Prospective evaluation of structure-based simulations reveal their ability to predict the impact of kinase mutations on inhibitor binding

Sukrit Singh*[1¶], Vytautas Gapsys*[2^], Matteo Aldeghi*[3†], David Schaller[1§4], Aziz M. Rangwala[5], Jessica B. White[6], Joseph P. Bluck[7], Jenke Scheen[8], William G. Glass[1], Jiaye Guo[1], Sikander Hayat[9], Bert L. de Groot[3], Andrea Volkamer[4,10], Clara D. Christ[7], Markus A. Seeliger[5¶], John D. Chodera[1¶]

*These authors contributed equally to this work
¶ Correspondence: sukrit.singh@choderalab.org, john.chodera@choderalab.org, markus.seeliger@stonybrook.edu

AFFILIATIONS:
[1] Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[2] Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, Beerse 2340, Belgium
[3] Computational Biomolecular Dynamics Group, Department of Theoretical and Computational Biophysics, Max Planck Institute for multidisciplinary sciences, D-37077 Göttingen, Germany
[4] In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
[5] Department of Pharmacological Sciences, Stony Brook University Medical School, Stony Brook, NY 11794, United States
[6] Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, NY 10065, United States
[7] Structural Biology & Computational Design, Research and Development, Pharmaceuticals, Bayer AG, 13342 Berlin, Germany
[8] Open Molecular Software Foundation, Davis, CA 95618, USA
[9] Department of medicine II, University Hospital Aachen, Aachen, Germany
[10] Data Driven Drug Design, Faculty of Mathematics and Computer Sciences, Saarland University, Saarbrücken, Germany

[§]Present address: Nuvisan ICB, Life Science Chemistry, Digital Müllerstraße 178, 13353 Berlin, Germany
^Present address: Computational Chemistry, Janssen Research & Development, Janssen Pharmaceutica N. V., Turnhoutseweg 30, B-2340 Beerse, Belgium
[†]Present address: Bayer Research and Innovation Center, 238 Main St, Cambridge, MA 02142, USA.

ABSTRACT: Small molecule kinase inhibitors are critical in the modern treatment of cancers, evidenced by the existence of over 80 FDA-approved small-molecule kinase inhibitors. Unfortunately, intrinsic or acquired resistance, often causing therapy discontinuation, is frequently caused by mutations in the kinase therapeutic target. The advent of clinical tumor sequencing has opened additional opportunities for precision oncology to improve patient outcomes by pairing optimal therapies with tumor mutation profiles. However, modern precision oncology efforts are hindered by lack of sufficient biochemical or clinical evidence to classify each mutation as resistant or sensitive to existing inhibitors. Structure-based methods show promising accuracy in retrospective benchmarks at predicting whether a kinase mutation will perturb inhibitor binding, but comparisons are made by pooling disparate experimental measurements across different conditions. We present the first prospective benchmark of structure-based approaches on a blinded dataset of in-cell kinase inhibitor affinities to Abl kinase mutants using a NanoBRET reporter assay. We compare NanoBRET results to structure-based methods and their ability to estimate the impact of mutations on inhibitor binding (measured as ΔΔG). Comparing physics-based simulations, Rosetta, and previous machine learning models, we find that structure-based methods accurately classify kinase mutations as inhibitor-resistant or inhibitor-sensitizing, and each approach has a similar degree of accuracy. We find that physics-based simulations are best suited to estimate ΔΔG of mutations that are distal to the kinase active site. To probe modes of failure, we investigate two clinically significant mutations poorly predicted by our methods, T315A and L298F, and find that starting configurations and protonation states significantly alter the accuracy of our predictions. Our experimental and computational measurements provide a benchmark for estimating the impact of mutations on inhibitor binding

affinity for future methods and structure-based models. These structure-based methods have potential utility in identifying optimal therapies for tumor-specific mutations, predicting resistance mutations in the absence of clinical data, and identifying potential sensitizing mutations to established inhibitors.

INTRODUCTION

The emergence of drug-resistant mutations represents a significant obstacle in the effective treatment of various diseases, including cancer. This is most recently exemplified by over 80 FDA-approved kinase inhibitors already having known resistant mutations.[1] These mutations can arise in the target proteins of therapeutic agents, rendering them less susceptible or completely resistant to the drugs' intended mechanisms of action. The ability to predict drug-resistant mutations in advance promises to guide precision medicine.[2–4] A number of susceptible and resistance mutations have been characterized for FDA-approved kinase inhibitors (e.g. OncoKB) but a much larger number of mutations in the target of therapy (the kinase domain) have been observed for which no data is available.[4–9]

Predicting the impact of mutations remains challenging due to the complex and diverse nature of resistance mechanisms. Kinase mutations may decrease drug-binding affinity or potency,[10–12] increase kinase activity,[13–16] tune inhibitor sensitivity profiles,[11,17,18] or any combination of these mechanisms, or other mechanisms involving additional cellular machinery.[19,20] Alternatively, mutations may shift the population of conformations towards alternative states, decreasing the population of drug-compatible conformations.[21–26] Some resistance mutations may even be compensatory, increasing activity of the kinase either through shifting towards increased propensity towards an active kinase,[27,28] or increasing the affinity for ATP.[14,22,29] Of these potential mechanisms, the most direct way that a mutation can cause drug resistance is by perturbing the kinase-inhibitor binding affinity.[10–12,30,31]

Mutation-induced changes in kinase-inhibitor binding affinity would lead to reduced drug potency. In some cases, a change in binding affinity upon mutation may indicate that a mutation *sensitizes* a protein to an inhibitor, offering new therapeutic strategies.[17,32,33] Thus, being able to measure the impact of a mutation on protein-inhibitor binding affinity offers new insights and avenues into circumventing drug resistance. Experimental methods such as mutagenesis studies and binding assays can provide critical biophysical insight but are often time-consuming, costly, and limited in throughput.[28,34,35] One approach to mechanistically characterizing drug-resistant mutations involves measuring the direct change in binding free energy (ΔΔG).[10–12,31,36]

Structure-informed methods show promise in predicting the impact of mutations on ΔΔG, and may help classify mutants as resistant or susceptible to kinase inhibitors. Existing benchmarks have only been retrospective.[10,12,31] Computational approaches have been previously shown to predict changes in binding affinity caused by mutations with an average error of 1.2 to 2.0 kcal/mol.[10,37] Rosetta-based methods predict ΔΔG using Monte Carlo methods to sample and optimize protein conformations, predicting mutational effects on protein structure and stability.[10,12,38] However Monte Carlo search methods are limited in their ability to sample conformational changes upon mutation and so may not capture impact of a mutation on the conformational landscape of a protein. Molecular dynamics (MD)-based methods leverage the power of atomistic simulations to explore the dynamics and energetics of protein-drug complexes.[23,39,40] By simulating the behavior of the system over time, molecular dynamics can capture the effects of mutations on the stability and dynamics of the binding pocket and the interactions with the drug.[23,33,41] However these methods require extensive simulation to appropriately sample larger clinically-relevant systems, limiting throughput.[39] Recently, machine learning (ML) methods have also gained prominence in predicting ΔΔG changes in binding affinity.[12,31,42] Such ML approaches utilize large datasets of experimentally determined binding affinities to train predictive models.[10,12] By learning the relationship between sequence, structure, and binding affinity, ML models can predict the impact of mutations on binding affinity with remarkable accuracy. However, these methods require large datasets of clinically identified mutations, and knowledge of their mechanistic impact to accurately predict mutational impact.[43,44] As such, ML methods require an existing dataset to train upon using consistent experimental measurements, which does not exist yet. Once such datasets become available, active learning loops can be deployed to cycle between experimental measurement, model training, and predictive evaluation to improve ML model prediction further.[45]

Alchemical methods, also called Free Energy Calculations (FECs), have emerged as powerful computational methods for predicting changes in binding affinity without the need for extensive experimental data.[31,36,37,39,46,47] These alchemical methods are used to estimate free energies of binding (ΔG) by estimating the energetic cost of atoms going from one thermodynamic configuration to another via tuning the strength of atomic interactions via a so-called alchemical transformation.[47] To study the impact of a mutation on drug binding, we estimate the energetic cost of transforming amino acids between wild-type (WT) and mutant residues in the presence and absence of a ligand (Supp. Fig. 1). These alchemical methods compute free energy differences (ΔG) using a so-called "transformation" between the wild-type and mutant amino acid atoms. This transformation is "alchemical" in that it scales, using the parameter λ, bonded and nonbonded interactions of an amino acid sidechain found in WT and mutant. Several alchemical approaches can be used to estimate ΔΔG changes associated with mutations. Replica exchange methods, such as Hamiltonian replica exchange molecular dynamics (RepEx, HREX, or REMD),[46,48–51] enable enhanced sampling of conformational space and provide insights into the thermodynamics of the binding process. Non-equilibrium switching methods utilize fast out-of-equilibrium transitions along the reaction coordinate to estimate free energy differences.[36,52,52–55] However, existing benchmarks have only been retrospective,[10,12,36,55,56] and the prospective capabilities of FECs require examination.

Previous FEC benchmarks derive from IC50 measurements across multiple types of biochemical or in cellular measurements,[18,24] ranging from qPCR detection, Ba/F3 activity assays, or mobility shift assays (gel shift).[30,57–60] In turn, there are many more sources of experimental variation that can affect the quality of benchmarking measurements.[61] Furthermore, variation in experimental contexts, from cells to in vitro mobility measurements, to variability in cell type for qPCR, may cause context-dependent variation that alters the measured impact of clinical kinase mutations.[61,62] Such protocol specific variability may not capture accurate cancer-like cellular contexts and environments, thus introducing noise into the benchmarking dataset.[63,64] Ideally, our ability to measure IC50 would use a consistent experimental readout that more directly and consistently reports on binding. There is a need for comparing against a dataset that uses a single type of measurement in a controlled rigorous manner.

