Supporting Information

Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation

Matteo Aldeghi[1], Vytautas Gapsys[1], Bert L. de Groot[1]*

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

*To whom correspondence should be addressed.

Email: bgroot@gwdg.de

Tel. +49-551-2012308

Fax. +49-551-2012302

Table of Contents

Methods

Dataset

The dataset of binding free energy differences and associated structural information was extracted from the Platinum database[1]. The database was filtered to retain only affinities that were determined by isothermal titration calorimetry (ITC) or surface plasmon resonance (SPR), and proline mutations were excluded. Proline mutations are technically challenging as they involve a bond-breaking perturbation[2] and as such were not supported in the original *pmx* library[3,4]. For calculations using Charmm force fields, also glycine mutations were excluded, as it is currently not possible to interpolate between different grid-based energy correction maps (CMAPs) in Gromacs. Because the contribution of the CMAP to $\partial H/\partial \lambda$ is neglected, and because the CMAPs of glycine and proline are different from those of other residues, mutations involving these two residues in Charmm would result in incorrect free energy differences. Other entries were excluded due to the poor quality of the structural data (e.g. only C-alpha atoms present) or the presence of non-standard amino acids in proximity to the ligand. The PDB-ID 3AQT was modified by replacing all selenomethionines with methionines. To obtain a distribution of $\Delta\Delta G$ that was not biased towards mutations unfavorable for binding, when X-ray structures of both wild type and mutant proteins were available for positive $\Delta\Delta G$ values, we took the mutant protein as being the "wild type" so to invert the sign of the affinity change. This resulted in a set of 134 experimental $\Delta\Delta G$ values used as reference. All information pertaining the benchmark set (PDB-IDs, ligands, affinities, temperature, pH, etc.) can be found in the Supporting Information.

System Setup

The structures of the protein-ligand complexes were taken from the PDB (PDB-IDs in Supporting Information). Missing loops in 3ZY2, 3H2K, and 3AQT were modelled using Sphinx[5]. Apo structures were generated simply by removing the ligand atoms. Crystallographic water molecules were removed. All mutant structures were generated using FoldX v4,[6] including the cases where experimental structural data for the mutant was available. Protein protonation states were assigned at experimental pH (Supporting Information) using the protein preparation tool in *HTMD* (v1.12)[7], which uses *propka* v3.1[8,9]. Ligand protonation states were assigned based on the most abundant species at the pH of interest using the *pKa* calculator in MarvinSketch (v17, ChemAxon)[10]. Proteins were modelled with the Amber99sb*-ILDN[11–13], Amber14sb[14], Charmm22*[15–17], Charmm36[18], and Charmm36m[19] force fields. The TIP3P water model was used.[20] Ligands were modelled with GAFF2 (v2.1)[21] via *AmberTools 16* and *acpype* (v2017-01-17),[22,23] and CGenFF (v3.0.1)[24] via *paramchem*[25,26]. With GAFF2, two charge models were tested: AM1-BCC[27,28] and restrained electrostatic potential (RESP)[29] charges. Geometry optimizations and molecular electrostatic potential calculations (ESP) were performed with Gaussian 09 (Rev. D.01), both at the HF/6-31G* level of theory. ESP points were sampled according to the Merz-Kollman scheme.[30,31] In addition, in the GAFF2/RESP models the σ-hole on halogen atoms was modelled as described by Kolar & Hobza[32].

The protein-ligand systems were solvated in a dodecahedral box with periodic boundary conditions and a minimum distance between the solute and the box of 12 Å. Sodium and chloride ions were added to neutralize the wild type systems at the concentration of 0.15 M. For the mutant systems, the same number of ions as in the wild type systems was added; i.e. the net charge of the wild type systems was always zero, while the net charge of the mutant systems was allowed to deviate from zero according to the mutation. During testing on a model system, we found that the finite size artefacts due

to the use of Ewald summation almost completely cancelled out in the two legs of the thermodynamic cycle used for the free energy calculations (Figure S13).

Since FoldX does not consider the presence of ligands when mutating the protein, clashes with the ligands in the mutated complexes are possible. If clashes (defined as protein heavy atoms within 1 Å from any ligand heavy atom) were present, an approach similar to the one reported for *alchembed*[33] was used: after 2,000 steepest descent steps, the ligand vdW interactions were switched on in 2,000 MD steps carried out with a 0.5 fs timestep, while using position restraints (1,000 kJ mol^{-1} nm^{-2}) on all heavy atoms. This procedure resolved all clashes encountered.

Free Energy Calculations

All simulations were carried out in Gromacs 2016.[34,35] 10,000 energy minimization steps were performed using a steepest descent algorithm. The systems were subsequently simulated for 100 ps in the isothermal-isobaric ensemble (NPT) with harmonic position restraints applied to all solute heavy atoms with a force constant of 1,000 kJ mol^{-1} nm^{-2}. Temperature was coupled using Langevin dynamics at the experimental target temperature, while pressure was coupled using the Berendsen weak coupling algorithm with a target pressure of 1 bar.[36–38] The particle mesh Ewald (PME) algorithm[39] was used for electrostatic interactions with a real space cut-off of 10 Å when using Amber force fields and 12 Å when using Charmm force fields, a spline order of 4, a relative tolerance of 10^{-5} and a Fourier spacing of 1.2 Å. The Verlet cut-off scheme with the potential-shift modifier was used with a Lennard-Jones interaction cut-off of 10 Å with Amber and 12 Å with Charmm force fields, and a buffer tolerance of 0.005 kJ/mol/ps.[40] All bonds were constrained with the P-LINCS algorithm.[41] For equilibration, 1 ns unrestrained MD simulations were then performed in the NPT ensemble with the

Parrinello-Rahman pressure coupling algorithm at 1 bar with a time constant of 2 ps.[42] Production simulations were then performed for up to 10 ns in length.

For each free energy calculation, multiple equilibrium simulation repeats were used (up to 10), so that the above described procedure (from system setup to minimization, equilibration, and production MD) was repeated multiple times for each ΔΔG estimate. From each equilibrium simulation, between 2 and 100 equally spaced frames were extracted as the starting configurations for the non-equilibrium part of the calculations. The non-interacting ("dummy") atoms needed to morph the wild-type residues into mutant ones were introduced at this stage with the *pmx* package[4], using the mutant structure proposed by FoldX as a template. The positions of the dummy atoms were minimized while freezing the rest of the system. These systems containing hybrid residues were then simulated for 10 ps to equilibrate velocities. Amino acid side chains were finally alchemically morphed at constant speed during non-equilibrium simulations of 20, 40, 50, 80, and 100 ps in length. The work values associated with each nonequilibrium transition were extracted using thermodynamic integration (TI) and then used to estimate the free energy differences with the Bennett's Acceptance Ratio (BAR).[43–45]

Point estimates of the free energy differences (Figure 1a: $\Delta G^{apo}_{WT\rightarrow MT}$ and $\Delta G^{holo}_{WT\rightarrow MT}$) were calculated with BAR after pooling all available forward and reverse work values coming from the non-equilibrium trajectories spawned from all equilibrium simulation repeats that were run for a free energy calculation. Uncertainties in $\Delta G^{apo}_{WT\rightarrow MT}$ and $\Delta G^{holo}_{WT\rightarrow MT}$ were estimated as standard errors ($\sigma_{\Delta G}$) by separately considering each equilibrium simulation and its related non-equilibrium trajectories as independent calculations (e.g. when 10 equilibrium simulations were used, 10 independent ΔG values could be obtained, and these were used to estimate $\sigma_{\Delta G}$). These uncertainties were then propagated to the final $\Delta\Delta G_{bind}$ estimate so to obtain the estimate of the standard error $\sigma_{\Delta\Delta G}$. We define the overall

precision of a free energy protocol as the RMS of the 134 $\sigma_{\Delta\Delta G}$ values (RMSσ) obtained across the whole set of mutations. In addition, we define the overall "reproducibility" of a free energy protocol as the RMSE between the two sets of $\Delta\Delta G_{bind}$ point estimates obtained by running the same protocol twice on the whole set of 134 mutations (using 5 or 10 equilibrium simulations and 150 or 300 non-equilibrium trajectories in both directions for each mutation).

Rosetta Calculations

Binding free energy changes were calculated in Rosetta using three different protocols: the *cartesian_ddg* protocol[46–48] (Rosetta v2017.36), the *coupled_moves* protocol[49] (Rosetta v2017.36), and the *flex_ddg* protocol[50] (Rosetta v2017.52). For the *cartesian_ddg* protocol, the biological assembly of the proteins was used, as it was done for the MD simulations. For the *coupled_moves* and *flex_ddg* protocols, the smallest number of chains needed to capture the protein-ligand interactions were kept. For instance, for a homo-tetramer in which the four chains bind four ligands separately, only one chain and one ligand were kept (e.g. streptavidin, PDB-ID 3RY2); for a homo-dimer binding a single ligand at the interface of the two chains, both protein chains were kept (e.g. HIV-1 protease, PDB-ID 2Q63); for a homo-trimer in which each of the three ligands binds at the interface of two chains, one ligand and two chains were kept (e.g. ACPS, PDB-ID 2JBZ). Ligand parameters were obtained with the *molfile_to_params.py* script provided with Rosetta (Supporting Information).

