Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches

Supporting Information

Matteo Aldeghi*[a], Vytautas Gapsys[a] and Bert L. de Groot†[a]

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

This PDF file includes:

- Methods
- Figures S1-S11
- Tables S1-S2

Other supporting materials for this manuscript include the following:

- Data S1: detailed information on the data set and numerical results for all calculations, provided as an Excel spreadsheet (XLSX).
- Data S2: input files pertaining to the molecular-dynamics-based free energy calculations and Rosetta calculations, provided as a compressed archive file (ZIP, 1.6 MB).
- Data S3: input files and *Jupyter* notebooks used for the training and testing of the machine learning model, provided as a compressed archive file (ZIP, 832 MB). This file can be downloaded from the following address: `https://doi.org/10.5281/zenodo.3350897`.

*maldegh@gwdg.de
†bgroot@gwdg.de

Contents

Methods

Dataset

A dataset containing 144 binding affinity changes ($\Delta\Delta G$) for eight TKIs due to point mutations in human Abl kinase is provided by Hauser *et al.*[1]. The authors also provide structure files of the Abl:TKI complexes. Six of these are structures resolved experimentally via X-ray crystallography (4WA9, 3UE4, 4XEY, 1OPJ, 3CS9, 3OXZ) and two were obtained via docking (referred to as DOK1 and DOK2). For calculations using Charmm force fields, glycine mutations were excluded, as it is currently not possible to interpolate between different grid-based energy correction maps (CMAPs) in Gromacs. Thus, the effective dataset size used in Charmm calculations was of 137 $\Delta\Delta G$ values rather than 144. The standard error associated with the $\Delta\Delta G$ values was taken to be 0.32 kcal/mol, based on the interlaboratory variability of the IC_{50} derived $\Delta\Delta G$ measurements.[1]. Details of the dataset can be found in Data S1.

System Setup

The structures of the Abl:TKI complexes were taken from Hauser *et al.*[1]. Apo structures were generated by discarding the ligand atoms. Crystallographic water molecules were retained. All mutant structures were generated using FoldX (v4)[2]. Protein protonation states were assigned at pH 7.4 with the protein preparation tool in HTMD (v1.12)[3], which uses PDB2PQR[4] and PROPKA v3.1[5, 6]. The protonation states of the ligands were kept as they were determined in Hauser *et al.*[1].

Proteins were modelled with the Amber99sb*-ILDN[7, 8, 9] (we abbreviate this as "A99"), Amber14sb[10] (A14), Charmm22*[11, 12, 13] (C22), and Charmm36[14] (C36) force fields. In addition, we also tested a modified version of Amber99sb*-ILDN in which the parameters for hydroxyl groups were adapted as described in Fennel *et al.*[15]; we refer to this force field as Amber99sb*-ILDN-DC (A99*dc*). The TIP3P water model[16] was used. Ligands were modelled with GAFF2 (v2.1)[17] via AmberTools 16 and CGenFF (v3.0.1)[18] via paramchem[19, 20]. In the GAFF2 models, restrained electrostatic potential (RESP)[21] charges were used. 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. Only 3 optimization steps were carried out to keep ligands' conformations close to the bound poses. ESP points were sampled according to the Merz-Kollman scheme[22, 23]. In addition, the σ-hole on halogen atoms was modelled as described by Kolář and Hobza[24]. All ligand parameters can be found in the input files in Data S2.

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.

Because FoldX does not consider the presence of ligands when mutating the protein, clashes with the ligands in the mutated complexes may occur. A clash was considered present if any protein heavy atom was within 1.5 Å of any ligand heavy atom. If one or more clashes were present, an approach similar to the one reported for *alchembed*[25] was used to resolve them: 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.

Free Energy Calculations

All simulations were carried out in Gromacs 2016[26, 27] on cluster nodes equipped with an Intel Xeon processor of Ivy Bridge (4 cores, e.g. E3-1270 v2) or Broadwell (10 cores, e.g. E5-2630 v4) architecture and a Nvidia GeForce GPU 10 series (GTX 1070, GTX 1080, or GTX 1080 Ti). Compute times for a single $\Delta\Delta G$ estimate with a representative node architecture are shown in Table 1.

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}. Equations of motion were integrated with a leap-frog integrator and a time-step of 2 fs. The temperature was coupled with the stochastic v-rescale thermostat of Bussi *et al.*[28] at the target temperature of 300 K. The pressure was controlled with the Berendsen weak coupling algorithm[29] at a target pressure of 1 bar. The particle mesh Ewald (PME) algorithm[30] 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[31] 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^{-1} ps^{-1}. All bonds were constrained with the P-LINCS algorithm[32]. For equilibration, 1 ns unrestrained MD simulations were then performed in the NPT ensemble with the Parrinello-Rahman pressure coupling algorithm[33] at 1 bar with a time constant of 2 ps. Production simulations were then performed for 3 ns. For the more expensive A99ℓ protocol, simulations of 5 ns were used.

For each free energy calculation, the above procedure for equilibrium simulations (from system setup to minimization, equilibration, and production MD) was repeated ten times on both the apo and complex states, of both wild-type and mutant Abl kinase, for each $\Delta\Delta G$ estimate. From each of these ten equilibrium simulations, 30 equally spaced frames were extracted as the starting configurations for the non-equilibrium part of the calculations, for a total of 300 non-equilibrium trajectories (in both directions, wild-type to mutant and mutant to wild-type) for each mutation. For A99ℓ, ten repeated equilibrium simulations were used for charge-conserving mutations, and twenty for charge-changing mutations; from these, a total of 400 frames for charge-conserving mutations, and 800 frames for charge-changing mutations, were extracted. The non-interacting ("dummy") atoms needed to morph the wild-type residues into mutant ones were introduced at this stage with the *pmx* package[34], 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 80 ps in length (100 ps were used for A99ℓ). The work values associated with each non-equilibrium transition were extracted using thermodynamic integration (TI)[35] and then used to estimate the free energy differences with the Bennett's Acceptance Ratio (BAR)[36, 37, 38].

Point estimates of the free energy differences (Figure S1: $\Delta G^{apo}_{WT\rightarrow MT}$ 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. 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. These uncertainties were then propagated to the final $\Delta\Delta G$ estimate to obtain the estimate of the standard error $\sigma_{\Delta\Delta G}$.

Rosetta Calculations

Binding free energy changes were calculated with Rosetta (v2017.52) using the *flex_ddg* protocol[39]. These calculations were carried out on cluster nodes equipped with an Intel Xeon processor of

Broadwell architecture (E5-2630 v4), using one CPU core per $\Delta\Delta G$ calculation. Ligand parameters were obtained with the `molfile_to_params.py` script provided with Rosetta. The REF2015 and beta_nov2016 (referred to as βNOV16) scoring functions were used. The final $\Delta\Delta G$ estimates were the average values of the generalized additive model obtained from 35 iterations of the protocol[39]. The command lines used for the Rosetta calculations and the input files can be found in Data S2.

Machine Learning

The machine learning (ML) model was built in *python* using the `ExtraTreesRegressor` class in the *scikit-learn* library[40]. This model uses ensembles of randomized decision trees[41] in a similar fashion to random forest. The input files and the code (as *Jupyter* notebooks) used to train and test the ML models are provided in Data S3. All computations pertaining the ML results were performed on a desktop machine equipped with an Intel Xeon processor of Broadwell architecture (E5-1630 v4). While here we describe the general procedure used to prepare the training dataset, to calculate and select features, and to test the models, the details needed to reproduce all results are found in the notebooks provided as part of the Supporting Information.