NanoBRET offers a high throughput in-cell approach to prospectively evaluate mutations and their impact on kinase-inhibitor binding.[65,66] This method relies on the principle of resonance energy transfer between a bioluminescent donor and a fluorescent acceptor. NanoBRET has proven valuable for high-throughput screening of mutations and their effects on ligand binding,[24,67,68] and allows for the cellular environment to influence kinase-inhibitor binding, providing an in-cell characterization of the change in affinity.[24,67] The consistency in experimental measurements allows us avoid multiple sources of experimental variation when benchmarking FEC results.[62] NanoBRET data has previously demonstrated its capacity to measure the impact of mutations on Abl kinase, even revealing new possible kinetic mechanisms of drug resistance.[24] Computational tools to identify these highly sensitive mutations *a priori* may reduce patient burden and increase treatment outcomes as we are able to more precisely treat patients who would benefit most from the use of sensitized drugs.

Here, we present the first prospective benchmark of structure-informed physical and machine learning models for the prediction of resistance/susceptibility, using a recently-described NanoBRET assay to measure the impact of clinical Abl mutations within cells, considering first- and second-generation TKIs, imatinib and dasatinib (Fig. 1 and 2).[24,60,69–71] We compare and evaluate the performance of structure-based methods to predict the impact of mutations, including replica exchange and non-equilibrium switching methods, Rosetta-based methods,[38] and a previously published Random Forest method (Fig. 3).[10,56] We apply these methods on clinically identified mutations of Abl kinase, and compare ΔΔG predictions against measured values from experimental nanoBRET measurements of Abl binding to imatinib and dasatinib. Furthermore, we show that high throughput biophysical methods like nanoBRET provide comparable measurements to other low-throughput methods, demonstrating the value of nanoBRET as a benchmark measurement for assessing ΔΔG accuracy. We further demonstrate the capacity of free energy calculations to independently predict accurate ΔΔG values even for distal mutations that are not near the binding site. Lastly, we also demonstrate that FECs are capable of improving ΔΔG by considering alternative protonation states. The integration of multiple computational methods allows for a comprehensive and reliable prediction of mutation effects, aiding in the design of effective therapeutic strategies and advancing precision medicine.

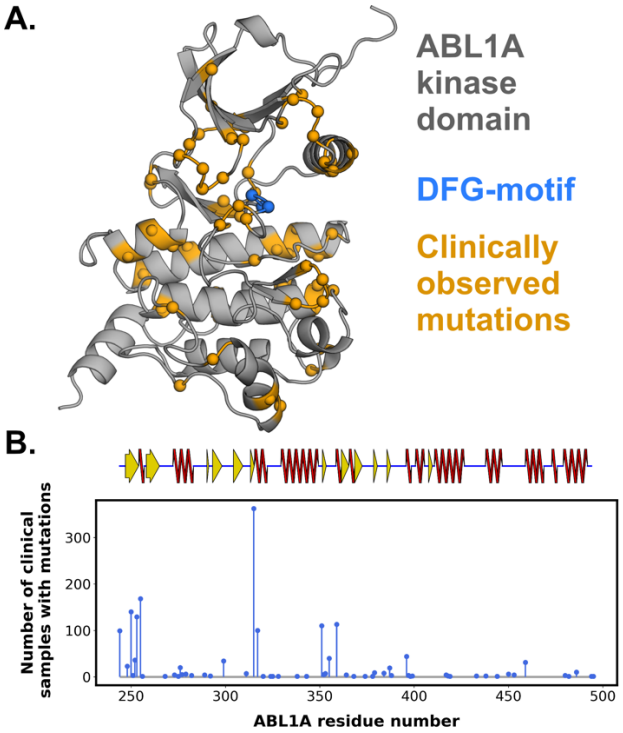


Figure 1. Clinically relevant ABL1A mutations occur throughout the kinase domain at varying frequencies. **A.** Structure of Abl kinase (PDB: 1OPJ), highlighting the positions of clinically derived Abl mutations (orange spheres) associated with resistance (defined by COSMIC or OncoKB). **B.** Lollipop plot denoting the number of clinical samples found in the COSMIC/OncoKB databases denoting the number of times a mutation occurs at each position of the kinase domain sequence. Secondary structure of the corresponding region is indicated (above) denoting whether the region is helical (red), a β-strand (yellow) or a loop (blue).

RESULTS AND DISCUSSION:

NanoBRET is consistent with previous experimental measurements of mutational impact to identify inhibitor-resistant and -sensitizing mutations.

To understand the clinical relevance of mutations within the ABL1A kinase domain, we gathered mutations observed in clinical settings using known databases and quantified mutation prevalence across clinical samples (Fig. 1).[5] While there are multiple possible mutations per position, the total number of clinical samples with a particular mutation can be represented using a lollipop plot (Fig. 1B). Mapping the distribution and frequency of clinical mutations to the structure of ABL1A kinase domain (Fig. 1A) also allows us to identify structural hotspots with high mutation propensity, or pinpoint hotspots where mutations are particularly prevalent in ABL1A. We note that mutations are distributed throughout the secondary structure and sequence of ABL1A, and while certain positions have many mutations, mutations do not appear to concentrate to any particular hotspot or secondary structural kinase element.

We obtain NanoBRET measurements that quantify the impact of 90 highly prevalent mutations from this clinical dataset (Fig. 2, SI data). NanoBRET allows for precise quantification of the changes in drug binding affinity caused by these mutations.[24,67,68] These NanoBRET measurements measure the degree of inhibitor-target engagement using bioluminescent probes that luminesce when bound to one another.[24,67,68] These engagement measurements can be converted into an IC50 by scanning engagement across a variety of titratable concentrations.[24] Our structural map demonstrates the broad distribution of mutated residues in relation to the drug binding site (Fig. 1A)

We can also assess the fold-change in IC50 for any mutant's imatinib- or dasatinib- affinity relative to wild type (WT). Mapping the maximal fold-change onto the three-dimensional structure of ABL1A kinase (Fig. 2A) allows us to visualize the range of impact mutations have on ligand binding. Mapping the maximum fold impact of a mutation at a specific position from NanoBRET onto the structure of Abl kinase highlights the broad distribution

of mutated residues relative to the inhibitor binding site (Fig. 2A). Notably, the color-coded mutations illustrate a spectrum of impacts, with some mutations causing increased inhibitor resistance, while others surprisingly lead to increased inhibitor sensitivity. The latter suggests that certain mutations, rather than conferring resistance, render the kinase more amenable to inhibition by existing drugs, thereby identifying potentially promising targets for therapeutic intervention. Mutations found to induce inhibitor sensitization may present promising targets for existing drugs, while others require additional treatment development.

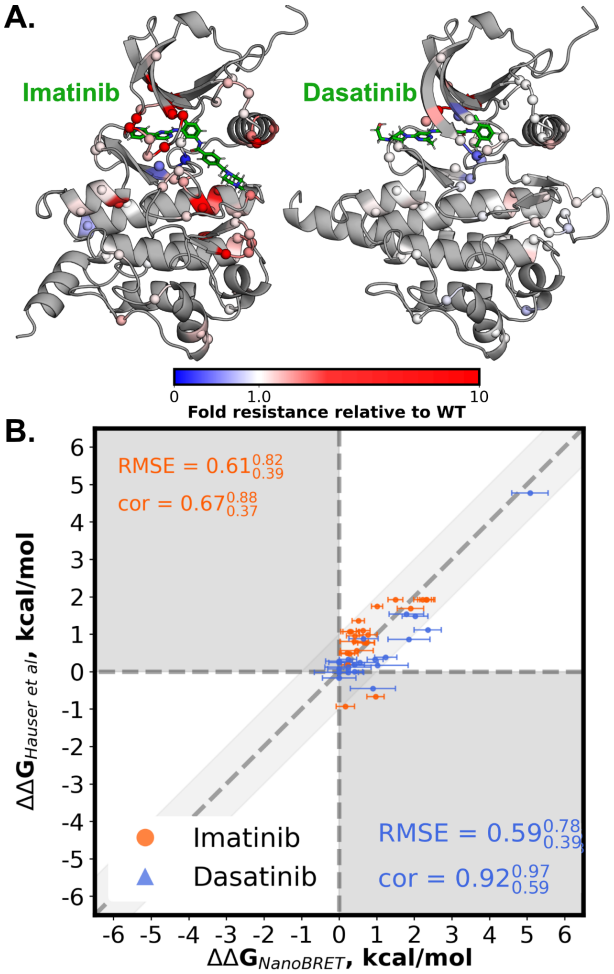


Figure 2: An intracellular NanoBRET assay for measuring the impact of clinical mutations on kinase inhibitor binding free energy is concordant with previously-reported IC50-derived data. A. Experimental IC50 values of Abl kinase binding to Imatinib (left, green) or Dasatinib (right, green) for a variety of mutations (spheres) normalized by the WT IC50 of each compound (color scale, bottom). This WT-based normalization demonstrates which mutations are resistant or sensitizing to either imatinib or dasatinib. **B.** Experimental ΔΔG values from the current work plotted against those from Hauser et al. A subset of 18 mutations common to the sets of Hauser et al. and the current work is considered. Data points for which the experimental values disagree by more than 1 kcal/mol are marked in a darker edge color. Root Mean Square Error (RMSE) and Pearson correlation (labeled "cor") between datasets noted for each mutation-ligand pair, alongside 95% confidence intervals.