In the *cartesian_ddg* protocol, the lowest scoring model of the protein-ligand complex after refinement with the *relax* command (command lines in Text S3) was used as input for the protocol. The final ΔΔG scores were obtained by averaging the results of the 50 iterations of the *cartesian_ddg* protocol. Five different scoring functions were used: *score12*, *talaris2013*, *talaris2014*,[51–54] *REF2015*,[48]

and *beta_nov2016*. In the *coupled_moves* and the *flex_ddg* protocols, only the *talaris2014*, *REF2015*, and *beta_nov2016* scoring functions were tested. The final ΔΔG scores for the *coupled_moves* protocol were obtained by taking the lowest energy score among the 20 iterations performed for each calculation. For the *flex_ddg* protocol, the final ΔΔG estimates were instead the average values of the generalized additive model obtained from 35 iterations of the protocol.[50] The command lines used for each Rosetta calculation are reported in Text S3.

Data Analysis

The accuracy of the calculations was evaluated using three performance measures: the root mean square error (RMSE), the Pearson correlation, and the area under the receiver operating characteristic curve (AUC-ROC). The uncertainty in the estimate of these measures was evaluated by bootstrap. Specifically, pairs of experimental and calculated ΔΔG values were resampled with replacement 10^5 times and, given that each experimental and calculated ΔΔG value has an associated standard error, a data point was sampled randomly by assuming a Gaussian distribution around the mean values. In this way, the uncertainty due to the imprecision of the calculations, due to the specific choice of dataset and due to variability within the dataset are reflected in the confidence intervals shown. From the 10^5 bootstrap measures obtained, we took the 2.5 and 97.5 percentile as the lower and upper bounds of the 95% confidence interval. For the Rosetta calculations, which did not return an estimate of the standard error, the parametric part of the bootstrap procedure was not performed. The experimental standard error for the ΔΔG values was taken to be 0.18 kcal/mol, based on the variance of the ΔG measurements found by the ABRF-MIRG'02 study for isothermal titration calorimetry (ITC)[55]. A bootstrap procedure was also used to obtain p-values for the differences between force fields and approaches. In this case, triplets of ΔΔG values were resampled with replacement together 10^5 times: ΔΔG values from experiment and from the two approaches to be compared. Both non-parametric and parametric bootstrap strategies were

used together as described above whenever possible. At each bootstrap iteration, the difference in the performance measure of interest (e.g. RMSE) between the two approaches to be compared was stored. In this way, at the end of the procedure, 10^5 bootstrap differences (e.g. Δ_{RMSE}) would have been collected. The fraction of differences crossing zero was multiplied by two so to provide a two-tailed p-value for the difference observed.[56] Data analysis was performed in *python* using the *numpy*, *scipy*, *pandas*, *scikit-learn, matplotlib,* and *seaborn* libraries.

Table S1. Reproducibility of free energy calculations by protein system. The table reports the RMSD between two repeated calculations performed with the Amber99sb*-ILDN/GAFF2 force field. Four protein systems were found to be particularly challenging, with RMSDs above 1 kcal/mol: aldose reductase, streptavidin, D7R4 protein, and RolR.

Protein	#ΔΔG	RMSE (kcal/mol)
Streptavidin	8	1.61
D7R4 Protein	11	1.50
Aldose reductase	26	1.27
RolR	3	1.11
HIV-1 Protease	9	0.90
O-FucT-1	9	0.89
Esterase LipA	6	0.86
HSP82	24	0.73
RtxA	10	0.64
KPA reductase	4	0.62
ACPS	4	0.60
BCKADE2	5	0.44
Anti-tumor lectin	4	0.40
Epsin	3	0.31
Exoenzyme C3	2	0.17
PMT	4	0.11
DHFR	2	0.01

RolR: bacterial transcriptional repressor RolR
O-FucT-1: Putative GDP-fucose protein o-fucosyltransferase 1
HSP82: ATP-dependent molecular chaperone HSP82
RtxA: RTX toxin RtxA
ACPS: holo-[acyl-carrier-protein] synthase
BCKADE2: Lipoamide acetyltransferase component of branched-chain alpha-keto acid dehydrogenase complex
PMT: Phosphoethanolamine N-methyltransferase
DHFR: dihydrofolate reductase type I.

Table S2. Rosetta results. Each evaluation metric is shown with its 95% confidence interval. "n.a.": "not applicable" is used for combination of Rosetta protocols and scoring functions that were not developed to return quantitative results.

Protocol	Scoring Function	#ΔΔG	Experimental ΔΔG Range (kcal/mol)	RMSE (kcal/mol)	Pearson	AUC-ROC
Cartesian_ddg	Score12	86	6.1	n.a.	$0.25^{0.43}_{0.03}$	$0.60^{0.72}_{0.46}$
		134	9.5	n.a.	$023.^{0.40}_{0.04}$	$0.59^{0.69}_{0.50}$
Cartesian_ddg	Talaris2013	86	6.1	n.a.	$0.23^{0.40}_{0.01}$	$0.59^{0.68}_{0.43}$
		134	9.5	n.a.	$0.35^{0.48}_{0.21}$	$0.65^{0.73}_{0.54}$
Cartesian_ddg	Talaris2014	86	6.1	n.a.	$0.23^{0.41}_{0.00}$	$0.59^{0.68}_{0.43}$
		134	9.5	n.a.	$0.34^{0.47}_{0.18}$	$0.63^{0.71}_{0.52}$
Cartesian_ddg	REF2015	86	6.1	$5.68^{6.72}_{4.61}$	$0.23^{0.41}_{0.03}$	$0.64^{0.73}_{0.48}$
		134	9.5	$5.98^{6.93}_{5.01}$	$0.33^{0.47}_{0.17}$	$0.64^{0.73}_{0.54}$
Cartesian_ddg	Beta_nov16	86	6.1	$6.36^{7.46}_{5.22}$	$0.23^{0.39}_{0.04}$	$0.58^{0.69}_{0.44}$
		134	9.5	$6.52^{7.58}_{5.43}$	$0.33^{0.46}_{0.17}$	$0.62^{0.72}_{0.53}$
Coupled_moves	Talaris2014	86	6.1	n.a.	$0.11^{0.26}_{-0.06}$	$0.55^{0.66}_{0.39}$
		134	9.5	n.a.	$0.25^{0.37}_{0.12}$	$0.58^{0.68}_{0.47}$
Coupled_moves	REF2015	86	6.1	n.a.	$0.08^{0.24}_{-0.09}$	$0.60^{0.70}_{0.43}$
		134	9.5	n.a.	$0.23^{0.34}_{0.11}$	$0.63^{0.72}_{0.52}$
Coupled_moves	Beta_nov16	86	6.1	n.a.	$0.16^{0.31}_{-0.02}$	$0.59^{0.69}_{0.43}$
		134	9.5	n.a.	$0.25^{0.35}_{0.14}$	$0.63^{0.72}_{0.52}$
Flex_ddg	Talaris2014	86	6.1	$1.04^{1.21}_{0.90}$	$0.36^{0.53}_{0.12}$	$0.52^{0.62}_{0.37}$
		134	9.5	$1.45^{1.68}_{1.25}$	$0.26^{0.41}_{0.08}$	$0.52^{0.61}_{0.41}$
Flex_ddg	REF2015	86	6.1	$1.05^{1.21}_{0.92}$	$0.33^{0.48}_{0.11}$	$0.55^{0.66}_{0.41}$
		134	9.5	$1.42^{1.62}_{1.25}$	$0.31^{0.44}_{0.15}$	$0.53^{0.62}_{0.42}$
Flex_ddg	Beta_nov16	86	6.1	$0.99^{1.16}_{0.86}$	$0.39^{0.54}_{0.16}$	$0.61^{0.71}_{0.47}$
		134	9.5	$1.46^{1.73}_{1.23}$	$0.25^{0.43}_{0.04}$	$0.56^{0.65}_{0.45}$

Figure S1. Chemical structures of the ligands present in the dataset. For each ligand, the residue name and the number of calculations it is present in are shown. The protonation states depicted were determined by RDKit and are just for visualization. The details of the protonation states used can be found in the input files provided in the Supporting Information.