Training Dataset

A dataset for training and validation, containing 484 entries, was created from the Platinum database[42]. From the whole database we excluded entries if they: (i) were not point-mutations; (ii) referred to PDB-IDs containing "broken ligand structures" as described in the database; (iii) referred to PDB-IDs in which the ligand was poorly resolved. The latter criterion was applied by matching PDB-IDs and ligand IDs to the Twilight database (`www.ruppweb.org/twilight`; v11-01-2018) and excluding entries if the real-space correlation coefficient (RSCC) was below 0.8.[43, 44] Additional manual curation was also performed. Details of this and how overall the training/validation dataset was created are provided in Data S3 (`0_filter_platinum_database.ipynb`).

Features and Feature Selection

A total of 128 features, which we thought being potentially informative for the prediction of affinity changes upon protein mutation, were calculated. More specifically, 18 ligand properties (e.g. molecular weight, calculated logP, number of rotatable bonds) were calculated with *RDKit* (v2018.09.1; `www.rdkit.org`). 21 properties describing the mutation environment (e.g. distribution of ligand and protein atoms around the mutation site, number of polar/apolar/charged residues in the binding pocket) were calculated with *Biopython* (v1.73; `www.biopython.org`). 13 features describing the change in the amino acid chemical nature were calculated using precomputed properties for each amino acid (e.g. change in side-chain volume, hydropathy, number of hydrogen bond donors). Among these there was also the change in folding free energy upon mutation as predicted by FoldX v4[2]. 6 features describing protein-ligand interactions (hydrogen bonds, hydrophobic contacts, salt bridges, π-stacking, cation-π interactions, and halogen bonds) were calculated with the Protein-Ligand Interaction Profiler (PLIP)[45]. The Vina binding score, along with 59 Vina features were calculated with AutoDock Vina[46] via scripts that are part of DeltaVina[47]. The latter tool, in conjunction with the molecular surface calculation library MSMS[48], was also used to calculate 10 pharmacophore-based solvent-accessible surface area (SASA) features[47]. Finally, the wild-type binding affinity was included. See Data S3 (`1_extract_features.ipynb`) for details on all features.

Feature selection was performed with a greedy algorithm using the *mlxtend* library[49]. We allowed the selection of any number of features, up to 40, which minimized the mean-squared error of 10-fold cross-validation on the Platinum dataset. The folds were built such that each of them would contain a unique set of proteins not present in the other folds. The same feature selection procedure was also adopted for the 8-fold nested cross-validation on the TKI dataset. See Data S3 (`2_train_test.ipynb`) for details on the feature selection procedure, and the training and testing of the ML model.

Data Analysis

The accuracy of the calculations was evaluated using three performance measures: the root-mean-square error (RMSE), the Pearson correlation (r), and the area under the precision-recall curve (AUPRC). The uncertainty in these measures was evaluated by bootstrap. Pairs of experimental and calculated $\Delta\Delta G$ values were resampled with replacement 10^5 times. For each bootstrap sample, RMSE, R, and AUPRC were calculated. From these 10^5 bootstrap measures, the 2.5 and 97.5 percentiles were taken as the lower and upper bounds of the 95% confidence interval.

A bootstrap procedure was also used to obtain p-values for the differences between approaches. In this case, triplets of $\Delta\Delta G$ values were resampled with replacement together 10^5 times: $\Delta\Delta G$ values from experiment and from the two approaches to be compared. At each bootstrap iteration, the difference in the performance measure of interest (e.g. RMSE) between the two computational approaches to be compared was stored. 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. Data analysis was performed in *python* using the *numpy*[50], *scipy*[51], *pandas*[52], *scikit-learn*[40], *matplotlib*[53], and *seaborn*[54] libraries.

Figures

Figure S1: Thermodynamic cycle used in the MD-based free energy calculations.

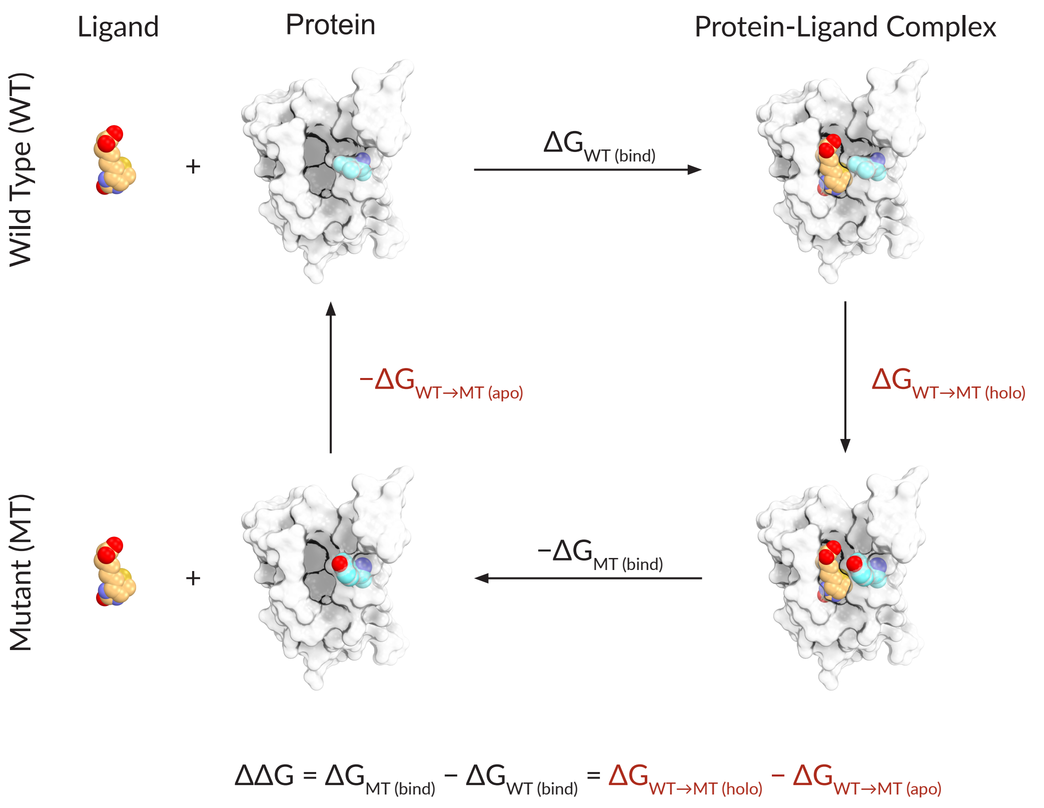

Ligand
Protein
Protein-Ligand Complex
Wild Type (WT)
Mutant (MT)
+
+
$\Delta G_{\text{WT (bind)}}$
$-\Delta G_{\text{WT}\rightarrow\text{MT (apo)}}$
$\Delta G_{\text{WT}\rightarrow\text{MT (holo)}}$
$-\Delta G_{\text{MT (bind)}}$
$\Delta\Delta G = \Delta G_{\text{MT (bind)}} - \Delta G_{\text{WT (bind)}} = \Delta G_{\text{WT}\rightarrow\text{MT (holo)}} - \Delta G_{\text{WT}\rightarrow\text{MT (apo)}}$

Figure S2: Accuracy of the $\Delta\Delta G$ estimates for all force fields tested with the MD-based free energy calculations. (a) Scatter plots of experimental versus calculated $\Delta\Delta G$ values. The identity line is shown as a dashed gray line. The four quadrants indicate the location of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the definition of resistant and susceptible mutations used[1]. Each $\Delta\Delta G$ estimate is color-coded according to its absolute error with respect to the experimental $\Delta\Delta G$ value. (b) Summary of the performance of the $\Delta\Delta G$ estimates across approaches in terms of RMSE, Pearson correlation, and AUPRC (point estimates and the 95% CIs are shown). Differences at three levels of significance are reported using labels within the chart. Based on these results, and on this dataset, these force fields qualitatively perform in the following order: $A99 = A99\ell > A14 > A99dc > C22 > C36$.

