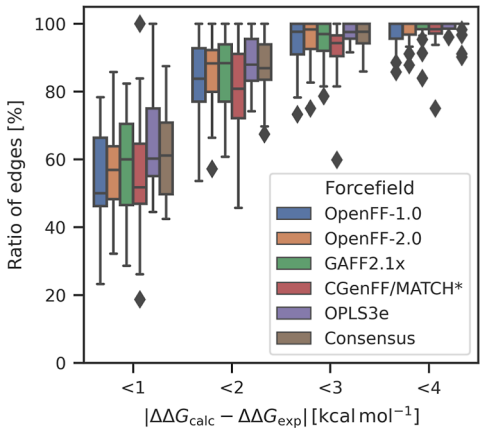


Figure 2. Ratio of calculated $\Delta\Delta G$ within various different absolute error thresholds compared to the experimental value for the different force fields. Box-and-whiskers show the distribution across the various targets. Each box illustrates the first and third quartiles with the median shown as a horizontal bar inside the box and the whiskers are at 1.5× (interquartile range) from the box edges.

of perturbations with absolute errors versus experiments below a certain threshold. Each box illustrates the distribution across the various targets first and third quartiles, with the median shown as a horizontal bar inside the box, and the whiskers extend up to the minimum (least performing target) and maximum (highest performing target), but at most up to 1.5× (interquartile range) from the box edges (with outliers shown as markers). We observed differences between the FFs in minimum, median, and maximum ratios. For a threshold of 1 kcal mol^{-1} from experiment, the median across targets is at 50% of edges for OpenFF-1.0 and 52% for CGenFF/MATCH*. This median ratio is notably higher for OpenFF-2.0 (57%), GAFF2.1x (60%), OPLS3e (60%), and the consensus approach (61%). Also, the trend of the ratio for the worst performing targets is similar. For the public FFs, the worse performing targets exhibit between 19 and 32% of edges within a 1 kcal mol^{-1} threshold. For OPLS3e and the consensus approach, this ratio is considerably higher at 44 and 42%, respectively.

These trends persist when looking at higher unsigned error thresholds of 2, 3, or 4 kcal mol^{-1}.

A strong target dependence of the accuracy of the results can be clearly seen. For OpenFF-1.0 and a threshold of <1 kcal mol^{-1} from the experiment (left blue box in Figure 2), only 23% of the edges agreed with the experiment within the threshold for the worst-performing target (SHP2). On the other hand, 78% of edges in the best-performing target (TNKS2) were correct considering the threshold. This difference between the worst- and best-performing targets can be reduced with the consensus approach, which seems to

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

correct for large outliers. Various reasons can lead to a disproportionate number of outliers for a few targets. One reason can be inaccuracies in the setup of the starting structures. This could be the wrong starting poses of the ligand, inadequate protein preparation, or unlikely protonation or tautomeric states, both in the ligand and in the protein. If all FFs show low performance for a specific target it suggests a common preparation error. The protein and ligands might be more flexible in certain targets, and the free energy estimate only converges if two or more conformational states are sampled sufficiently. Thus, more sampling or even enhanced sampling would be needed to adequately model such a target. Some targets have ligand sets with more difficult perturbations. For example, charge changes, charge redistribution, or the creation/annihilation of large moieties like cyclohexyl groups are difficult perturbations, which either would require longer sampling times, or are even better treated with absolute binding free energy approaches.[96,97] Some targets might feature certain chemical moieties, which are not adequately described by the respective FF. The use of inadequate parameters may explain why the use of OPLS3e leads to fewer outliers, as the use of custom parameters describes specific chemistries better than a general FF.[75,95,98] Finally, the experimental results might not be entirely suitable for comparing to calculated binding free energies.[10] The MD calculations may not mimic the exact experimental conditions (temperature, ion concentrations, and cosolvents), or the assay may only have limited correlation with the binding free energy that is targeted in the RBFE calculations. But this has a limited impact when comparing the different FFs, as they are all compared to the same data.

Accuracy of Predicted ΔG. Figure 3 shows the trend in significant differences between FFs changes when comparing accuracy in terms of back-calculated absolute binding free energies ΔG. In this analysis, in terms of RMSE to experimental measurement, OPLS3e still significantly outperforms OpenFF-2.0 and CGenFF/MATCH*; however, its difference to OpenFF-1.0 and GAFF2.1x is no longer significant (Figure 3a,b). The consensus approach outperforms the individual open source FFs, similarly as it was for the $\Delta\Delta G$ comparison.

We also compared FF predictions in terms of their ability to correctly rank binders based on their ΔG values by using Kendall's τ_K correlation coefficient (τ_K). This measure again reveals the same two variants outperforming the others—OPLS3e and the consensus approach. While the pattern of significant differences between FFs is rather complex (Figure 3d), the differences are small in magnitude, showing that each of the FFs can be trusted to yield a compound ranking of similar quality. The Supporting Information, Figures S.5−S.8 illustrate aggregated statistics based on ΔG per target and across all targets for all the FFs, including the consensus approaches. The corresponding values can be found in the Supporting Information, Tables S.6−S.9. Additionally, correlation plots are provided for OpenFF-2.0, CGenFF/MATCH*, GAFF2.1x, OPLS3e, and the consensus approach in the Supporting Information, Figures S.9−S.12.

Determinants of the Prediction Accuracy. There are numerous underlying causes for the differences in accuracy in addition to the small molecule FF, e.g., sampling, specifics of the calculation procedure, and initial system setup. In the analysis in Figure 4, we attempted to elucidate the main

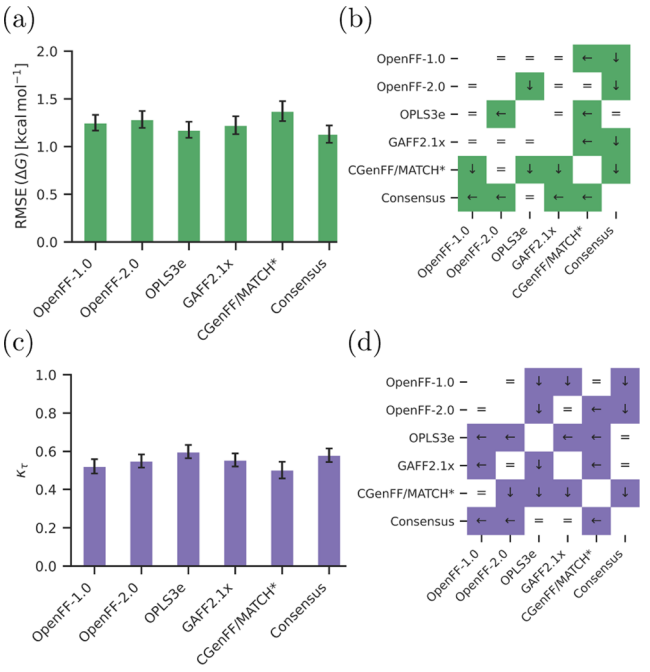


Figure 3. Comparison of ΔG values of the ligands obtained from calculations using the five force fields OpenFF-1.0, OpenFF-2.0, GAFF2.1x, CGenFF/MATCH*, and OPLS3e and the consensus approach. (a) RMSE comparison across all targets and 598 ligands. (b) Illustrations of significance of differences between the different sets. (c) Comparison of ranking metric τ_K across all targets and 598 ligands. (d) Illustrations of significance of differences between the different sets. Colors denote the different metrics (green for RMSE and purple for τ_K). In panels (b) and (d), a white matrix element with an equal sign ("=") means that the differences between the two force fields are statistically insignificant. Colored matrix element means there is a significant difference considering a 95% confidence interval. Arrow in a colored matrix element points at the force field, which has the lower error (either left or down).

determinants underlying $\Delta\Delta G$ prediction accuracy related to the convergence of an alchemical perturbation.

