ACS central science

Research Article

Cite This: ACS Cent. Sci. 2018, 4, 1708−1718

http://pubs.acs.org/journal/acscii

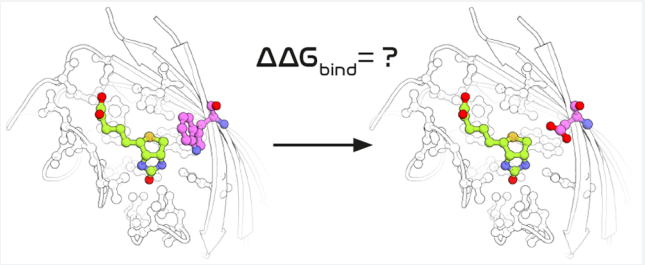
# Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation

Matteo Aldeghi, Vytautas Gapsys, and Bert L. de Groot*

Computational Biomolecular Dynamics Group, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

Supporting Information

**ABSTRACT:** The design of proteins with novel ligand-binding functions holds great potential for application in biomedicine and biotechnology. However, our ability to engineer ligand-binding proteins is still limited, and current approaches rely primarily on experimentation. Computation could reduce the cost of the development process and would allow rigorous testing of our understanding of the principles governing molecular recognition. While computational methods have proven successful in the early stages of the discovery process, optimization approaches that can quantitatively predict ligand affinity changes upon protein mutation are still lacking. Here, we assess the ability of free energy calculations based on first-principles statistical mechanics, as well as the latest Rosetta protocols, to quantitatively predict such affinity changes on a challenging set of 134 mutations. After evaluating different protocols with computational efficiency in mind, we investigate the performance of different force fields. We show that both the free energy calculations and Rosetta are able to quantitatively predict changes in ligand binding affinity upon protein mutations, yet the best predictions are the result of combining the estimates of both methods. These closely match the experimentally determined $\Delta\Delta G$ values, with a root-mean-square error of 1.2 kcal/mol for the full benchmark set and of 0.8 kcal/mol for a subset of protein systems providing the most reproducible results. The currently achievable accuracy offers the prospect of being able to employ computation for the optimization of ligand-binding proteins as well as the prediction of drug resistance.




## INTRODUCTION

Ligand-binding proteins play essential roles in living organisms, with interactions between small organic molecules and proteins triggering a multitude of signal transduction processes.[1−3] Given their high affinity and selectivity, nontoxicity, and biodegradability, the design of proteins with novel ligand-binding functions also holds great potential for application in biomedicine and biotechnology.[4−10] Fast computational approaches that rely on mixed physics- and knowledge-based potentials, such as Rosetta,[11,12] have already proven successful in the early stages of the discovery process. For instance, Tinberg et al.[13] engineered protein binders to the steroid digoxigenin by first designing a minimal ligand-binding shell, then searching the protein data bank[14] (PDB) for suitable scaffolds, and finally optimizing the designs experimentally. Among the 17 computationally designed proteins, two were binding to digoxigenin in the micromolar range. After experimental optimization of the most promising design, a protein with sub-nanomolar affinity for the steroid was found. A similar approach has also been employed for the design of artificial enzymes by using a model of the transition state as the target ligand.[15−18] Other design approaches that have been proposed involve ligand docking to known protein structures[19] or *de novo* design of short protein sequences using a combination of docking, molecular dynamics (MD), and Monte Carlo simulations.[20]

However, our ability to engineer ligand-binding proteins (e.g., biosensors and enzymes) is still limited, and current approaches rely heavily on experimentation, in particular at the optimization stage.[21−24] The limitations of Rosetta have often been ascribed to a limited treatment of backbone flexibility and lack of explicit solvation, so that computational approaches that tackle these challenges may provide a more accurate estimation of ligand affinity changes upon protein mutation.[4,22,25] Free energy calculations based on first-principles statistical mechanics that make use of nonphysical (i.e., *alchemical*) pathways in a thermodynamic cycle have become increasingly popular in small molecule drug discovery for the optimization of lead compounds[26,27] and have now started being used prospectively by the pharmaceutical industry.[28] Recently, alchemical free energy calculations have also shown promise for the prediction of protein thermostability and drug resistance.[24,29−33] These calculations naturally take into account the full flexibility of the protein−ligand complex and the discrete nature of the solvent, and return the exact binding affinity changes given the energy function (i.e., force field) used. These calculations could be incorporated into protein design pipelines; however, their quantitative performance for

Received: October 5, 2018
Published: December 13, 2018






DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

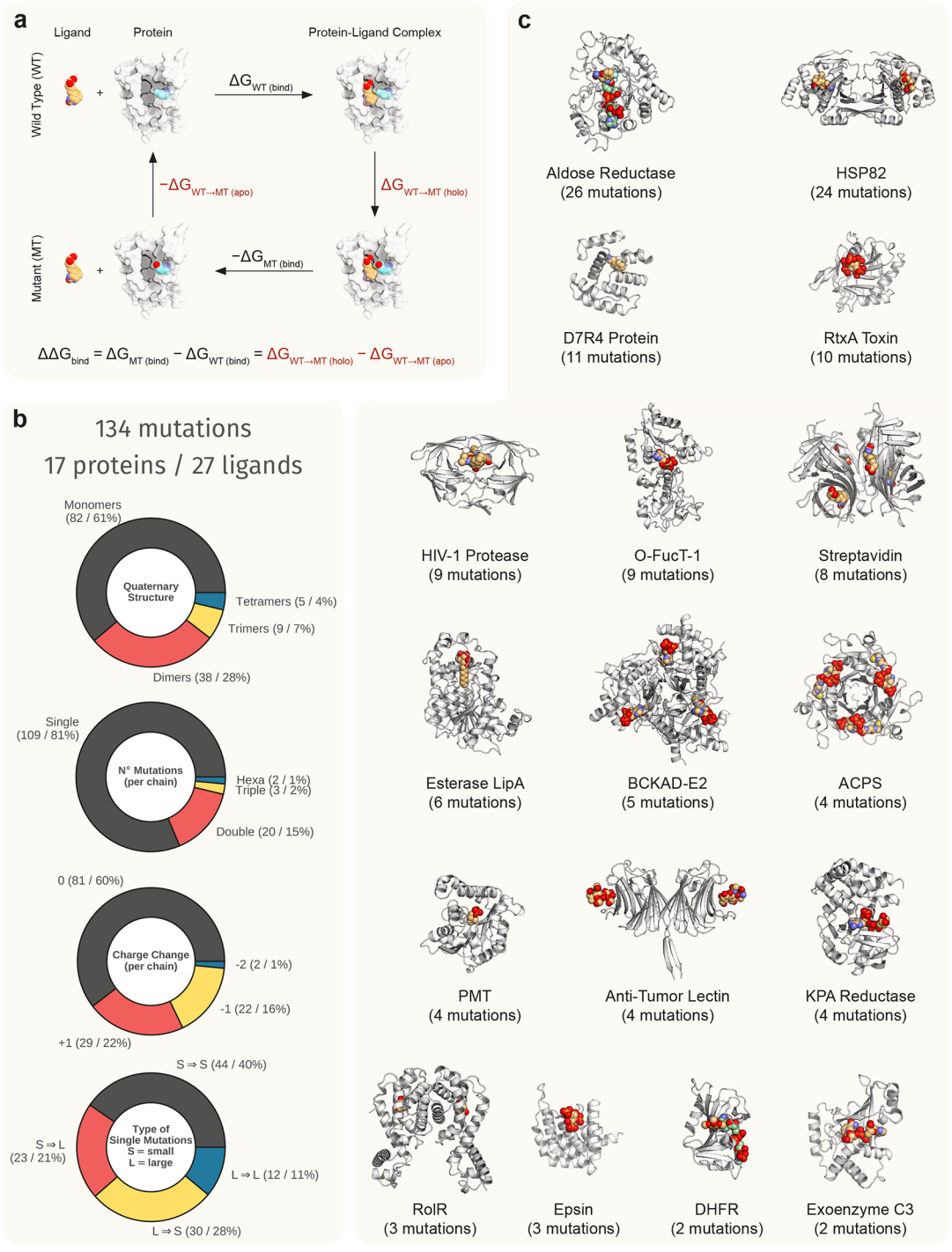



**Figure 1.** Overview of the benchmark data set studied. (a) Thermodynamic cycle showing the quantity to be predicted ($\Delta\Delta G_{bind}$); the free energy differences estimated via alchemical free energy calculations are highlighted in red. (b) Statistics of the data set about the protein−ligand systems and type of mutations considered. (c) Cartoon representation of the 17 protein systems present in the data set, with the number of affinity changes upon mutation reported. Ligands and cofactors are represented by spheres.

the prediction of ligand-binding affinity changes upon protein mutation is largely unclear at this stage.