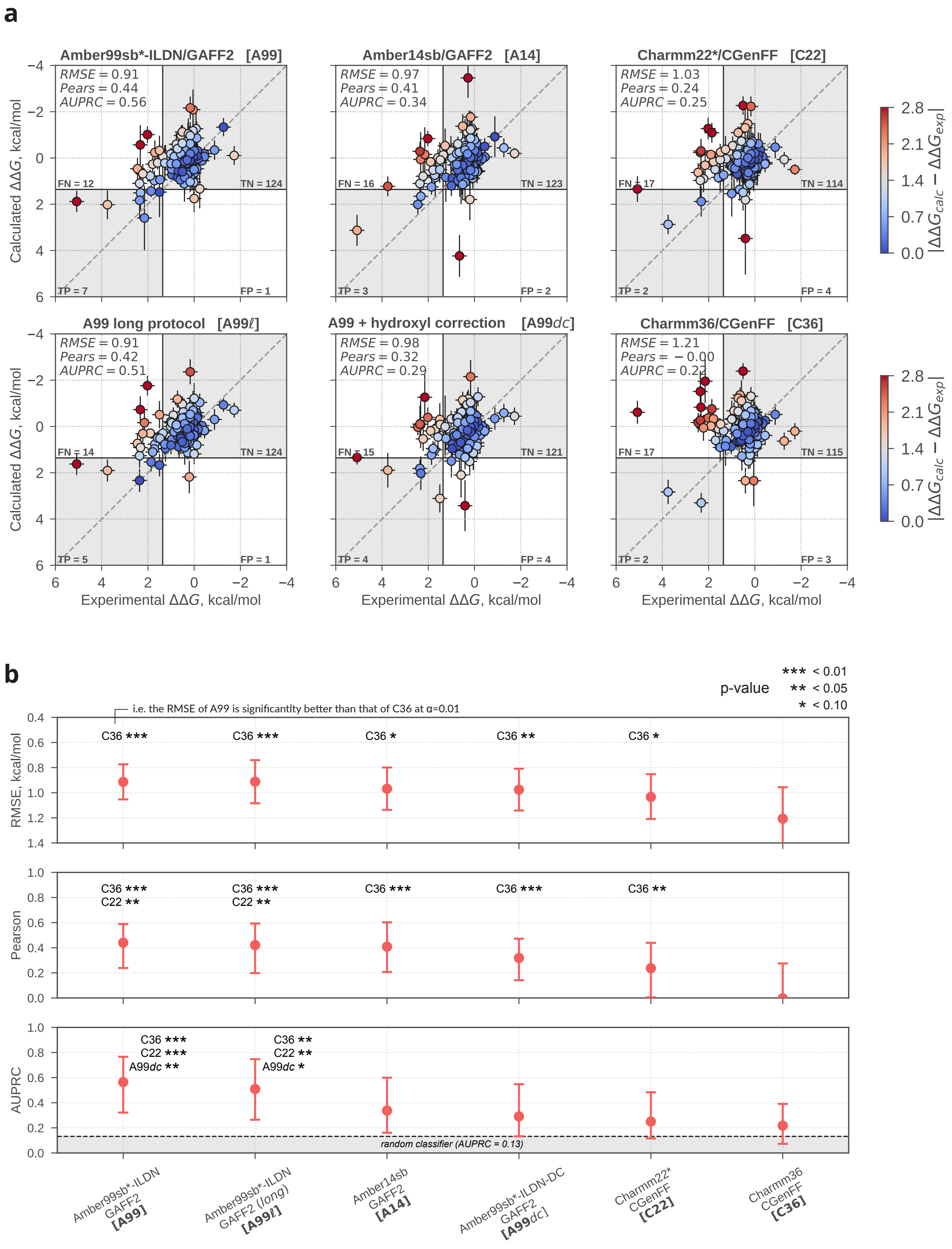


Figure S3: Effect of halogen bond modeling on Charmm calculations. Shown are scatter plots of experimental versus calculated $\Delta\Delta G$ values for the estimates obtained with the Charmm22*/CGenFF force field (C22, left), and for the same force field when modelling halogen bonds with an additional σ-hole particle (C22σ, middle) as described by Gutiérrez *et al.*[55]. As only bosutinib, dasatinib, and gefitinib contain halogen-bonding atoms (Cl, Br, I), the $\Delta\Delta G$ estimates concerning these inhibitors only are plotted. The identity line is shown as a dashed gray line. Each $\Delta\Delta G$ estimate is color-coded according to its absolute error with respect to the experimental $\Delta\Delta G$ value. The performance of the estimates in terms of RMSE, Pearson correlation, and AUPRC (point estimates and the 95% bootstrapped confidence intervals) are shown on each plot. The scatter plot on the right shows the agreement between C22 and C22σ estimates. Overall, the addition of σ-hole particles to model halogen bonding did not significantly improve the results on this dataset. "RMSE": root mean square error; "AUPRC": area under the precision-recall curve; "AUROC": area under the receiver operating characteristic curve.

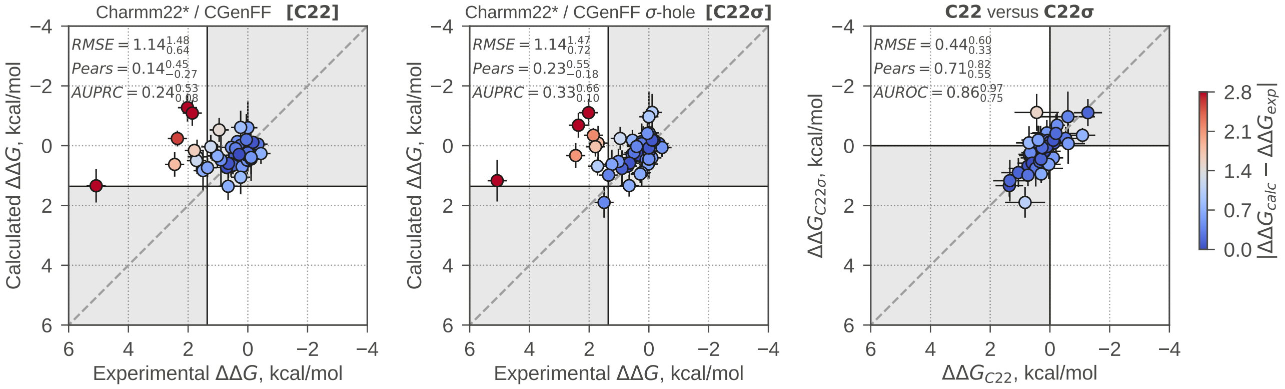
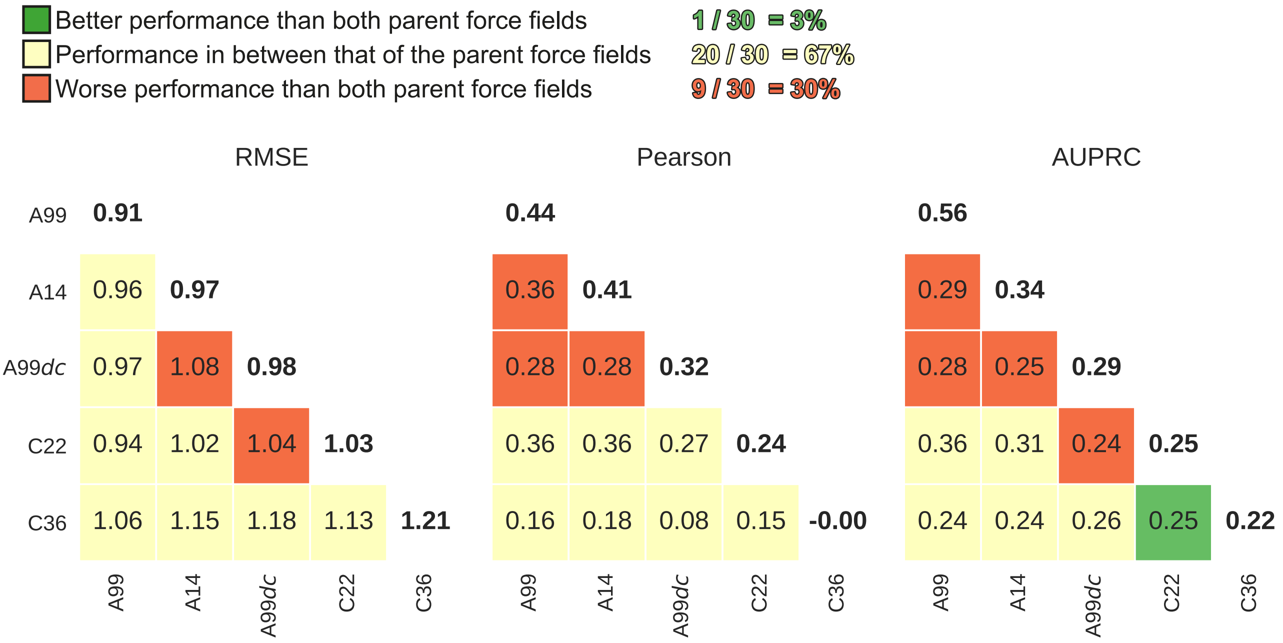