In particular, we noticed that larger calculated $\Delta\Delta G$ values are associated with a larger error (Figure 4f). Namely, the alchemical approach can be expected to become less accurate when the predicted change in free energy of binding is large. This effect is in turn explained by the difficulty in converging such perturbations: predicted large free energy differences correlate with the lack in convergence of the estimates (Figure 4d). While there are many factors influencing the convergence of an alchemical perturbation, we observed that a simple count of heavy atoms that need to be introduced/annihilated shows a low, but statistically significant correlation (Pearson's r = 0.08, p-value <0.01) with the absolute error (Figure 4c) and larger correlation with the convergence measure (Figure 4a). Similar trends as for the heavy atom count can be seen in the Supporting Information for the counts of rotatable bonds (Figure S.22), counts of rings (Figure S.23), changes or positions of the formal charges (Figure S.24), and the LOMAP score[26] (Figure S.25). In Figure 4, we used the $\Delta\Delta G$ values and convergence metric α of the simulations using OpenFF-2.0. Although the edges might show different levels of convergence between the FFs (Figure S.13), overall we found that the ratios of converged simulations differ insignificantly among OpenFF-1.0, OpenFF-2.0, GAFF2.1x, and CGenFF/MATCH* (Figure S.14). Moreover, we

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

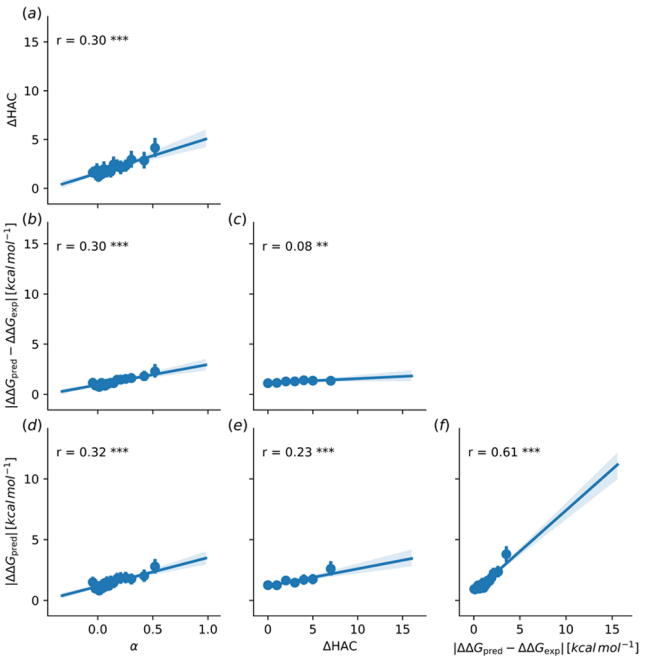


Figure 4. Visualization of pairwise relationships between the change in number of heavy atoms in the end states, the absolute error between experimental and calculated values $|\Delta\Delta G_{pred} - \Delta\Delta G_{exp}|$, the calculated relative free energies $\Delta\Delta G_{pred}$ (OpenFF-2.0), and the average convergence measure α[92] (averaged over three solvent and three complex simulation legs). Subplots show linear regression plots between the respective properties. Pearson's correlation coefficient is given in the graph together with its *p*-value indicated as stars (one, two, or three stars for the confidence level of <0.05, <0.01, and <0.001, respectively). For illustration purposes, the data was binned into 20 bins and their average with standard deviation are shown as dot with error bars. Regression was performed on the original data. Panels a, c, d, and f mark the trends described in the text. More detailed illustration of this figure is shown in the Supporting Information, Figure S.17.

observed the same trends as described above for OpenFF-2.0 for the results calculated with OpenFF-1.0 (Figure S.16), GAFF2.1x (Figure S.18), and CGenFF/MATCH* (Figure S.19).

All in all, this simple trace through the dependencies in the data already reveals some of the determinants limiting the accuracy of our predictions. For larger perturbations, the calculation convergence suffers, thus reducing the agreement between the prediction and experiment. It is important to note, however, that the identified signal is noisy, i.e., not every large perturbation will be inaccurate and not all well converged simulations will yield perfect binding free energy predictions. The identified determinants for prediction accuracy are only general trends in a complex picture.

In addition to these factors, the accuracy of the prediction will also be influenced by the technical setup of the calculation procedure. For example, it has been observed that even file conversion by different software packages may introduce artifacts in molecular structure.[88] Also, combining small molecule FFs with disparate charge models will have an effect on the prediction accuracy.[99,100] Differences between simulation packages[101] and free energy protocols[102] will influence the sampling and, subsequently, the final free energy estimates. Considering the limited sampling used in the standard free energy calculation protocols, the starting structure quality often affects the prediction accuracy.[12−14]

OpenFF Improvement. *Nonconverged Results Are Less Accurate.* The difference between the set of all results and the converged set is illustrated in Figure 5a as histograms of deviations between experimental and calculated values (see the Methods Section for details about the convergence criteria). Whereas all edges consisting of converged and nonconverged perturbations show a large standard deviation of 1.72 kcal mol^{-1}, the filtered set of 850 converged edges has a reduced standard deviation of 1.35 kcal mol^{-1}, while the remaining 278 not converged edges are enriched in outliers resulting in a larger standard deviation of 2.54 kcal mol^{-1}. The convergence criteria can therefore be used to flag calculations, which are likely to have larger errors without prior knowledge of experimental results.

Figure 5b,d and Table 2 compare three OpenFF versions by means of RMSE between calculated and experimental $\Delta\Delta G$ values for results obtained on a subset of 551 perturbations (of which 340 are converged) in eight different targets. While the intermediate version OpenFF-1.2 did not show an improvement over OpenFF-1.0, OpenFF-2.0 significantly improved in accuracy compared to the previous OpenFF-1.2 (Figure 5c). This trend holds both for all edges and the converged set of edges.