Here, we assess the ability of alchemical free energy calculations to quantitatively predict ligand binding affinity changes upon protein mutation on a challenging set of 134 mutations across 17 proteins and 27 ligands. We adopt an approach that calculates the nonequilibrium work associated with the alchemical transformation of protein side-chains (Figure 1a), using *pmx*[34] to build the hybrid structures and topologies. The computational efficiency of different setup protocols is first evaluated, and then the performance of multiple force fields is tested. We find that investing approximately equal amounts of simulation time in the equilibrium and nonequilibrium parts of the calculations results in the most efficient protocols, and that given a fixed amount of computational resources it is beneficial to average results from multiple force fields in a consensus approach. When compared to the experimental data, the free energy



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

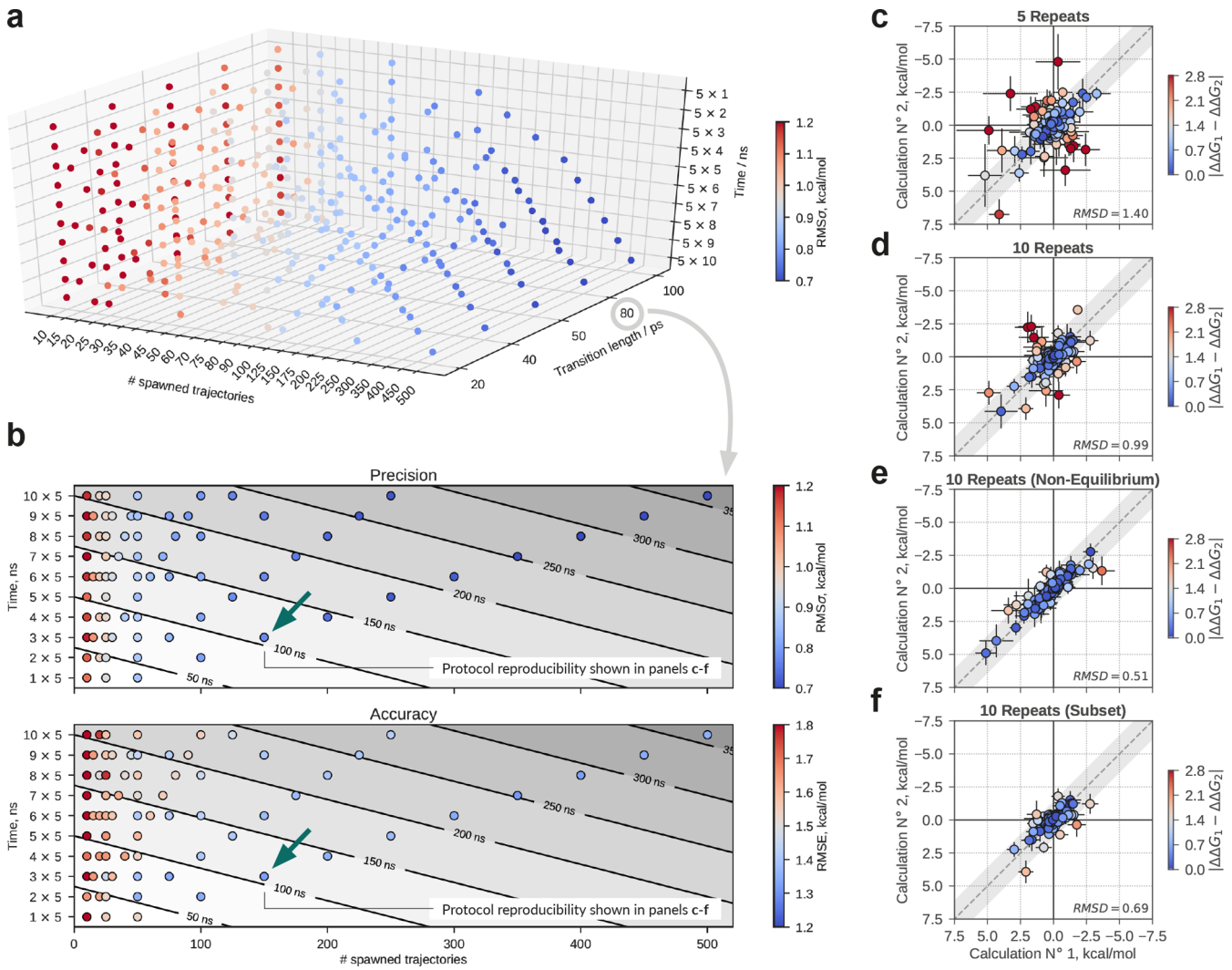



**Figure 2.** Calibration of the nonequilibrium free energy protocol. (a) Space of protocol setup parameters tested. The three axes indicate the length of the equilibrium simulations (five repeats of 1−10 ns), the number of nonequilibrium trajectories spawned from the equilibrium simulations (from 10 to 500), and their length (from 20 to 100 ps). Each mark represents a specific combination of the above three variables, with the color indicating the overall precision of the calculations (RMS$\sigma$). Equivalent plot color-coded by accuracy (RMSE) in Figure S3. (b) Scatter plots showing the overall precision and accuracy of different setup protocols that used nonequilibrium trajectories of 80 ps. Isolines indicate the computational cost (in simulation time) for one $\Delta\Delta G$ estimate. A green arrow indicates the protocol that was chosen for further calculations. (c) Reproducibility of the calculations. The scatter plot shows the agreement between two sets of $\Delta\Delta G$ estimates. For the second estimate, both the equilibrium and nonequilibrium parts of the calculations were repeated. On the bottom-right corner of the plot, the RMSD between the repeated calculations is shown. (d) Reproducibility of the calculations when increasing the number of independent equilibrium simulations to 10. Also in this case, both the equilibrium and nonequilibrium parts of the calculations were repeated. (e) Reproducibility of the nonequilibrium part of the calculations. In this case, two sets of nonequilibrium transitions were started from the same equilibrium simulations. (f) Reproducibility of the calculations (both equilibrium and nonequilibrium) for a subset of the data with four challenging protein systems excluded.

calculations achieve root-mean-square errors (RMSEs) as low as ∼1.3 kcal/mol, and ∼1.0 kcal/mol when excluding a few systems posing convergence challenges. Finally, we compare the MD results to those obtained with multiple Rosetta protocols, representing state-of-the-art approaches for protein design.[4,13,15−18,21] We find that a Rosetta protocol (*flex_ddg*)[35] recently proposed for studying protein−protein binding performs well also for protein−ligand binding. Simple averaging of Rosetta and MD results produced an improved performance for most combinations of force fields and Rosetta scoring functions, with the best combinations returning RMSEs of ∼1.2 kcal/mol for the full data set and ∼0.8 kcal/mol for a subset of well-converged results.

## RESULTS

To rigorously test the performance of the calculations, we first built a benchmark set consisting of 134 mutations across 17 proteins and 27 ligands (Figure 1b−c; ligands shown in Figure S1) from the Platinum[36] database (see Methods in the SI). Each mutation has an associated experimental binding free energy difference ($\Delta\Delta G$) determined by isothermal titration calorimetry (ITC; 119 data points) or surface plasmon resonance (SPR; 15 data points), which we aim to compute starting from the X-ray structure of the wild type protein. The $\Delta\Delta G$ values have a range of 9.5 kcal/mol and are normally distributed (Figure S2) with a mean of 0.2 kcal/mol and standard deviation of 1.5 kcal/mol. Overall, this is a diverse and challenging benchmark set that contains large and flexible


DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

ligands, proteins with different folds, some also containing cofactors, and many charge-changing and small-to-large/large-to-small mutations. The results should thus provide a realistic average performance of the calculations across different protein−ligand systems. First, we evaluate how the setup of the calculations and their overall computational cost affect the precision and accuracy of the predictions so to be able to choose an efficient protocol for further calculations. Second, we test and compare multiple force fields. And finally, we compare the performance of the free energy calculations to that of different Rosetta protocols.