We compare these NanoBRET measurements with previously published work on ABL1 mutations and their impact on imatinib and dasatinib binding affinity (Fig. 2B).[31,57,59,60] By comparing the IC50 of wild-type and mutant ABL1A binding to a known inhibitor, imatinib or dasatinib, we estimate the change in free energy of binding (ΔΔG); the NanoBRET protocol using tracer concentrations at the K_M ensures that significant changes in IC50 will be largely driven by changes in inhibitor binding affinity. The ΔΔG values from previous works can be correlated directly against these ΔΔG measurements (Fig. 2B). The comparison of ΔΔG values obtained from NanoBRET with those from prior experimental data exhibits a noteworthy consistency for both Imatinib and Dasatinib. This consistency is quantitatively supported by low RMSE values and decent correlation coefficients (RMSE=0.61 kcal/mol for imatinib, and 0.59 kcal/mol for dasatinib), indicating the accuracy of our

NanoBRET assay. It is worth noting that our NanoBRET measurements are derived from single-shot experiments, ensuring a high-throughput collection protocol. As such, we estimate error associated with each NanoBRET from the goodness-of-fit for each individual mutation (Fig. 2B, SI data). This provides a nuanced estimate of the experimental uncertainty despite the lack of statistical replicates. In conclusion, low RMSE and correlation scores not only reinforce the consistency of the NanoBRET measurements with previously published results but also underscore NanoBRET's utility as a tool for the high-throughput screening of kinase mutations.

Free energy calculations prospectively predict the change in affinity with RMSE of 1 kcal/mol and provide a robust classification metric for predicting a resistant or sensitizing mutation.

While NanoBRET is a robust high throughput tool to assess mutations and their impact on inhibitor binding, it would be particularly useful, and cost-effective, for computational approaches to *prospectively* parse mutations as resistant, sensitizing, or neutral relative to any particular inhibitor. By forecasting the impact of genetic alterations on drug efficacy, computational predictions promise a preemptive tailoring of therapeutic strategies.

Prospective testing is crucial to verify the applicability of computational models. Retrospective analyses may inadvertently incorporate biases from known outcomes and contain outdated information and biases influence benchmarking comparisons. A prospective approach evaluates the predictive power of models in novel, untested scenarios, where tools are employed in standard default fashions without any biased use of the tool to a known outcome. This forward-looking methodology is essential for validating the robustness of predictive algorithms and ensuring their utility in clinical settings.

To effectively classify mutations as inhibitor-resistant or -sensitizing, a variety of computational and machine learning approaches were prospectively tested against NanoBRET. These methods include free energy calculations using two different sampling strategies: Hamiltonian Replica Exchange (RepEx or HREX),[46,48–51] and Non-Equilibrium perturbations (NEQ),[12,31,36,46] Rosetta flex_ddg,[12,38] and a machine learning model utilizing Random Forests (called "ML" here).[10,12] Each method offers a unique computational strategy to predict the impact of mutations on protein stability and drug binding (Fig. 3). Briefly, alchemical free energy methods estimate the impact of a mutation on drug binding by estimating the thermodynamic cost of transforming one amino acid to another in the presence and absence of a ligand. These alchemical methods compute free energy differences (ΔG) using a so-called "transformation" between the wild-type and mutant amino acid atoms. This transformation is "alchemical" in that it scales, using the parameter λ, bonded and nonbonded interactions of an amino acid sidechain found in WT and mutant Abl kinase (Fig. 3) such that $\lambda=0$ represents a complete WT Abl and $\lambda=1$ represents the mutant; values of $0<\lambda<1$ are a scaled combination of the two sets of interactions. During a transformation, the sidechain of one amino acid (WT) is "phased out" while another amino acid chain (the mutant) is "phased in" (Fig. 3).[37,46,47,72] In computing the free energy (ΔG) in both holo and apo conditions, a thermodynamic cycle is constructed, from which we subtract the apo ΔG from the holo ΔG to arrive at the thermodynamic impact of mutation upon drug-binding (ΔΔG) (Supp. Fig. 1).[47] The Rosetta flex_ddg protocol models the structural and energetic effects of mutations by generating mutant structures and extensively sampling possible rotamers. The protocol then iteratively refines ΔΔG across multiple iterations using the Rosetta scoring function to generate averaged predictions that account for likely rotamers of both WT and mutant amino acid.[38]

AI/ML approaches that directly assess the impact of mutations on drug binding have the potential to reduce the computational cost of predicting drug resistant mutations. In fact, many powerful models have been published that accurately predict the impact of mutations on protein stability.[73–76] While stability is correlated with drug binding, it would be particularly useful if models provided direct readouts of the impact upon drug binding. In this work we will focus on prior published models that directly predict ΔΔG upon mutation of drug binding; Specifically, we will focus on a prior published Random Forest model to assess Abl kinase mutations.[10,12] A Random Forest model uses an ensemble of randomized decision trees to predict the change in binding affinity (ΔΔG) of drugs upon point mutations in the human Abl kinase.[10,12,77] The model was trained on a curated dataset of 144 ΔΔG values for eight tyrosine kinase inhibitors, incorporating a diverse set of features that describe ligand properties, mutation environments, amino acid changes, protein-ligand interactions, docking

scores, and solvent accessibility.[31,57,59,60] Feature selection was performed using a greedy algorithm to identify the most predictive features.

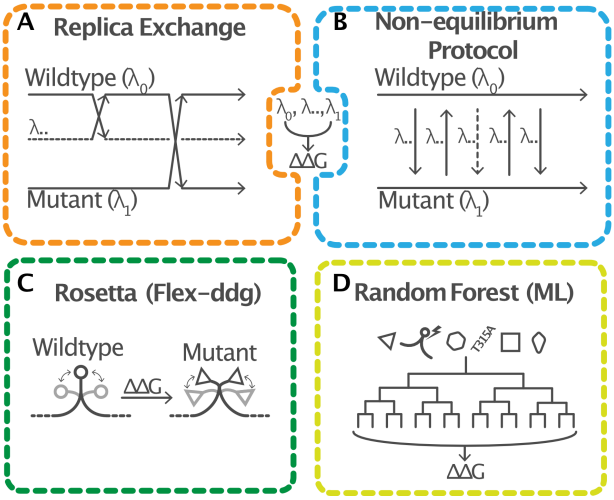


Figure 3. Overview of computational methods used to prospectively assess estimation of ΔΔG upon mutation. Physics-based alchemical methods (top) estimate the impact of mutation on drug binding (ΔΔG) by generating an alchemical transformation between wild-type (WT) and mutant constructs using a lambda parameter sampled using **A.** Replica Exchange methods and **B.** Non-equilibrium protocols. **C.** The Rosetta flex–ddg protocol considers rotameric positions of both WT and mutant amino acid side chains, using the Rosetta scoring function to compute the energetic impact of both constructs. **D.** Random forest models generated using an ensemble of decision trees are trained on prior assay measurements and structural data to generate ΔΔG value predictions.[31]

Comparing each method against known mutations that have been previously studied highlights the capacity of each method to reasonably predict the impact of mutation on imatinib or dasatinib affinity. To test the prospective accuracy of these methods, we compared their predictions against experimental NanoBRET measurements for mutations in a manner similar to previous studies (Fig. 4).[31,57,59,60] This comparative analysis demonstrates that while each approach has varying degrees of success, they collectively exhibit a reasonable predictive capacity for the same set of mutations.

Interestingly, all approaches appear to consistently do better at predicting the impact of mutations on dasatinib binding than imatinib binding, highlighted by the lower RMSE for dasatinib vs. the RMSE of imatinib across all the approaches (Fig. 4). This is further supported by the degree of correlation between calculated changes in affinity and NanoBRET measurements in the datasets. The discrepancy in predictive capability between imatinib and dasatinib may be related to the degree of conformational sampling required to accurately sample imatinib binding vs. dasatinib binding. Previous work has shown that to characterize the complete ABL1 imatinib binding, a conformational transition in ABL1A must occur.[78–80] However, efforts to characterize the full ensemble of structural transitions on binding have been limited by sampling, and there may be additional ABL1-Imatinib configurations.[26,81,82] No such conformational transition or degree of sampling is needed to sample dasatinib binding.[78] As such, there may be currently unseen states of the ABL1-Imatinib bound complex that must be considered in addition to the singular crystal that was used as a starting configuration here. The differences between dasatinib and imatinib are consistent across the different MD force fields as well as Rosetta. Thus, it may be that the degree of conformational sampling required to achieve predictively convergent calculations within 1 kcal/mol differs between the two ligands.