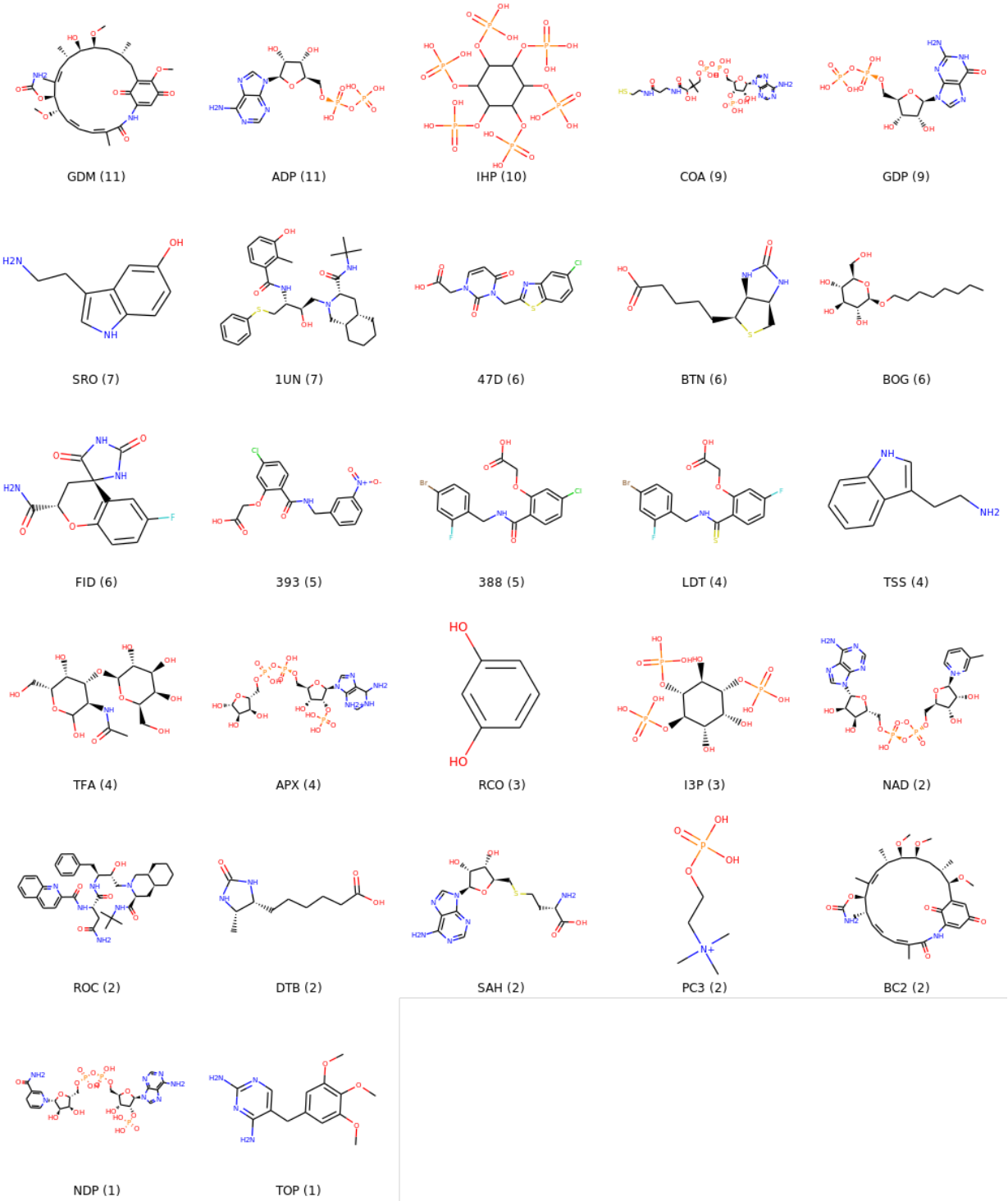


Figure S2. Distribution of experimental ΔΔG values present in the dataset. All ΔΔG values are shown as a rug plot, along with their histogram and kernel density estimate (KDE), and a Gaussian fit of the data.

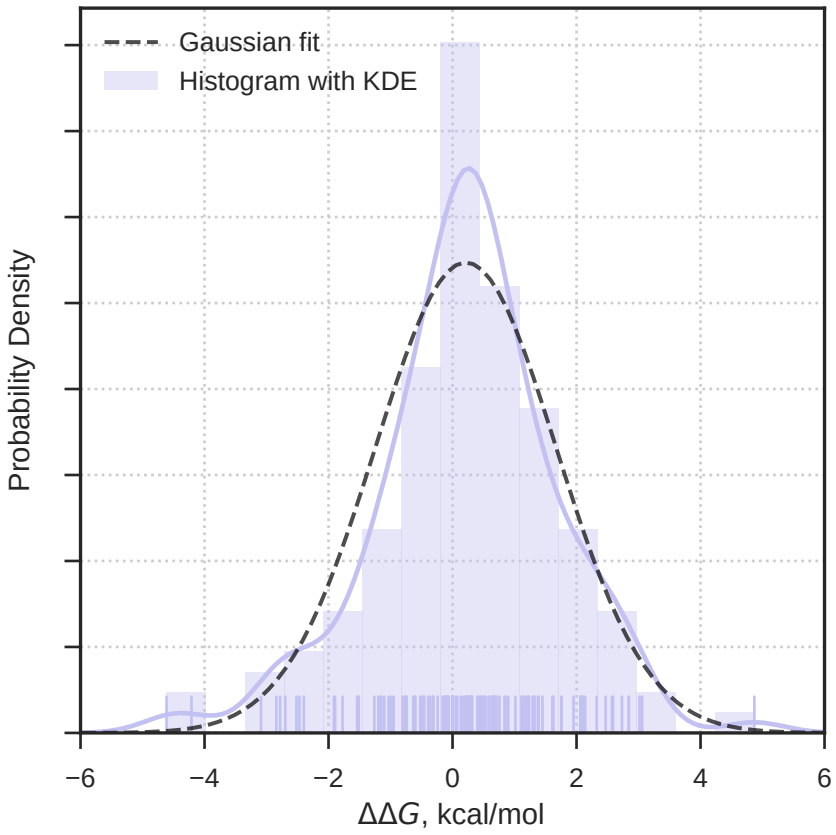


Figure S3. Space of protocol setup parameters tested. The three axes indicate the length of the equilibrium simulations (5 repeats of 1 to 10 ns), the number of nonequilibrium trajectories spawned from the equilibrium simulations (from 10 to 500), and their length (from 20 to 100 ps). Each mark represents a specific combination of the above three variables, with the color indicating the overall accuracy (RMSE) of each calculation.

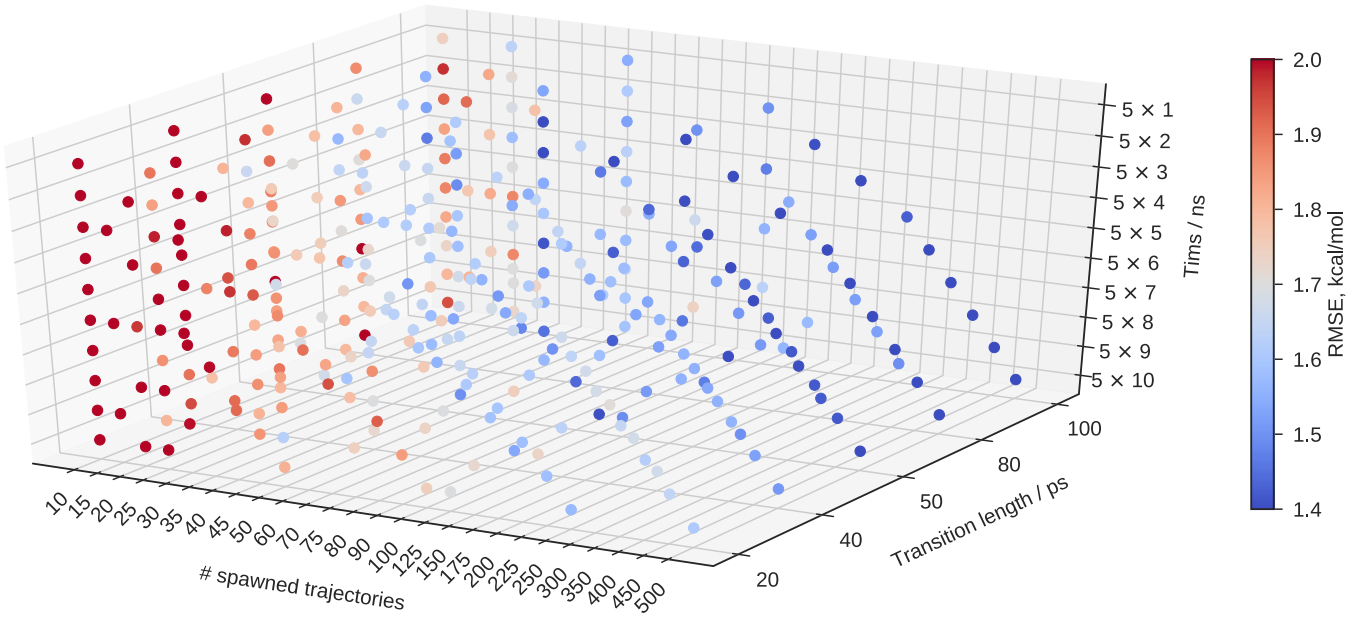


Figure S4. Effect of the nonequilibrium transition length on the overall precision (RMSσ) and accuracy (RMSE) of the calculations. The RMSσ and RMSE of all protocols tested (Figure 2a and Figure S3) have been pooled for each transition length and shown here as swarm and box plots. The boxes show the first, second, and third quartiles of the data, while the whiskers are up to 1.5 times the interquartile range. Looking at the median values, while the precision seems to keep improving with slower transitions, the accuracy seems to have reached its optimum at transition lengths of 80 ps.

