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[*,†]

†*Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, Beerse B-2340, Belgium*

‡*Computational Biomolecular Dynamics Group, Max Planck Institute for Multidisciplinary Sciences, Am Fassberg 11, 37077 Göttingen, Germany*

¶*Department of Chemistry, University of California, Irvine*

§*Department of Pharmaceutical Sciences, University of California, Irvine*

E-mail: gtresade@its.jnj.com

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 is nowadays a popular approach for 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 accuracies comparable to OPLS3e. While Parsley and Sage are performing comparable 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 cannot only be attributed to the force field parameters, but is also dependent of input preparation, and sampling convergence of the calculations. Especially large perturbations and non-converged simulations lead to less accurate predictions. The input structures, Gromacs force field files as well as the analysis python notebooks are available on github.

KEYWORDS: Open Force Field, Parsley, Sage, OPLS3e, OPLS, GAFF, CGenFF, Small Molecule Force Field, Binding Affinity, Free Energy, pmx, FEP+, Benchmark, Open Source Software, Open Science, Drug Discovery, Biomolecular Simulation.

Introduction

Prioritizing the synthesis of compounds by means of computationally predicted binding affinities has become one of the central strategies in small molecule drug discovery. 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 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, thus reducing the computational cost of free energy calculations as well as a beneficial cancellation of errors arising from calculating the separate solvation and protein legs for similar ligands. 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.[1]

Due to tremendous algorithmic advances and continuous increase in computational power in the last decades, these calculations are nowadays frequently utilized. However, the calculations are still costly (≈ 10$ per relative free energy difference when using commercial software, while open source solutions can be cheaper[2]). The accuracy with respect to experimental affinities is typically in the range of $1-2\,\mathrm{kcal\,mol^{-1}}$ [3–8] with best performing cases arguably capable of approaching experimental accuracy.[9] When comparing to experiment, there are mainly four sources of error encountered in binding free energy calculations: system setup, force field parameters, sampling time and experimental uncertainty. Firstly, the set-up of the system has a significant impact on the prediction accuracy. This includes the exact chemical composition of the system, consisting of protein, ligand, solvent, potential ions and co-factors. All the molecules need to be in their correct tautomeric and charge state. Also, the initial coordinates of all atoms will strongly affect the results, as well as the simulation parameters mimicking the experimental conditions. 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.[10] 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 force field parameters used in classical mechanics simulations.

Another source of error in free energy estimates comes from finite sampling. Current computational power allows reaching microsecond simulation timescales, 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, water positions in the binding site. Therefore, the limited sampling does not always ensure proper representation of the thermodynamic ensemble underlying the modelled system. This issue may be minimized by employing enhanced sampling protocols[11] such as replica exchange,[12,13] metadynamics/local elevation,[14,15] umbrella sampling or well designed sampling (MC) moves.[16]

Finally, uncertainty in the experimental measurements for the reference data limits the achievable prediction accuracy.[17] Typically, one compares the result of calculations to the experimentally measured bioactivity data, which itself has error 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 the experiment did not measure the same observable (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.[18]

While some analyses suggest that the sampling, force field and experimental errors might contribute in a quantitatively similar manner,[19] generally, the magnitude of each source of error is unknown and will likely be case dependent. In the current work we concentrate on quantifying force field related errors by comparing six small molecule molecular mechanics force fields in a benchmark of relative protein-ligand binding free energy calculations. For each force field we obtained to 1116 $\Delta\Delta$G estimates across 22 protein targets. The large and diverse set of systems allows comparing not only distinct force field families - GAFF, CGenFF, OPLS, OpenFF - but also different versions of OpenFF: v1.0, v1.2, v2.0. With OpenFF presenting a novel direction in force field development,[20–22] here, we demonstrate the ability of this force field to deliver high accuracy binding free energy predictions.

Methods

Data set

The employed benchmark data set is listed in Supplementary Information Table S.1. A total of 22 protein targets, 598 ligands and 1116 alchemical perturbations was considered.

In order to compare to other calculations we selected benchmark sets from previously published literature. Eight datasets originate from Wang et al.[3] 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.[5] Furthermore, we included protein–ligand systems that have appeared in various other FEP studies: GALECTIN-3[23] PDE2,[24] PDE10,[25] ROS1[26] and two additional BACE datasets.[27–30] To keep our results as comparable as possible to prior calculations, we used the exact same input coordinates of the prepared systems as were previously used in the studies of Gapsys et al.,[31] Schindler et al.[5] and Perez Benito et al.[26] The input structures are provided in the protein-ligand-benchmark repository.[32]

Calculation details

pmx/GROMACS non-equilibrium switching approach

The prepared protein and ligand structures were parameterized using the corresponding force field parameters (see below). The remainder of the preparation and the simulation protocol followed the non-equilibrium TI protocol from the study of Gapsys et al.[31] and is summarized in the following. For each perturbation, hybrid coordinates and topologies were generated from the physical end state ligand coordinates and topologies using pmx.[33] 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 inter-atomic 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, disallowing mapping heavy atoms to hydrogens. The obtained mapping is used to create hybrid structures and topologies following a single topology approach.

The two branches of thermodynamic cycle were prepared for simulation: ligand in water and ligand bound to the protein. The systems were placed in a dodecahedral box with the minimal distance of 1.5 nm to the box wall. The solutes were solvated with the TIP3P[34] water and sodium and chloride ions were added to neutralize the system and reach 150 mM salt concentration.

Amber99sb*ILDN[35–37] force field was used to parameterize the proteins for the simulations with OpenFF and GAFF2.1x ligand force fields. The ion parameters for these simulations were taken from Joung & Cheatham.[34] The Charmm36m[38] protein force field was used in combination with the MATCH/CGenFF ligand parameters.

To calculate relative free energy differences, firstly, 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. From the generated trajectories, the first 2.256 ns 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, ending with the fast alchemical transitions was repeated 3 times. 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, and which happens to be the default in the commercial FEP+ software and used in many published studies.