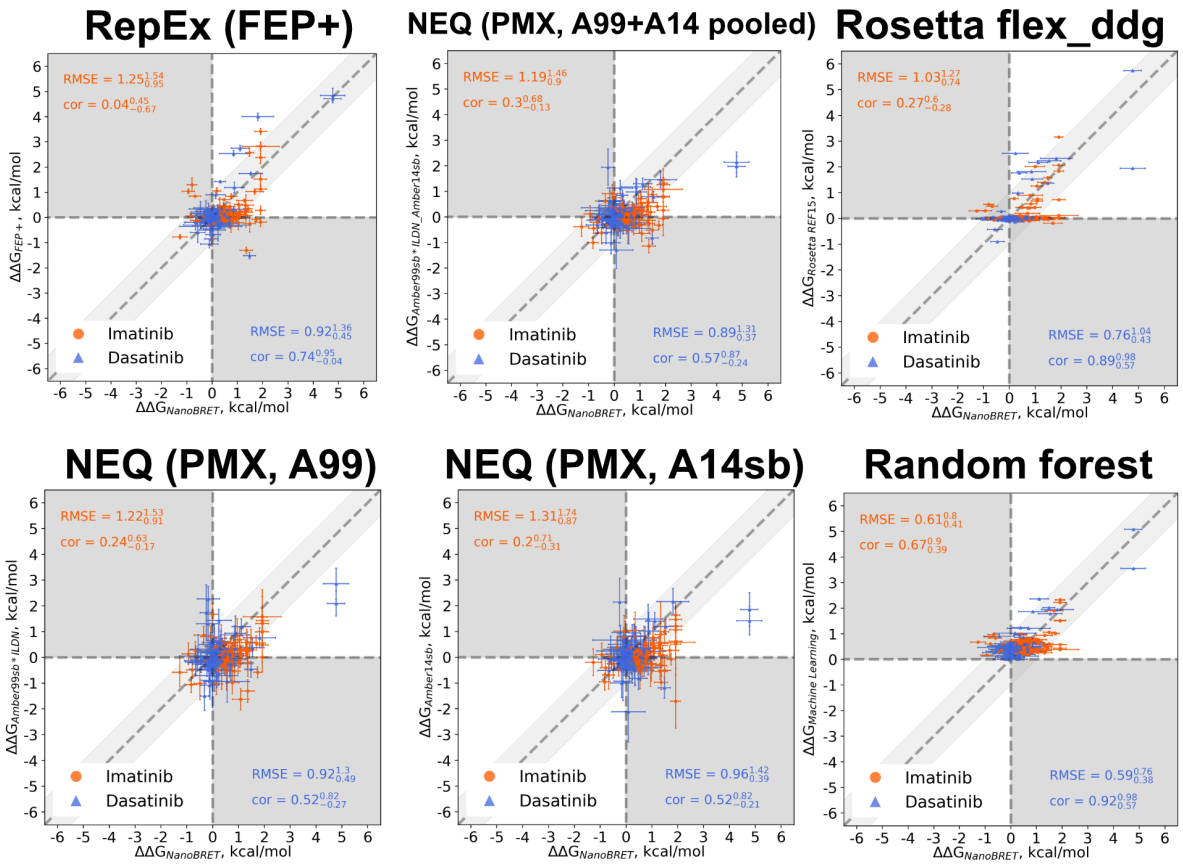


Figure 4. Prospective computational methods have comparable summary statistics and performance in predicting the impact of mutation on binding. Scatterplots showing the prospective ability of different computational methods and their ability to predict the ΔΔG of either imatinib (orange) or dasatinib (blue) binding. Root Mean Square Error (RMSE) and Pearson correlation (labeled "cor") are provided in the top left (imatinib) and bottom right corners (dasatinib). These comparisons are all done relative to the same ΔΔG measurements collected using NanoBRET. Prospective methods shown are Replica Exchange using FEP+ (top left), Nonequilibrium switching using PMX using Amber99 force field (bottom left), the Amber14sb force field (bottom middle), and the resultant prediction taken from pooling the work values from both force fields (top middle). Rosetta's flex_ddg (top right) and a random forest model trained on prior data (bottom right) are also shown.

Marking the 1 kcal/mol boundary as an acceptable margin of error and degree of significance for perturbative mutations, we can categorize mutations as either resistant, sensitizing, or neither (Supp Fig. 2). Previous results show that most reproducible estimations occur with an RMSE of up to 1 kcal/mol.[12] To then capture the most reproducible predictions, we consider a mutation to be correctly assigned as resistant if both the NanoBRET and computational approach predicted an increase in ΔΔG by ΔΔG > +1 kcal/mol for a single inhibitor. Conversely, a mutation is considered to be correctly assigned as sensitizing if both NanoBRET and computational predictions identify drug-binding affinity is decreased by ΔΔG > -1 kcal/mol. We can visualize these classifications as "quadrants" in our correlation analysis, generating truth tables (confusion matrices) for each computational method (Supp Fig. 1).[31] This matrix allows us to evaluate the performance of each computational approach in correctly classifying mutations and compare that performance statistically against a baseline classifier.

From these truth tables, we show that computational approaches classify mutations as resistant or sensitizing better than a baseline classifier would, shown using a precision recall curve (Fig. 5, Supp. Table 1). A precision-recall curve graphically represents model performance in contexts where the classifications have different populations and frequencies. This allows us to evaluate the sensitivity and specificity of our measurement and model. Precision, known as positive predictive value, measures the ratio of the predictive model's true positives to all the positive predictions. Recall, also defined as sensitivity, measures the model's ability to identify all true positives from the data. This allows us to identify the best threshold for the predictive model, and helps us establish whether our classifier is performing better than random behavior. A higher area

under the curve (AUC) indicates a better model performance. We compare the performance of each computational approach as a classifier to a baseline classifier (Fig. 5, Supp. Table 1), where the baseline classifier reports the performance if everything was labeled as statistically positive, or in our case, a resistant or sensitizing mutation. For the 1 kcal/mol cutoff, all approaches perform better than a baseline classifier, having a positive distance above the baseline as shown by the positive distance above the baseline (Supp. Table 1).

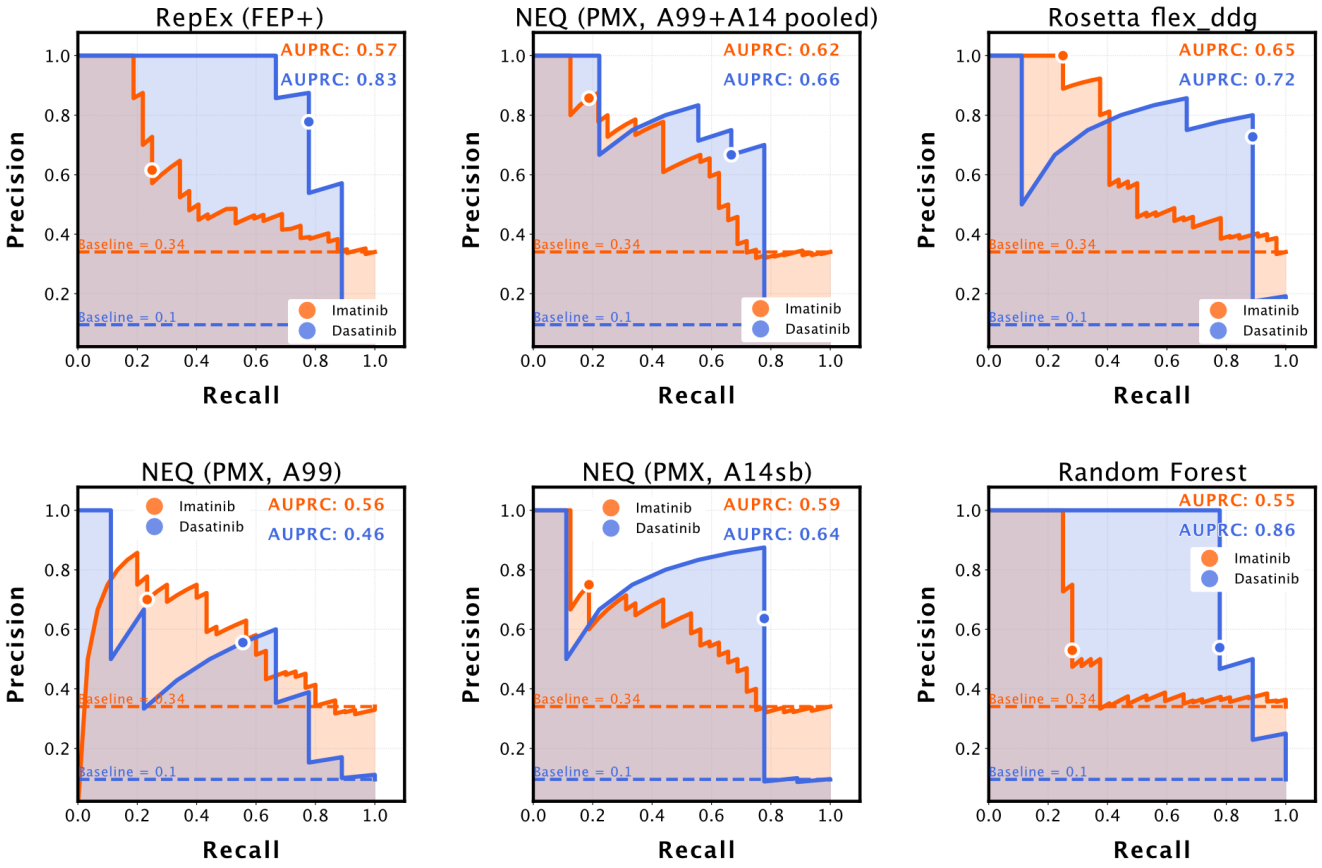


Figure 5. Precision-Recall Curves for each method to demonstrate their ability as classifiers is better than random, and that they are all similar in performance. Precision is defined as the fraction of mutations that are classified as resistant that are actually resistant, while recall is the fraction of mutations that are classified as resistant. The area under the precision recall curve (AUPRC) is computed for both imatinib (orange) and dasatinib (blue). The singular point on the curves corresponds to the precision (x) and recall (y) value for mutations with $|\Delta\Delta G| > 1$ kcal/mol, and are computed for imatinib (orange) and dasatinib (blue), and compared against the performance of a random classifier (dashed line).

The distance from the baseline at 1 kcal/mol acts as a useful measure to compare both the performance of each approach against one another as well as between the two different drugs. At the 1 kcal/mol margin, it is worth noting that the Random Forest model is closest in performance to the baseline classifier (Supp. Table 1), indicating that it performed relatively worse than the other methods as a classification tool compared to a statistical baseline. The larger number of false positives (mutations that computationally are not predicted to reduce the affinity but experimentally do) indicates that the Random Forest model was over-assigning resistant/sensitizing classifications. Consistent with RMSE metrics, we find that imatinib is much harder to classify relative to baseline than dasatinib for many of the tools, with each approach having a much greater distance to baseline when classifying dasatinib than imatinib. Our Random Forest model is particularly bad at classifying imatinib mutations relative to dasatinib, with its overall performance being worse than the physics based alchemical method.