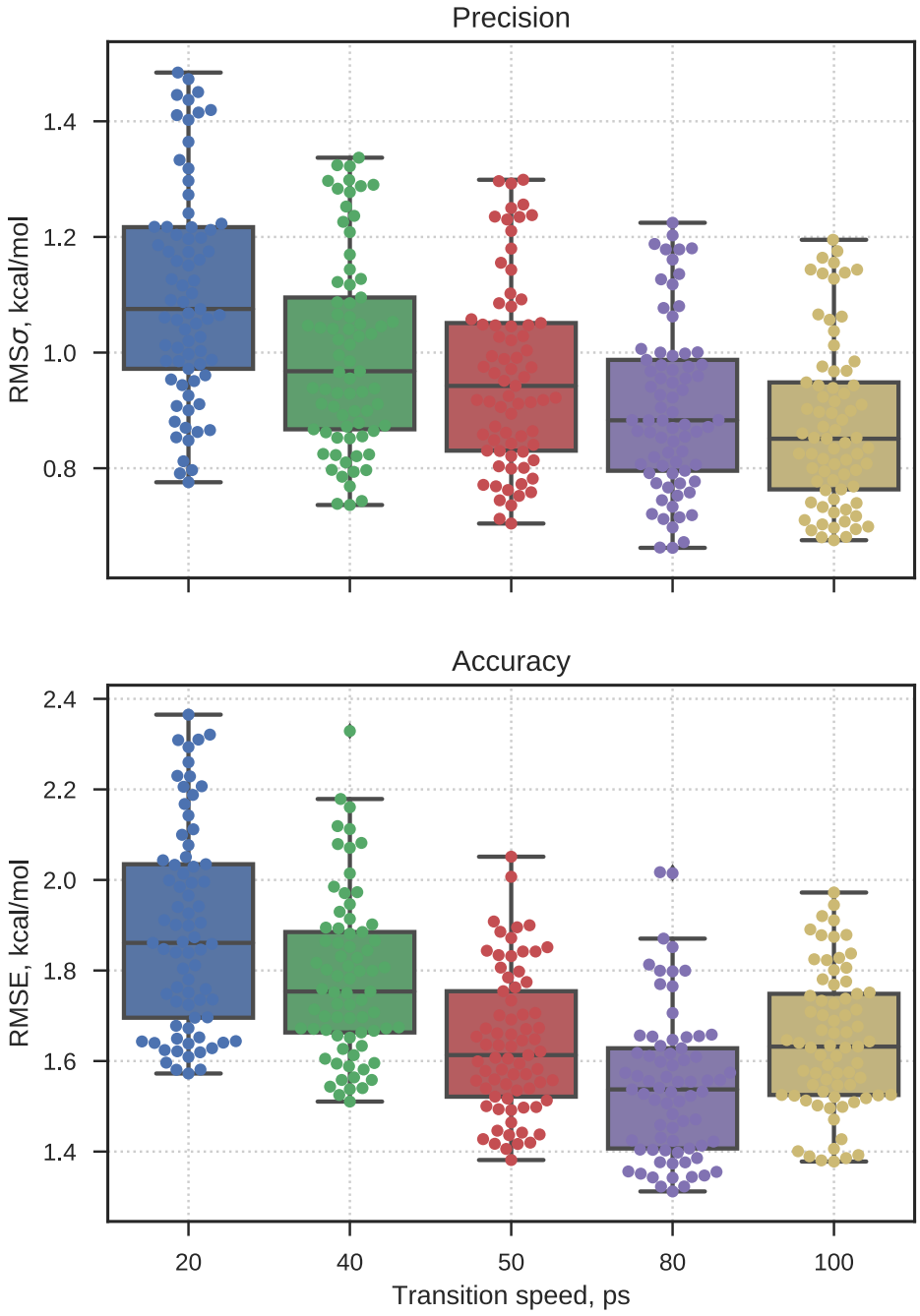


Figure S5. Reproducibility of the calculations and ion placement. a) Scatter plot showing the results for all 134 ΔΔG values obtained in two repeated sets of calculations. The uncertainties shown are one standard error of the mean. The RMSE obtained for the two repeats was large (1.57 kcal/mol). Part of the issue was found to be due to ion placement: in the first set of calculations, some structural water molecules were replaced by ions during system setup, while this did not happen in the second set of calculations. Consequently, ion placement biased some of the results in the two different repeats in a way that was not accounted for by the standard error (all equilibrium simulations were originally started from the same system configuration). b) Aldose reductase starting structure shown as an example. A buried water molecule was replaced by a chloride ion (green sphere), biasing the equilibrium sampling of the calculation.

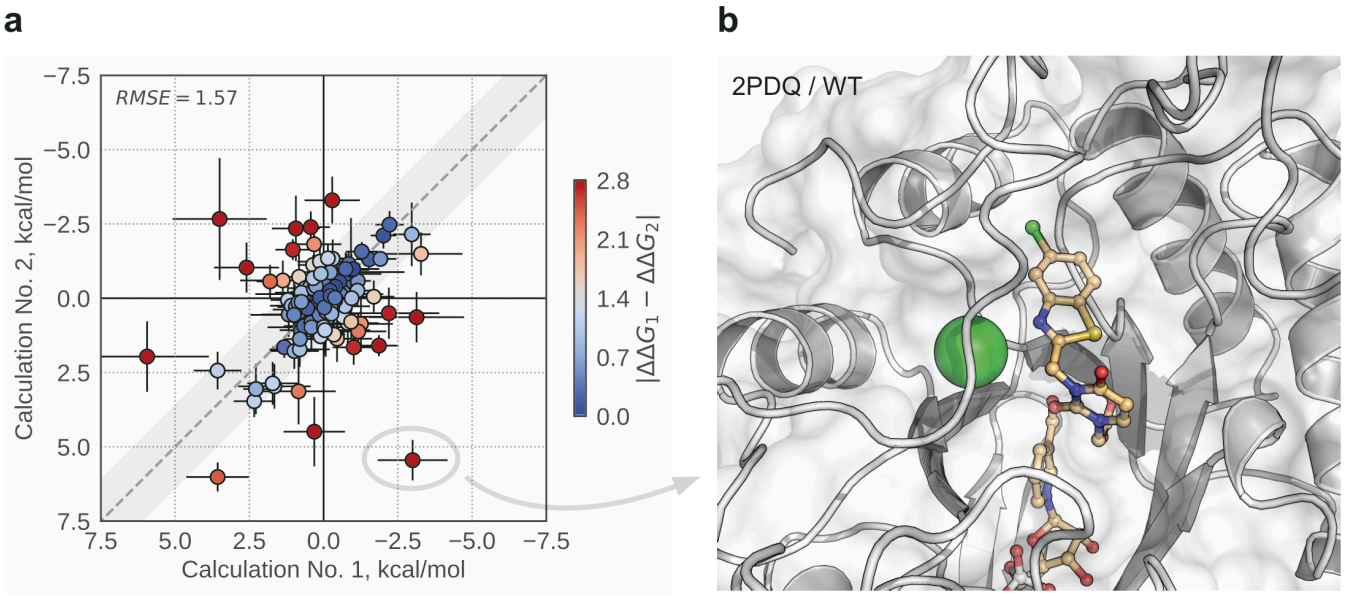


Figure S6. Performance of free energy calculations when using X-ray structures of the apo states for five protein systems: aldose reductase (PDB-ID 1ADS), streptavidin (3RY1), HIV-1 protease (1HHP), RolR (3AQS), and ACPS (2JCA). The results shown are for these systems only (total of 50 ΔΔG values). Performance is summarized for the force fields tested in terms of RMSE, Pearson correlation, and AUC-ROC (point estimates and 95% CIs are shown). Full-circle markers refer to the free energy calculations that used the X-ray structures of the protein-ligand complexes only, while empty-circle markers refer to the free energy calculations that also used X-ray structures of the apo proteins. Overall, the use of apo structures did not result in a significant improvement of the calculations' performance for this dataset.