The simulation temperature was kept at 298 K by means of the stochastic dynamics integrator with the friction of 0.5 ps^{-1}. The protein-ligand calculation data collected from[6] used molecular dynamics integrator in combination with the velocity rescaling thermostat[39] with the time constant of 0.1 ps. The pressure was controlled by means of the Parrinello-Rahman barostat[40] with the time constant of 5 ps keeping pressure at 1 bar. Electrostatic interactions were treated by means of the Particle Mesh Ewald (PME)[41,42] with the direct space cutoff of 1.1 nm, relative strength of interactions at the cutoff of 10^{-5}, 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. Non-bonded interactions during the alchemical transitions were softened. The functional form of the softcore potential described in[43] (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[44] was used with the parameters $\alpha = 0.3$ and $\sigma = 0.25$ nm. The bonds were constrained by means of LINCS algorithm.[45]

From the alchemical transitions, work values were collected and free energy differences were calculated based on the Crooks Fluctuation Theorem[46] using maximum likelihood estimator.[47]

Free energy perturbation using FEP+

The free energy calculations using Schrodinger's FEP+[3] were retrieved from published results and the calculation details can be found therein.[6,26,48] The calculations use the same input structures as available in the reference dataset as well as the same alchemical perturbations.[49] The automated Schrodinger protocol with default settings was employed for the simulation, *i.e.* 5 ns simulation time, 12-24 λ points per perturbation, Hamiltonian replica exchange and the replica exchange solute tempering protocol. The protein and ligands were parameterized using the OPLS3e force field with custom parameters.[50] The results for targets BACE, BACE (HUNT), BACE (P2), CDK2, GALECTIN, JNK1, MCL1, P38, PDE2, PTP1B, THROMBIN and TYK2 were retrieved from reference 6. Reference 48 is the source of the results of targets CDK8, CMET, EG5, HIF2A, PFKFB3, SHP2, SYK, and TNKS2. Finally, the results of targets PDE10 and ROS1 were taken from reference 26.

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 force fields version used for each system in the Supplementary Table S.1.

Open Force Field Open Force Field (OpenFF) parameters were used in 3 different versions (Parsley v1.0.0[21] and v1.2.1 and Sage v.2.0.0[51]). The OpenFF toolkit 0.8.4[20,52] was used to parameterize the ligands with AM1BCC charges.[53,54] In the following the three force fields are named OpenFF-1.0, OpenFF-1.2 and OpenFF-2.0, without the final patch number of the release.

GAFF2.1x GAFF parameters were assigned by means of Antechamber[55] and ACPYPE.[56] AM1-BCC partial charge model was used. Offset charges on chlorine and bromine were added according to the rules described in 57. We specify the force field as ”GAFF2.1x” as results across the dataset are pulled from two different studies, with some systems using GAFF2.1,[31] and a later study using GAFF2.11.[58] Table S.1 lists the exact force field used for each target.

CGenFF/MATCH* Small molecule parameterization with the CGenFF[59] force field was performed by assigning atom types with the MATCH[60] 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[61,62] 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.[63] Throughout the manuscript we refer to this parameterization as CGenFF/MATCH* to mark that a several different tools were employed in the parameterization procedure which may lead to differences in assigned parameters depending on the atom-typing, generalized force field version and even structure converter used.[64]

OPLS3e The Schrodinger force field OPLS3[65] and OPLS3e[50] was used in the FEP+ results presented, which were taken from published sources.[5,26,31] Table S.1 lists the source of the results for each target. For simplicity, we labelled all the FEP+ results in the plots and tables as ”OPLS3e”.

Consensus approach For the consensus approach ”Consensus”, the results were averaged over one repeat each from OpenFF-2.0, GAFF2.1x and CGenFF/MATCH*. This sums up to the same sampling time as the results from the single force fields.

Two alternative consensus approaches were calculated, which are only presented in the Supplementary Information. The first 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 manuscript can be followed or reproduced with the python notebooks available at `https://github.com/dfhahn/protein-ligand-benchmark-analysis`.[?]

Calculation of $\Delta\Delta G$ and ΔG values

For the relative binding free energy ($\Delta\Delta G$) values we used the raw values without any cycle closure correction as they reflect better potential shortcomings of force fields. For the pmx results, we calculated the $\Delta\Delta G$ 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 (ΔG), we calculated the MLE estimate with the package arsenic[66] for $\Delta\Delta G$ values coming both from FEP+ and pmx.

Metrics

The performance of the calculations employing different force fields are 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 ($\Delta\Delta G$) compared to the experimental values. These were calculated for the single target sets and the whole set of 1116 edges.

For the final binding free energies of ligands (ΔG), the node-based RMSE and MUE was 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 $\tau_{\mathrm{K,overall}}$ considering the whole dataset, we calculated the weighted average of the Kendall's τ_{K} of all individual targets.

$$\tau_{\mathrm{K,overall}} = \frac{1}{N} \sum_{\mathrm{targets}} N_{\mathrm{target}} \tau_{\mathrm{K,target}}, \tag{1}$$

where N is the sum of all considered ligands across targets, N_{target} is the number of ligands of a target and $\tau_{\mathrm{K,target}}$ 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 Supplementary Information.

Error Calculation

If not stated otherwise, all results are given with 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 experiment, we calculated the significance by bootstrapping using a confidence interval of 95%.

Convergence criteria for perturbations

To discriminate the error of force field 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 non-equilibrium sampling. α is $-1 \leq \alpha \leq 1$ and is described in more detail in reference 67, Equation 5. The second criterion is the standard deviation of the $\Delta\Delta G$ values $\sigma(\Delta\Delta G)$ over the three repeats. For a perturbation to be considered converged, both requirements $\alpha < 0.8$ and $\sigma(\Delta\Delta G) < 1.5\mathrm{kcal\,mol^{-1}}$ 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 force field parameters involved in the perturbations were identified, *i.e.* only parameters which are 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[68] as it is suited to correlate binary vectors (parameter either used or not used in edges). The obtained correlation matrix between parameters was then clustered with spectral clustering[69] 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

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

where $\mathrm{RMSE}_{\mathrm{FF}}(p)$ is the $RMSE$ between predicted $\Delta\Delta G$ with force field 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 latest OpenFF forcefield OpenFF-2.0 (Sage) to exemplify accuracy achievable with the open source force field. 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 deviation from the segment center (white background). Based on the $\Delta\Delta G$ values of the edges, an RMSE of $1.7^{1.9}_{1.6}\,\mathrm{kcal\,mol^{-1}}$ ($\mathrm{MUE} = 1.2^{1.3}_{1.1}\,\mathrm{kcal\,mol^{-1}}$) was obtained. This is in line with current industry standards.[48]

Overall, the open source force fields performed comparably 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 (Figures 1a and 1b). The obtained RMSE values from experiment are for GAFF2.1x is $1.7^{2.0}_{1.5}$, 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 force fields significantly outperformed each of the open source force fields considered separately (RMSE of $1.5^{1.6}_{1.4}$). The OPLS3e force field shows a significantly lower RMSE of $1.3^{1.4}_{1.3}$ kcal mol^{-1} when averaged over all ΔΔG used in this work.