Taken together, these results highlight the capacity of the employed computational methods to act prospectively as classifiers within an average accuracy of 0.81 for resistant or sensitizing mutations (Supp. Table 1). Modern computational approaches act as useful classifiers for sorting clinically observed mutations

into useful, potentially actionable categories. These approaches appeared very similar to one another using RMSE and correlative measures, with no clear indicator of relative performance. Precision-recall curves, the area underneath them, and comparing relative to a baseline classifier allows us to better compare between methods and across inhibitors to identify improvements in performances and challenging drugs/mutations that may require further analysis. However, the above findings highlight that it is necessary to analyze datasets and prospective predictions analysis beyond using summary statistics and correlative measures when evaluating benchmarks.

Alchemical methods can detect impact of distal mutations upon ligand binding

While each of these methods may be similarly capable of predicting the impact of mutations on binding, summary statistics and analysis may mask issues when considering prospective benchmarking predictions. In turn, it is important to consider each mutation's prediction individually to observe similarities or performance improvement. For example, the similarity of results shown from the classifier model may mask a degree of per-residue and per-mutation variance that results from a larger systematic issue.

By looking at the maximum possible improvement possible by simulation per mutation, we find that, while predictions are within ~1 kcal/mol experiment, free energy calculations generally offer more improved predictions at any position in the sequence. By subtracting the least accurate non-free energy calculation prediction (Random Forest/ML and Rosetta) from the most accurate alchemical prediction, we compute the maximum amount that physics-based simulations are capable of making better predictions (Supp. Fig. 3). A more negative improvement score thus conveys that simulations were much closer to the answer than non-free-energy methods were, while a positive score indicates the reverse. We find that, for any single mutation, free energy calculations provide a small but consistent margin of improvement, at most 0.5 kcal/mol, relative to alternative methods such as Rosetta or ML-based methods (Supp. Fig. 3).

Potential improvements offered by simulation-based methods such as alchemical physics-based simulations is further emphasized when considering allosteric mutations (i.e. distant residues relative to the ligand binding). Identifying distal mutations that perturb binding free energy remains a challenge for modern computational methods but has great potential to help prospectively identify allosteric mutations and cryptic pockets.[23,40,41,83–86] However identifying significant distal mutations, and prospectively analyzing them, remains a challenge due to their relative infrequency in clinical datasets that identify significant mutations with known impact.

We consider how distal a mutation is relative to the ligand binding site by computing the distance from the center of mass of a residue to the center of mass of the inhibitor. We first note that experimental ΔΔG from NanoBRET experiments show many distal mutations impact inhibitor binding, with residues up to 30 Å away having an impact on ΔΔG of imatinib binding (Fig. 6A). However, we note that many residues have large error bars due to the singlicate measurements. A significant alteration to ΔΔG would be much clearer if a mutation with ΔΔG is non-zero while its error bars are much smaller (Supp. Fig. 4-6). Most residues with significant impact on inhibitor binding are in closer proximity to the binding site, with only a few distal mutations having significant impact (Fig. 6A, Supp. Fig. 5,6).

By projecting the predicted ΔΔG or the deviation of predicted ΔΔG from experiment (Fig. 6B, Supp. Fig. 6), we note that both Rosetta and the Random forest (ML) methods appear to have systematic biases in their predictions (Fig. 6B); This is especially obvious for Rosetta and ML methods when looking directly at the ΔΔG predicted values as a function of distance. Both approaches appear to flatten out and center around a single value, with Rosetta predictions returning ΔΔG=0 kcal/mol beyond a certain distance from the active site. Similarly, the random forest model makes predictions for residues within the binding pocket of imatinib beyond 1 kcal/mol, collapsing to predictions around a ΔΔG~0.5 kcal/mol upon moving further away from the drug binding pocket. This systematic error is likely the result of in-built distance cutoffs in the computational approaches. By enforcing a cutoff to make mutation ΔΔG predictions computationally tractable, a mutation's local environment is mainly considered. As a result, while a mutation may have some impact on its local and global environment, constraining only to local effects will ignore any impact a distal mutation may have on inhibitor binding affinity.

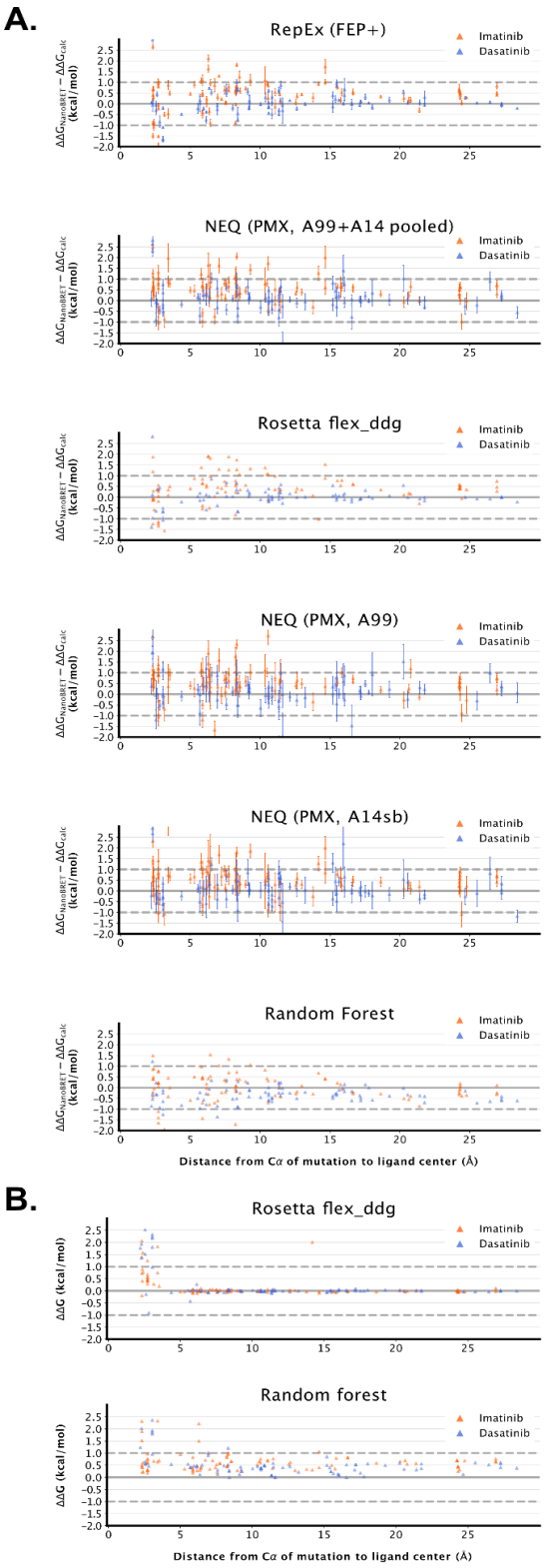


Figure 6. Physics-based methods are capable of estimating ΔΔG for distal mutations, while default parameters for Rosetta and ML-based methods collapse at a distance. A. The deviation from predicted ΔΔG to experiment is plotted for imatinib (orange) and dasatinib (blue) as a function of the distance from the residue's Cα carbon to the center of mass of the ligand in the crystal structure. Values for shown (top to bottom) for Replica Exchange using FEP+, PMX predictions taken from pooling the work values from both Amber99 and Amber14sb, Rosetta's flex_ddg protocol, Nonequilibrium switching using PMX using Amber99 force field, followed by Nonequilibrium switching using the Amber14sb force field, and a random forest model trained on prior data **B.** Predicted ΔΔG values for shown (top to bottom) for Rosetta's flex_ddg protocol and the random forest model trained on prior data, indicating the model predicts within a small range of values for distal mutations.

Conversely, we note that free energy calculation predictions appear to deviate from experiment more randomly, but remain within 1 kcal/mol even at a distance (Fig. 6A, Supp. Fig. 4). Consistent with experimental results, residues predicted to have the strongest impact on ΔΔG are generally closer to the active site. Importantly, even at a distance, the vast majority of computed deviation from experimental values appears to be below 1 kcal/mol (Fig. 6A). However, due to the error within the NanoBRET experiments, it remains unclear if this low margin of error is computationally driven, or if these mutations themselves do not impart a very large ΔΔG upon inhibitor binding, reducing the possible margin of error. Regardless, alchemical methods appear capable of predicting and considering the impact of distal mutations.

Structural considerations on a per-mutation basis allow us to more broadly assess the prospective ability of computational methods to predict the impact of mutation on an inhibitor's binding affinity. The error bars within our experimental results also further highlight the importance of considering experimental error as well as computational error when considering predictions.

Retrospective analyses of clinically significant mutations emphasize the importance of starting configurations, protonation states, and sampling.

While each method appears to predict the impact of each mutation to a consistent rate, based on summary statistics and AUPRC findings, it is interesting to also note the consistent points of failure. Certain clinically notable mutations such as T315A, a known clinical mutation that causes imatinib resistance,[16,87,88] are inaccurately predicted in a consistently poor manner. In fact, PMX based NEQ protocols are unable to predict even the correct sign of T315A (Supp. Table 2). The consistent failing to accurately predict the ΔΔG of certain mutations suggests a systematic error across all methods that results in this inaccurate prediction.

There are many possible sources of error that could systematically bias ΔΔG estimations. Indeed, considerable effort has gone into characterizing the source of errors that occur in ΔΔG predictions from replica exchange alchemical transformations.[46] These previous efforts identified many potential system specific structural features that might be slow to converge in a replica exchange sampling schema. Consistently, there may be structural features that are slow to converge that are the basis of some errors in the alchemical transformations. However, other ΔΔG estimations presented here do not necessarily use physics-based simulations, instead utilize other information, sampling schemes, or alternative datasets to generate their predictions. As a result, there are only a few common systematic sources of biases due to the variable parameters and defaults used in each estimation. The consistency in deviation from NanoBRET across multiple approaches may also hint at discrepancies between experimental measurements of target engagement and the direct binding affinity estimated via computation (Supp. Table 2).