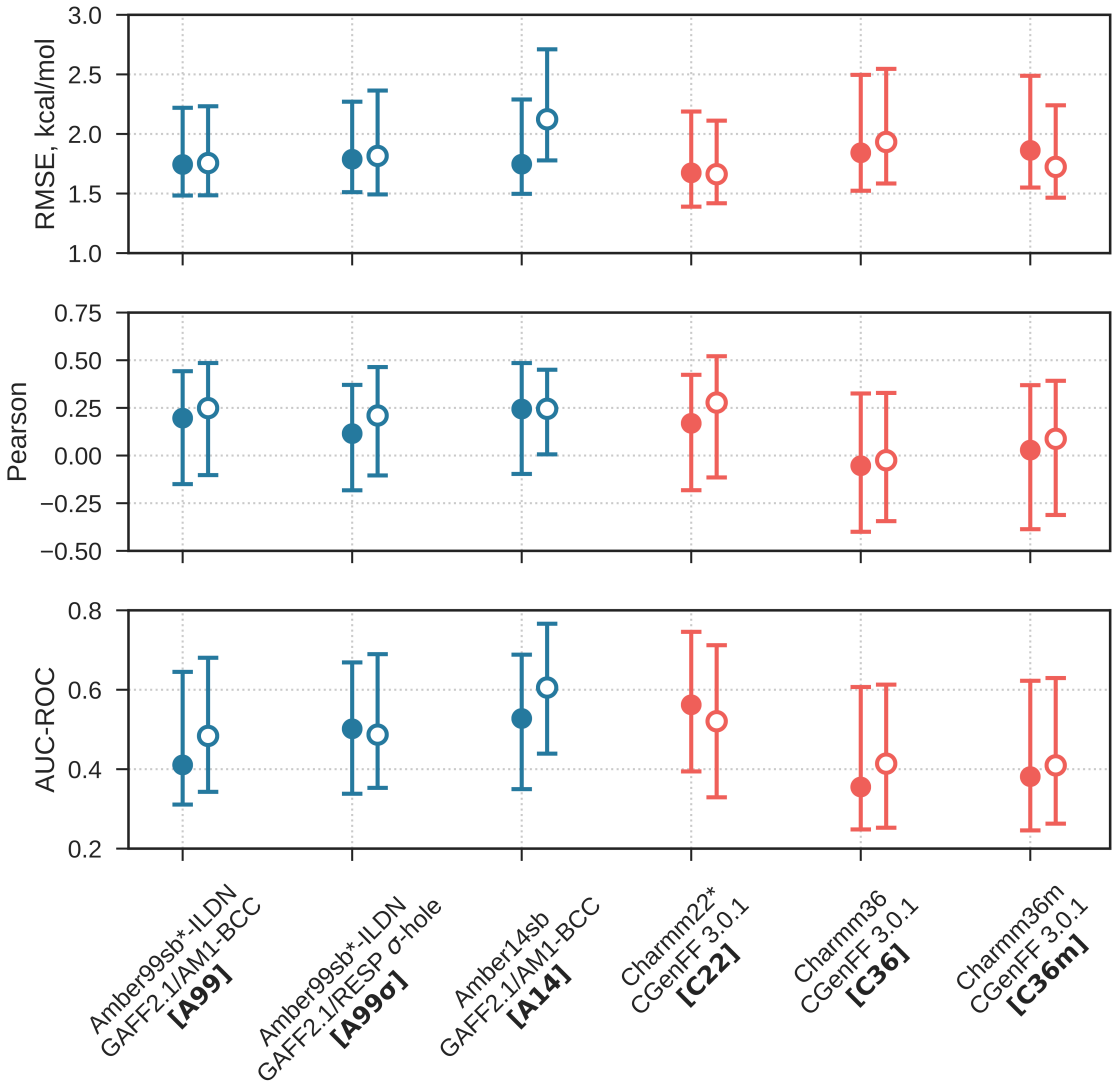


Figure S7. Analysis of the performance of the force field consensus approach. On the diagonal of the matrices shown are the performances achieved by the parent force fields using all simulation data available. The off-diagonal elements show instead the performance of the consensus results obtained by averaging the results of the parent force fields when using half of the simulation data from each parent. Cells are color-coded depending whether performance of the consensus approach is better, in between, or worse than the performance of the two parent force fields. Note that when combining Charmm and Amber force fields, a subset of the full dataset is considered that excludes glycine mutations (see Methods). Standard errors in the final percentages shown were determined by bootstrap.

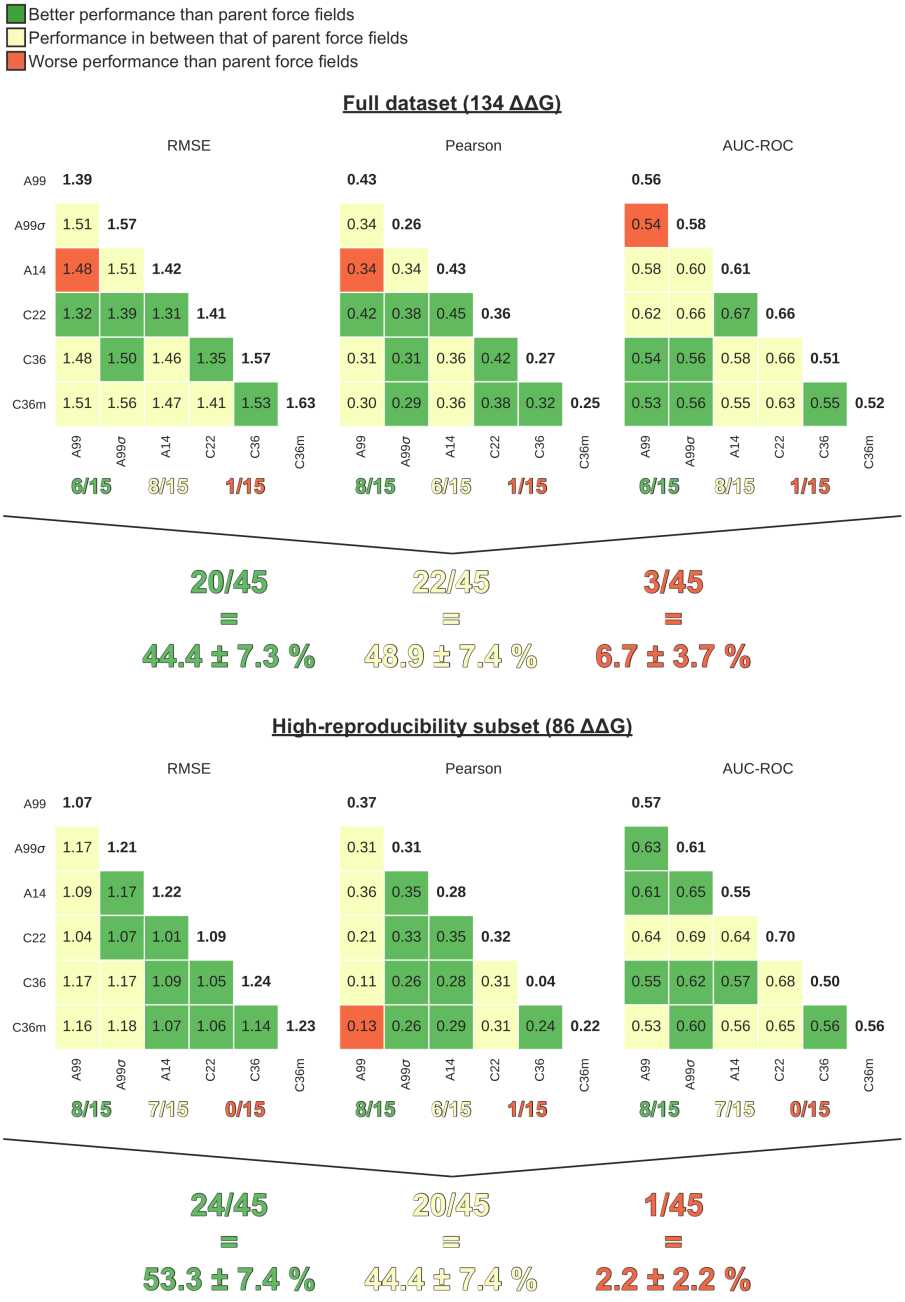


Figure S8. Analysis of the performance of combined Rosetta and MD consensus approach. Each cell shows the performance the consensus results obtained by averaging the results from Rosetta (*flex_ddg* protocol) and the free energy calculations. Cells are color-coded depending whether performance of the consensus approach is better, in between, or worse than the performance of Rosetta or MD alone. Note that when combining Charmm and Amber force fields, a subset of the full dataset is considered that excludes glycine mutations (see Methods). Standard errors in the final percentages shown were determined by bootstrap.

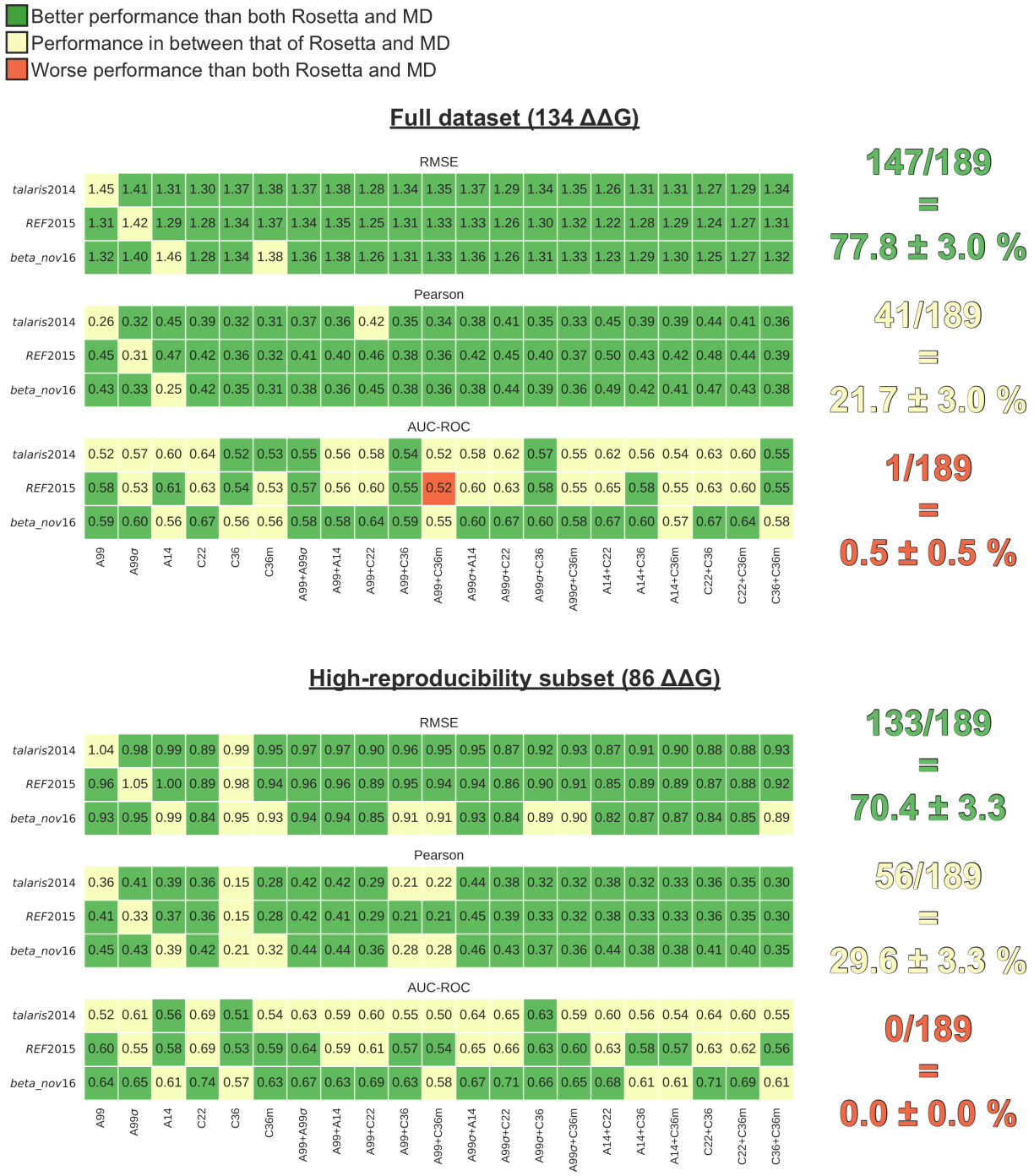


Figure S9. Overview of the performance of three machine learning approaches. The evaluation measures on the top-left of the plots refer to the high-reproducibility subset, while those on the bottom-right to the full dataset.