Table 1 and Figure 1d list the per-target accuracy reached by each force field 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.

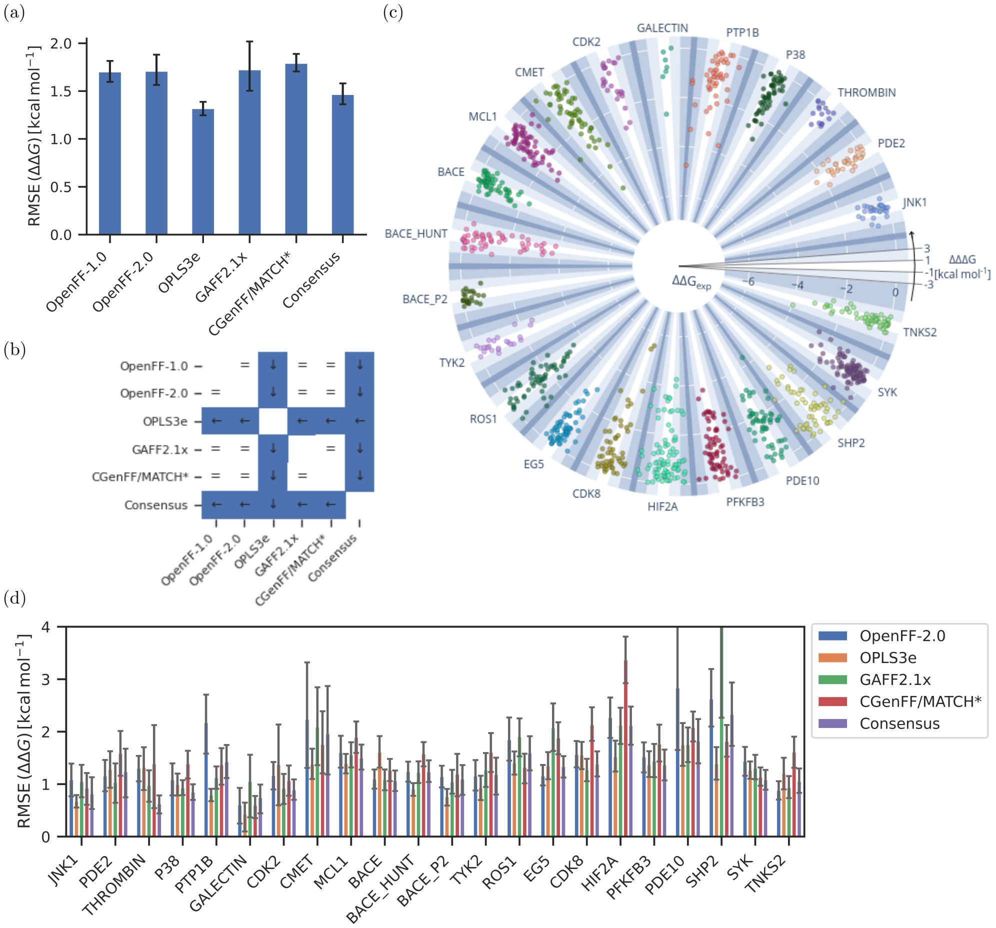

Figure 1: *Comparison of $\Delta\Delta G$ values of the perturbations obtained from calculations using the five force fields OpenFF-1.0, OpenFF-2.0, GAFF2.1x, CGenFF/MATCH* and OPLS3e and the Consensus approach.* (a) Overall RMSE comparison across all targets and all 1116 perturbations. (b) Illustration of significant differences between pairs of force fields. A white matrix element with an equal sign "=" means that the differences between the two force fields are statistically insignificant. A colored matrix element means there is a significant difference considering a 95% confidence interval. The arrow in the 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. The radial distance denotes the experimental $\Delta\Delta G_{\mathrm{exp}}$. The deviation of the calculation from experiment is shown on the angular axis as deviation from the segment center (white background). The 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. The lower and upper bound of the 95% confidence interval are given as error bars. The corresponding graph with mean unsigned error (MUE) instead of RMSE can be found in the Supplementary Information, Figure S.2.

Although the aggregated RMSE statistics overall (Figure 1a) or per-target (Figure 1d) do not show a significant difference between the public force fields, the differences become

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 root mean squared error (RMSE) of the $\Delta\Delta G$ values of the perturbations.* 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 value for the respective force field. The upper and lower bound of the 95% confidence interval are given as sub- and superscript. All values are in kcal mol^{-1}. The corresponding table with mean unsigned error (MUE) instead of RMSE can be found in the Supplementary Information, Figure S.2.

	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	$1.1^{1.5}_{0.8}$	$1.1^{1.5}_{0.9}$	$1.6^{2.0}_{1.2}$	$1.3^{1.6}_{0.9}$	$1.0^{1.2}_{0.7}$	$1.1^{1.5}_{0.8}$

more apparent by looking at number of outliers. Figure 2 shows the ratio of perturbations with absolute errors versus experiment 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 1.5x(interquartile range) from the box edges (with outliers shown as markers). We observed differences between the force fields 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 are similar. For the public force fields, 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 experiment (left blue box in Figure 2), only 23% of the edges agreed with experiment within the threshold for the least performing target (SHP2). On the other hand, 78% of edges in the best performing target (TNKS2) were correct considering the threshold. This difference between least and best performing target can be reduced with the Consensus approach, which seems to 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 wrong starting poses of the ligand, inadequate protein preparation or unlikely protonation or tautomeric states, both in ligand and in protein. If all force fields 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 moves 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. Some targets might feature certain chemical moieties which are not adequately described by the respective force field. The use of inadequate parameters could explain why the use of OPLS3e leads to less outliers, as the use of custom parameters describes specific chemistries better than a general force field. Finally, the experimental results might be unsuitable. The MD calculations may not mimic the exact experimental conditions (temperature, ion concentrations, co-solvents) or the assay may only have limited correlation with the binding free energy that is targeted in the RBFE calculations. But this has no impact when comparing the different force fields, as they are all compared to the same data.