To retrospectively analyze these predictions in an open-source, interpretable manner at scale, we employ the open source free energy tool, Perses.[46] Perses allows us to assess multiple sampling strategies, considering both replica exchange and nonequilibrium methods used by FEP+ and PMX, respectively, using a single unified implementation (Supp. Figure 7-8).[72,89,90] We can rapidly generate NEQ estimations using Perses in a massively parallel manner by deploying alchemical simulations of mutations on Folding@home.[39,91] Folding@home is a distributed computing platform where molecular simulations are run in discrete work units. The asynchronous nature of each work unit makes Folding@home particularly suited to evaluate many mutations in parallel.[37,92] Indeed, prior computational work has leveraged Folding@home's massively parallel nature to evaluate mutations and potential inhibitors in an embarrassingly parallelizable manner using NEQ sampling; each work unit can run an individual replicate of a singular forward and reverse transformation.[37,92,93] These individual "cycles" of NEQ replicates can then be aggregated to generate ΔΔG estimations (Supp. Figure 6-7). To obtain detailed assessments of the source of biases within our estimations, we focus on emblematic examples of clinically relevant mutations that were inaccurately predicted. Specifically, we consider two mutations: the perturbative "gatekeeper mutation" T315A mutation with known clinical significance to give rise to imatinib resistance,[16,87,88] and the known stabilizing mutation L298F that improves imatinib binding affinity and increases sensitivity (Fig. 7, Supp. Table 2-4).[24] Identifying sources of errors spanning both these mutations, one perturbative and one stabilizing, can suggest generalized sources of errors that span different mechanistic impacts.

Perses retrospective ΔΔG predictions are consistent with prospective predictive findings, showing that it is a useful platform for benchmarking and testing our results (Supp. Fig. 7-8). Perses predictions appear to be consistent in summary statistics with findings from both RepEx and NEQ (Supp. Fig. 7). Perses predictions appear to behave similarly to other approaches when considered as a classifier (Supp. Fig. 7), returning equivalent or better precision-recall statistics and F1 scores (Supp. Fig. 7). Retrospective calculations using Perses also appear to capture the impact of distal mutations by FECs with a similar degree of consistency to other free energy methods (Supp. Fig. 8). Deviation of predictions by Perses from experiments across distances also appears similar (Supp. Fig. 8); This consistency with other free energy predictions demonstrates that Perses provides an appropriate platform to conduct detailed testing of high error estimations. Importantly, improvements made in Perses ΔΔG predictions would generalize and apply to other platforms and tools as well.

In the case of T315A, we find that the systematic failure in ΔΔG prediction is corrected by alteration of the starting configuration used to run these free energy calculations (Fig. 7, Supp Table 2-3). In our estimation attempts with T315A in the presence of imatinib, we find that ΔΔG estimations are highly inaccurate; Both physics-based sampling methods, RepEx and NEQ, fail to predict the correct sign of ΔΔG upon the T315A transformation (Fig. 7, Supp Table 2-3). Given this broad inability across multiple methods to accurately predict the correct sign of ΔΔG, the source of consistent error is likely related to some shared input or parameter that drives this error. In turn, we investigated the starting structure provided for Abl-imatinib for running ΔΔG predictions, since that is a common starting point for each of these protocols.

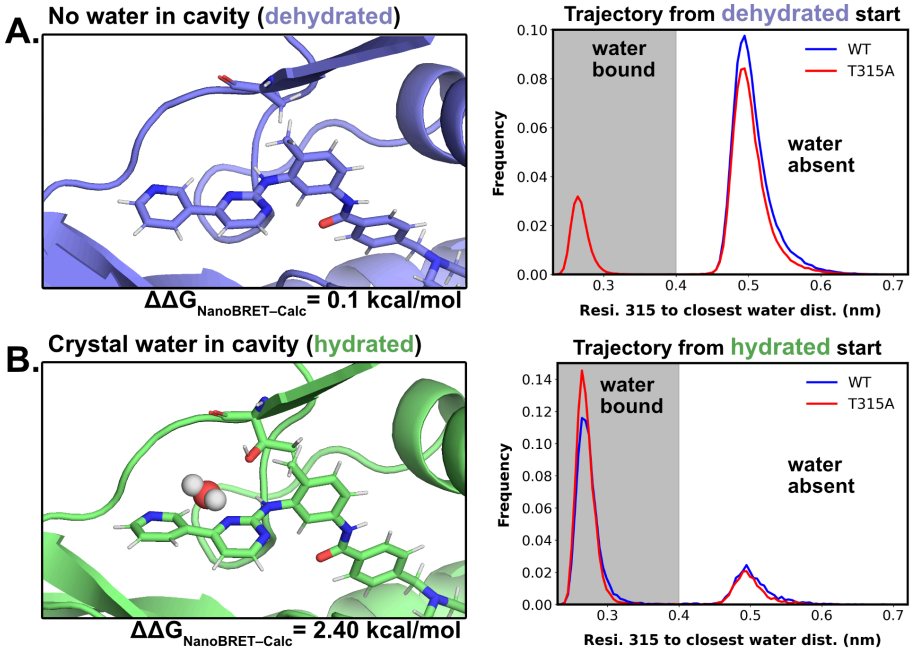


Figure 7. Starting configuration details such as water placement, play a major role in ΔΔG accuracy. A. Structure of a dehydrated T315 residue (left, blue) that yields a low error in ΔΔG prediction between physics-based simulations from Perses and PMX (below image). Equilibrium MD trajectories (top right) started from this dehydrated structure for both WT (red) and the T315A mutant (blue) show that the closest water molecule rarely comes within 4 Å of the residue (grey region), largely keeping the residue away from water (white region). **B.** Conversely, structures started from a hydrated structure (bottom left, green), where the closest water molecule is within 4 Å of residue 315, yields highly inaccurate ΔΔG predictions relative to NanoBRET (below image). Consistently, equilibrium MD trajectories (bottom right) started from the hydrated structure for both WT (red) and T315A (blue) show that the water largely remains bound (grey region) to the molecule for the same number of steps that the alchemical free energy predictions are run for, with little water dissociation occurring (white region).

We observe a water molecule adjacent to T315 that we conjecture is having nonbonded interaction with the polar side chain of threonine (Fig. 7B). However, upon mutation to alanine, one would expect that the water would exit that interacting region due to the lack of interaction. We investigate the exchange of water within this pocket by measuring the distance between residue 315 and the closest water molecule at each frame (Fig. 7).

Based on these structures, we define a hydrated starting structure of Abl kinase as hydrated if the closest water molecule is within 0.4 nm of residue 315. Investigating the hydration status of residue 315 in Perses-NEQ trajectories reveals that the closest water molecule never dissociates about 5 Å away from the binding site (Fig. 7). To test whether this hydrated state is consistently present at equilibrium, we ran equilibrium MD simulation from the hydrated starting structure with both T315 and A315 (Supp. Fig. 9). We observe that the closest water molecule rarely dissociates away from residue 315 (Supp. Fig. 9). While a few transitions are observed where the water molecule dissociates from residue 315 (Supp. Fig. 9), we hypothesize that this lack of water dissociation upon the T315A mutation may be a source of error.

Consistently, we find that ΔΔG estimations that start from dehydrated structures, where the closest water molecule to residue 315 is deleted, improves our predictions (Fig. 7A). Using these dehydrated structures as starting configurations in both Perses and PMX structures, the sign of our predictions aligns with experimental results and reduces the deviation from experiment to 0.10 kcal/mol (Fig. 7A). Consistently, trajectories starting from a dehydrated starting structure only rarely observe a water molecule come within 4 Å of the mutated T315A construct, but a water molecule still does associate with the wild type T315A simulations (Supp. Fig. 9). Interestingly, both simulations starting from dehydrated structures have a degree of mixing; water in the dehydrated trajectories are both interacting with T315 and dissociating (Fig. 7, Supp. Fig. 9). Overall, these findings indicate to us that in the absence of an appropriate starting structure, sufficient sampling of water configurations would enable accurate predictions. As a result, we recommend ensuring sufficient sampling using established metrics.[94,95] As well as using a long enough switching time both to ensure appropriate structural relaxation to enable equilibration of solvent molecules.

For both T315A and L298F, we find that considering alternative protonation states improves ΔΔG estimations (Supp. Table 2-4). Due to resource limitations and default parameters, prospective methods only consider a singular protonation state, imatinib with 0 net charge, which we call imatinib+0. However, imatinib can also exist at physiological pH conditions with a +1 charge, a species we call imatinib+1. Dasatinib does not have multiple likely titratable states at physiological pH however, and so only one species needs to be considered.

For T315A and L298F, prospective methods struggle to obtain consistent ΔΔG predictions for imatinib+0 (Supp. Table 2, 4). Some methods such as FEP+ estimate the correct sign of ΔΔG upon the L298F mutation relative to the experimental NanoBRET measurement (Supp. Table 4). However, prospective evaluation of T315A is much more difficult for all three prospective methods, FEP+, PMX with A99 force field, and PMX with the A14 force field (Supp Table. 2, 4). Prospective ΔΔG measurements are unable to get the correct ΔΔG sign when predicting the impact of the T315A mutation on imatinib+0 binding (Supp table 2, 4).