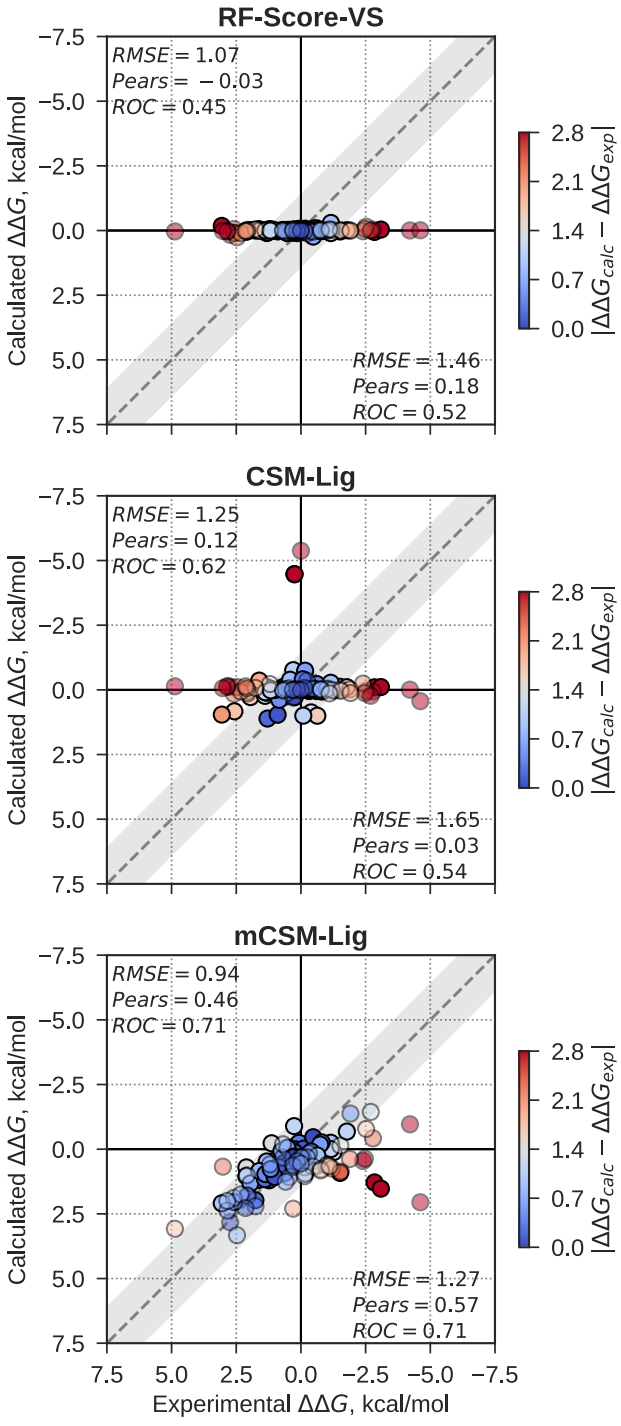


Figure S10. Free energy calculation performance by protein system. Only proteins with at least 5 data points for both Amber and Charmm force fields are shown. a) Swarm and box plots of the performance of the six force fields tested across different protein systems. The boxes show the first, second, and third quartiles of the data, while the whiskers are up to 1.5 times the interquartile range. b) Example scatter plots for the A99 force field.

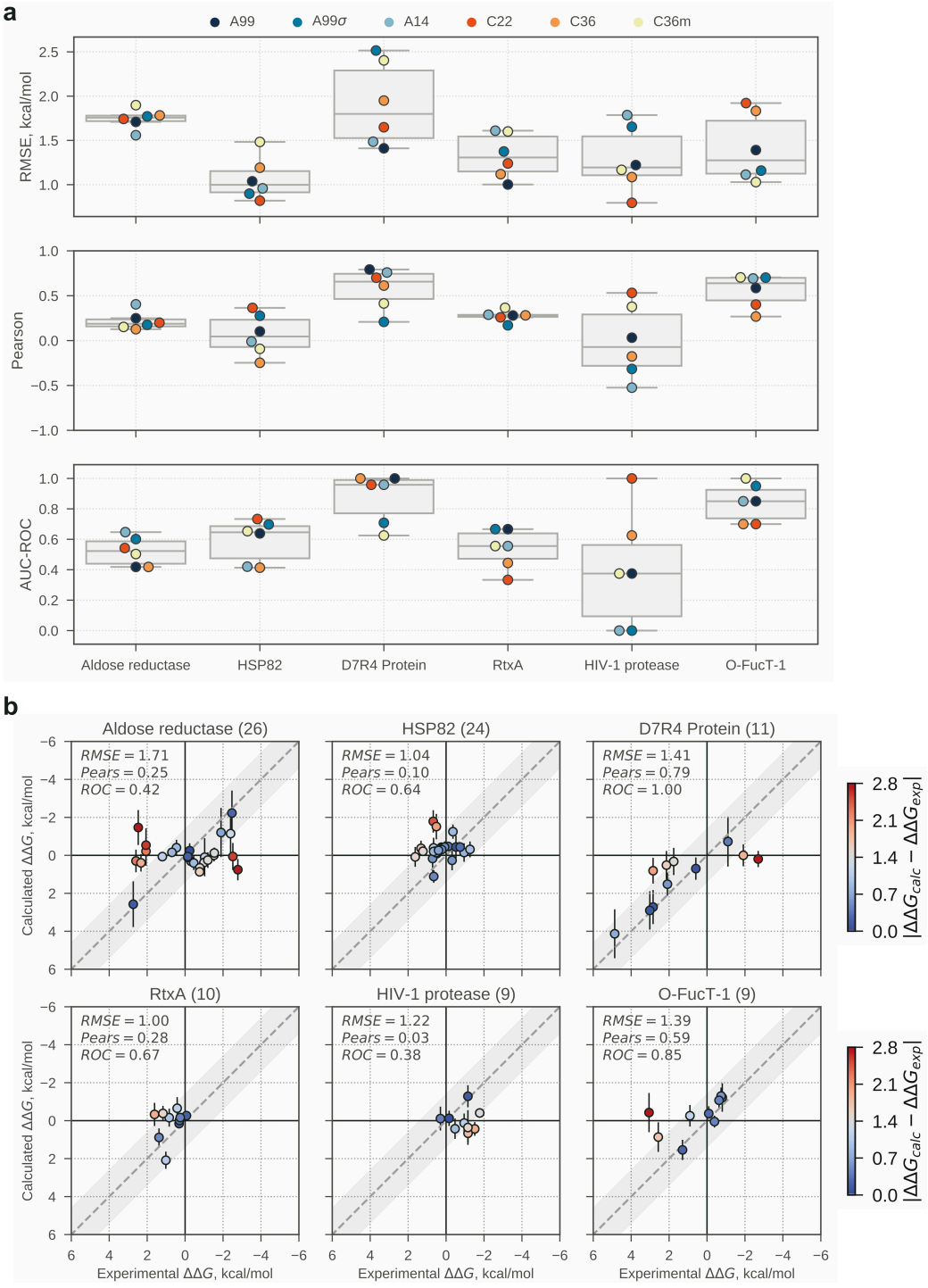


Figure S11. Precision (standard error: σ_{calc}) and accuracy (absolute error compared to experimental values) of the calculations by number of concurrent mutations. Data for all 6 force fields were pooled together. Each mark represents one ΔΔG estimate. The boxes show the first, second, and third quartiles of the data, while the whiskers are up to 1.5 times the interquartile range. Both precision and accuracy seem to deteriorate when performing multiple mutations at once.