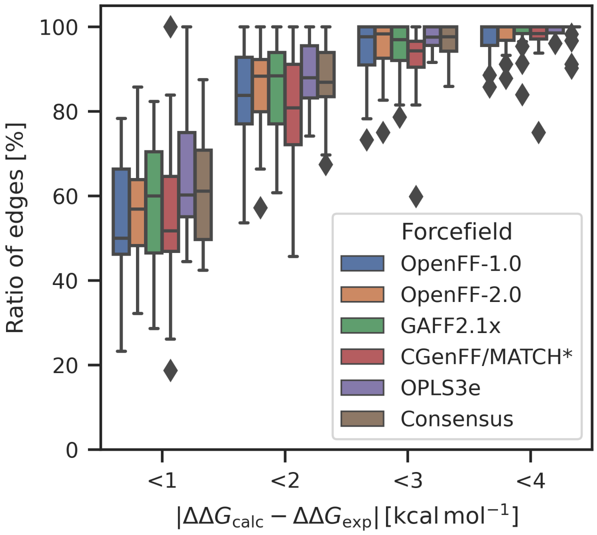

Figure 2: *The ratio of calculated ΔΔG within various different absolute error thresholds compared to the experimental value for the different force fields force fields.* The box-and-whiskers show the distribution across the various targets. Each box illustrates the first and third quartiles with the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

Accuracy of predicted ΔG Figure 3 shows the trend in significant differences between force fields changes when comparing accuracy in terms of back-calculated absolute binding free energies ΔG (Figure 3). 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. The Consensus approach outperforms the individual open source force fields, similarly as it was for the ΔΔG comparison.

We also compared force field predictions in terms of their ability to correctly rank binders based on their ΔG values by using the 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 force fields is rather complex (Figure 3d), the differences are small in magnitude, showing that each of the force fields can be trusted to yield a compound ranking of similar quality. The Supplementary Information, Figures S.5-S.8 illustrate aggregated statistics based on ΔG per target and across all targets for all the force fields, including the consensus approaches. The corresponding values can be found in the Supplementary Information, Tables S.6-S.8. Additionally, correlation plots are provided for OpenFF-2.0, CGenFF/MATCH*, GAFF2.1x, OPLS3e and the Consensus approach in Supplementary 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 force field, e.g. sampling, specifics of the calculation procedure, initial system setup. In the analysis in Figure 4 we attempted to elucidate the main determinants underlying ΔΔG prediction accuracy related to the convergence of an alchemical perturbation.

In particular, we noticed that larger calculated ΔΔG values are associated with a larger error (second row, third column in Figure 4). 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 (second row, fourth column). While there are many factors influencing convergence of an alchemical perturbation, we observed that a simple count of heavy atoms that need to be introduced/annihilated correlates slightly with the absolute error (third row, last column) and well with the convergence measure (fourth row, last column). Similar trends can be seen

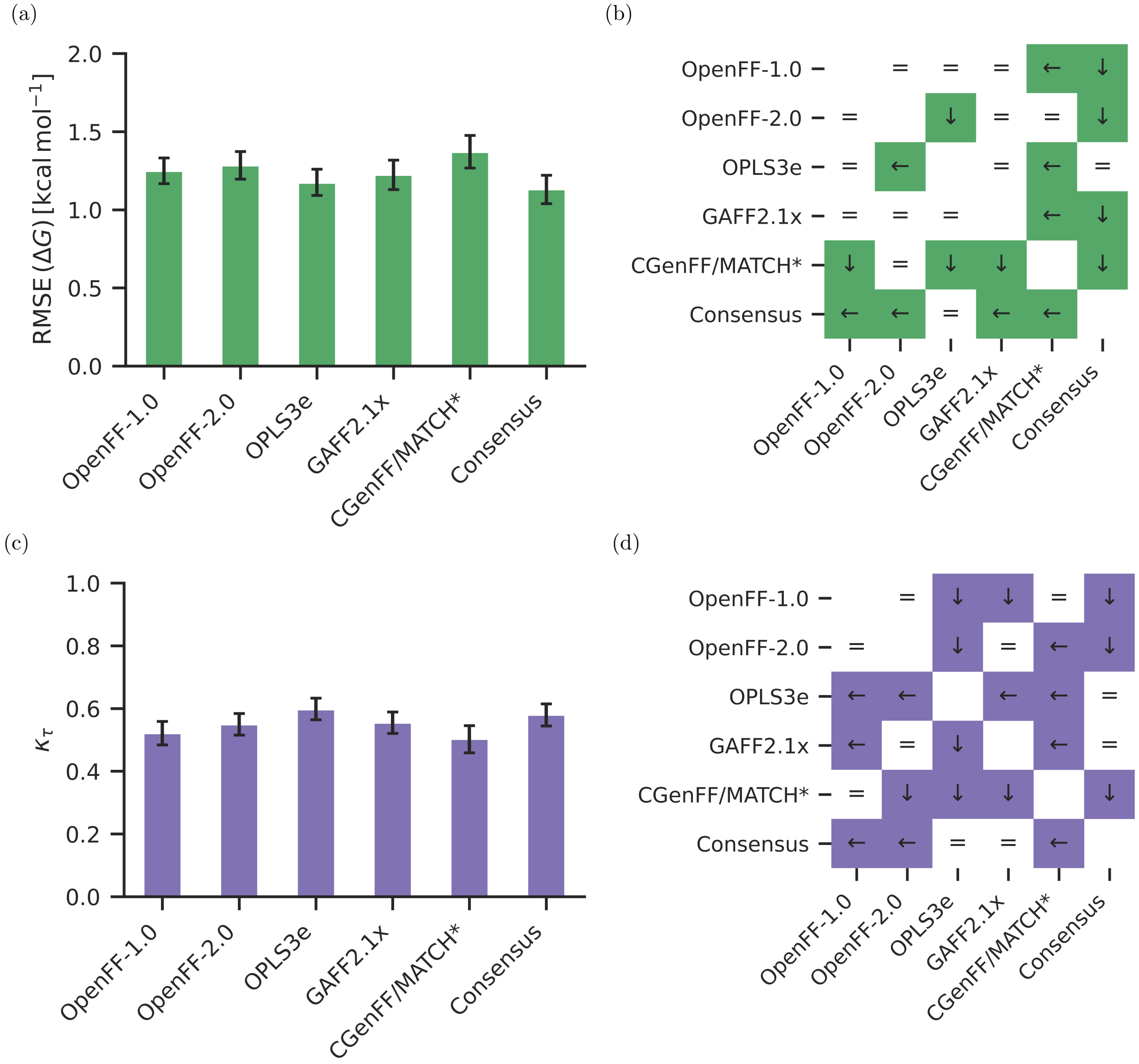

Figure 3: *Comparison of ΔG values of the ligands obtained from calculations using the five force fields OpenFF-1.0, OpenFF-2.0, GAFF2.1x, CGenFF/MATCH* and OPLS3e and the Consensus approach.* (a) RMSE comparison across all targets and 598 ligands. (b) Illustrations of significance of differences between the different set. (c) Comparison of ranking metric τ_{K} across all targets and 598 ligands. (d) Illustrations of significance of differences between the different sets. The 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. A colored matrix element means there is a significant difference considering a 95% confidence interval. The arrow in the blue matrix element points at the force field which has the lower error (either left or down).

in the Supplementary Information for the counts of rotatable bonds (Figure S.15), counts of rings (Figure S.16), changes or positions of the formal charges (Figure S.17) and the LOMAP score[70] (Figure S.18).