We find that considering both protonation states of imatinib improves prediction capabilities. Running Perses RepEx methods or NEQ on Folding@home, we estimate ΔΔG for both T315A and L298F mutations for both imatinib protonation states. Using previously established approaches to consider the bulk ΔΔG given multiple protonation states,[46] we compute a ΔΔG estimation in the presence of “bulk” imatinib, considering both protonation states by their propensity to exist at a given pH. Considering both protonation species generates more accurate ΔΔG estimations for T315A (Supp. Table 2), accurately predicting the correct sign of ΔΔG for both species of imatinib and the bulk estimated value. Despite Dasatinib only having a single protonation state, prospective methods obtain much more accurate ΔΔG values for L298F than for T315A (Supp. Table 3). The inaccuracy of ΔΔG predictions in the T315A predictions can be explained by an improper starting configuration with a trapped water molecule adjacent to residue 315, as described above (Fig. 7, Supp. Fig. 7-9). Previous work has shown that considering the protonation state of titratable residues can improve sources of error in ΔΔG estimations.[96,97] In contrast, we find that changing the protonation state of titratable amino acids does not improve ΔΔG accuracy relative to experiment (Supp. Table 5). However, previous work has highlighted the importance of protonation states in titratable residues in ABL1A kinase,[46,79,96] indicating that while protonation state may be important for kinase activity, it may have less of an impact on direct inhibitor binding in the context of these two inhibitors. Overall, our work in light of previous findings highlights the importance of considering the protonation state of both titratable amino acids as well as of the ligand when predicting the impact of mutations.

Overall, using Perses to retrospectively analyze emblematic mutations that are challenging to estimate, we show that there are multiple potential sources in the starting configuration of a system that can introduce large systematic errors in estimations of mutation ΔΔG. These starting configuration errors can arise mainly in two ways: protonation state and through under-sampling of water interaction networks. Protonation states of both ligand and titratable amino acids must be considered when comparing with experiment, as it is important to consider both sources of variation in protonation states. This includes considering both tautomers as well as altered charge species of ligands and amino acids. We also show that appropriately sampling water networks is critical, as alchemical transformations can perturb these water networks without giving them time to appropriately relax and equilibrate. This can in turn introduce large errors in ΔΔG estimation. This can be remedied by longer sampling times, larger sampling windows, and ensuring appropriate time to relax between non-equilibrium alchemical transformation steps.

CONCLUSIONS: In this work we highlight the ability of computational physics-based methods to prospectively estimate the thermodynamic cost of mutating an amino acid to predict the ΔΔG of ligand binding upon mutation. Using two different inhibitors of Abl kinase, imatinib and dasatinib, we show that NanoBRET measurements provide a consistent reproducible measurement for experimentally measuring ΔΔG for multiple constructs without having to rely on pooling measurements from a variety of sources and accruing multiple points of bias and error. This demonstrates that NanoBRET is a suitable tool to provide benchmarks for computational predictions of mutation-driven impact, and that this dataset will provide an initial test bed for the benchmarking of future predictive methods (SI data). We evaluate multiple prospective physics- and structure-based methods, alchemical free energy methods sampling via both replica exchange and nonequilibrium switching, Rosetta's flex_ddg protocol, and a Random Forest model trained entirely on prior data. We show that all methods provide reasonably similar prediction accuracy, but that physics-based simulation provides improved minimum prediction accuracy on a per-residue level. Importantly, we emphasize that all structure-based and physics-based methods can reasonably propose whether or not a mutation is significantly resistant or sensitizing with an average accuracy of 0.81. Within this range of accuracy, these methods can act as suitable classifiers to prospectively predict whether a mutation is resistant or sensitizing in the absence of prior data. We also show that physics-based simulations are best able to capture the impact of distal mutations on ΔΔG of inhibitor binding, presumably due to their ability to capture the impacts of mutation on the dynamics of the kinase. Physics-based simulations are also more able to capture long-range electrostatic effects at greater distances, some approaches such as Rosetta and Random Forest models do not consider due to default cutoffs in their implementations. Lastly, we consider emblematically difficult mutations to predict and show that considering either 1) alternative starting configurations to better sample water-interaction networks, or 2) multiple protonation states and charge species, can improve ΔΔG prediction accuracy. By sharing these datasets, both experimental and computational predictions, in an open-source manner, we provide an initial test benchmark for future computational methods to evaluate their ability to predict the impact of resistant and sensitizing mutations. We hope these datasets highlight the utility of these methods and their capacity to classify the mechanistic impact of mutations in data-poor regimes where prior biochemical data is absent.

METHODS:

Lollipop plot generation of clinical variants. All variants were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) ABL1 entry (ENST00000318560) as of 7/28/2023.[5] Counts were aggregated for the 94 variants assessed in this study.

Prospective Free energy calculations:

Free energy calculations with FEP+. System preparation and free energy calculations were performed using Schrödinger Maestro Suite Version 2020-3 (Schrödinger Release 2020-3: Maestro, Prime, FEP+; Schrödinger, LLC: New York, 2020).[98] System preparation used the protein preparation wizard within Maestro and resembled Hauser et. al. but differed in some aspects which will be pointed out below.[31] Residue numbering was adjusted such that the threonine gatekeeper of Abl kinase obtained the residue number 315 to match common practice. As in Hauser et. al. chain B of PDB structure 1OPJ was used as input for the calculations of imatinib in complex with Abl.[71] However, the complete chain as present in the PDB file was used. Solvent exposed serine 336 (355 in PDB file) was mutated to asparagine using Maestro. Termini were capped and all water molecules present in the PDB structure were kept. No loops were missing and needed modeling. Imatinib was modeled as positively charged by the protein preparation wizard. For some of the calculations,

imatinib was manually neutralized. For the simulations of dasatinib in complex with Abl PDB structure 4XEY was used.[99] Specifically, Chain A of the PDB structure (4XEY) was used, but no homology modelling was done to add on unresolved N- and C-terminal residues. Both imatinib and dasatinib ligand parameters are described by the OPLS forcefield.[100] The preparation wizard neutralized aspartates 381 and 421. Input structures for imatinib and dasatinib can be found as a part of the supplementary material and are available online (https://osf.io/s6ktq/). The effect of residue mutation on ligand affinity was calculated using FEP+ with default settings i.e. using the muVT ensemble, a simulation time of 5ns per lambda window and 12, 16, or 24 lambda windows for standard, core hopping, and charge changing perturbations, respectively. Each run was performed in triplicates using different random number seeds. Error estimates correspond to standard deviations over the three runs.

Curation and preparation of structures for Nonequilibrium protocols with PMX (NEQ). Structures of the ABL1A-inhibitor complexes were used as previously described by drawing up on previously published crystal structures (4XEY and 1OPJ for dasatinib and imatinib, respectively).[31] Apo structures were generated by discarding ligand atoms, and crystallographic water molecules were retained. All mutant structures were generated using FoldX v4.[101] Amino acid protonation states were set at pH 7.4 using PDB2PQR and PROPKA v3.1 via the HTMD protein preparation tool (v1.12).[102–105] Ligand protonation states were kept previously described in Hauser et al by default.[31] Proteins were described using the Amber99sb*-ILDN (shown as “A99”) and Amber14sb (shown as “A14sb”) force fields.[106–109] The TIP3P water model was used.[111] Ligands parameters were created using GAFF2 (v2.1) via AmberTools 16[112] where charges were described using restrained electrostatic potential (RESP).[113] Gaussian 09 (Rev D.01) was used to conduct geometry optimizations and molecular electrostatic potential (ESP) calculations using the HF/6-31G* level of theory. Three optimization steps were used to ensure the ligand conformation remained similar to the kinase-bound poses. ESP points were sampled according to the Merz-Kollman scheme.[114,115] Halogen atom σ-holes were modeled as previously described by Kolář and Hobza.[116] All ligand parameters can be found in the input files in (https://osf.io/s6ktq/). Protein-ligand systems were solvated in a dodecahedral box with periodic boundary conditions with a minimum padding distance 12 Å from the protein system to the edge of the box. Sodium and chloride ions were added to neutralize the wild-type (WT) system at a concentration of 0.15 M NaCl. For the mutants, 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. Clashes may also be present in this initial structure as FoldX does not consider the presence of ligands when inserting mutations in the protein. Clashes were 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 alchembed was used to resolve them: 2000 steepest descent minimization steps were conducted, after which, the ligand vdW interactions were switched on for 2000 additional steps in using the MD integrator steps carried out with a 0.5 femtosecond time step with position restraints (at 1000 kJ mol^{-1} nm^{-2}) on all heavy atoms.[117]

Non-equilibrium protocol (NEQ) with PMX.
All simulations were carried out on Gromacs 2016 on Intel Xeon processors with Ivy Bridge (4 cores, E3-1270 v2) or Broadwell (10 cores, E5-2630 v4) architectures and NVIDIA GeForce GPUs (GTX 1070, GTX 1080, or GTX 1080 Ti).[118,119] Energy minimization was carried out using a steepest descent algorithm for 10,000 steps. 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 1000 kJ mol^{-1} nm^{-2}. Equations of motion were integrated with a leap-frog integrator and a time-step of 2 femtoseconds (fs). The temperature was coupled with the stochastic v-rescale thermostat at the target temperature of 300 K.[120] The pressure was controlled with the Berendsen weak coupling algorithm at a target pressure of 1 bar.[121] The particle mesh Ewald (PME) algorithm was used for electrostatic interactions with a real space cut-off of 10 Å when using Amber force fields, a spline order of 4, a relative tolerance of 10^{-5}, and a Fourier spacing of 1.2 Å.[122] Verlet cut-off schemes with the potential-shift modifier was used with a Lennard-Jones interaction cut-off of 10 Å, and a buffer tolerance of 0.005 kJ mol^{-1} ps^{-1}.[123] All bonds were constrained with the P-LINCS algorithm.[124] For equilibration, unrestrained MD simulations were then performed for 1ns in the NPT ensemble with the Parrinello-Rahman barostat at 1 bar with a time constant of 2 ps.[125] Production simulations were then performed for 3 ns for A14 and 5 ns for A99.