Figure S4: Results obtained with a consensus force field approach in which results from pairs of force fields are averaged while considering half of the simulated time for each of the two[56]. On the diagonal of the matrices are the performances achieved by the parent force fields using all simulation data available. The off-diagonal elements show the performance of the consensus results obtained by averaging the results of the parent force fields (abbreviations as defined in Table S1) when using half of the simulation data from each parent (i.e. the first five equilibrium simulation repeats and the associated non-equilibrium trajectories). Cells are color-coded depending whether performance of the consensus approach was better, in between, or worse than the performance of the two parent force fields.



🟩 Better performance than both parent force fields 1 / 30 = 3%

🟨 Performance in between that of the parent force fields 20 / 30 = 67%

🟥 Worse performance than both parent force fields 9 / 30 = 30%

RMSE

	A99	A14	A99*dc*	C22	C36
A99	**0.91**				
A14	🟨 0.96	**0.97**			
A99*dc*	🟨 0.97	🟥 1.08	**0.98**		
C22	🟨 0.94	🟨 1.02	🟥 1.04	**1.03**	
C36	🟨 1.06	🟨 1.15	🟨 1.18	🟨 1.13	**1.21**

Pearson

	A99	A14	A99*dc*	C22	C36
A99	**0.44**				
A14	🟥 0.36	**0.41**			
A99*dc*	🟥 0.28	🟥 0.28	**0.32**		
C22	🟨 0.36	🟨 0.36	🟨 0.27	**0.24**	
C36	🟨 0.16	🟨 0.18	🟨 0.08	🟨 0.15	**-0.00**

AUPRC

	A99	A14	A99*dc*	C22	C36
A99	**0.56**				
A14	🟥 0.29	**0.34**			
A99*dc*	🟥 0.28	🟥 0.25	**0.29**		
C22	🟨 0.36	🟨 0.31	🟥 0.24	**0.25**	
C36	🟨 0.24	🟨 0.24	🟨 0.26	🟩 0.25	**0.22**

Figure S5: Performances of the methods by change in net charge upon mutation. (a) Scatter plots of experimental versus calculated $\Delta\Delta G$ values. The identity is shown as a dashed gray line. Each $\Delta\Delta G$ estimate is color-coded according to charge change associated with the mutation. Error bars are omitted for clarity. (b) Distribution of absolute errors ($|\Delta\Delta G_{calc} - \Delta\Delta G_{exp}|$) by net charge change and approach used. (c) RMSE for the $\Delta\Delta G$ estimates involving charge-conserving and charge-changing mutations (point estimates from the original samples and 95% bootstrapped confidence intervals are shown). Overall, no significant dependence of the results on the net charge change of the system was observed.

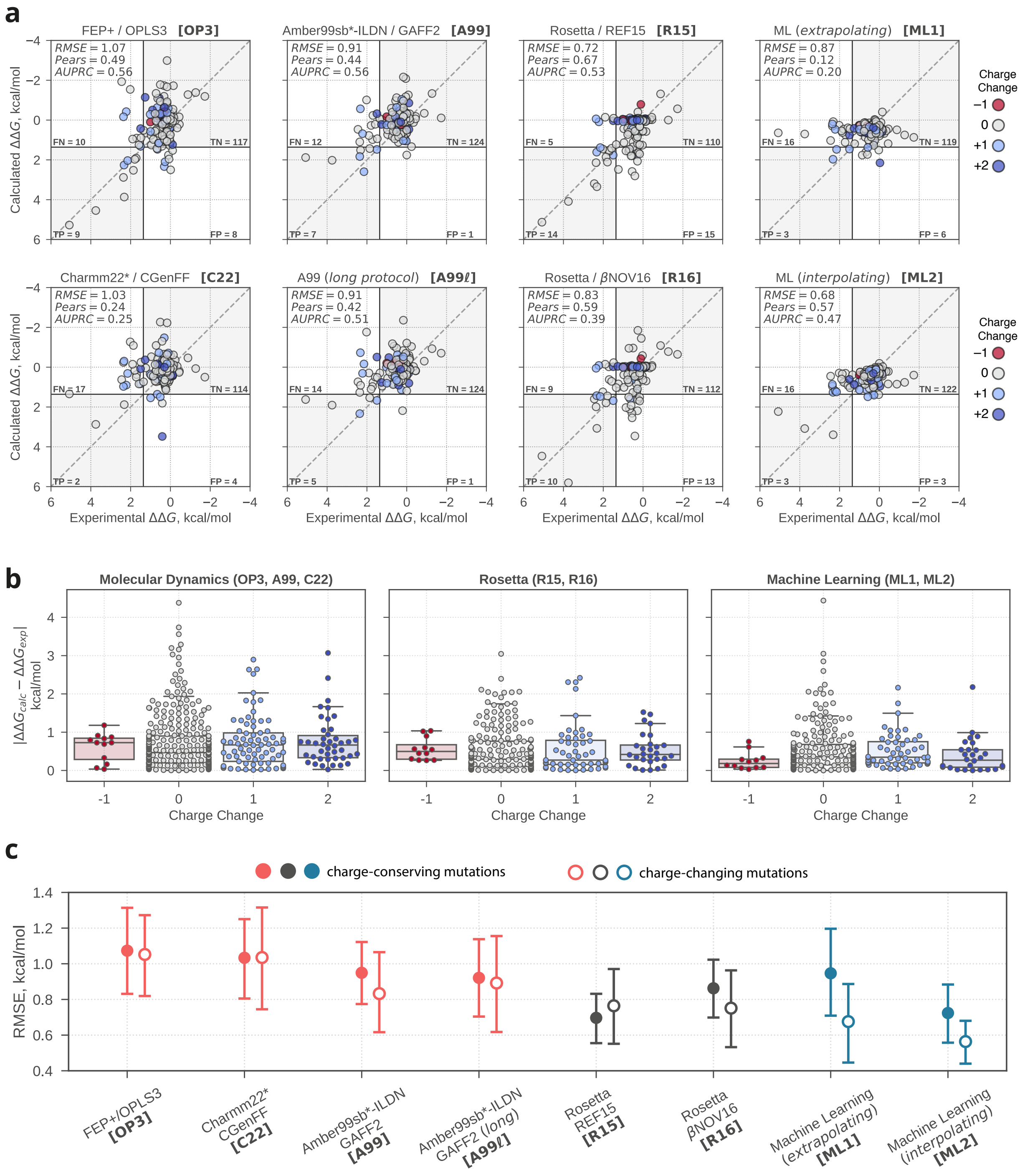
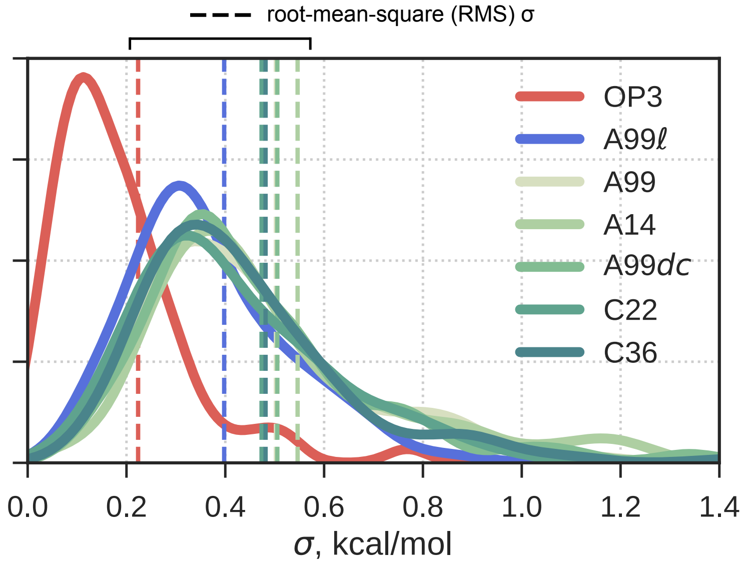