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 prediction. The identified determinants for prediction accuracy are only general trends in a complex picture.

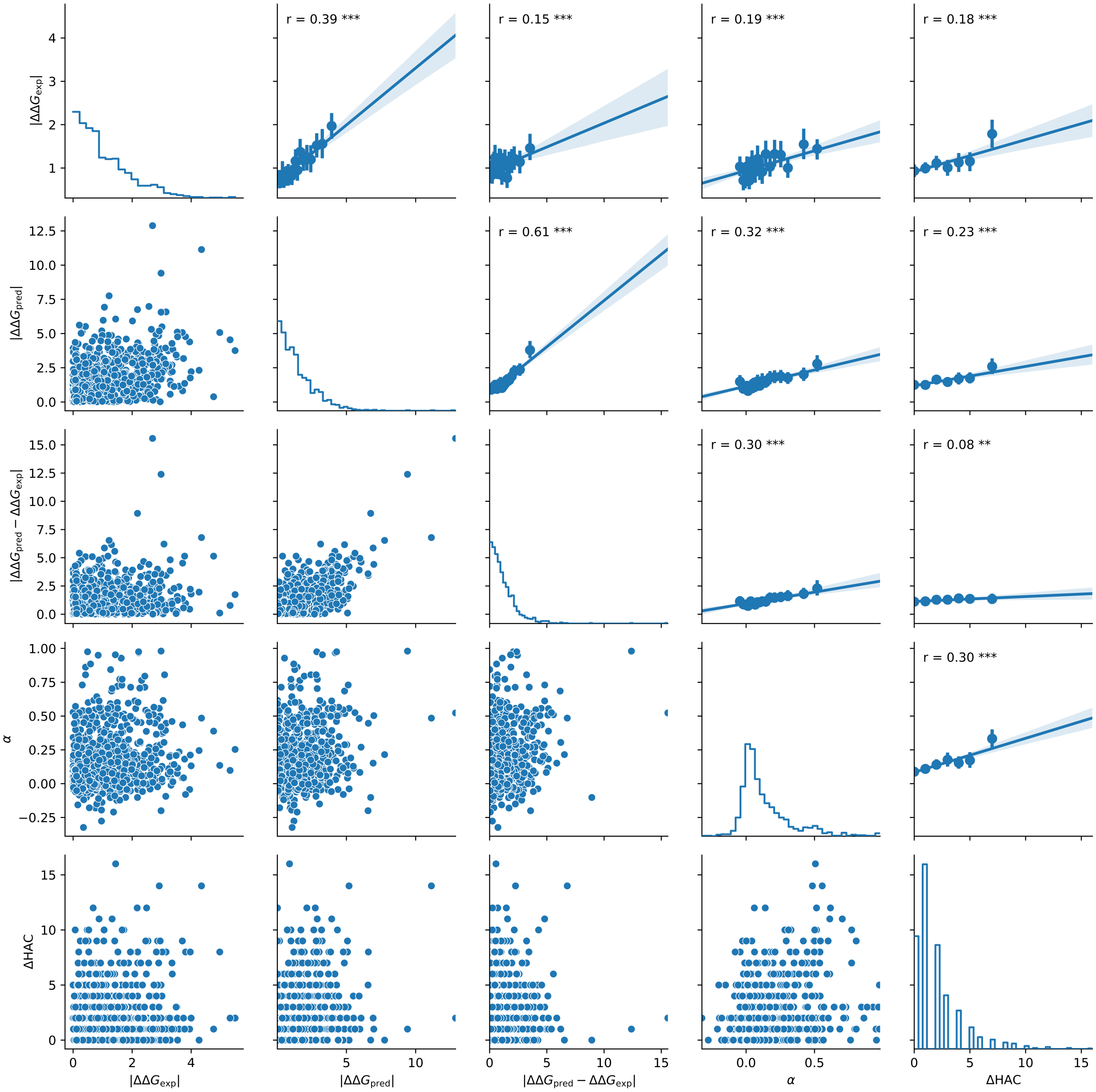

Figure 4: *Visualization of pairwise relationships between, the experimental relative free energies $\Delta\Delta G_{\mathrm{exp.}}$, the calculated relative free energies $\Delta\Delta G_{\mathrm{pred}}$ (OpenFF-2.0), the absolute error between experimental and calculated values $|\Delta\Delta G_{\mathrm{pred}} - \Delta\Delta G_{\mathrm{exp.}}|$, the average convergence measure α*[67] *(averaged over three solvent and three complex simulation legs), the change in number of heavy atoms in the end states.* The lower left triangle shows the original 1116 datapoints. The upper right triangle plots show linear regression plots. For illustration purposes, the data was binned into 20 bins and their average with standard deviation are shown as dot with error bars. The regression was performed on the original data. The diagonal shows histograms of the respective properties.

OpenFF force field improvement

Non-converged 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 Section for details about the convergence criteria). Whereas all edges consisting of converged and non-converged 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 2.54kcal mol^{-1}. The convergence criteria can therefore be used to flag calculation which are likely to have larger errors without prior knowledge of experimental results.