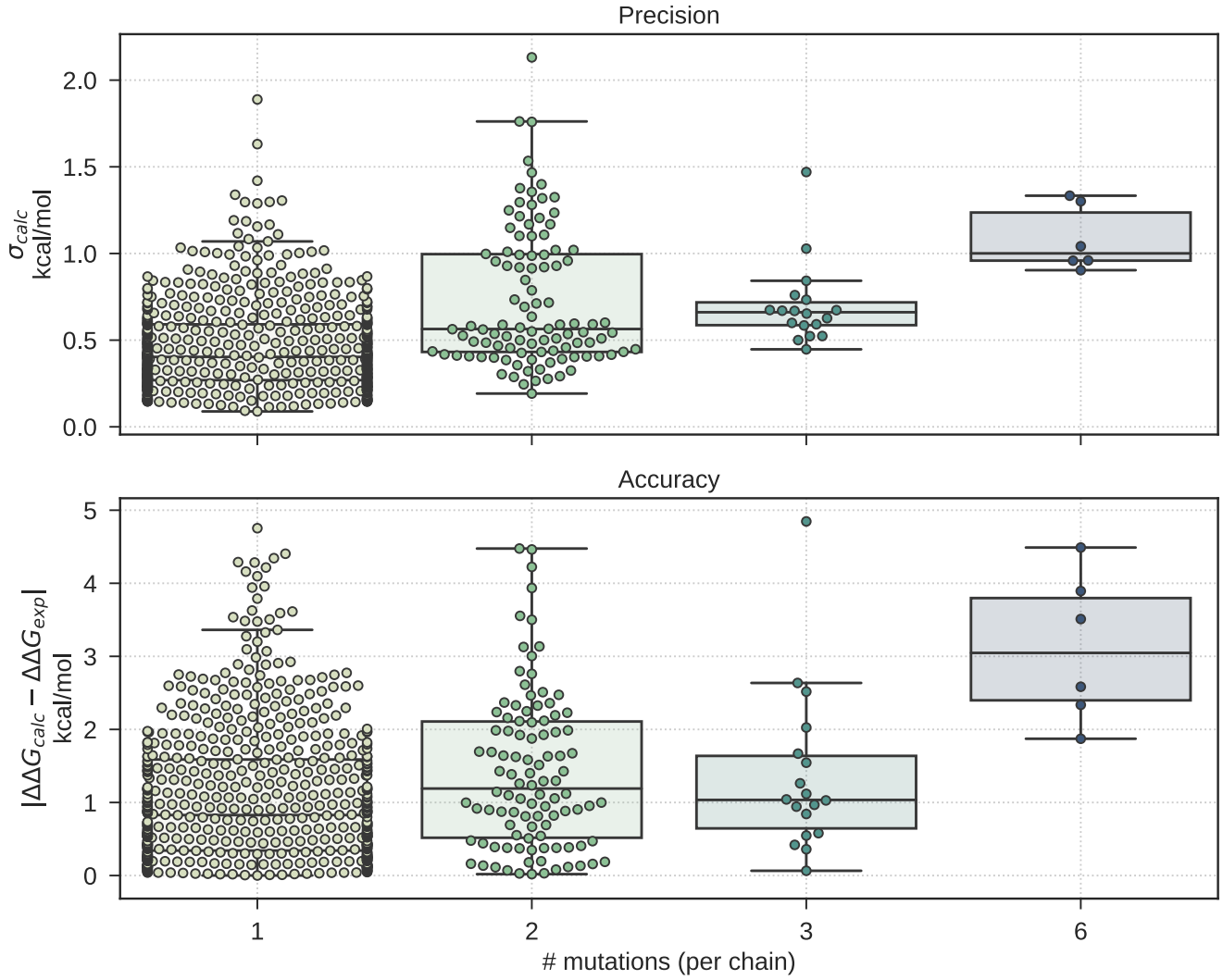


Figure S12. Precision (standard error: σ_{calc}) and accuracy (absolute error compared to experimental values) of the calculations by change in net charge of the system upon mutation. Data for all 6 force fields were pooled together. Each mark represents one ΔΔG estimate. The boxes show the first, second, and third quartiles of the data, while the whiskers are up to 1.5 times the interquartile range. While charge-changing mutations might be more challenging then charge-conserving ones overall, it is hard to discern a clear trend that would be suggestive of simulation artefacts dominating the calculation errors.

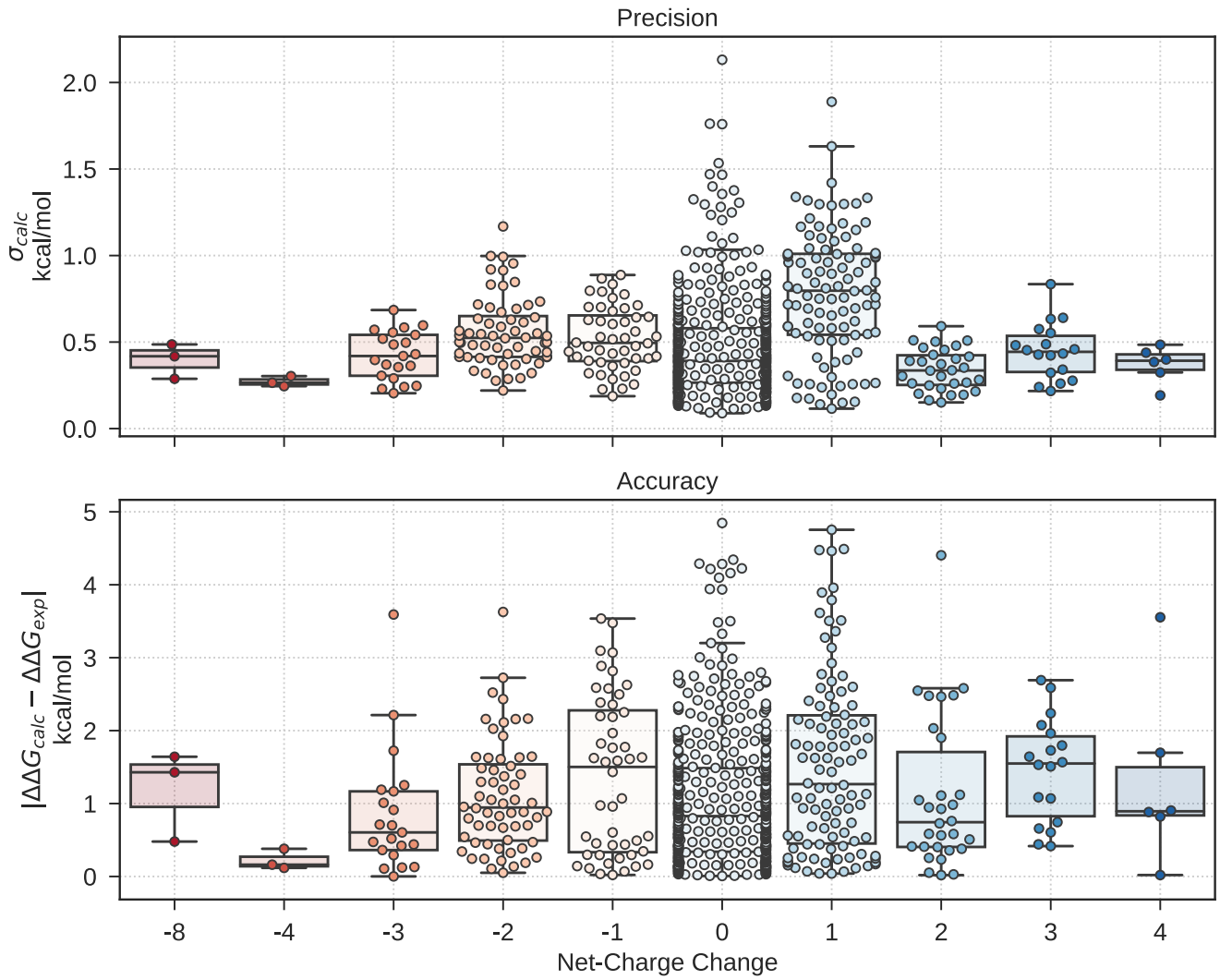
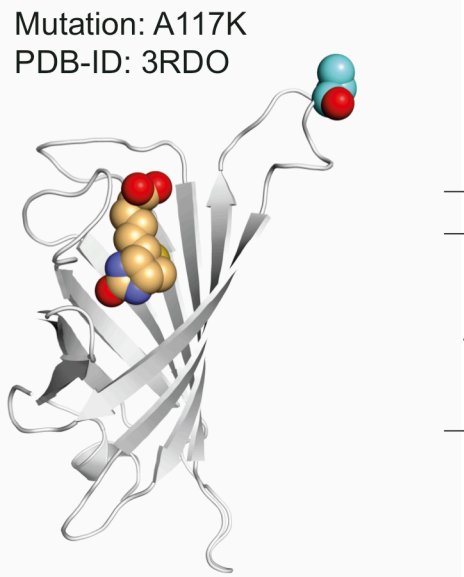


Figure S13. Effect of finite size effects on charge-changing mutation results. We used the A117K mutation in streptavidin (PDB-ID 3RDO) as a test case, due to the large finite size effects noticed when running the calculation at two different box sizes (edge of the cubic box of 6 and 12 nm). To aid convergence, we kept the protein and ligand atoms frozen during the simulations, with the exception of the residue being mutated. Despite the large size effects noticed for the ΔG calculations in the complex and apo states, these almost exactly cancelled out in the final ΔΔG values, with a barely significant difference of 0.18 ± 0.09 kcal/mol (mean and standard error). Uncertainties are one standard error obtained from ten independent calculations.



All values in kcal/mol

	6 nm	12 nm	Difference
ΔG Apo	−6.77 ± 0.05	−8.81 ± 0.04	−2.04 ± 0.06
ΔG Complex	−7.06 ± 0.04	−8.91 ± 0.05	−1.85 ± 0.06
ΔΔG	−0.29 ± 0.06	−0.11 ± 0.07	−0.18 ± 0.09

Text S1: note about root-mean-square error

The root-mean-square function (RMS) is homogenous and scales linearly with the size of the errors ($\boldsymbol{e}$), such that $c \cdot RMS(\boldsymbol{e}) = RMS(c\boldsymbol{e})$, with c being a constant:

$$RMS(\boldsymbol{e}) = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N}}$$

$$c\sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} c^2 e_i^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (ce_i)^2}{N}}$$

Where e is the "error" or "deviation" (e.g. the difference between ΔΔG values of two repeated calculations), and N the number of samples. In the context of the RMSEs shown in Figure 2 and measuring reproducibility, this means that halving the RMSE can also be interpreted as halving the errors/deviations between each repeated ΔΔG estimate: $RMS(\boldsymbol{e_1}) = \frac{RMS(\boldsymbol{e_2})}{2} = RMS\left(\frac{\boldsymbol{e_2}}{2}\right)$.