Effect of Force Field Parameter Change from OpenFF-1.0 to OpenFF-2.0. In Figure 6a, we highlight FF parameter changes between two OpenFF versions, 1.0 and 2.0, and their effect on the predicted free energy accuracy for the cases where the effect is statistically significant. In these cases, various other factors influencing the accuracy like starting conformations and convergence cannot be the cause for the difference; therefore, it is more likely that the underlying reason is the FF parameters. For example, an ester group is described by its angle (OpenFF code a15), bond (b20), improper (i2), and torsion (t107, t110) parameters, which were modified between the OpenFF releases. Altogether, the RMSE between the predicted and experimental $\Delta\Delta G$ for the perturbations of the ester groups drops by 0.5 kcal mol^{-1} when going from OpenFF-1.0 to OpenFF-2.0 (Figure 6a). An example for a perturbation involving an ester group is shown in Figure 6b: in this case, the new OpenFF-2.0 parameters led to a reduction in the error of $\Delta\Delta G$ by 1.1 kcal mol^{-1}.

Similar trends are observed for the other significant changes in FF parameters: the predicted free energy difference is more accurate for the modified parameters. The largest improvement in this analysis was observed for the changes in the hydroxyl group bound to a sp^2 carbon involving the bond (b18) and torsion (t106) parameters. Figure 6c illustrates a case where this improvement resulted in 1.3 kcal mol^{-1} increase in free energy calculation accuracy.

There are only a few parameter groups that result in decreased $\Delta\Delta G$ prediction accuracy for OpenFF-2.0 compared to OpenFF-1.0. Namely, changes in parameters describing sulfur-containing groups like thioethers (a38, b51) or sulfonamides (t145, t148) and torsions (t13 and t14) describing cyclopropyl groups appear to have a detrimental effect on binding affinity accuracy.

The improvement of free energy results related to parameter changes is remarkable as the parameters were designed on the condensed phase and QM properties of small molecules. We show that improving the latter properties also has a positive

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

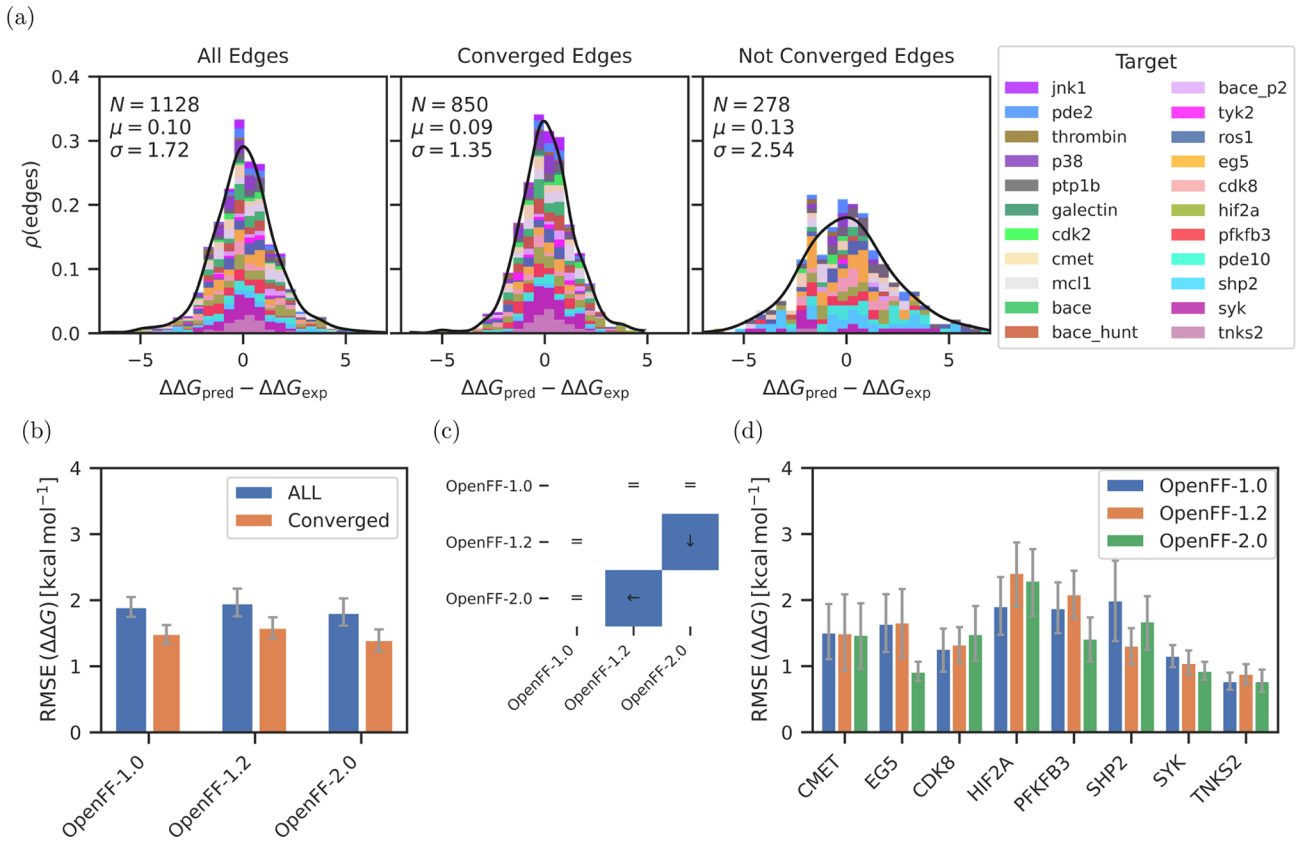


Figure 5. Comparison of the three force fields OpenFF-1.0, OpenFF-1.2, and OpenFF-2.0 based on the $\Delta\Delta G$ values. Panel (a) shows the absolute error distributions between experimental and calculated $\Delta\Delta G$ using OpenFF-2.0 for three sets of edges. First set in the left subpanel contains all edges, the second set in the center contains only converged edges, and the third set in the right contains the not converged edges (which is the difference set between the first and second set). See the Methods Section for more details about the convergence criteria. Different colors denote the different targets and the black line is a normal distribution fitted to the data. Text in the panel lists the number of edges N, the center μ, and the standard deviation σ of the normal distribution. Panel (b) shows the RMSE across all edges of 8 targets for the three force fields of the OpenFF family. Blue bars correspond to all edges and the orange bars only to the converged ones. Panel (c) illustrates significant differences between the force field sets shown in panel (b). White matrix element with an equal sign ("=") means that the differences between the two force fields are statistically insignificant. Blue matrix element denotes a significant difference considering a 95% confidence interval. Arrow in a blue matrix element points at the force field, which has the lower error. Panel (d) shows the RMSE of the $\Delta\Delta G$ values per target for the three force fields OpenFF-1.0, OpenFF-1.2, and OpenFF-2.0. Lower and upper bound of the 95% confidence interval are given as error bars. All values are in kcal mol^{-1}.