**Calibration of the Free Energy Protocol.** In nonequilibrium free energy calculations, there are three main setup variables affecting the total amount of simulation time and, thus, computational cost:

(a) the length of the equilibrium simulations of the end-states (the bound and unbound forms of the wild type and mutant protein; Figure 1a);

(b) the total number of nonequilibrium trajectories that are spawned from the equilibrium simulations; and

(c) the length of these nonequilibrium trajectories, which is proportional to the speed of the alchemical transformation.

In this first part of the study, we wanted to test how these three variables affect the precision and accuracy of the free energy protocol (the Amber99sb*-ILDN/GAFF2 force field was used at this stage). Thus, we tested protocols with five repeated equilibrium simulations between 1 and 10 ns in length (always after 1 ns of equilibration), a total number of nonequilibrium trajectories between 10 and 500, and five different transition lengths of 20, 40, 50, 80, and 100 ps (Figure 2a). The work values associated with each nonequilibrium transition were extracted using thermodynamic integration (TI)[37] and then used to estimate the free energy differences with the Bennett's acceptance ratio (BAR)[38,39] relying on the Crooks Fluctuation Theorem.[40−42] The computationally cheapest protocol required 20.8 ns of simulation time, and the most expensive 400 ns. Accuracy was quantified as the RMSE between calculated and experimental free energies for the whole set of 134 mutations. Precision was quantified as the root-mean-square (RMS) of the standard errors ($\sigma$) of the predicted $\Delta\Delta G$s, with $\sigma$ being estimated from the five independent simulation repeats. In principle, the precision also determines how reproducible the results are. However, this relies on accurate estimates of the uncertainty for each predicted $\Delta\Delta G$ value. Thus, as a stricter test of the reproducibility of our chosen protocol, we defined and quantified "reproducibility" as the root-mean-square deviation (RMSD) between the $\Delta\Delta G$ values obtained from two independent sets of calculations, where each set comprised five equilibrium runs and 150 nonequilibrium trajectories in both directions considering the whole set of 134 mutations.

The more expensive protocols tended to return more precise (Figure 2a) and accurate (Figure S3) results. As there seemed to be no more benefits in terms of accuracy when increasing transition lengths from 80 to 100 ps (Figure S4), we focused on the 80 ps data. Figure 2b shows the behavior of the precision and accuracy of the calculations when using protocols with different equilibrium simulation lengths and number of nonequilibrium trajectories. We observed that, given a fixed amount of simulation time, exchanging equilibrium for nonequilibrium sampling (i.e., moving across the isolines of computational cost) resulted in both improved precision and accuracy. The most effective protocols involved investing about the same amount of simulation time in the equilibrium and nonequilibrium part of the calculations. In addition, the precision and accuracy of the predictions improved quickly from 10 to 100 nonequilibrium trajectories, after which further improvements came at a higher cost. Surprisingly, on average, we did not see a strong association between the length of the equilibrium trajectories and the accuracy or precision of the calculations.

On the basis of this analysis, one of the cheaper protocols investing about half of the simulation time in nonequilibrium sampling was chosen for further testing, and specifically the protocol using 150 nonequilibrium trajectories of 80 ps in length, spawned from five equilibrium simulations of 3 ns (equivalent to 108 ns of total simulation time per $\Delta\Delta G$ calculation; Figure 2b). As a strict test of reproducibility, we repeated all calculations (including building the simulation systems) with this protocol and measured the RMSD between the $\Delta\Delta G$ values obtained from two independent sets of calculations (Figure 2c−f). Random ion placement during system setup was found to negatively affect reproducibility (Figure S5); therefore, we first updated our protocol to resolve this issue: ions were not allowed to be placed in the vicinity of the protein, and each equilibrium simulation was started from a different ion configuration. With this precaution, the RMSD between two repeated calculations was 1.40 kcal/mol (Figure 2c), which was above the target RMSD of 1 kcal/mol we wanted to reach.

To improve the precision and reproducibility of the calculations, we resorted to a fourth setup variable: the number of equilibrium simulation repeats (initially set to five). This variable was not originally screened systematically because, assuming that each equilibrium simulation is independent, the precision of each $\Delta\Delta G$ estimate (and the RMSD between two sets of repeated calculations) should drop with the square root of the number of simulation repeats, which is also faster than when adjusting the other three setup variables studied (Figure 2b). By using 10 repeated equilibrium simulations and 300 nonequilibrium trajectories (i.e., keeping the same ratio of equilibrium/nonequilibrium sampling, and thus also doubling the cost per $\Delta\Delta G$ calculation to 216 ns), a reproducibility RMSD of ∼1 kcal/mol was achieved (Figure 2d), as expected with independent simulation repeats. To further investigate the source of uncertainty, we repeated only the nonequilibrium part of the calculations from the same set of equilibrium simulations. A comparison of the results obtained with these two runs (Figure 2e) revealed that about half (∼0.5 kcal/mol) of the reproducibility RMSD could be attributed to uncertainty in the nonequilibrium sampling (Text S1). Therefore, also based on this observation, it appears that on average it is reasonable to invest approximately equal computation effort in the equilibrium and nonequilibrium parts of the calculations.

While we have been discussing protocol performance in average terms (across all protein−ligand systems), different protein systems present different sampling challenges. Indeed, different degrees of reproducibility were observed across systems, with four protein systems being found to be particularly challenging (Text S2 and Table S1). Excluding these systems, the RMSD between two repeated sets of calculations decreased further to ∼0.7 kcal/mol (Figure 2f). From here on, we will show the results for the full set of 17