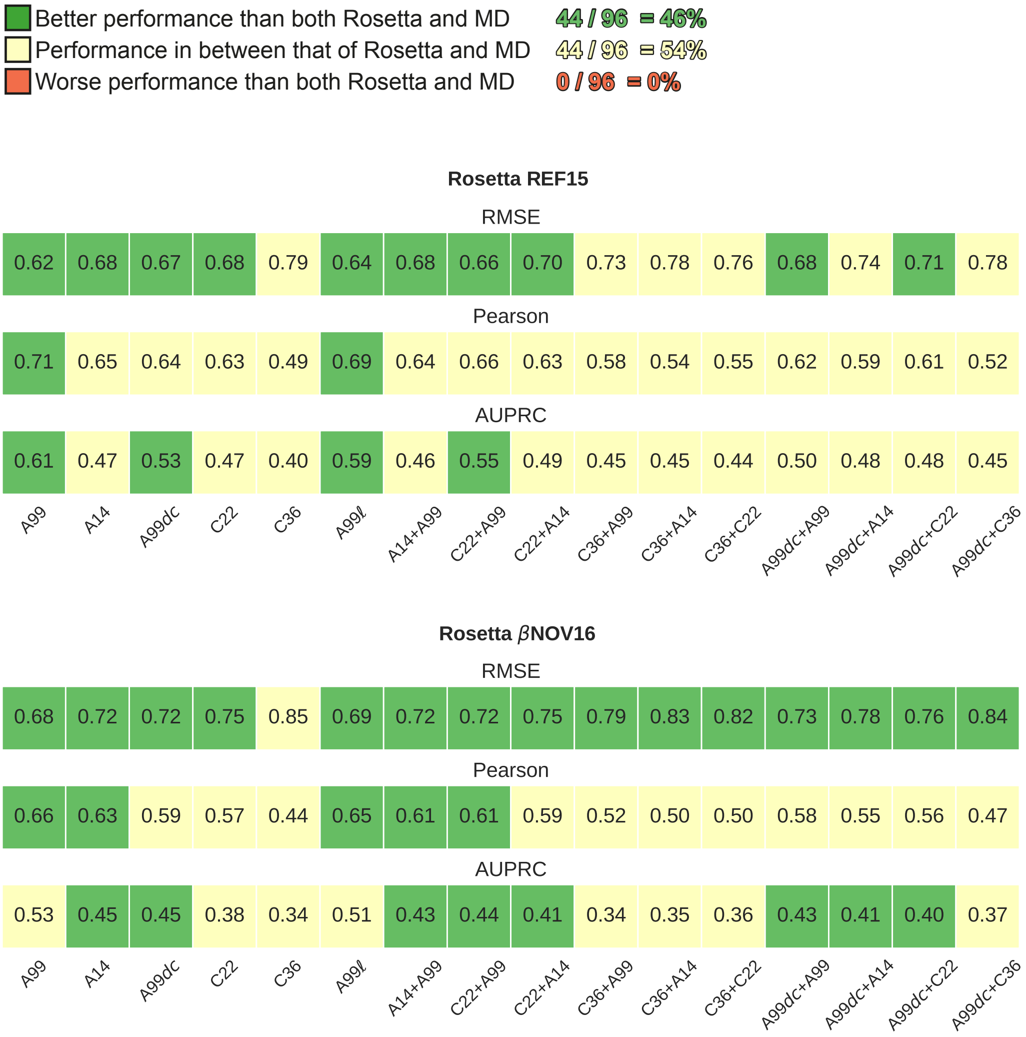
Figure S6: Uncertainty of the MD-based free energy calculations. Shown is the Gaussian kernel density estimate of the standard errors (σ) for the 144 $\Delta\Delta G$ values obtained with the different force fields and protocols tested. Vertical dashed lines mark the root-mean-square of the σ values (numerical results reported in the table at the bottom). The data for OP3 are from the three repeated calculations in Hauser *et al.*[1]. A99, A14, A99*dc*, C22, and C36 (shades of green) return similar distributions of σ values (RMSσ between 0.47 and 0.55 kcal/mol). The more expensive A99ℓ protocol (blue) returned slightly more precise estimates (RMSσ = 0.40 kcal/mol), whereas OP3 (red) achieved considerably higher precision (RMSσ = 0.22 kcal/mol). Comparing the overall accuracy (RMSE) of the $\Delta\Delta G$ estimates to their overall precision (RMSσ), one notes that these tend to be more precise than they are accurate, roughly by a factor of two for the non-equilibrium calculations (A99, A14, A99*dc*, C22, C36) and by a factor of five for the equilibrium OP3 calculations.



	OP3	A99ℓ	A99	A14	A99*dc*	C22	C36
RMSE[a]	1.07	0.91	0.91	0.97	0.98	1.03	1.21
RMSσ[a]	0.22	0.40	0.50	0.55	0.51	0.47	0.48
RMSE/RMSσ	4.77	2.29	1.82	1.77	1.93	2.18	2.51

[a] values in kcal/mol

Figure S7: Results obtained with a combined Rosetta and MD consensus approach[56]. Each cell shows the performance of the consensus results obtained by averaging the results from Rosetta (REF15 scoring function at the top, βNOV16 function at the bottom) and the free energy calculations (abbreviations as defined in Table S1; the use of two abbreviations, e.g. "A14+C22", denotes a consensus force field result). Cells are color-coded depending on whether the performance of the consensus approach was better, in between, or worse than the performance of Rosetta or MD alone.