Text S2: Protein systems posing reproducibility challenges

Obtaining reproducible results (as measured by the RMSD between two repeated calculations) was found to be particularly challenging (RMSD > 1 kcal/mol) for four protein systems: aldose reductase, streptavidin, D7R4 protein, and RolR. Aldose reductase and streptavidin might have been expected to provide precision and thus reproducibility challenges. Aldose reductase presents a flexible binding site with little ordered secondary structure, where most ligands bind with an induced-fit mechanism.[57,58] For streptavidin, only 3 (out of 8) calculations involved single-mutants, 2 involved hexa-mutants also affecting the quaternary structure of the protein, and 6 involved the mutation of Ser52, a residue affecting the structure and flexibility of a loop at the top of the binding pocket.[59] Thus, in these two systems one might have easily foreseen sampling issues given the information available. On the other hand, it would have been hard to foresee the reproducibility challenges posed by RolR and D7R4 proteins without repeated calculations, as there is no clear indication that sampling could be an issue in these two systems. These two proteins have a large proportion of charge-changing mutations, which often require more sampling to converge. However, other systems that in our dataset too have a large proportion of charge-changing mutations (e.g. Epsin, ACPS, BCKADE2, Anti-tumor lectin) did not present the same reproducibility challenges.

Text S3: Rosetta sample commands

Shown are example commands using the *ref2015* scoring function. For *talaris* scoring functions, the `restore_talaris_behavior` flag was added; for the *beta_nov16* scoring function, the `corrections:beta_nov16` flag was added.

Cartesian_ddg

```
relax.linuxgccrelease -s
input.pdb -use_input_sc -constrain_relax_to_start_coords-nstruct
50 -relax:coord_constrain_sidechains -relax:ramp_constraints
false -relax:cartesian -relax:min_type
lbfgs_armijo_nonmonotone -score:weights ref2015_cart -extra_res_fa
ligand.params


cartesian_ddg.linuxgccrelease -s relaxed.pdb -ddg:mut_file
mut_list.txt -ddg:iterations 50 -optimization:default_max_cycles
200 -bbnbr 1 -relax:min_type lbfgs_armijo_nonmonotone -fa_max_dis 9.0
-score:weights ref2015_cart -ddg:dump_pdbs false -extra_res_fa
ligand.params
```

Coupled_moves

```
coupled_moves.linuxgccrelease -s input.pdb -resfile
myresfile.dat -database
~/rosetta_src_2017.36.59679_bundle/main/database/ -mute
protocols.backrub.BackrubMover -extra_res_fa ligand.params  -ex1 -ex2
-extrachi_cutoff 0 -nstruct 20 -coupled_moves:mc_kt
0.6 -coupled_moves:ntrials 1000 -coupled_moves:initial_repack
true -coupled_moves:ligand_mode true -coupled_moves:fix_backbone false
-coupled_moves:bias_sampling true -coupled_moves:bump_check
true -coupled_moves:ligand_weight 1.0 -score:weights ref2015
```

Flex_ddg

```
rosetta_scripts.linuxgccrelease    -s    input.pdb    -parser:protocol
ddG-backrub.xml        -in:file:fullatom        -ignore_zero_occupancy
false -ex1 -ex2 -extra_res_fa ligand.params -score:weights ref2015
```

A sample ddG-backrub.xml file is attached in the Supporting Information.

References

(1) 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.

(2) Liu, S.; Wang, L.; Mobley, D. L. Is Ring Breaking Feasible in Relative Binding Free Energy Calculations? *J. Chem. Inf. Model.* **2015**, *55* (4), 727–735.

(3) Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98* (10), 2309–2316.

(4) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Pmx: Automated Protein Structure and Topology Generation for Alchemical Perturbations. *J. Comput. Chem.* **2015**, *36* (5), 348–354.

(5) Marks, C.; Nowak, J.; Klostermann, S.; Georges, G.; Dunbar, J.; Shi, J.; Kelm, S.; Deane, C. M. Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve Protein Loop Prediction. *Bioinformatics* **2017**, *33* (9), 1346–1353.

(6) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33* (SUPPL. 2), W382–W388.

(7) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12* (4), 1845–1852.

(8) Olsson, M. H. M.; SØndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537.

(9) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of p K a Values. *J. Chem. Theory Comput.* **2011**, *7* (7), 2284–2295.

(10) ChemAxon. MarvinSketch. **2017**, *v17.13*.

(11) 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 Struct. Funct. Bioinforma.* **2006**, *65* (3), 712–725.

(12) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113* (26), 9004–9015.

(13) 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 Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950–1958.

(14) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(15) MacKerell, a D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(16) MacKerell, a D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields. *J. Comp. Chem.* **2004**, *25* (11), 1400–1415.

(17) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100* (9), L47–L49.

(18) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D.

Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone φ, ψ and Side-Chain χ 1 and χ 2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273.
(19) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2016**, *14* (1), 71–73.
(20) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.
(21) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.
(22) Case, D. A.; Betz, R. M.; Cerutti, D. S.; T.E. Cheatham, I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; J. Kaus, A. K.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Kollman, P. A. AMBER 16. University of California, San Francisco 2016.
(23) da Silva, A.; Vranken, W. ACPYPE - AnteChamber PYthon Parser InterfacE. *BMC Res. Notes* **2012**, *5* (1), 367.
(24) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31* (4), 671–690.
(25) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154.
(26) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52* (12), 3155–3168.
(27) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21* (2), 132–146.
(28) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641.
(29) Bayly, C. C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. a. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
(30) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11* (4), 431–439.
(31) Chandra, S. U.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5* (2), 129–145.
(32) Kolář, M.; Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8* (4), 1325–1333.
(33) Jefferys, E.; Sands, Z. A.; Shi, J.; Sansom, M. S. P.; Fowler, P. W. Alchembed: A Computational Method for Incorporating Multiple Proteins into Complex Lipid Geometries. *J. Chem. Theory Comput.* **2015**, *11* (6), 2743–2754.
(34) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.
(35) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to

Supercomputers. *SoftwareX* **2015**, *2*, 1–7.
(36) Goga, N.; Rzepiela, A. J.; de Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **2012**, *8* (10), 3637–3649.
(37) Van Gunsteren, W. F.; Berendsen, H. J. C. A Leap-Frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1* (3), 173–185.
(38) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
(39) 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* (19), 8577–8593.
(40) Páll, S.; Hess, B. A Flexible Algorithm for Calculating Pair Interactions on SIMD Architectures. *Comput. Phys. Commun.* **2013**, *184* (12), 2641–2650.
(41) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116–122.
(42) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.
(43) Crooks, G. E. Path-Ensemble Averages in Systems Driven Far from Equilibrium. *Phys. Rev. E* **2000**, *61* (3), 2361–2366.
(44) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245–268.
(45) 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.
(46) H., K. E.; Andrew, L.; David, B. Role of Conformational Sampling in Computing Mutation-induced Changes in Protein Structure and Stability. *Proteins Struct. Funct. Bioinforma.* **2010**, *79* (3), 830–838.
(47) 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.
(48) 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.
(49) 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), e1004335.
(50) 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.
(51) Yifan, S.; Michael, T.; Andrew, L.; James, T.; David, B. Structure-guided Forcefield Optimization. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (6), 1898–1909.
(52) Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19* (6), 844–858.
(53) Leaver-Fay, A.; O'Meara, M. J.; Tyka, M.; Jacak, R.; Song, Y.; Kellogg, E. H.; Thompson, J.; Davis, I. W.; Pache, R. A.; Lyskov, S.; Gray, J. J.; Kortemme, T.; Richardson, J. S.; Havranek, J. J.; Snoeyink, J.; Baker, D.; Kuhlman, B. Chapter Six - Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement. In *Methods in Protein Design*; Keating, A. E. B.

T.-M. in E., Ed.; Academic Press, 2013; Vol. 523, pp 109–143.
(54) O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput.* **2015**, *11* (2), 609–622.
(55) Myszka, D. G.; Abdiche, Y. N.; Arisaka, F.; Byron, O.; Eisenstein, E.; Hensley, P.; Thomson, J. A.; Lombardo, C. R.; Schwarz, F.; Stafford, W.; M. L. Doylej. The ABRF-MIRG'02 Study: Assembly State, Thermodynamic, and Kinetic Analysis of an Enzyme/Inhibitor Interaction. *J. Biomol. Technol.* **2003**, *14* (4), 247–269.
(56) Aldeghi, M.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Statistical Analysis on the Performance of Molecular Mechanics Poisson-Boltzmann Surface Area versus Absolute Binding Free Energy Calculations: Bromodomains as a Case Study. *J. Chem. Inf. Model.* **2017**, *57* (9), 2203–2221.
(57) Sotriffer, C. A.; Krämer, O.; Klebe, G. Probing Flexibility and "Induced-Fit" Phenomena in Aldose Reductase by Comparative Crystal Structure Analysis and Molecular Dynamics Simulations. *Proteins Struct. Funct. Genet.* **2004**, *56* (1), 52–66.
(58) Steuber, H.; Heine, A.; Podjarny, A.; Klebe, G. Merging the Binding Sites of Aldose and Aldehyde Reductase for Detection of Inhibitor Selectivity-Determining Features. *J. Mol. Biol.* **2008**, *379* (5), 991–1016.
(59) Magalhães, M. L. B.; Czekster, C. M.; Guan, R.; Malashkevich, V. N.; Almo, S. C.; Levy, M. Evolved Streptavidin Mutants Reveal Key Role of Loop Residue in High-Affinity Binding. *Protein Sci.* **2011**, *20* (7), 1145–1154.