Figure 5b compares three OpenFF versions by means of RMSE values of $\Delta\Delta G$ values compared to experimental 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 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 force field 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. For example, an ester group has an angle (OpenFF code a15), bond (b20), improper (i2), and torsion (t107, t110) changes. Altogether, the RMSE between the predicted and experimental $\Delta\Delta$G for the perturbations of ester groups drops by 0.5 kcal mol^{-1}. 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 force field parameters: the predicted free energy difference is more accurate for the modified parameters. The largest

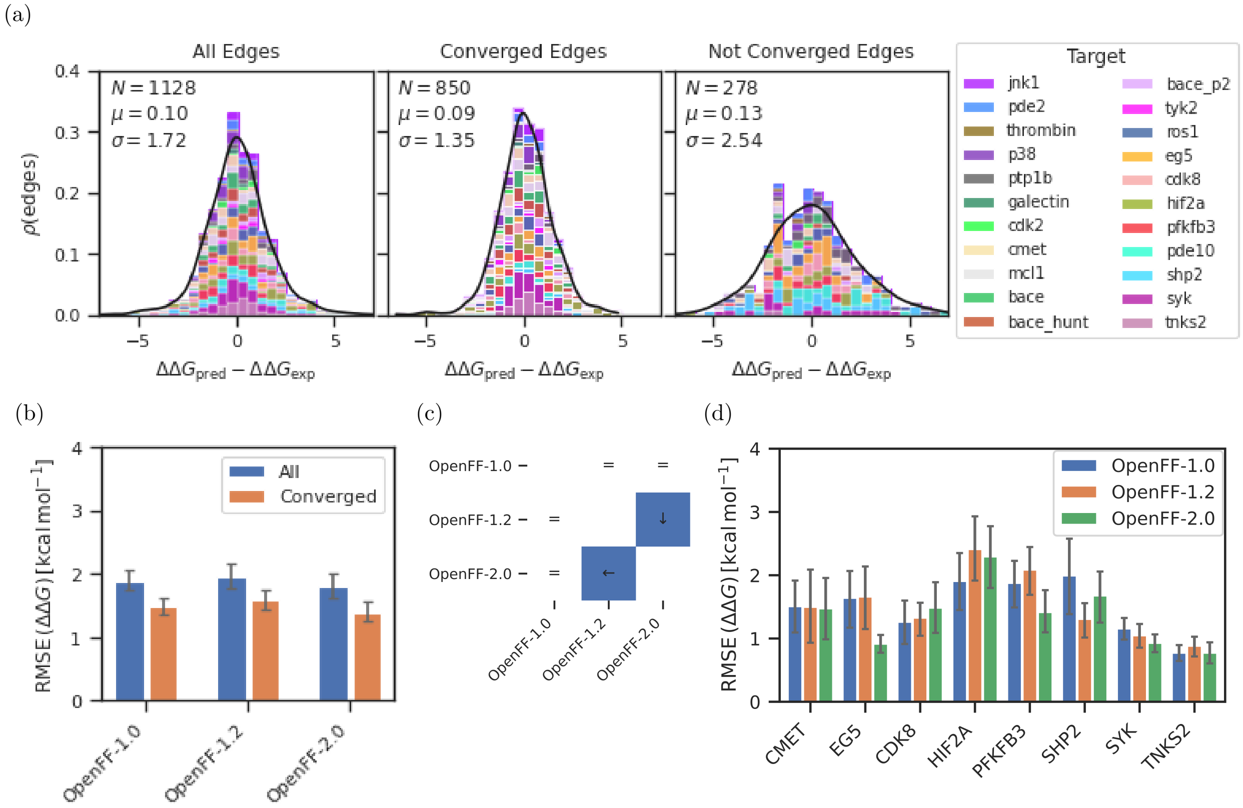

Figure 5: *Comparison of the three force fields OpenFF-1.0, OpenFF-1.2, and OpenFF-2.0 based on the $\Delta\Delta G$ values.* Panel (a) shows the absolute error distributions between experimental $\Delta\Delta G$ and calculated $\Delta\Delta G$ using OpenFF-2.0 for three sets of edges. The first set 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 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. The 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 root mean square error (RMSE) across all edges of 8 targets for the three force fields of the OpenFF family. The blue bars concern all edges, the orange bars only the converged ones. Panel (c) illustrates significant differences between the force field sets shown in Panel b. A white matrix element with an equal sign "=" means that the differences between the two force fields are statistically insignificant. The colored matrix element means there is a significant difference considering a 95% confidence interval. The arrow in the 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. The lower and upper bound of the 95% confidence interval are given as error bars. All values are in kcal mol^{-1}.

improvement in this analysis was observed for the changes in the hydroxyl group bound to a sp2 carbon involving the bond (b18) and torsion (t106) parameters. Figure 6c illustrates

Table 2: *Comparison of the three force fields OpenFF-1.0, OpenFF-1.2, and OpenFF-2.0 based on the root mean squared error (RMSE) of the $\Delta\Delta G$ values of the converged perturbations.* 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 value for the respective force field. The upper and lower and upper bound of the 95% confidence interval are given as sub- and superscript. All values are in kcal mol^{-1}. The values are illustrated in Figures 5b and 5d.

	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	$1.2^{1.3}_{1.0}$	$1.0^{1.2}_{0.9}$	$0.9^{1.1}_{0.8}$
TNKS2	51	$0.8^{0.9}_{0.6}$	$0.9^{1.0}_{0.7}$	$0.8^{1.0}_{0.6}$

a case where this improvement resulted in 1.3 kcal mol^{-1} increase in free energy calculation accuracy.

There are only several parameter sets that result in a decreased $\Delta\Delta G$ prediction accuracy for OpenFF-2.0 compared to OpenFF-1.0. Namely, changes in parameters describing sulphur-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.

Conclusions

On a set of 598 ligands each binding to one of 22 targets, we showed that the public force fields 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 RMSE of relative binding free energies $\Delta\Delta G$ (perturbations) and the RMSE and Kendall's