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

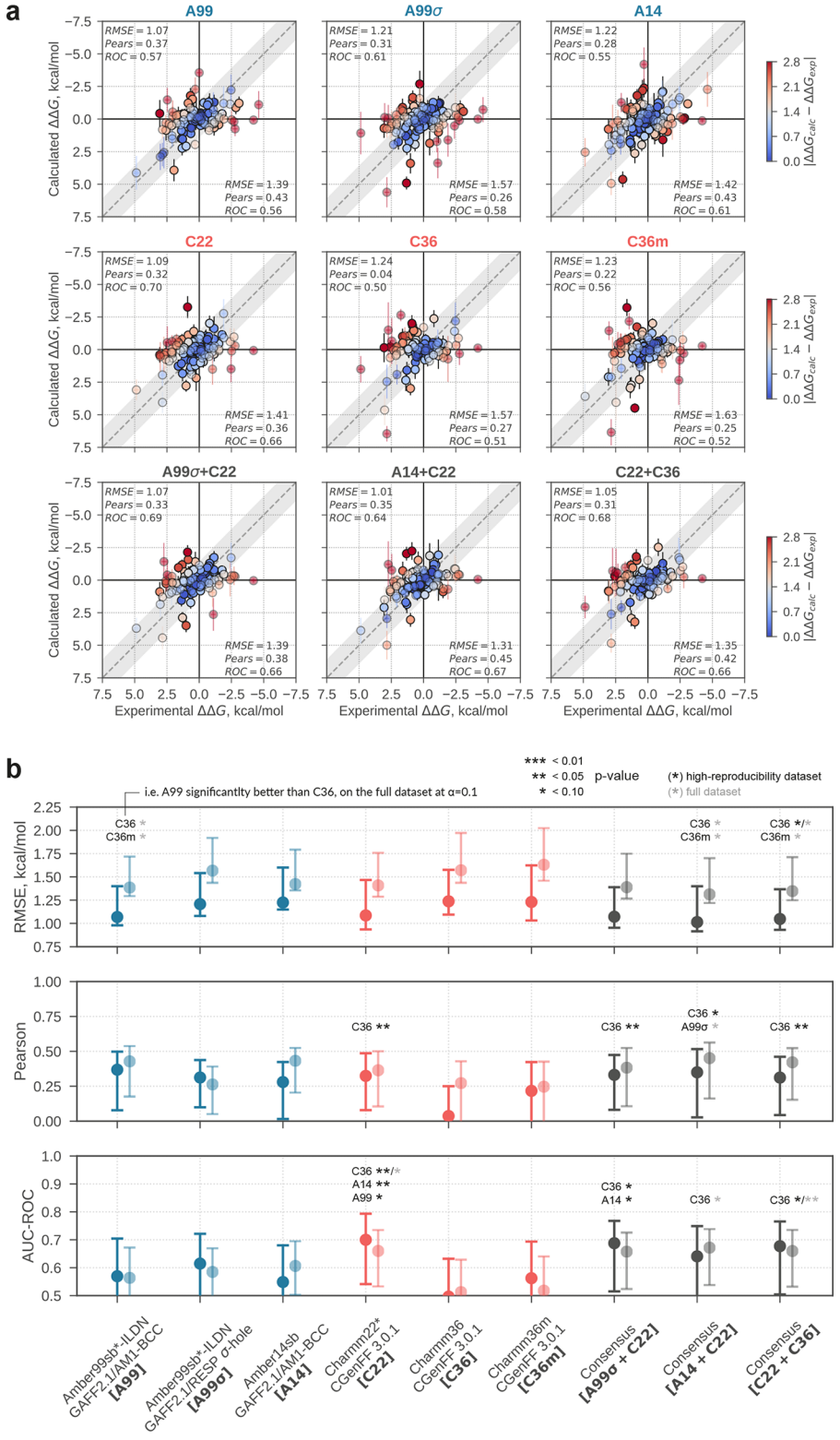



**Figure 3.** Performance of the free energy calculations with different force fields and force field combinations. (a) Scatter plots of experimental versus calculated $\Delta\Delta G$ values. The identity line is shown as a dashed gray line, while the shaded area indicates the region where $\Delta\Delta G$ estimates are within 1.4 kcal/mol of experiment (i.e., within a 10-fold error in $K_d$ change at 300 K). The performance for the high-reproducibility subset of the data is reported at the top-left of each plot, while the performance for the whole data set is shown at the bottom-right. Color-coding is used to indicate the error of each individual $\Delta\Delta G$ estimate. (b) Summary of the performance of the calculations across force fields in terms of RMSE, Pearson correlation, and AUC-ROC (point estimates and the 95% CIs are shown). Differences at three levels of significance are reported using labels within the chart: e.g., a "C36 *" label above the RMSE mark of A99 indicates that the RMSE of A99 is significantly lower (i.e., agreement with experiment is better) than that of C36 at $\alpha$ = 0.10. Marks in solid colors refer to the high-reproducibility subset, while marks in semitransparent colors refer to the full data set.



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

Table 1. Summary Statistics of the ΔΔG Estimates[a]

| abbr. | protein FF | ligand FF | #ΔΔG | experimental ΔΔG range (kcal/mol) | RMSE (kcal/mol) | Pearson | AUC-ROC |
|---|---|---|---|---|---|---|---|
| A99 | Amber99sb*-ILDN | GAFF2.1/AM1-BCC | 86 | 6.1 | $1.07^{1.40}_{0.98}$ | $0.37^{0.50}_{0.08}$ | $0.57^{0.70}_{0.45}$ |
| | | | 134 | 9.5 | $1.39^{1.72}_{1.29}$ | $0.43^{0.54}_{0.17}$ | $0.56^{0.67}_{0.48}$ |
| A99σ | Amber99sb*-ILDN | GAFF2.1/RESP | 86 | 6.1 | $1.21^{1.54}_{1.08}$ | $0.31^{0.44}_{0.10}$ | $0.61^{0.72}_{0.48}$ |
| | | | 134 | 9.5 | $1.57^{1.92}_{1.43}$ | $0.26^{0.39}_{0.05}$ | $0.58^{0.67}_{0.47}$ |
| A14 | Amber14sb | GAFF2.1/AM1-BCC | 86 | 6.1 | $1.22^{1.60}_{1.15}$ | $0.28^{0.42}_{0.01}$ | $0.55^{0.68}_{0.43}$ |
| | | | 134 | 9.5 | $1.42^{1.79}_{1.35}$ | $0.43^{0.52}_{0.20}$ | $0.61^{0.69}_{0.50}$ |
| C22 | Charmm22* | CGenFF 3.0.1 | 75 | 4.8 | $1.09^{1.47}_{0.93}$ | $0.32^{0.49}_{0.08}$ | $0.70^{0.79}_{0.54}$ |
| | | | 117 | 9.1 | $1.41^{1.76}_{1.29}$ | $0.36^{0.50}_{0.11}$ | $0.66^{0.73}_{0.53}$ |
| C36 | Charmm36 | CGenFF 3.0.1 | 75 | 4.8 | $1.24^{1.58}_{1.09}$ | $0.04^{0.25}_{-0.22}$ | $0.50^{0.63}_{0.36}$ |
| | | | 117 | 9.1 | $1.57^{1.97}_{1.44}$ | $0.27^{0.43}_{-0.01}$ | $0.51^{0.63}_{0.42}$ |
| C36m | Charmm36m | CGenFF 3.0.1 | 75 | 4.8 | $1.23^{1.62}_{1.03}$ | $0.22^{0.42}_{-0.08}$ | $0.56^{0.69}_{0.43}$ |
| | | | 117 | 9.1 | $1.63^{2.03}_{1.46}$ | $0.25^{0.43}_{-0.03}$ | $0.52^{0.64}_{0.43}$ |
| A99σ + C22 | Consensus [A99σ + C22] | | 75 | 4.8 | $1.07^{1.39}_{0.95}$ | $0.33^{0.47}_{0.08}$ | $0.69^{0.77}_{0.51}$ |
| | | | 117 | 9.1 | $1.39^{1.75}_{1.26}$ | $0.38^{0.52}_{0.11}$ | $0.66^{0.73}_{0.52}$ |
| A14 + C22 | Consensus [A14 + C22] | | 75 | 4.8 | $1.01^{1.40}_{0.92}$ | $0.35^{0.52}_{0.03}$ | $0.64^{0.75}_{0.49}$ |
| | | | 117 | 9.1 | $1.31^{1.70}_{1.22}$ | $0.45^{0.56}_{0.16}$ | $0.67^{0.74}_{0.54}$ |
| C22 + C36 | Consensus [C22 + C36] | | 75 | 4.8 | $1.05^{1.37}_{0.93}$ | $0.31^{0.46}_{0.04}$ | $0.68^{0.77}_{0.50}$ |
| | | | 117 | 9.1 | $1.35^{1.71}_{1.25}$ | $0.42^{0.52}_{0.16}$ | $0.66^{0.73}_{0.53}$ |
| ROS | Rosetta (*flex_ddg/nov16*) | | 86 | 6.1 | $0.99^{1.16}_{0.86}$ | $0.39^{0.54}_{0.17}$ | $0.61^{0.71}_{0.47}$ |
| | | | 134 | 9.5 | $1.46^{1.73}_{1.23}$ | $0.25^{0.43}_{0.04}$ | $0.56^{0.65}_{0.45}$ |
| RMD | Rosetta + MD [ROS and A14 + C22] | | 75 | 4.8 | $0.82^{1.03}_{0.74}$ | $0.44^{0.60}_{0.11}$ | $0.68^{0.76}_{0.51}$ |
| | | | 117 | 9.1 | $1.23^{1.46}_{1.10}$ | $0.49^{0.59}_{0.24}$ | $0.67^{0.73}_{0.54}$ |

[a]For each set of calculations, the statistics based on both the high-reproducibility and full datasets are shown on the first and second line, respectively. Results obtained with Charmm force fields are based on slightly smaller datasets (#ΔΔG) as glycine mutations were excluded for these force fields (see Methods in the SI). "abbr.": "FF": force field; "RMSE": root mean square error; "AUC-ROC": area under the receiver operating characteristic curve.

protein systems (134 mutations) as well as for the high-reproducibility subset that excludes the four protein systems mentioned above (86 mutations). Also note that from here on we will discuss the accuracy for only one of the two sets of calculations performed with the Amber99sb*-ILDN/GAFF2 force field, as the two sets showed comparable accuracy with respect to experiment (RMSE of $1.07^{1.40}_{0.98}$ kcal/mol versus $1.08^{1.37}_{1.01}$ kcal/mol for the high-reproducibility subset, and of $1.39^{1.72}_{1.29}$ kcal/mol versus $1.50^{1.83}_{1.38}$ kcal/mol for the full data set).