Table 2. Comparison of the Three Force Fields OpenFF-1.0, OpenFF-1.2, and OpenFF-2.0 Based on the RMSE of the $\Delta\Delta G$ Values of the Converged Perturbations[a]

	N	RMSE [kcal mol^{-1}]		
		OpenFF 1.0	OpenFF 1.2	OpenFF 2.0
ALL	320	$1.5^{1.6}_{1.4}$	$1.5^{1.7}_{1.4}$	$1.4^{1.6}_{1.2}$
CDK8	27	$1.3^{1.6}_{0.9}$	$1.3^{1.6}_{1.1}$	$1.4^{1.9}_{1.1}$
CMET	35	$1.5^{2.0}_{1.1}$	$1.5^{2.1}_{0.9}$	$1.4^{1.9}_{1.0}$
EG5	29	$1.6^{2.1}_{1.2}$	$1.6^{2.1}_{1.2}$	$0.9^{1.1}_{0.8}$
HIF2A	45	$1.8^{2.3}_{1.5}$	$2.4^{2.9}_{1.9}$	$2.3^{2.8}_{1.8}$
PFKFB3	42	$1.9^{2.2}_{1.5}$	$2.0^{2.4}_{1.7}$	$1.4^{1.7}_{1.1}$
SHP2	17	$1.9^{2.5}_{1.4}$	$1.3^{1.6}_{1.0}$	$1.7^{2.1}_{1.3}$
SYK	74	$1.7^{2.1}_{1.3}$	$1.0^{1.2}_{0.9}$	$0.9^{1.1}_{0.8}$
TNKS2	51	$0.8^{0.9}_{0.6}$	$0.9^{1.0}_{0.7}$	$0.8^{1.0}_{0.6}$

[a]Each row represents a target set (or "all" for all target sets combined) with a specified number N of perturbations followed by the RMSE between experimental and calculated values for the respective FF. The upper and lower bounds of the 95% confidence interval are given as sub- and superscript. All values are in kcal mol^{-1}. The values are illustrated in Figure 5b,d.

and significant effect on the downstream free energy of binding calculation results.

CONCLUSIONS

On a set of 598 ligands each binding to one of 22 targets, we showed that the public FFs OpenFF-1.0 (Parsley), OpenFF-2.0 (Sage), GAFF2.1x, and CGenFF/MATCH* are performing comparably based on aggregated statistics across the whole dataset, both in terms of the RMSE of relative binding free energies $\Delta\Delta G$ (perturbations) and the RMSE and Kendall's tau of binding free energies ΔG. The proprietary FF OPLS3e performs significantly better, but a consensus approach based on Sage, GAFF2.1x, and CGenFF/MATCH* is similarly accurate based on ΔG regarding the RMSE and Kendall's τ. There is a clear target dependence, which can be attributed to input preparation, protein (binding pocket) flexibility, chemistry of ligands, and difficulty of perturbations (in terms of heavy atom changes). While Parsley and Sage are performing comparably based on aggregated statistics across the whole dataset, there are differences in terms of outliers. A parameter analysis revealed that improved parameters lead to significant improvement in the accuracy of affinity predictions

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

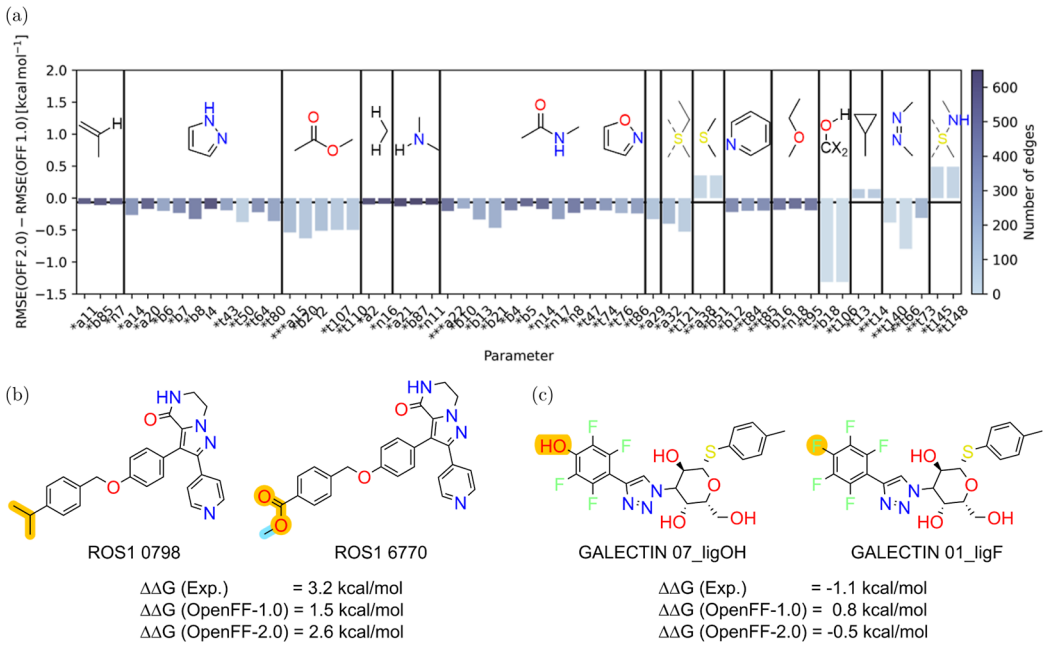


Figure 6. Analysis of parameter differences between OpenFF-1.0 and OpenFF-2.0. Panel (a) shows the RMSE difference between OpenFF-1.0 and OpenFF-2.0 for subsets of converged edges, where a certain parameter is perturbed (*x*-axis) and the difference between OpenFF-1.0 and OpenFF-2.0 is significant (CI 95%). Stars (*) in front of the parameter label denote how much the parameter changed between OpenFF-1.0 and OpenFF-2.0 (3 stars denote the largest change). Horizontal black line denotes the insignificant difference (−0.06 kcal mol^{-1}) for the whole set of perturbations. Vertical bars separate groups of bars with high correlations, i.e., they are usually employed concurrently in perturbations. Chemical structure shows an example substructure where each group of parameters is employed. Panel (b) shows an example perturbation where an ester function (third group in panel (a)) is introduced. Free energy prediction improved from $\Delta\Delta G$ = 1.5 kcal mol^{-1} (OpenFF-1.0) to $\Delta\Delta G$ = 2.6 kcal mol^{-1} (OpenFF-2.0) with an experimental value of $\Delta\Delta G$ = 3.2 kcal mol^{-1}. Panel (c) shows an example perturbation where an aromatic hydroxy group (fourth group from the right in panel (a)) is morphed into a fluorine atom. Free energy prediction improved from $\Delta\Delta G$ = 0.8 kcal mol^{-1} (OpenFF-1.0) to $\Delta\Delta G$ = −0.5 kcal mol^{-1} (OpenFF-2.0) with an experimental value of $\Delta\Delta G$ = −1.1 kcal mol^{-1}. In panels (b) and (c), the perturbed atoms and bonds are highlighted in orange, whereas annihilated atoms and bonds are highlighted in cyan.

on more than 50 subsets of the dataset involving those parameters, while six subsets involving certain parameters showed lower accuracy. Thus, we can show that there is a considerable improvement of successive OpenFF versions.