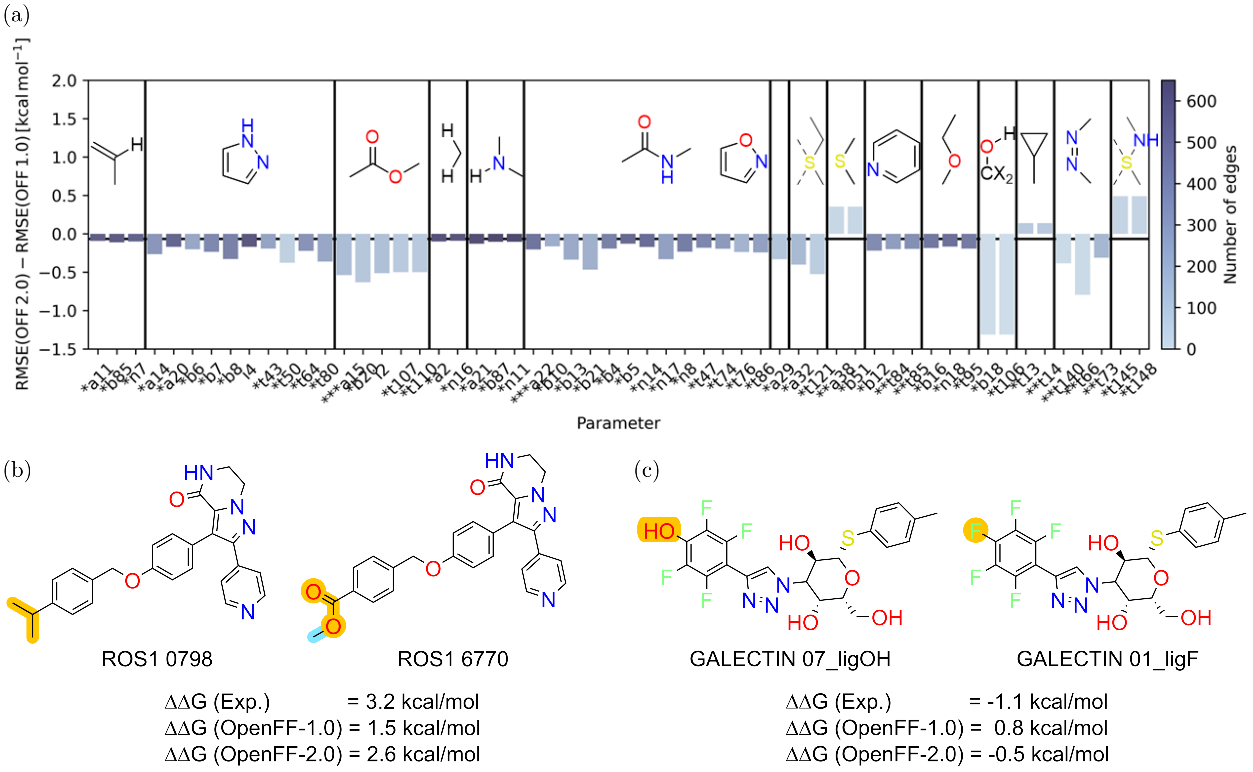

Figure 6: *Analysis of parameter differences between OpenFF-1.0 and OpenFF-2.0.* Panel (a) shows the RMSE difference between OpenFF-1.0 and OpenFF-2.0 for subsets of converged edges, where a certain parameter is perturbed (x-axis) and the difference between OpenFF-1.0 and OpenFF-2.0 is significant (CI 95%). The stars (*) in front of the parameter label denotes how much the parameter changed between OpenFF-1.0 and OpenFF-2.0 (3 stars denote the largest change). The horizontal black line denotes the insignificant difference ($-0.06\,\mathrm{kcal\,mol^{-1}}$) for whole set of perturbations. The vertical bars separate groups of bars with high correlations, i.e. they are usually employed concurrently in perturbations. The chemical structure shows an example substructure where each group of parameters is employed. Panel b shows an example perturbation where an ester function (1st group in Panel a) is introduced. The free energy prediction improved from $\Delta\Delta G = 1.5\,\mathrm{kcal\,mol^{-1}}$ (OpenFF-1.0) to $\Delta\Delta G = 2.6\,\mathrm{kcal\,mol^{-1}}$ (OpenFF-2.0) at an experimental value of $\Delta\Delta G = 3.2\,\mathrm{kcal\,mol^{-1}}$. Panel c shows an example perturbation where an aromatic hydroxy function (4th group in Panel a) is annihilated. The free energy prediction improved from $\Delta\Delta G = 0.8\,\mathrm{kcal\,mol^{-1}}$ (OpenFF-1.0) to $\Delta\Delta G = -0.5\,\mathrm{kcal\,mol^{-1}}$ (OpenFF-2.0) at an experimental value of $\Delta\Delta G = -1.1\,\mathrm{kcal\,mol^{-1}}$. In panels b and c, the perturbed atoms and bonds are highlighted in orange, whereas annihilated atoms and bonds are highlighted in cyan.

tau of binding free energies ΔG. The proprietary force field OPLS3e performs significantly better, but a Consensus approach based on Sage, GAFF2.1x, and CGenFF/MATCH* is similarly accurate. There is a clear target dependence which can be attributed to input preparation, protein (binding pocket) flexibility, chemistries of ligands, and difficulty of perturbations (in terms of heavy atom changes). While Parsley and Sage are performing comparable 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 on subsets of the dataset involving those parameters. Thus, we can show that there is a considerable improvement of successive OpenFF forcefield 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 force field versions. However, for this to be successful, further work would be valuable to reduce the influence of other (non force field parameter) sources of errors like large or difficult perturbations, inadequate input preparation or insufficient sampling.

Acknowledgement

The authors thank Lorenzo d'Amore (ORCID: 0000-0001-9145-6457), Marjolein Crabbe, Benjamin Ries (ORCID: 0000-0002-0945-8304), and Chris Bayly (ORCID: 0000-0003-2245-1956) and John Chodera (ORCID: 0000-0003-0542-119X) for their insightful discussions. We thank the Open Force Field Consortium and Initiative for scientific support, and the Molecular Sciences Software Institute (MolSSI) and the Open Molecular Software Foundation (OMSF) for their support of the Open Force Field Initiative.

Supporting Information Available

The Supplementary Information lists details about the employed target set, shows additional graphs and tables containing aggregated statistics and correlations with experiment in greater detail, and shows various properties of the simulated perturbations. Additional information can be found in the associated github repository, `https://github.com/dfhahn/protein-`

`ligand-benchmark-analysis`.,? which includes analysis code, information about the data set and additional figures.

References

(1) 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]. *Living Journal of Computational Molecular Science* **2020**, *2*.

(2) 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. *Journal of Chemical Information and Modeling* **2022**, *62*, 1691–1711.

(3) Wang, L. et al. 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. *Journal of the American Chemical Society* **2015**, *137*, 2695–2703.

(4) Song, L. F.; Lee, T.-S.; Zhu, C.; York, D. M.; Merz, K. M. Using AMBER18 for Relative Free Energy Calculations. *Journal of Chemical Information and Modeling* **2019**, *59*, 3128–3135.

(5) Schindler, C. E. M. et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *Journal of Chemical Information and Modeling* **2020**, *60*, 5457–5474.