**Accuracy of the Calculations and Force Field Comparison.** After calibrating the free energy protocol and assessing the reproducibility of the calculations, we evaluated the accuracy of multiple contemporary force fields from the Amber and Charmm families (Figure 3 and Table 1). The agreement between calculations and experiments was quantified using three performance measures: the RMSE, the Pearson correlation coefficient, and the area under the receiver operating characteristics curve (AUC-ROC). These measures capture different relationships between the calculated and experimental data: the RMSE measures the deviation between calculated and experimental values such that 68% of calculated ΔΔGs are within one RMSE of the experimental ones; the Pearson correlation coefficient quantifies the linearity of the relationship between calculated and experimental ΔΔGs; the AUC-ROC gauges instead the performance of the approach as a binary classifier. We show the 95% confidence intervals (CIs) of these measures in the relevant figures and tables. The CI was obtained via bootstrap, by random resampling with replacement the data set while at the same time stochastically resampling each ΔΔG estimate assuming a Gaussian distribution (see Methods in the SI). In such a way, the CI incorporates the imprecision of each estimate as well as the uncertainty due to the specific choice/availability of data set. Significant differences were estimated in a similar fashion, such that small *p*-values are indicative of performance differences that are unlikely to have been the result of the specific choice of data set or the imprecision of the estimates.

Figure 3a compares experimental and calculated ΔΔG values for the six force fields tested (abbreviations are explained in Table 1 and Figure 3b): three from the Amber (A99, A99σ, A14) and three from the Charmm (C22, C36, C36m) family. Figure 3b shows the performance of these force fields on the high-reproducibility subset of the data (solid color) and on the full data set (semitransparent color), according to the three performance measures described above. The performance is also reported in numerical format in Table 1. Overall, aside from some exceptions, the different force fields performed similarly, with the lowest RMSE reaching ∼1.1 kcal/mol, but moderate Pearson correlation (up to ∼0.4) and AUC-ROC (up to ∼0.70). RMSEs were between ∼1.4−1.6 kcal/mol for the whole data set, and around ∼1.1−1.2 kcal/mol for the high-reproducibility subset. The Pearson correlations achieved were between 0.22 and 0.43, with the exception of C36, which had a correlation of 0.04 for the reduced data set. C36 and C36m were also the only two force fields that did not achieve a correlation significantly (at $\alpha = 0.05$) different from zero. In terms of AUC-ROC, the C22 force field stood out as the only one achieving statistical difference from randomness (AUC-ROC of 0.50), with values of 0.66 and 0.70 for the full and the high-reproducibility data set, respectively. On average, the Amber force fields achieved AUC-ROC values slightly below 0.6, while the other Charmm force fields (C36 and C36m) just above 0.5. Note that in these results, the simulations of the apo states were initiated from crystal structures of the protein−ligand complexes. We investigated the effect of starting the simulations from crystal structures of the apo state, where



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

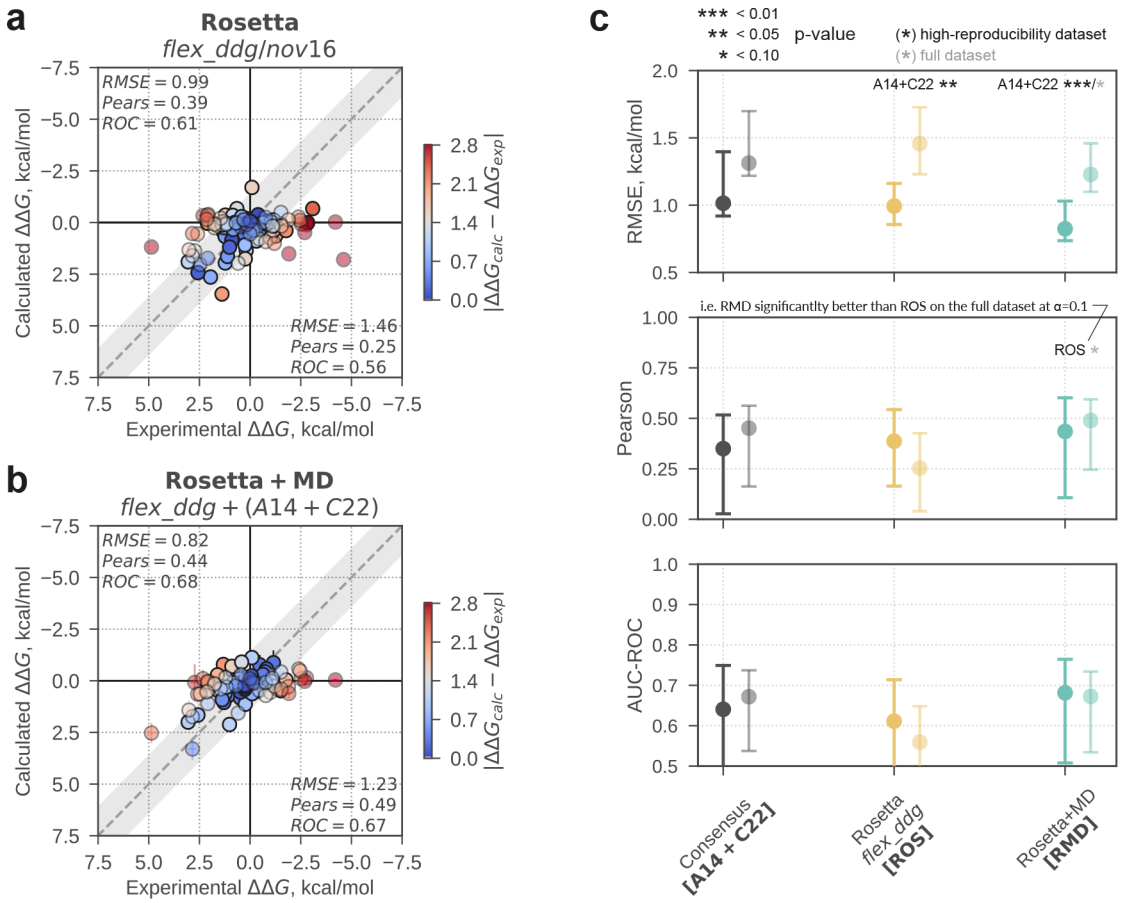



**Figure 4.** Performance of Rosetta protocols. (a) Experimental versus calculated affinity changes for the *flex_ddg/nov16* protocol. The identity line is shown as a dashed gray line, while the shaded area indicates the region where $\Delta\Delta G$ estimates are within 1.4 kcal/mol of experiment (i.e., within a 10-fold error in $K_d$ change at 300 K). The performance for the high-reproducibility subset of the data is reported at the top-left of the plot, while the performance for the whole data set is shown at the bottom-right. Color-coding is used to indicate the error of each individual $\Delta\Delta G$ estimate. (b) Experimental versus calculated affinity changes for the consensus results combining the *flex_ddg/nov16* results with the free energy calculations results A14 + C22. (c) Summary of the performance of the Rosetta calculations in terms of RMSE, Pearson correlation, and AUC-ROC (point estimate and 95% CIs are shown). Differences at three levels of significance are reported using labels within the chart: e.g., a "A14 + C22 **" label above the RMSE mark of ROS indicates that the RMSE of ROS is significantly lower than that of A14 + C22 at $\alpha = 0.05$. Marks in solid colors refer to the high-reproducibility subset, while marks in semitransparent colors refer to the full data set.

available. However, we did not observe a significant overall improvement (Figure S6).

As it has been previously observed that combining results from different force fields in a consensus approach can result in improved performance,[29,43] we explored this simply by averaging the results for pairs of force fields. To consider equivalent computational costs, consensus results were built by averaging $\Delta\Delta G$ estimates obtained with half of the simulations run for each force field (i.e., the first five equilibrium repeats). Then, we compared these consensus results with those of the two parent force fields obtained with the full amount of simulation data (i.e., all 10 equilibrium repeats). Effectively, this exercise answers the following practical question: given the chosen amount of computer time per calculation (in our case, 216 ns), is it more likely to achieve better accuracy by investing all sampling in a single force field or by investing half of the sampling in two separate force fields? Considering all possible combinations of the six force fields considered here, and the three performance measures chosen, we find that it is statistically sensible to use a consensus approach that employs two force fields (Figure S7). In particular, we observed that in ∼44% of the cases the accuracy of the consensus results (in terms of RMSE, Pearson, and AUC-ROC) was better than that of the two parent force fields; in ∼49% of the cases, the performance was in between that of the two parent force fields, and only in ∼7% of cases it was worse than both. When excluding the systems posing convergence challenges, the above percentages further improved to ∼53%, ∼44%, and ∼2%.