In the future, such a parameter analysis can be used to identify potentially problematic parameters, which can then be investigated and improved for next FF versions. Indeed, this study also allowed us to identify parameters in well converged but inaccurate perturbations, along with further calculations, this provides future investigation and possible avenues for FF improvement. However, for this to be successful, further work would be valuable to reduce the influence of other (non FF parameter) sources of errors like large or difficult perturbations, inadequate input preparation, or insufficient sampling.

■ ASSOCIATED CONTENT

Data Availability Statement

The used input structures, molecular dynamics topologies, and experimental data are provided as a Zenodo record (https://zenodo.org/records/10782775)[103] or can be retrieved from the protein−ligand benchmark GitHub repository, release 0.2.1 (https://github.com/openforce-field/protein-ligand-benchmark/).[59] Simulation input was prepared, and trajectories were analyzed with pmx, which is freely available on GitHub (https://github.com/deGrootLab/pmx).[60] Corresponding input parameter files and workflow scripts used to prepare the simulations are also available on GitHub (https://github.com/dfhahn/pmx) or as a Zenodo record.[104,105] The simulations were run with Gromacs (http://gromacs.org/).[106] The analysis code and code to create figures are provided in the protein−ligand benchmark analysis repository, release 0.3.0 (https://github.com/dfhahn/protein-ligand-benchmark-analysis).[90]

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00417.

Details about the employed target set, additional graphs and tables containing aggregated statistics and correlations with experiment in greater detail, and analysis of various properties of the simulated perturbations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

David F. Hahn − *Computational Chemistry, Janssen Research & Development, Beerse 2340, Belgium;* orcid.org/0000-0003-2830-6880; Email: dhahn3@its.jnj.com

Authors

Vytautas Gapsys − *Computational Biomolecular Dynamics Group, Max Planck Institute for Multidisciplinary Sciences, Göttingen 37077, Germany; Computational Chemistry, Janssen Research & Development, Beerse 2340, Belgium;* orcid.org/0000-0002-6761-7780

Bert L. de Groot − *Computational Biomolecular Dynamics Group, Max Planck Institute for Multidisciplinary Sciences, Göttingen 37077, Germany;* orcid.org/0000-0003-3570-3534

David L. Mobley − *Department of Chemistry and Department of Pharmaceutical Sciences, University of California, Irvine, California 92697, United States;* orcid.org/0000-0002-1083-5533

Gary Tresadern − *Computational Chemistry, Janssen Research & Development, Beerse 2340, Belgium;* orcid.org/0000-0002-4801-1644

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00417

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Lorenzo d'Amore, Marjolein Crabbe, Benjamin Ries, Chris Bayly, and John Chodera for their insightful discussions. DLM appreciates financial support from the National Institutes of Health (R01GM132386). We thank the OpenFF Consortium and Initiative for their scientific support, and the Molecular Sciences Software Institute (MolSSI) and the Open Molecular Software Foundation (OMSF) for their support of the OpenFF Initiative.

■ REFERENCES

(1) Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4153−4169.

(2) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical Free Energy Calculations for Drug Discovery. *J. Chem. Phys.* **2012**, *137*, 230901.

(3) Mey, A. S.; Allen, B. K.; Bruce Macdonald, H. E.; Chodera, J. D.; Hahn, D. F.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *LiveCoMS* **2020**, *2*, 18378.

(4) Kutzner, C.; Kniep, C.; Cherian, A.; Nordstrom, L.; Grubmüller, H.; de Groot, B. L.; Gapsys, V. GROMACS in the Cloud: A Global Supercomputer to Speed Up Alchemical Drug Design. *J. Chem. Inf. Model.* **2022**, *62*, 1691−1711.

(5) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.

(6) Song, L. F.; Lee, T.-S.; Zhu, C.; York, D. M.; Merz, K. M. Using AMBER18 for Relative Free Energy Calculations. *J. Chem. Inf. Model.* **2019**, *59*, 3128−3135.

(7) Schindler, C. E. M.; Baumann, H.; Blum, A.; Böse, D.; Buchstaller, H.-P.; Burgdorf, L.; Cappel, D.; Chekler, E.; Czodrowski, P.; Dorsch, D.; Eguida, M. K. I.; Follows, B.; Fuchß, T.; Grädler, U.; Gunera, J.; Johnson, T.; Jorand Lebrun, C.; Karra, S.; Klein, M.; Knehans, T.; Koetzner, L.; Krier, M.; Leiendecker, M.; Leuthner, B.; Li, L.; Mochalkin, I.; Musil, D.; Neagu, C.; Rippmann, F.; Schiemann, K.; Schulz, R.; Steinbrecher, T.; Tanzer, E.-M.; Unzue Lopez, A.; Viacava Follis, A.; Wegener, A.; Kuhn, D. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J. Chem. Inf. Model.* **2020**, *60*, 5457−5474.

(8) Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L. Large Scale Relative Protein Ligand Binding Affinities Using Non-Equilibrium Alchemy. *Chem. Sci.* **2020**, *11*, 1140−1152.

(9) Kuhn, M.; Firth-Clark, S.; Tosco, P.; Mey, A. S. J. S.; Mackey, M.; Michel, J. Assessment of Binding Affinity via Alchemical Free-Energy Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 3120−3130.

(10) Ross, G. A.; Lu, C.; Scarabelli, G.; Albanese, S. K.; Houang, E.; Abel, R.; Harder, E. D.; Wang, L. The Maximal and Current Accuracy of Rigorous Protein-Ligand Binding Free Energy Calculations. *Commun. Chem.* **2023**, *6*, 222.

(11) Tresadern, G.; Tatikola, K.; Cabrera, J.; Wang, L.; Abel, R.; Van Vlijmen, H.; Geys, H. The Impact of Experimental and Calculated Error on the Performance of Affinity Predictions. *J. Chem. Inf. Model.* **2022**, *62*, 703−717.

(12) Shih, A. Y.; Hack, M.; Mirzadegan, T. Impact of Protein Preparation on Resulting Accuracy of FEP Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 5287−5289.

(13) Cappel, D.; Jerome, S.; Hessler, G.; Matter, H. Impact of Different Automated Binding Pose Generation Approaches on Relative Binding Free Energy Simulations. *J. Chem. Inf. Model.* **2020**, *60*, 1432−1444.

(14) Suruzhon, M.; Bodnarchuk, M. S.; Ciancetta, A.; Viner, R.; Wall, I. D.; Essex, J. W. Sensitivity of Binding Free Energy Calculations to Initial Protein Crystal Structure. *J. Chem. Theory Comput.* **2021**, *17*, 1806−1821.

(15) Hahn, D. F.; Bayly, C. I.; Macdonald, H. E. B.; Chodera, J. D.; Gapsys, V.; Mey, A. S. J. S.; Mobley, D. L.; Benito, L. P.; Schindler, C. E. M.; Tresadern, G.; Warren, G. L. Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]. *Living J. Comput. Mol. Sci.* **2022**, *4*, 1497.