(6) 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. *Chemical Science* **2020**, *11*, 1140–1152.

(7) 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. *Journal of Chemical Information and Modeling* **2020**, *60*, 3120–3130.

(8) Ross, G.; Lu, C.; Scarabelli, G.; Albanese, S.; Houang, E.; Abel, R.; Harder, E.; Wang, L. *The maximal and current accuracy of rigorous protein-ligand binding free energy calculations*; preprint, 2022.

(9) 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. *Journal of Chemical Information and Modeling* **2022**, *62*, 703–717.

(10) 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. 2021.

(11) 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. *Journal of chemical theory and computation* **2020**, *16*, 1630–1645.

(12) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *The Journal of Physical Chemistry B* **2011**, *115*, 9431–9438.

(13) 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. *Journal of Chemical Theory and Computation* **2013**, *9*, 1282–1293.

(14) 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. *Journal of Chemical Theory and Computation* **2013**, *9*, 1334–1346.

(15) 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. *Journal of Chemical Theory and Computation* **2023**, *19*, 1805–1817.

(16) 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. *The Journal of Physical Chemistry B* **2018**, *122*, 5579–5598.

(17) 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. *Journal of Chemical Information and Modeling* **2022**,

(18) Brown, S. P.; Muchmore, S. W.; Hajduk, P. J. Healthy skepticism: assessing realistic model performance. *Drug discovery today* **2009**, *14*, 420–427.

(19) 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. *Angewandte Chemie International Edition* **2016**, *55*, 7364–7368.

(20) 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., et al. Escaping atom types in force fields using direct chemical perception. *Journal of chemical theory and computation* **2018**, *14*, 6076–6092.

(21) Qiu, Y.; Smith, D. G.; Boothroyd, S.; Jang, H.; Hahn, D. F.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C. D., et al. Development and Benchmarking of Open

Force Field v1. 0.0—the Parsley Small-Molecule Force Field. *Journal of Chemical Theory and Computation* **2021**, *17*, 6262–6280.

(22) Boothroyd, S. et al. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *Journal of Chemical Theory and Computation* **2023**, *19*, 3251–3275.

(23) Delaine, T. et al. Galectin-3-Binding Glycomimetics That Strongly Reduce Bleomycin-Induced Lung Fibrosis and Modulate Intracellular Glycan Recognition. *ChemBioChem* **2016**, *17*, 1759–1770.

(24) 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 Medicinal Chemistry Letters* **2014**, *5*, 1049–1053.

(25) 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. *Journal of Medicinal Chemistry* **2014**, *57*, 4196–4212.

(26) Pérez-Benito, L.; Casajuana-Martin, N.; Jiménez-Rosés, M.; van Vlijmen, H.; Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *Journal of Chemical Theory and Computation* **2019**, *15*, 1884–1895.

(27) 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. *Journal of Medicinal Chemistry* **2010**, *53*, 1146–1158.

(28) Hunt, K. W. et al. 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. *Journal of Medicinal Chemistry* **2013**, *56*, 3379–3403.

(29) Ciordia, M.; Pérez-Benito, L.; Delgado, F.; Trabanco, A. A.; Tresadern, G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. *Journal of Chemical Information and Modeling* **2016**, *56*, 1856–1871.

(30) 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. *Journal of Chemical Theory and Computation* **2017**, *13*, 1439–1453.

(31) 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. *Chemical Science* **2020**, *11*, 1140–1152.

(32) Hahn, David F., protein-ligand-benchmark-analysis: Release 0.3.0. 2023; `https://zenodo.org/record/8283717`.

(33) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *Journal of Computational Chemistry* **2015**, *36*, 348–354.

(34) 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.

(35) 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.

(36) 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.

(37) 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.

(38) 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. *Nature methods* **2017**, *14*, 71–73.

(39) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(40) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(41) 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.

(42) 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.

(43) Gapsys, V.; Seeliger, D.; de Groot, B. L. New soft-core potential function for molecular dynamics based alchemical free energy calculations. *Journal of Chemical Theory and Computation* **2012**, *8*, 2373–2382.

(44) 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. *Chemical physics letters* **1994**, *222*, 529–539.

(45) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

(46) Crooks, G. E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60*, 2721–2726.

(47) 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.

(48) Schindler, C. et al. *Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects*; Preprint, 2020.

(49) Hahn, D. F.; Wagner, J. openforcefield/protein-ligand-benchmark: 0.2.0 Addition of new targets. 2021; `https://zenodo.org/record/5679599`.

(50) 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. *Journal of Chemical Theory and Computation* **2019**, *15*, 1863–1874.

(51) Boothroyd, S. et al. *Development and Benchmarking of Open Force Field 2.0.0 — the Sage Small Molecule Force Field*; preprint, 2022.

(52) Wagner, J. et al. openforcefield/openforcefield: 0.8.3 Major bugfix release. 2021; `https://zenodo.org/record/4429313`.

(53) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of computational chemistry* **2000**, *21*, 132–146.

(54) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **2002**, *23*, 1623–1641.

(55) 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*.

(56) Sousa Da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **2012**, *5*, 367.

(57) Ibrahim, M. A. A. Molecular mechanical study of halogen bonding in drug discovery. *Journal of Computational Chemistry* **2011**, *32*, 2564–2574.

(58) 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. *Journal of chemical information and modeling* **2022**, *62*, 1172–1177.

(59) 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. *Journal of Computational Chemistry* **2009**, NA–NA.

(60) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247–260.

(61) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *Journal of Chemical Information and Modeling* **2012**, *52*, 3144–3154.

(62) 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. *Journal of Chemical Information and Modeling* **2012**, *52*, 3155–3168.

(63) 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. *Bioorganic & Medicinal Chemistry* **2016**, *24*, 4812–4825.

(64) Orr, A. A.; Sharif, S.; Wang, J.; MacKerell Jr, A. D. Preserving the Integrity of Empirical Force Fields. *Journal of Chemical Information and Modeling* **2022**, *62*, 3825–3831.

(65) Harder, E. et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation* **2016**, *12*, 281–296.

(66) Bruce Macdonald, H. E. Openforcefield/openff-arsenic. Open Force Field Initiative, 2020.

(67) Hahn, A. M.; Then, H. Measuring the convergence of Monte Carlo free-energy calculations. *Physical Review E* **2010**, *81*, 041117.

(68) Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **1975**, *405*, 442–451.

(69) Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2000**, *22*, 888–905.

(70) 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. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 755–770.