🟩 Better performance than both Rosetta and MD 44 / 96 = 46%
□ Performance in between that of Rosetta and MD 44 / 96 = 54%
🟥 Worse performance than both Rosetta and MD 0 / 96 = 0%

Rosetta REF15

	A99	A14	A99dc	C22	C36	A99ℓ	A14+A99	C22+A99	C22+A14	C36+A99	C36+A14	C36+C22	A99dc+A99	A99dc+A14	A99dc+C22	A99dc+C36
RMSE	0.62	0.68	0.67	0.68	0.79	0.64	0.68	0.66	0.70	0.73	0.78	0.76	0.68	0.74	0.71	0.78
Pearson	0.71	0.65	0.64	0.63	0.49	0.69	0.64	0.66	0.63	0.58	0.54	0.55	0.62	0.59	0.61	0.52
AUPRC	0.61	0.47	0.53	0.47	0.40	0.59	0.46	0.55	0.49	0.45	0.45	0.44	0.50	0.48	0.48	0.45

Rosetta βNOV16

	A99	A14	A99dc	C22	C36	A99ℓ	A14+A99	C22+A99	C22+A14	C36+A99	C36+A14	C36+C22	A99dc+A99	A99dc+A14	A99dc+C22	A99dc+C36
RMSE	0.68	0.72	0.72	0.75	0.85	0.69	0.72	0.72	0.75	0.79	0.83	0.82	0.73	0.78	0.76	0.84
Pearson	0.66	0.63	0.59	0.57	0.44	0.65	0.61	0.61	0.59	0.52	0.50	0.50	0.58	0.55	0.56	0.47
AUPRC	0.53	0.45	0.45	0.38	0.34	0.51	0.43	0.44	0.41	0.34	0.35	0.36	0.43	0.41	0.40	0.37

Figure S8: Scatter plot of $\Delta\Delta G$ estimates obtained with a consensus score by averaging the results of R15 and A99. Shown are the experimental versus calculated $\Delta\Delta G$ values. The identity is shown as a dashed gray line. The four quadrants indicate the location of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the definition of resistant and susceptible mutations used[1] and the threshold $\Delta\Delta G_{calc} > 1.36$ kcal/mol. Each $\Delta\Delta G$ estimate is color-coded according to its absolute error with respect to the experimental $\Delta\Delta G$ value. The point estimate and 95% confidence interval (x_{lower}^{upper}) for the performance measures used (RMSE, Pearson correlation, and AUPRC) are shown.

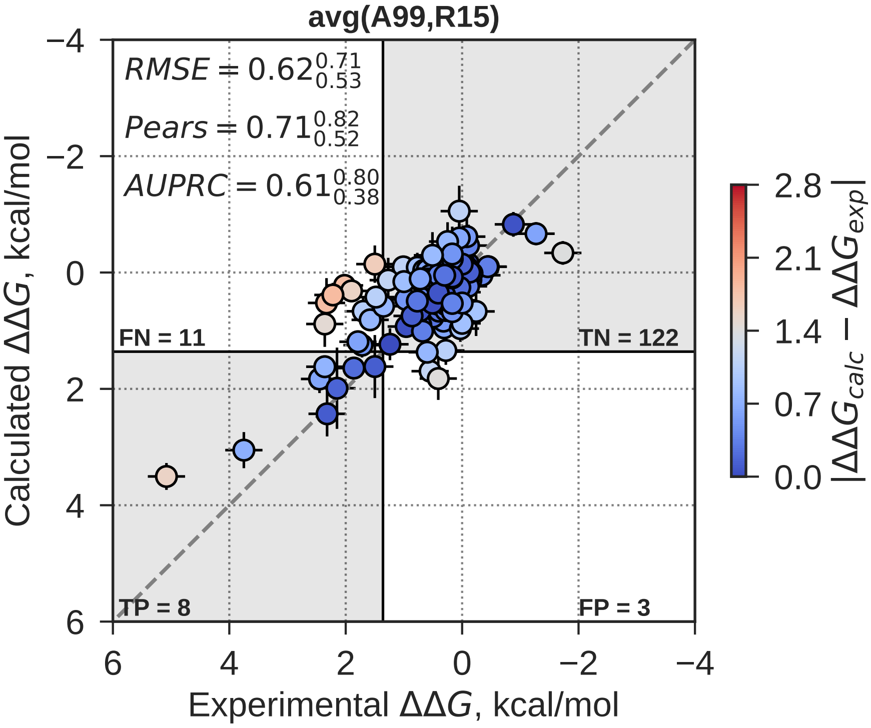


Figure S9: Scatter plot of $\Delta\Delta G$ estimates obtained with mCSM-Lig[57]. Shown are the experimental versus calculated $\Delta\Delta G$ values. The identity is shown as a dashed gray line. The four quadrants indicate the location of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the definition of resistant and susceptible mutations used[1] and the threshold $\Delta\Delta G_{calc} > 1.36$ kcal/mol. Each $\Delta\Delta G$ estimate is color-coded according to its absolute error with respect to the experimental $\Delta\Delta G$ value. The point estimate and 95% confidence interval (x^{upper}_{lower}) for the performance measures used (RMSE, Pearson correlation, and AUPRC) are shown.

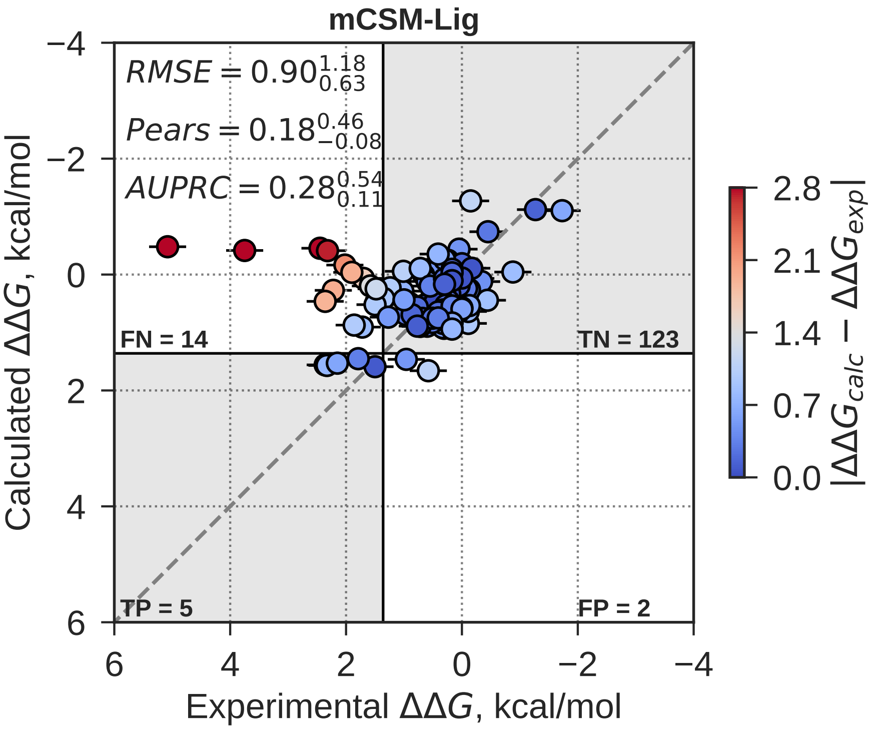


Figure S10: Precision-recall curves in order of decreasing area under the curve (AUPRC). Abbreviations used are as defined in Table 1 and S1; "max(A99,R15)" refers to the consensus approach in which, for each mutation, the most positive $\Delta\Delta$G estimate among A99 and R15 was selected, and "avg(A99,R15)" to the consensus approach in which the $\Delta\Delta$G estimates of A99 and R15 were averaged. The expected curve for a random estimator is shown as a dashed black line (baseline with AUPRC of 0.13). The precision and recall when considering the threshold of $\Delta\Delta G_{calc} > 1.36$ kcal/mol is marked by a purple circle on the curves. The circle in pink indicates the precision and recall when classifying the largest 15% $\Delta\Delta G_{calc}$ as resistant. The exact numerical values for precision and recall under these conditions are shown in Table S2.