(16) Hahn, D. F.; König, G.; Hünenberger, P. H. Overcoming Orthogonal Barriers in Alchemical Free Energy Calculations: On the Relative Merits of λ-Variations, λ-Extrapolations, and Biasing. *J. Chem. Theory Comput.* **2020**, *16*, 1630−1645.

(17) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115*, 9431−9438.

(18) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *J. Chem. Theory Comput.* **2013**, *9*, 1282−1293.

(19) Itoh, S. G.; Okumura, H. Replica-Permutation Method with the Suwa−Todo Algorithm beyond the Replica-Exchange Method. *J. Chem. Theory Comput.* **2013**, *9*, 570−581.

(20) Hahn, D. F.; Hünenberger, P. H. Alchemical Free-Energy Calculations by Multiple-Replica λ-Dynamics: The Conveyor Belt Thermodynamic Integration Scheme. *J. Chem. Theory Comput.* **2019**, *15*, 2392−2419.

(21) Hansen, N.; Hünenberger, P. H.; van Gunsteren, W. F. Efficient Combination of Environment Change and Alchemical Perturbation within the Enveloping Distribution Sampling (EDS) Scheme: Twin-System EDS and Application to the Determination of Octanol−Water Partition Coefficients. *J. Chem. Theory Comput.* **2013**, *9*, 1334−1346.

(22) Hsu, W.-T.; Piomponi, V.; Merz, P. T.; Bussi, G.; Shirts, M. R. Alchemical Metadynamics: Adding Alchemical Variables to Metadynamics to Enhance Sampling in Free Energy Calculations. *J. Chem. Theory Comput.* **2023**, *19*, 1805−1817.

(23) Gill, S. C.; Lim, N. M.; Grinaway, P. B.; Rustenburg, A. S.; Fass, J.; Ross, G. A.; Chodera, J. D.; Mobley, D. L. Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *J. Phys. Chem. B* **2018**, *122*, 5579−5598.

(24) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J. Chem. Theory Comput.* **2018**, *14*, 6127−6138.

(25) Gapsys, V.; de Groot, B. L. On the Importance of Statistics in Molecular Simulations for Thermodynamics, Kinetics and Simulation Box Size. *Elife* **2020**, *9*, No. e57589.