In Figure 3 and Table 1 we report the data for three of the consensus results (A99$\sigma$ + C22, A14 + C22, and C22 + C36). The RMSEs for all these three consensus pairs were the same or better than the best RMSE achieved by a single force field (A99), despite not including the data from this force field. In terms of Pearson correlation, the consensus results compare favorably to those obtained with a single force field, with values between 0.31 and 0.45. In particular, all three consensus combinations retained the significant difference over C36, including C22 + C36. The AUC-ROC of the consensus results approaches closely that of C22, being equal or above 0.64 in all cases. Even C22 + C36 shows good performance (0.68/0.66) compared to most single force fields, despite C36 having the worst AUC-ROC among all force fields tested (0.50/0.51).



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

**Performance of Rosetta and Combined Protocols.** To compare the results of the free energy calculations to a different approach, and to explore the complementarities of different methods, we tested the performance of Rosetta as it currently is the gold-standard tool for protein design. In total, we tested 11 different combinations of Rosetta protocols and scoring functions (Table S2).[12,35,44,45] However, here we discuss only the results of the best performing protocol (*flex_ddg*[35]) with one of the latest Rosetta scoring functions (*beta_nov16,* or here simply "*nov16*"). This protocol was recently proposed for the prediction of changes in protein−protein affinity upon mutation,[35] and we adapted it to calculate changes in protein−ligand affinity. Despite not being originally intended for this purpose, the protocol performed well and comparably to the more rigorous free energy calculations (Figure 4a), achieving a RMSE of 0.99 and 1.46 kcal/mol for the high-reproducibility and the full data sets, respectively. For the more reproducible subset of the data, also the Pearson correlation was competitive to MD (0.39), whereas for the full data set it had a lower performance (0.25). The AUC-ROC was slightly inferior to that of the free energy calculations using consensus force-fields (0.61 and 0.56 for the high-reproducibility and the full data sets, respectively), however, not significantly so (Figure 4c).

Given that the results obtained with *flex_ddg* were on the same scale (in terms of energy units) as those obtained with MD, we combined the information from both approaches via simple averaging, similarly to what was done previously for different force fields. While many forms of regression could be used to result in optimal performance on this data set, we focused on the simplest approach to avoid overfitting. The *flex_ddg* protocol was tested with three scoring functions in total (*talaris2014*, *REF2015*, and *beta_nov16*; see Table S2), and free energy calculations with six force fields. When all possible consensus force field results are considered, there are 63 ways the Rosetta and MD data can be averaged and evaluated with each of the three performance measures employed here (Figure S8). It was found that the performance of these Rosetta + MD consensus results improves upon both parent data in the majority of instances (Figure S8). Specifically, in ∼78% of cases the consensus results were better than both the MD and the Rosetta results, in ∼22% they were in between, while only in ∼0.5% were they worse than both. In Figure 4b−c we show the consensus results derived from the *flex_ddg/nov16* Rosetta protocol and the A14 + C22 free energy calculations as an example (numerical results in Table 1). In this case, the RMSE for the high-reproducibility data set drops well below the 1 kcal/mol mark (0.82 kcal/mol), while also the best RMSE for the full data set is obtained (1.23 kcal/mol). Pearson correlation is also the highest among all results for both data sets (0.44 and 0.49). Finally, the AUC-ROC is comparable to that of the best results obtained via free energy calculations (0.68 and 0.67).

In addition to MD and Rosetta, we also tested three machine learning algorithms: two trained to predict ligand binding affinity (*RF-Score-VS*[46] and *CSM-Lig*[47]) and one specifically trained to predict changes in binding affinity upon protein mutation (*mCSM-Lig*[48]). Perhaps unsurprisingly, the two algorithms trained to predict affinities performed poorly when applied to the different task of predicting binding affinity changes, as they seemed to be insensitive to a single or few protein mutations. Conversely, the *mCSM-Lig* algorithm (based on Gaussian process regression) performed well, similarly to the results obtained here by combining the MD and Rosetta data (Figure S9). However, this algorithm was trained on the same data here used for testing (Platinum database[36]), so this is not an independent validation of the approach. In contrast, the physics-based models are trained on simple physical properties such that this data set tests whether they can extrapolate to more complex properties. While machine learning in general, and *mCSM-Lig* in particular, are certainly promising avenues for the fast estimation of changes in ligand-binding affinity upon protein mutation, other tests on new data sets will be needed to evaluate the performance of such algorithms.

## DISCUSSION

In this work we tested multiple approaches for the estimation of ligand binding affinity changes upon protein mutation on a data set of 134 $\Delta\Delta G$ values across 17 proteins and 27 ligands. Given the diverse and challenging nature of the test set, we believe the results shown provide a representative picture of the average performance for the methods tested. Free energy calculations that employed 216 ns per estimate were used to test the accuracy of six different modern force fields. It was shown that combining the results of two force fields in a consensus approach provides better performance than investing all the simulation time in a single force field. In such a way, the free energy calculations achieved RMSEs with respect to experiment as low as ∼1.3 kcal/mol when considering the full data set, and as low as ∼1.0 kcal/mol when excluding four protein systems posing reproducibility challenges. Overall, this performance is in line with what was observed in previous studies.[31,32,49] In particular, Hauser et al.[31] used similar calculations to estimate the effect of Abl kinase mutations on inhibitor affinity in the context of drug resistance: using a data set of 144 $\Delta\Delta G$ values across eight inhibitors, spanning a range of ∼6 kcal/mol, they obtained a RMSE of 1.1 kcal/mol.

From these calculations, additional observations that can be useful for the practical application of the methodology can be made.

(a) The most efficient free energy protocols invested approximately equal amounts of simulation time in the equilibrium and nonequilibrium parts of the calculations.

(b) Random ion placement within the simulation box was found to be detrimental to reproducibility in some instances, as internal water molecules might be replaced by ions, and ions placed in buried protein cavities could bias equilibrium sampling. Despite its simplistic nature, our approach to exclude water molecules directly around the protein to be replaced by ions, and then allow for a short equilibration, was sufficient to solve the reproducibility issue. More sophisticated approaches that place ions in electrostatic potential minima, or a more rigorous treatment of salt conditions,[50] could also obviate the same issue while allowing for starting ion configurations closer to equilibrium.

(c) The fact that we did not observe an improvement of the results when initiating the simulations of the apo states from apo-state X-ray crystal structures might suggest that, usually, the structure of the complex provides a reasonable starting structure also for the apo state.

(d) While the performance reported is representative of an average across protein−ligand systems, the calculations achieve different performances for specific systems



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

(Figure S10). When focusing on a specific target, it is still recommended to validate the methodology against experimental data whenever possible.

(e) The more mutations carried out in a single calculation, the less precise and accurate the estimates were (Figure S11).

(f) We did not observe a trend for which the net-charge change of the simulation box correlated with the accuracy of the predictions (Figure S12), supporting the hypothesis that, for this specific application/setup, finite size artifacts due to the use of Ewald summation methods[51] mostly cancel out between the two legs of the thermodynamic cycle such that other sources of error dominate the final errors observed (see also Methods and Figure S13).

In this study we also found that a recently proposed Rosetta protocol (*flex_ddg*) can quantitatively predict changes in ligand affinity upon protein mutation (RMSE of ∼1.0 and ∼1.5 kcal/mol for the reduced and full data sets, respectively). However, we note that there is room for improving the performances of both the free energy and Rosetta calculations. For instance, here we carried out a straightforward adaptation of the *flex_ddg* protocol that does not sample different ligand conformations; additional adjustments accounting for ligand flexibility might result in higher performance. The efficiency of the nonequilibrium free energy protocol can also be improved by, e.g., considering nonlinear paths for the alchemical transformation or more suitable soft-core potentials.[52] We also note that in this study we used a fairly short equilibration time of 1 ns for all systems, while some may need longer times to equilibrate.[53] In fact, here we studied multiple protein−ligand systems without in-depth knowledge of any of them. When focusing on a specific system of interest, the user is also likely to have prior information that might be used to improve upon the quality and modeling (e.g., on protonation states, or changes in ligand binding poses). Refinement of force field parameters for the ligand molecules could also result in improved accuracy.[54−56] In fact, general advances in the potential energy and scoring functions used,[12,57] and in the quality and speed of sampling,[58−60] can be expected to result in corresponding improvements in the convergence and accuracy of the calculations here described.