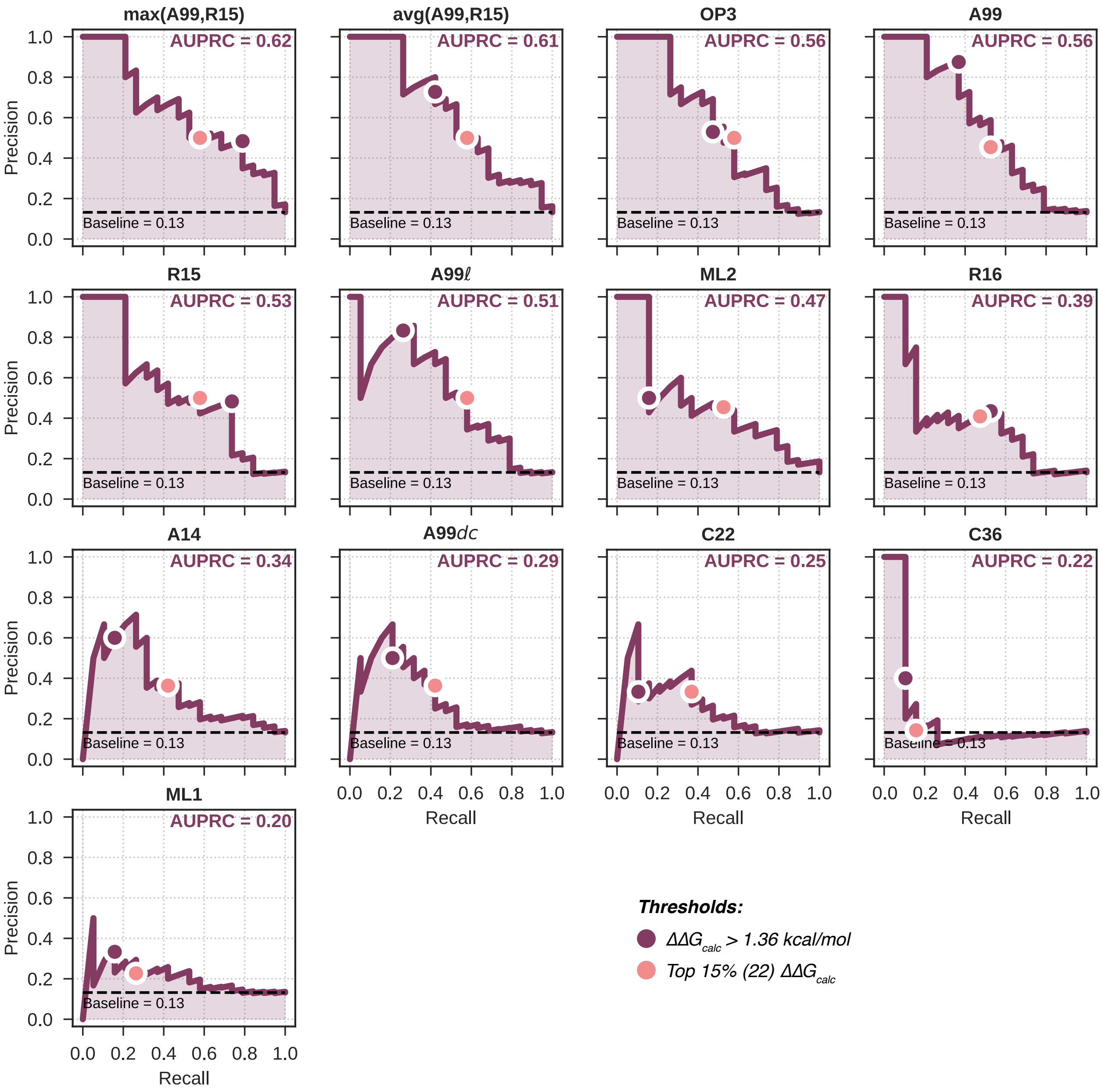
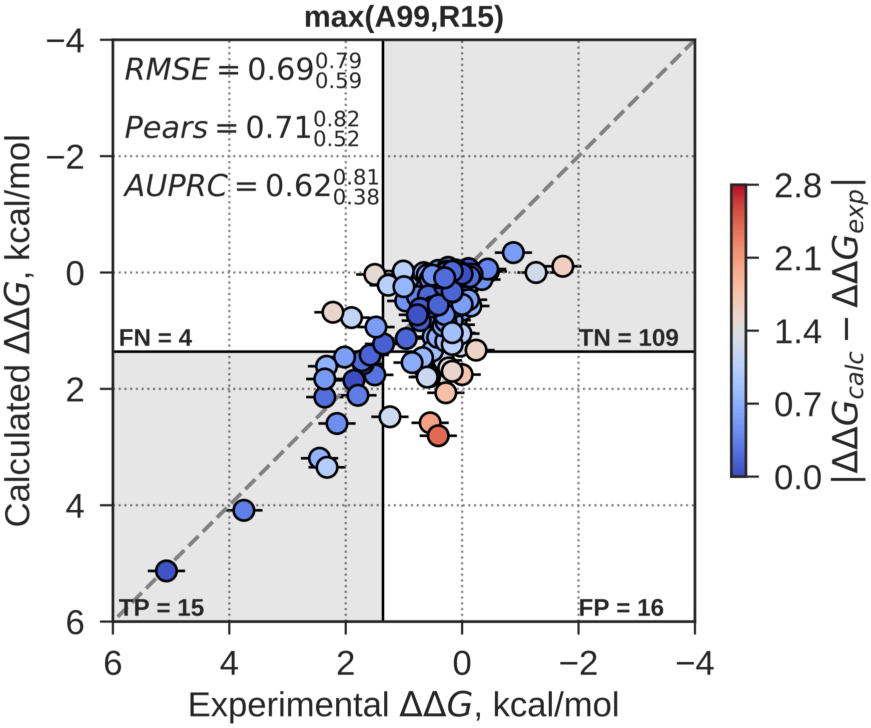


Figure S11: Scatter plot of $\Delta\Delta G$ estimates obtained with a consensus score by taking the more positive $\Delta\Delta G$ value among those of A99 and R15. Shown are the experimental versus calculated $\Delta\Delta G$ values. The identity is shown as a dashed gray line. The four quadrants indicate the location of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the definition of resistant and susceptible mutations used[1] and the threshold $\Delta\Delta G_{calc} > 1.36$ kcal/mol. Each $\Delta\Delta G$ estimate is color-coded according to its absolute error with respect to the experimental $\Delta\Delta G$ value. The point estimate and 95% confidence interval (x_{lower}^{upper}) for the performance measures used (RMSE, Pearson correlation, and AUPRC) are shown.



Tables

Table S1: Summary of the MD-based free energy calculations. For the performance measures, the point estimates and their 95% confidence intervals are shown. “Abbr.”: abbreviation; “RMSE”: root mean square error; “AUPRC”: area under the precision-recall curve.