(26) Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. Lead Optimization

Mapper: Automating Free Energy Calculations for Lead Optimization. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 755−770.
(27) Xu, H. Optimal Measurement Network of Pairwise Differences. *J. Chem. Inf. Model.* **2019**, *59*, 4720−4728.
(28) Pitman, M.; Hahn, D. F.; Tresadern, G.; Mobley, D. L. To Design Scalable Free Energy Perturbation Networks, Optimal Is Not Enough. *J. Chem. Inf. Model.* **2023**, *63*, 1776−1793.
(29) Gapsys, V.; Seeliger, D.; de Groot, B. L. New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 2373−2382.
(30) Naden, L. N.; Pham, T. T.; Shirts, M. R. Linear Basis Function Approach to Efficient Alchemical Free Energy Calculations. 1. Removal of Uncharged Atomic Sites. *J. Chem. Theory Comput.* **2014**, *10*, 1128−1149.
(31) König, G.; Ries, B.; Hünenberger, P. H.; Riniker, S. Efficient Alchemical Intermediate States in Free Energy Calculations Using λ-Enveloping Distribution Sampling. *J. Chem. Theory Comput.* **2021**, *17*, 5805−5815.
(32) Tsai, H.-C.; Lee, T.-S.; Ganguly, A.; Giese, T. J.; Ebert, M. C.; Labute, P.; Merz, K. M.; York, D. M. AMBER Free Energy Tools: A New Framework for the Design of Optimized Alchemical Transformation Pathways. *J. Chem. Theory Comput.* **2023**, *19*, 640−658.
(33) Baumann, H. M.; Dybeck, E.; McClendon, C. L.; Pickard, F. C.; Gapsys, V.; Pérez-Benito, L.; Hahn, D. F.; Tresadern, G.; Mathiowetz, A. M.; Mobley, D. L. Broadening the Scope of Binding Free Energy Calculations Using a Separated Topologies Approach. *J. Chem. Theory Comput.* **2023**, *19*, 5058−5076.
(34) Christ, C. D.; van Gunsteren, W. F. Enveloping Distribution Sampling: A Method to Calculate Free Energy Differences from a Single Simulation. *J. Chem. Phys.* **2007**, *126*, 184110.
(35) Sidler, D.; Cristòfol-Clough, M.; Riniker, S. Efficient Round-Trip Time Optimization for Replica-Exchange Enveloping Distribution Sampling (RE-EDS). *J. Chem. Theory Comput.* **2017**, *13*, 3020−3030.
(36) Perthold, J. W.; Oostenbrink, C. Accelerated Enveloping Distribution Sampling: Enabling Sampling of Multiple End States while Preserving Local Energy Minima. *J. Phys. Chem. B* **2018**, *122*, 5030−5037.
(37) Ries, B.; Normak, K.; Weiß, R. G.; Rieder, S.; Barros, E. P.; Champion, C.; König, G.; Riniker, S. Relative Free-Energy Calculations for Scaffold Hopping-Type Transformations with an Automated RE-EDS Sampling Procedure. *J. Comput.-Aided Mol. Des.* **2022**, *36*, 117−130.
(38) Ge, Y.; Wych, D. C.; Samways, M. L.; Wall, M. E.; Essex, J. W.; Mobley, D. L. Enhancing Sampling of Water Rehydration on Ligand Binding: A Comparison of Techniques. *J. Chem. Theory Comput.* **2022**, *18*, 1359−1381.
(39) Ekberg, V.; Samways, M. L.; Misini Ignjatović, M.; Essex, J. W.; Ryde, U. Comparison of Grand Canonical and Conventional Molecular Dynamics Simulation Methods for Protein-Bound Water Networks. *ACS Phys. Chem. Au* **2022**, *2*, 247−259.
(40) Gracia Carmona, O.; Gillhofer, M.; Tomasiak, L.; De Ruiter, A.; Oostenbrink, C. Accelerated Enveloping Distribution Sampling to Probe the Presence of Water Molecules. *J. Chem. Theory Comput.* **2023**, *19*, 3379−3390.
(41) Ben-Shalom, I. Y.; Lin, Z.; Radak, B. K.; Lin, C.; Sherman, W.; Gilson, M. K. Accounting for the Central Role of Interfacial Water in Protein−Ligand Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2020**, *16*, 7883−7894.
(42) Bergazin, T. D.; Ben-Shalom, I. Y.; Lim, N. M.; Gill, S. C.; Gilson, M. K.; Mobley, D. L. Enhancing Water Sampling of Buried Binding Sites Using Nonequilibrium Candidate Monte Carlo. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 167−177.
(43) Ross, G. A.; Bodnarchuk, M. S.; Essex, J. W. Water Sites, Networks, And Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **2015**, *137*, 14930−14943.
(44) Ross, G. A.; Russell, E.; Deng, Y.; Lu, C.; Harder, E. D.; Abel, R.; Wang, L. Enhancing Water Sampling in Free Energy Calculations with Grand Canonical Monte Carlo. *J. Chem. Theory Comput.* **2020**, *16*, 6061−6076.
(45) Melling, O. J.; Samways, M. L.; Ge, Y.; Mobley, D. L.; Essex, J. W. Enhanced Grand Canonical Sampling of Occluded Water Sites Using Nonequilibrium Candidate Monte Carlo. *J. Chem. Theory Comput.* **2023**, *19*, 1050−1062.
(46) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy Skepticism: Assessing Realistic Model Performance. *Drug Discovery Today* **2009**, *14*, 420−427.
(47) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chem., Int. Ed.* **2016**, *55*, 7364−7368.
(48) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076−6092.
(49) Qiu, Y.; Smith, D. G. A.; Boothroyd, S.; Jang, H.; Hahn, D. F.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C. D.; Rizzi, A.; Tjanaka, B.; Tresadern, G.; Lucas, X.; Shirts, M. R.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Mobley, D. L.; Wang, L.-P. Development and Benchmarking of Open Force Field v1.0.0—The Parsley Small-Molecule Force Field. *J. Chem. Theory Comput.* **2021**, *17*, 6262−6280.
(50) Boothroyd, S.; Behara, P. K.; Madin, O. C.; Hahn, D. F.; Jang, H.; Gapsys, V.; Wagner, J. R.; Horton, J. T.; Dotson, D. L.; Thompson, M. W.; Maat, J.; Gokey, T.; Wang, L.-P.; Cole, D. J.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Shirts, M. R.; Mobley, D. L. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theory Comput.* **2023**, *19*, 3251−3275.
(51) Delaine, T.; Collins, P.; MacKinnon, A.; Sharma, G.; Stegmayr, J.; Rajput, V. K.; Mandal, S.; Cumpstey, I.; Larumbe, A.; Salameh, B. A.; Kahl-Knutsson, B.; van Hattum, H.; van Scherpenzeel, M.; Pieters, R. J.; Sethi, T.; Schambye, H.; Oredsson, S.; Leffler, H.; Blanchard, H.; Nilsson, U. J. Galectin-3-Binding Glycomimetics That Strongly Reduce Bleomycin-Induced Lung Fibrosis and Modulate Intracellular Glycan Recognition. *ChemBioChem* **2016**, *17*, 1759−1770.
(52) Buijnsters, P.; De Angelis, M.; Langlois, X.; Rombouts, F. J. R.; Sanderson, W.; Tresadern, G.; Ritchie, A.; Trabanco, A. A.; VanHoof, G.; Roosbroeck, Y. V.; Andrés, J. I. Structure-Based Design of a Potent, Selective, and Brain Penetrating PDE2 Inhibitor with Demonstrated Target Engagement. *ACS Med. Chem. Lett.* **2014**, *5*, 1049−1053.
(53) Bartolomé-Nebreda, J. M.; Delgado, F.; Martín-Martín, M. L.; Martínez-Viturro, C. M.; Pastor, J.; Tong, H. M.; Iturrino, L.; Macdonald, G. J.; Sanderson, W.; Megens, A.; Langlois, X.; Somers, M.; Vanhoof, G.; Conde-Ceide, S. Discovery of a Potent, Selective, and Orally Active Phosphodiesterase 10A Inhibitor for the Potential Treatment of Schizophrenia. *J. Med. Chem.* **2014**, *57*, 4196−4212.
(54) Pérez-Benito, L.; Casajuana-Martin, N.; Jiménez-Rosés, M.; van Vlijmen, H.; Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *J. Chem. Theory Comput.* **2019**, *15*, 1884−1895.
(55) Malamas, M. S.; Erdei, J.; Gunawan, I.; Turner, J.; Hu, Y.; Wagner, E.; Fan, K.; Chopra, R.; Olland, A.; Bard, J.; Jacobsen, S.; Magolda, R. L.; Pangalos, M.; Robichaud, A. J. Design and Synthesis of 5,5′-Disubstituted Aminohydantoins as Potent and Selective Human β-Secretase (BACE1) Inhibitors. *J. Med. Chem.* **2010**, *53*, 1146−1158.
(56) Hunt, K. W.; Cook, A. W.; Watts, R. J.; Clark, C. T.; Vigers, G.; Smith, D.; Metcalf, A. T.; Gunawardana, I. W.; Burkard, M.; Cox, A. A.; Geck Do, M. K.; Dutcher, D.; Thomas, A. A.; Rana, S.; Kallan, N. C.; DeLisle, R. K.; Rizzi, J. P.; Regal, K.; Sammond, D.; Groneberg, R.; Siu, M.; Purkey, H.; Lyssikatos, J. P.; Marlow, A.; Liu, X.; Tang, T. P. Spirocyclic β-Site Amyloid Precursor Protein Cleaving Enzyme 1 (BACE1) Inhibitors: From Hit to Lowering of Cerebrospinal Fluid (CSF) Amyloid β in a Higher Species. *J. Med. Chem.* **2013**, *56*, 3379−3403.