It is important to put the performances of the calculations in the context of their computational cost. With the free energy calculations, each $\Delta\Delta G$ estimate took between 2 and 5 days on a node with 20 CPU threads and 1 GPU (Intel Xeon E5-2630 v4; GTX 1080 Ti), depending on the size of the system (from ∼30 000 to ∼100 000 atoms). With the *flex_ddg* protocol, each $\Delta\Delta G$ estimate took up to a day on a single CPU core. It thus emerges that Rosetta and in particular the *flex_ddg* protocol is likely the most appropriate starting point for a campaign aimed designing ligand-binding proteins or anticipating drug resistance. However, we argue that alchemical free energy calculations are a complementary approach that brings additional value at the optimization stage of the design process. First, the best free-energy-based consensus approaches (e.g., A14 + C22) did return better RMSE, Pearson, and AUC-ROC performance than Rosetta when considering the full data set (Table 1 and Figure 4c), despite the differences not being significant at the $\alpha$ = 0.10 level. This might suggest that the challenges faced by MD simulations are also present, and possibly to a larger extent, in the Rosetta calculations when considering highly flexible systems. Second, the best performance was obtained only when combining MD and Rosetta results (RMSEs < 1.0 kcal/mol on the high-reproducibility data set; Figure S8). Third, nonequilibrium free energy calculations provide not only a $\Delta\Delta G$ estimate, but also extensive sampling of the four end states of interest (apo and holo simulations for the wild type and mutant proteins). These equilibrium simulations in explicit solvent can be further analyzed and used to either apply additional filters or to rationalize experimental results, as was done for instance by Privett et al.[25] and Kiss et al.[61]

Thus, in summary, a possible integrated protocol might include Rosetta calculations for an initial larger screen, followed by refinement of the most promising results via the incorporation of free energy calculations, which would allow for a higher predictive power as well as provide dynamical insight into the effect of the mutations. On the basis of the results obtained here, combining Rosetta and free energy calculation results via simple averaging can achieve RMSEs below 1 kcal/mol when considering protein systems providing well-converged results.

## CONCLUSION

We have shown how both rigorous nonequilibrium free energy calculations based on MD simulations, as well as the latest Rosetta protocols, are able to quantitatively predict changes in ligand binding affinity upon protein mutations. In particular, the best predictions, which were the result of combining the estimate from both methods, closely matched the experimentally determined $\Delta\Delta G$ values, with RMSE of ∼1.2 kcal/mol for the full benchmark set and of ∼0.8 kcal/mol for the set of protein systems showing well-converged results. As these calculations can find direct application for the design of ligand-binding proteins and the prediction of drug resistance, the present results confirm the potential of both physics-based and knowledge-based computational approaches for complementing experimentation in the engineering of biological macromolecules and the design of more robust small molecule drugs.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.8b00717.

Methods, Figures S1−S13, Tables S1−S2, and Text S1−S3 (PDF)

Detailed information on the data set, and numerical results of the free energy and Rosetta calculations provided as an Excel spreadsheet (XLSX)

Input files pertaining to the free energy and Rosetta calculations provided as a compressed archive file (ZIP)

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: bgroot@gwdg.de. Tel. +49-551-2012308. Fax. +49-551-2012302.

### ORCID

Matteo Aldeghi: 0000-0003-0019-8806

Vytautas Gapsys: 0000-0002-6761-7780

Bert L. de Groot: 0000-0003-3570-3534



DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

M.A. was supported by a Postdoctoral Research Fellowship of the Alexander von Humboldt Foundation. V.G. was supported by the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union (Contract H2020-EINFRA-2015-1-675728). M.A. would like to thank Michal H. Kolar for help with the modelling of σ-holes on halogens, and Colin A. Smith, Noah Ollikainen, and Kyle Barlow for support on the use of Rosetta.

## ■ REFERENCES

(1) Beato, M.; Chávez, S.; Truss, M. Transcriptional Regulation by Steroid Hormones. *Steroids* **1996**, *61* (4), 240−251.
(2) Ronnett, G. V.; Moon, C. G Proteins and Olfactory Signal Transduction. *Annu. Rev. Physiol.* **2002**, *64* (1), 189−222.
(3) Missale, C.; Nash, S. R.; Robinson, S. W.; Jaber, M.; Caron, M. G. Dopamine Receptors: From Structure to Function. *Physiol. Rev.* **1998**, *78* (1), 189−225.
(4) Yang, W.; Lai, L. Computational Design of Ligand-Binding Proteins. *Curr. Opin. Struct. Biol.* **2017**, *45*, 67−73.
(5) de Wolf, F. A.; Brett, G. M. Ligand-Binding Proteins: Their Potential for Application in Systems for Controlled Delivery and Uptake of Ligands. *Pharmacol. Rev.* **2000**, *52* (2), 207−236.
(6) Pakulska, M. M.; Miersch, S.; Shoichet, M. S. Designer Protein Delivery: From Natural to Engineered Affinity-Controlled Release Systems. *Science (Washington, DC, U. S.)* **2016**, *351* (6279), aac4750.
(7) Rogers, J. K.; Taylor, N. D.; Church, G. M. Biosensor-Based Engineering of Biosynthetic Pathways. *Curr. Opin. Biotechnol.* **2016**, *42*, 84−91.
(8) Feng, J.; Jester, B. W.; Tinberg, C. E.; Mandell, D. J.; Antunes, M. S.; Chari, R.; Morey, K. J.; Rios, X.; Medford, J. I.; Church, G. M.; Fields, S.; Baker, D. A General Strategy to Construct Small Molecule Biosensors in Eukaryotes. *eLife* **2015**, *4*, No. e10606.
(9) Taylor, N. D.; Garruss, A. S.; Moretti, R.; Chan, S.; Arbing, M. A.; Cascio, D.; Rogers, J. K.; Isaacs, F. J.; Kosuri, S.; Baker, D.; Fields, S.; Church, G. M.; Raman, S. Engineering an Allosteric Transcription Factor to Respond to New Ligands. *Nat. Methods* **2016**, *13* (2), 177−183.
(10) Zhou, L.; Bosscher, M.; Zhang, C.; Özçubukçu, S.; Zhang, L.; Zhang, W.; Li, C. J.; Liu, J.; Jensen, M. P.; Lai, L.; He, C. A Protein Engineered to Bind Uranyl Selectively and with Femtomolar Affinity. *Nat. Chem.* **2014**, *6* (3), 236−241.
(11) Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. E. A.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popović, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **2011**, *487* (C), 545−574.
(12) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031−3048.
(13) Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.; Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard, B. L.; Baker, D. Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. *Nature* **2013**, *501* (7466), 212−216.
(14) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **2003**, *10* (12), 980−980.
(15) Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; Dechancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. a; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008**, *453* (7192), 190−195.
(16) Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St. Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **2010**, *329* (5989), 309−313.
(17) Jiang, L.; Althoff, E. a; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. De Novo Computational Design of Retro-Aldol Enzymes. *Science* **2008**, *319* (5868), 1387−1391.
(18) Richter, F.; Blomberg, R.; Khare, S. D.; Kiss, G.; Kuzin, A. P.; Smith, A. J. T.; Gallaher, J.; Pianowski, Z.; Helgeson, R. C.; Grjasnow, A.; Xiao, R.; Seetharaman, J.; Su, M.; Vorobiev, S.; Lew, S.; Forouhar, F.; Kornhaber, G. J.; Hunt, J. F.; Montelione, G. T.; Tong, L.; Houk, K. N.; Hilvert, D.; Baker, D. Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **2012**, *134* (39), 16197−16206.
(19) Povarova, N. V.; Bozhanova, N. G.; Sarkisyan, K. S.; Gritcenko, R.; Baranov, M. S.; Yampolsky, I. V.; Lukyanov, K. A.; Mishin, A. S. Docking-Guided Identification of Protein Hosts for GFP Chromophore-like Ligands. *J. Mater. Chem. C* **2016**, *4* (14), 3036−3040.
(20) Hong Enriquez, R. P.; Pavan, S.; Benedetti, F.; Tossi, A.; Savoini, A.; Berti, F.; Laio, A. Designing Short Peptides with High Affinity for Organic Molecules: A Combined Docking, Molecular Dynamics, And Monte Carlo Approach. *J. Chem. Theory Comput.* **2012**, *8* (3), 1121−1128.
(21) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem., Int. Ed.* **2013**, *52* (22), 5700−5725.
(22) Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R. E.; Grütter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D. Precision Is Essential for Efficient Catalysis in an Evolved Kemp Eliminase. *Nature* **2013**, *503* (7476), 418−421.
(23) Schreier, B.; Stumpp, C.; Wiesner, S.; Höcker, B. Computational Design of Ligand Binding Is Not a Solved Problem. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (44), 18491−18496.
(24) Fowler, P. W.; Cole, K.; Gordon, N. C.; Kearns, A. M.; Llewelyn, M. J.; Peto, T. E. A.; Crook, D. W.; Walker, A. S. Robust Prediction of Resistance to Trimethoprim in Staphylococcus Aureus. *Cell Chem. Biol.* **2018**, *25* (3), 339−349.
(25) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. a.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative Approach to Computational Enzyme Design. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (10), 3790−3795.
(26) 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* (7), 2695−2703.
(27) Steinbrecher, T. B.; Dahlgren, M.; Cappel, D.; Lin, T.; Wang, L.; Krilov, G.; Abel, R.; Friesner, R. A.; Sherman, W. Accurate Binding Free Energy Predictions in Fragment Optimization. *J. Chem. Inf. Model.* **2015**, *55* (11), 2411−2420.
(28) Kuhn, B.; Tichý, M.; Wang, L.; Robinson, S.; Martin, R. E.; Kuglstatter, A.; Benz, J.; Giroud, M.; Schirmeister, T.; Abel, R.; Diederich, F.; Hert, J. Prospective Evaluation of Free Energy