Abbr.	**Force Field**	**RMSE (kcal/mol)**	**Pearson**	**AUPRC**
A99	Amber99sb*-ILDN[7, 8, 9] and GAFF[17] v2.1	$0.91^{1.05}_{0.77}$	$0.44^{0.59}_{0.24}$	$0.56^{0.77}_{0.32}$
A99ℓ	Amber99sb*-ILDN[7, 8, 9] and GAFF[17] v2.1	$0.91^{1.08}_{0.74}$	$0.42^{0.59}_{0.20}$	$0.51^{0.75}_{0.26}$
A14	Amber14sb[10] and GAFF[17] v2.1	$0.97^{1.14}_{0.80}$	$0.41^{0.60}_{0.21}$	$0.34^{0.60}_{0.16}$
A99*dc*	Amber99sb*-ILDN-DC[7, 8, 9, 15] and GAFF[17] v2.1	$0.98^{1.14}_{0.81}$	$0.32^{0.47}_{0.14}$	$0.29^{0.55}_{0.13}$
C22	Charmm22*[11, 12, 13] and CGenFF[18] v3.0.1	$1.03^{1.21}_{0.85}$	$0.24^{0.44}_{0.01}$	$0.25^{0.48}_{0.11}$
C36	Charmm36[14] and CGenFF[18] v3.0.1	$1.21^{1.46}_{0.96}$	$-0.00^{0.28}_{-0.28}$	$0.22^{0.39}_{0.07}$

Table S2: Binary classification performance of the methods tested. The threshold $\Delta\Delta G_{calc} > 1.36$ kcal/mol was used to classify mutations as resistant. In addition, the recall and precision of the results when classifying the largest 15% $\Delta\Delta G_{calc}$ as resistant are shown; this shows what the recall and precision are when a fixed number (22) of top-scoring mutations (i.e. those predicted to be most resistant, with largest $\Delta\Delta G_{calc}$) is selected. Abbreviations for the methods tested are used as defined in Tables 1 and S1.

	$\Delta\Delta G_{calc} > 1.36$ kcal/mol					**Top 15%**	
Approach	**Recall**	**Precision**	**Balanced Accuracy**	**F1 Score**	**Matthews Correlation Coefficient**	**Recall**	**Precision**
OP3	0.47	0.53	0.70	0.50	0.43	0.58	0.50
A99	0.37	0.88	0.68	0.52	0.53	0.53	0.45
A99ℓ	0.26	0.83	0.63	0.40	0.43	0.58	0.50
A14	0.16	0.60	0.57	0.25	0.26	0.42	0.46
A99*dc*	0.21	0.50	0.59	0.30	0.26	0.42	0.36
C22	0.11	0.33	0.54	0.16	0.12	0.37	0.33
C36	0.11	0.40	0.54	0.17	0.15	0.16	0.14
R15	0.74	0.48	0.81	0.58	0.52	0.58	0.50
R16	0.53	0.43	0.71	0.48	0.39	0.47	0.41
ML1	0.16	0.33	0.55	0.21	0.15	0.26	0.23
ML2	0.16	0.50	0.57	0.24	0.23	0.53	0.45
avg(A99,R15)	0.42	0.73	0.70	0.53	0.51	0.58	0.50
max(A99,R15)	0.79	0.48	0.83	0.60	0.54	0.58	0.50

References

[1] 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.

[2] 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* (W1), W382–W388.

[3] 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.

[4] Klebe, G.; Li, H.; Jensen, J. H.; Nielsen, J. E.; Czodrowski, P.; Dolinsky, T. J.; Baker, N. A. PDB2PQR: Expanding and Upgrading Automated Preparation of Biomolecular Structures for Molecular Simulations. *Nucleic Acids Res.* **2007**, *35* (W1), W522–W525.

[5] Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7* (2), 525–537.

[6] 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 pKa Values. *J. Chem. Theory Comput.* **2011**, *7* (7), 2284–2295.

[7] Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **2006**, *65* (3), 712–725.

[8] 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.

[9] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* **2010**, *78* (8), 1950–1958.

[10] 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.

[11] Mackerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher, W.; Roux, B.; Schlenkrich, M.; Smith, J.; 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.

[12] Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulation. *J. Comput. Chem.* **2004**, *25* (11), 1400–1415.

[13] 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.

[14] 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 phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273.

[15] Fennell, C. J.; Wymer, K. L.; Mobley, D. L. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and its Role in Small Molecule Hydration. *J. Phys. Chem. B* **2014**, *118* (24), 6438–6446.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11* (4), 431–439.

[23] Chandra, S. U.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5* (2), 129–145.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126* (1), 014101.

[29] 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.

[30] 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.

[31] Páll, S.; Hess, B. A Flexible Algorithm for Calculating Pair Interactions on SIMD Architectures. *Comput. Phys. Commun.* **2013**, *184* (12), 2641–2650.

[32] Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (1), 116–122.

[33] Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.

[34] 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.

[35] Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3* (5), 300–313.

[36] Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245–268.

[37] Crooks, G. E. Path-Ensemble Averages in Systems Driven Far From Equilibrium. *Phys. Rev. E* **2000**, *61* (3), 2361–2366.

[38] 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.

[39] 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.

[40] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[41] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63* (1), 3–42.

[42] Pires, D. E.; 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.

[43] Pozharski, E.; Weichenberger, C. X.; Rupp, B. Techniques, Tools and Best Practices for Ligand Electron-Density Analysis and Results from their Application to Deposited Crystal Structures. *Acta Crystallogr. D* **2013**, *69* (2), 150–167.

[44] Weichenberger, C. X.; Pozharski, E.; Rupp, B. Visualizing Ligand Molecules in Twilight Electron Density. *Acta Crystallogr. F* **2013**, *69* (2), 195–200.

[45] Salentin, S.; Schreiber, S.; Haupt, V. J.; Schroeder, M.; Adasme, M. F. PLIP: Fully Automated Protein–Ligand Interaction Profiler. *Nucleic Acids Res.* **2015**, *43* (W1), W443–W447.

[46] Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461.

[47] Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38* (3), 169–177.

[48] Sanner, M. F.; Olson, A. J.; Spehner, J.-C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38* (3), 305–320.

[49] Raschka, S. MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack. *J. Open Source Softw.* **2018**, *3* (24), 638.

[50] Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9* (3), 10–20.

[51] Jones, E.; Oliphant, T.; Peterson, P.; et al. SciPy: Open Source Scientific Tools for Python, **2001**.

[52] McKinney, W. Data Structures for Statistical Computing in Python. In van der Walt, S.;

Millman, J., eds., *Proceedings of the 9th Python in Science Conference.* 51–56.

[53] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95.

[54] Waskom, M.; Botvinnik, O.; O'Kane, D.; Hobson, P.; Lukauskas, S.; Gemperline, D. C.; Augspurger, T.; Halchenko, Y.; Cole, J. B.; Warmenhoven, J.; de Ruiter, J.; Pye, C.; Hoyer, S.; Vanderplas, J.; Villalba, S.; Kunter, G.; Quintero, E.; Bachant, P.; Martin, M.; Meyer, K.; Miles, A.; Ram, Y.; Yarkoni, T.; Williams, M. L.; Evans, C.; Fitzgerald, C.; Brian; Fonnesbeck, C.; Lee, A.; Qalieh, A. mwaskom/seaborn: v0.8.1 (September 2017), **2017**.

[55] Gutiérrez, I. S.; Lin, F.-Y.; Vanommeslaeghe, K.; Lemkul, J. A.; Armacost, K. A.; Brooks, C. L.; MacKerell, A. D. Parametrization of Halogen Bonds in the CHARMM General Force Field: Improved Treatment of Ligand–Protein Interactions. *Bioorg. Med. Chem.* **2016**, *24* (20), 4812–4825.

[56] Aldeghi, M.; Gapsys, V.; de Groot, B. L. Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation. *ACS Cent. Sci.* **2018**, *4* (12), 1708–1718.

[57] Pires, D. E.; 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.