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076

(57) Ciordia, M.; Pérez-Benito, L.; Delgado, F.; Trabanco, A. A.; Tresadern, G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. *J. Chem. Inf. Model.* **2016**, *56*, 1856−1871.
(58) Keränen, H.; Pérez-Benito, L.; Ciordia, M.; Delgado, F.; Steinbrecher, T. B.; Oehlrich, D.; van Vlijmen, H. W. T.; Trabanco, A. A.; Tresadern, G. Acylguanidine Beta Secretase 1 Inhibitors: A Combined Experimental and Free Energy Perturbation Study. *J. Chem. Theory Comput.* **2017**, *13*, 1439−1453.
(59) Hahn, D. F.; Wagner, J. R. Protein-Ligand Benchmark Dataset for Free Energy Calculations. 2022. https://zenodo.org/record/6600875 (accessed April 23, 2024).
(60) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *J. Comput. Chem.* **2015**, *36*, 348−354.
(61) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020−9041.
(62) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65*, 712−725.
(63) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.
(64) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78*, 1950−1958.
(65) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71−73.
(66) Goga, N.; Rzepiela, A. J.; De Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 3637−3649.
(67) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(68) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(69) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An *N*·log(*N*) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(70) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(71) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; Van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222*, 529−539.
(72) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(73) Crooks, G. E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60*, 2721−2726.
(74) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Phys. Rev. Lett.* **2003**, *91*, 140601.
(75) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863−1874.
(76) Wagner, J.; Mobley, D. L.; Thompson, M.; Bannan, C.; Chodera, J.; Rizzi, A.; Gokey, T.; Dotson, D.; Rodríguez-Guerra, J.; Camila; Bayly, C.; Horton, J.; Behara, P.; Lim, N. M.; Boothroyd, S.; Lim, V.; Sasmal, S.; Smith, D.; Wang, L.-P.; Zhao, Y. openforcefield/openforcefield: 0.8.3 Major Bugfix Release. 2021. https://zenodo.org/record/4429313 (accessed April 23, 2024).
(77) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(78) 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.
(79) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Antechamber: An Accessory Software Package for Molecular Mechanical Calculations. *J. Am. Chem. Soc.* **2001**, *222*, 2001.
(80) Sousa Da Silva, A. W.; Vranken, W. F. ACPYPE—AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **2012**, *5*, 367.
(81) Ibrahim, M. A. A. Molecular Mechanical Study of Halogen Bonding in Drug Discovery. *J. Comput. Chem.* **2011**, *32*, 2564−2574.
(82) Gapsys, V.; Hahn, D. F.; Tresadern, G.; Mobley, D. L.; Rampp, M.; de Groot, B. L. Pre-Exascale Computing of Protein−Ligand Binding Free Energies with Open Source Software for Drug Design. *J. Chem. Inf. Model.* **2022**, *62*, 1172−1177.
(83) 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.* **2009**, *31*, 671−690.
(84) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(85) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144−3154.
(86) 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.
(87) Soteras Gutiérrez, I.; Lin, F.-Y.; Vanommeslaeghe, K.; Lemkul, J. A.; Armacost, K. A.; Brooks, C. L.; MacKerell, A. D. Parametrization of Halogen Bonds in the CHARMM General Force Field: Improved Treatment of Ligand−Protein Interactions. *Bioorg. Med. Chem.* **2016**, *24*, 4812−4825.
(88) Orr, A. A.; Sharif, S.; Wang, J.; MacKerell, A. D., Jr. Preserving the Integrity of Empirical Force Fields. *J. Chem. Inf. Model.* **2022**, *62*, 3825−3831.
(89) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.
(90) Hahn, D. F. Protein-Ligand-Benchmark-Analysis: Release 0.3.0. 2023. https://zenodo.org/record/8283717 (accessed April 23, 2024).
(91) Macdonald, H. B.; Hahn, D. F.; Henry, M.; Chodera, J.; Dotson, D.; Glass, W.; Pulido, I. openforcefield/openff-arsenic: v0.2.1. 2022. https://zenodo.org/record/6210305 (accessed April 23, 2024).
(92) Hahn, A. M.; Then, H. Measuring the Convergence of Monte Carlo Free-Energy Calculations. *Phys. Rev. E* **2010**, *81*, 041117.
(93) Matthews, B. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.
(94) Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888−905.
(95) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J. Chem. Theory Comput.* **2021**, *17*, 4291−4300.
(96) Aldeghi, M.; Bluck, J. P.; Biggin, P. C. Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide. In *Computational Drug Discovery and Design*; Gore, M., Jagtap, U. B.,

Eds.; *Methods in Molecular Biology*; Springer New York: New York, NY, 2018; Vol. *1762*; pp 199−232..
(97) Khalak, Y.; Tresadern, G.; Aldeghi, M.; Baumann, H. M.; Mobley, D. L.; de Groot, B.; Gapsys, V. Alchemical Absolute Protein-Ligand Binding Free Energies for Drug Design. *Chem. Sci.* **2021**, *12*, 13958−13971.
(98) Horton, J. T.; Boothroyd, S.; Wagner, J.; Mitchell, J. A.; Gokey, T.; Dotson, D. L.; Behara, P. K.; Ramaswamy, V. K.; Mackey, M.; Chodera, J. D.; Anwar, J.; Mobley, D. L.; Cole, D. J. Open Force Field BespokeFit: Automating Bespoke Torsion Parametrization at Scale. *J. Chem. Inf. Model.* **2022**, *62*, 5622−5633.
(99) He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J. A Fast and High-Quality Charge Model for the Next Generation General AMBER Force Field. *J. Chem. Phys.* **2020**, *153*, 114502.
(100) Kashefolgheta, S.; Oliveira, M. P.; Rieder, S. R.; Horta, B. A.; Acree, W. E., Jr.; Hünenberger, P. H. Evaluating Classical Force Fields Against Experimental Cross-Solvation Free Energies. *J. Chem. Theory Comput.* **2020**, *16*, 7556−7580.
(101) Shirts, M. R.; Klein, C.; Swails, J. M.; Yin, J.; Gilson, M. K.; Mobley, D. L.; Case, D. A.; Zhong, E. D. Lessons Learned from Comparing Molecular Dynamics Engines on the SAMPL5 Dataset. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 147−161.
(102) Rizzi, A.; Jensen, T.; Slochower, D. R.; Aldeghi, M.; Gapsys, V.; Ntekoumes, D.; Bosisio, S.; Papadourakis, M.; Henriksen, N. M.; de Groot, B. L.; Cournia, Z.; Dickson, A.; Michel, J.; Gilson, M. K.; Shirts, M. R.; Mobley, D. L.; Chodera, J. D. The SAMPL6 SAMPLing Challenge: Assessing the Reliability and Efficiency of Binding Free Energy Calculations. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 601−633.
(103) Gapsys, V.; Hahn, D. Protein Ligand Benchmark Dataset to "Current State of Open Source Force Fields in Protein-Ligand Binding Affinity Predictions. 2024. https://zenodo.org/doi/10.5281/zenodo.10782775 (accessed April 23, 2024).
(104) Gapsys, V.; Aldeghi, M.; Michielssens, S.; Seeliger, D.; Hahn, D. F. dfhahn/pmx: pmx for Protein-Ligand Dataset. 2023. https://zenodo.org/doi/10.5281/zenodo.10057847 (accessed April 23, 2024).
(105) Gapsys, V.; Hahn, D. Workflow and Simulation Input Files to "Current State of Open Source Force Fields in Protein-Ligand Binding Affinity Predictions. 2024. https://zenodo.org/doi/10.5281/zenodo.10495732 (accessed April 23, 2024).
(106) Abraham, M.; Alekseenko, A.; Bergh, C.; Blau, C.; Briand, E.; Doijade, M.; Fleischmann, S.; Gapsys, V.; Garg, G.; Gorelov, S.; Gouaillardet, G.; Gray, A.; Irrgang, M. E.; Jalalypour, F.; Jordan, J.; Junghans, C.; Kanduri, P.; Keller, S.; Kutzner, C.; Lemkul, J. A.; Lundborg, M.; Merz, P.; Miletić, V.; Morozov, D.; Páll, S.; Schulz, R.; Shirts, M.; Shvetsov, A.; Soproni, B.; Van Der Spoel, D.; Turner, P.; Uphoff, C.; Villa, A.; Wingbermühle, S.; Zhmurov, A.; Bauer, P.; Hess, B.; Lindahl, E. *GROMACS 2023.2 Manual*; Zenodo Version Number: 2023.2 2023.

https://doi.org/10.1021/acs.jcim.4c00417
J. Chem. Inf. Model. 2024, 64, 5063−5076