DOI: 10.1021/acscentsci.8b00717
ACS Cent. Sci. 2018, 4, 1708−1718

Calculations for the Prioritization of Cathepsin L Inhibitors. *J. Med. Chem.* **2017**, *60* (6), 2485−2497.
(29) 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* (26), 7364−7368.
(30) Steinbrecher, T.; Zhu, C.; Wang, L.; Abel, R.; Negron, C.; Pearlman, D.; Feyfant, E.; Duan, J.; Sherman, W. Predicting the Effect of Amino Acid Single-Point Mutations on Protein Stability—Large-Scale Validation of MD-Based Relative Free Energy Calculations. *J. Mol. Biol.* **2017**, *429* (7), 948−963.
(31) Hauser, K.; Negron, C.; Albanese, S. K.; Ray, S.; Steinbrecher, T.; Abel, R.; Chodera, J. D.; Wang, L. Predicting Resistance of Clinical Abl Mutations to Targeted Kinase Inhibitors Using Alchemical Free-Energy Calculations. *Commun. Biol.* **2018**, *1* (1), 70.
(32) Bastys, T.; Gapsys, V.; Doncheva, N. T.; Kaiser, R.; de Groot, B. L.; Kalinina, O. V. Consistent Prediction of Mutation Effect on Drug Binding in HIV-1 Protease Using Alchemical Calculations. *J. Chem. Theory Comput.* **2018**, *14* (7), 3397−3408.
(33) Hayes, R. L.; Vilseck, J. Z.; Brooks, C. L. Approaching Protein Design with Multisite $\lambda$ Dynamics: Accurate and Scalable Mutational Folding Free Energies in T4 Lysozyme. *Protein Sci.* **2018**, *27*, 1910−1922.
(34) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Automated protein structure and topology generation for alchemical perturbations. *J. Comput. Chem.* **2015**, *36*, 348−354.
(35) Barlow, K. A.; Ó Conchúir, S.; Thompson, S.; Suresh, P.; Lucas, J. E.; Heinonen, M.; Kortemme, T. Flex DdG: Rosetta Ensemble-Based Estimation of Changes in Protein−Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **2018**, *122* (21), 5389−5399.
(36) Pires, D. E. V; Blundell, T. L.; Ascher, D. B. Platinum: A Database of Experimentally Measured Effects of Mutations on Structurally Defined Protein−Ligand Complexes. *Nucleic Acids Res.* **2015**, *43* (D1), D387−D391.
(37) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300−313.
(38) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245−268.
(39) 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* (14), 140601.
(40) Crooks, G. E. Path-Ensemble Averages in Systems Driven Far from Equilibrium. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2000**, *61* (3), 2361−2366.
(41) Crooks, G. E. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.* **1998**, *90* (5/6), 1481−1487.
(42) Crooks, G. E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1999**, *60* (3), 2721−2726.
(43) Gapsys, V.; de Groot, B. L. Alchemical Free Energy Calculations for Nucleotide Mutations in Protein−DNA Complexes. *J. Chem. Theory Comput.* **2017**, *13* (12), 6275−6289.
(44) Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12* (12), 6201−6212.
(45) Ollikainen, N.; de Jong, R. M.; Kortemme, T. Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-Design of Protein-Ligand Specificity. *PLoS Comput. Biol.* **2015**, *11* (9), No. e1004335.
(46) Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. *Sci. Rep.* **2017**, *7*, 46710.
(47) Pires, D. E. V.; Ascher, D. B. CSM-Lig: A Web Server for Assessing and Comparing Protein−Small Molecule Affinities. *Nucleic Acids Res.* **2016**, *44* (W1), W557.
(48) Pires, D. E. V.; Blundell, T. L.; Ascher, D. B. MCSM-Lig Quantifying the Effects of Mutations on Protein-Small Molecule Affinity in Genetic Disease and Emergence of Drug Resistance. *Sci. Rep.* **2016**, *6* (1), 29575.
(49) Rizzo, R. C.; Wang, D.-P.; Tirado-Rives, J.; Jorgensen, W. L. Validation of a Model for the Complex of HIV-1 Reverse Transcriptase with Sustiva through Computation of Resistance Profiles. *J. Am. Chem. Soc.* **2000**, *122* (51), 12898−12900.
(50) Ross, G. A.; Rustenburg, A. S.; Grinaway, P. B.; Fass, J.; Chodera, J. D. Biomolecular Simulations under Realistic Macroscopic Salt Conditions. *J. Phys. Chem. B* **2018**, *122* (21), 5466−5486.
(51) Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H. Calculating the Binding Free Energies of Charged Species Based on Explicit-Solvent Simulations Employing Lattice-Sum Methods: An Accurate Correction Scheme for Electrostatic Finite-Size Effects. *J. Chem. Phys.* **2013**, *139*, 184103.
(52) 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* (7), 2373−2382.
(53) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2016**, *12* (4), 1799−1805.
(54) Huang, L.; Roux, B. Automated Force Field Parameterization for Non-Polarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *J. Chem. Theory Comput.* **2013**, *9* (8), 3543−3556.
(55) Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L. Characterization of Biaryl Torsional Energetics and Its Treatment in OPLS All-Atom Force Fields. *J. Chem. Inf. Model.* **2013**, *53* (5), 1191−1199.
(56) Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Predictions of Ligand Selectivity from Absolute Binding Free Energy Calculations. *J. Am. Chem. Soc.* **2017**, *139* (2), 946−957.
(57) 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* (1), 281−296.
(58) Sivak, D. A.; Chodera, J. D.; Crooks, G. E. Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Phys. Rev. X* **2013**, *3* (1), 11007.
(59) Fass, J.; Sivak, D.; Crooks, G.; Beauchamp, K.; Leimkuhler, B.; Chodera, J. Quantifying Configuration-Sampling Error in Langevin Simulations of Complex Molecular Systems. *Entropy* **2018**, *20* (5), 318.
(60) 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* (30), 9431−9438.
(61) Kiss, G.; Röthlisberger, D.; Baker, D.; Houk, K. N. Evaluation and Ranking of Enzyme Designs. *Protein Sci.* **2010**, *19* (9), 1760−1773.


DOI: 10.1021/acscentsci.8b00717
*ACS Cent. Sci.* 2018, 4, 1708−1718