For ΔΔG estimations, the above procedure for equilibrium simulations was repeated ten times on both the apo and complex states of both wild-type and mutant to estimate ΔΔG. From each of these ten equilibrium simulations, 30 equally spaced frames were extracted to serve as starting configurations for the non-equilibrium protocol, generating a total of 300 non-equilibrium trajectories. There were 150 trajectories going from wild-type to mutant ("forward") and 150 trajectories going from mutant to wild-type ("reverse") for each mutation. For the A99 protocol, ten repeated equilibrium simulations were used for charge-conserving mutations, and twenty for charge-changing mutations; from these, 400 frames were extracted for charge-conserving mutations, and 800 frames were extracted for charge-changing mutations. The non-interacting ("dummy") atoms for morphing wild-type residues into mutants were introduced via pmx, using the mutant structure proposed by FoldX as a template.[72] Positions of the dummy atoms were then minimized while freezing the rest of the system. These systems, now containing "hybrid" residues, were then simulated for 10 ps to equilibrate velocities.

Finally, amino acid side chains were alchemically morphed at constant speed during non-equilibrium simulations of 80 ps in length for A14 and 100 ps for A99. Work values associated with each non-equilibrium transition were extracted using thermodynamic integration (TI) and then used to estimate the free energy differences with the Bennett's Acceptance Ratio (BAR).[52,126,127] Point estimates of the free energy differences ($\Delta G_{apo,\ forward}$ and $\Delta G_{holo,\ forward}$) were calculated with BAR after pooling all available forward and reverse work values coming from the nonequilibrium trajectories. Uncertainties in $\Delta G_{apo,forward}$ and $\Delta G_{holo,\ forward}$ were estimated as standard errors (σΔG) by considering each equilibrium simulation and the resulting non-equilibrium trajectories as independent calculations. Uncertainties were then propagated to the final ΔΔG estimate to obtain the estimate of the standard error σΔΔG.

Machine learning model deployment:

The machine learning (ML) model was built in Python using the ExtraTreesRegressor class in the scikit-learn library, following the approach similar to Aldeghi et al., with variations in dataset splitting applied to feature selection procedures.[12]

Training dataset curation. The dataset described in Hauser et al. was used for training the model.[31] This dataset contains 144 binding affinity changes (ΔΔG) for eight tyrosine kinase inhibitors (TKIs) due to point mutations in human Abl kinase. 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).[87,99,128–130] Models for mutant apo structures were generated using FoldX (v4). Structures of the mutant complexes were obtained by maintaining the ligand coordinates from the WT structures.

Features and feature selection. A total of 128 features were calculated and considered as inputs of the model: 18 ligand properties (e.g., molecular weight, calculated logP, number of rotatable bonds) were calculated with RDKit (v2018.09.1; https://www.rdkit.org), and 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),[131] 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 features, we also include the change in folding free energy upon mutation as predicted by FoldX v4. Six 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).[132] The Vina binding score, along with 59 Vina features were calculated with AutoDock Vina via scripts that are part of DeltaVina.[133] The latter tool, in conjunction with the molecular surface calculation library MSMS, was also used to calculate 10 pharmacophore-based solvent-accessible surface area (SASA) features.

Feature selection was performed with a greedy algorithm using the `mlxtend` library. We allowed the selection of any number of features up to a maximum of 40, which minimized the mean-squared error (MSE) of 10-fold cross-validation on the training set of 144 published ΔΔG values. The 10 folds were created such that each fold would contain a unique set of mutations. Our feature selection procedure selects those features that would allow the model to better extrapolate to previously unseen point mutations. From this protocol, 10 features

were selected: 1. Change in hydrogen bond acceptor SASA between the WT and mutant complexes, 2. Change in halogen SASA, 3. Number of ligand-mutated residue atom pairs within 2 Å of each other (reporting on whether the mutation might introduce steric clashes), 4. Gain/loss of cation-pi interactions and salt bridges between the WT and mutant complexes, 5. Change in number of aromatic rings for the protein residue, 6. Change in the number of hydrogen bond acceptor atoms for the protein residue, 7. Maximal distance between ligand and WT residue atoms, 8, 9. The 25th and 50th percentiles of the values of the angles between each ligand atom and the beta and alpha carbons of the protein residue (a measure of the side chain orientation with respect to the ligand), 10. number of rotatable bonds in the ligand.

The model was then trained on the full set of 144 ΔΔG values from Hauser et al. using these features.[31] This model was used to predict the ΔΔG values associated with the ABL1A kinase mutations described in this manuscript. The machine learning (ML) model was built in python using the ExtraTreesRegressor class in the scikit-learn library.[134] This model uses ensembles of randomized decision trees in a similar fashion to random forest.[77] The input files and the code (as Jupyter notebooks) used to train and test the ML models are provided online (https://osf.io/s6ktq/). All computations pertaining to the ML results were performed on a desktop machine equipped with an Intel Xeon processor of Broadwell architecture (E5-1630 v4). Splits were done by mutation. We split the Hauser dataset in 10 folds, such that each fold would have a different set of mutations, with the idea of encouraging the selection of feature that provide good extrapolation performance to new mutations (since we knew the ligands would be imatinib and dasatinib, but the mutations would be new).

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 ΔΔG values were resampled with replacement 105 times. For each bootstrap sample, RMSE, R, and AUPRC were calculated. From these 105 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 ΔΔG values were resampled with replacement together 105 times: ΔΔG 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, 105 bootstrap differences (e.g. ΔRMSE) would have been collected. The fraction of differences crossing zero was multiplied by two to provide a two-tailed p-value for the difference observed. Data analysis was performed in python using the numpy, scipy, pandas, scikit-learn, matplotlib, and seaborn libraries.[135–139] All comparative analysis and summary statistics do not consider the mutations included in the dataset used to train the model, to reduce bias caused by data leakage.

Rosetta prediction of mutation impact using flex_ddg. Using the above protocol in the structure preparation and data curation of the PMX calculations, Rosetta binding free energy changes were calculated with Rosetta (v2017.52) using the flex_ddg protocol.[38] These calculations were carried out on cluster nodes equipped with an Intel Xeon processor of S4 Broadwell architecture (E5-2630 v4), using one CPU core per ΔΔG 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 ΔΔG estimates were the average values of the generalized additive model obtained from 35 iterations of the protocol. The command lines used for the Rosetta calculations and the input files can be found online (https://osf.io/s6ktq/).

NanoBRET affinity assay. Bioluminescence resonance energy transfer (BRET) is used as a proximity-based measure of drug binding to kinase targets in live HEK293T cells.[140] BRET is observed between a NanoLuciferase (nLuc) tag on the full-length protein kinase and a tracer molecule (a BODIPY fluorophore attached to an ATP-competitive inhibitor scaffold). Upon binding of either imatinib or dasatinib to the kinase, the tracer is displaced to reduce BRET in a dose-dependent manner. The Abl NanoBRET affinity and residence time assay data used in this manuscript was collected as previously described by Lyczek et al.[24] Briefly, full-length Abl was cloned in frame with an N-terminal NanoLuc fusion, mutagenized, and used to transiently transfect HEK293T cells at a density of 2x105 cells/mL for twenty hours. Transfected cells were incubated with BRET Kinase Tracer K4 (Promega) at the previously measured Tracer IC50 and serially diluted

imatinib and dasatinib in OptiMEM media without phenol red. The system was allowed to equilibrate for two hours at 37ºC and 5% CO2. The 3X Complete Substrate and Inhibitor solution was prepared by mixing NanoBRET Nano-Glo Substrate (Promega) and Extracellular NanoLuc Inhibitor (Promega) into OPTI-MEM. Tracer compound was added to cells using a liquid dispenser at the tracer IC50 concentration for each mutant,[24] followed by addition of serially diluted inhibitor using an Echo 550 and incubated for 2 hours. Afterwards, 3X Complete Substrate and Inhibitor were added to cells and luminescence was measured at 450 nm (donor emission). BRET was measured at multiple serially diluted concentrations of either imatinib or dasatinib and 650 nm (acceptor emission) in a PHERAstar plate reader (BMG Labtech). Background-corrected BRET ratios (610 nm/450 nm) were determined by subtracting the BRET ratios of samples from the BRET ratios in the absence of tracer and inhibitor. BRET ratios were plotted as a function of inhibitor concentration and graphed using GraphPad Prism (v9). IC50 determination was done via curve fitting to:

$$Y = Bottom + \left(\frac{Top - Bottom}{1 + \left(\frac{IC_{50}}{X}\right)^{HillSlope}} \right)$$

where the $HillSlope$ describes the steepness of the curve (fixed to -1.0), and Top and $Bottom$ describe the upper and lower plateaus in the units of the Y-axis, X is the ligand concentration, and Y is the BRET ratio. Apparent affinity values for each compound/mutant pair were calculated using the Cheng-Prusoff equation.[141]

Retrospective analysis with Perses:

Structure preparation and setup for alchemical transformations within Perses. Perses calculations were performed with structures of human wild type ABL1 in complex with imatinib and dasatinib prepared using functionality of OEChem and Spruce from the OpenEye Toolkits 2021.1.1 implemented in the open-source framework KinoML.[11] PDB entries 2HYY (chain C) and 2GQG (chain A) were chosen as the ligands of interest were co-crystalized and due to the high quality score reported by KLIFS.[69,142–144] Unresolved side chains of both structures were modeled with Spruce (OpenEye). The phosphorylated tyrosine of 2GQG at position 393 was altered to a standard tyrosine residue using Spruce. Missing residues of 2HYY, D276 and T389-D391 were built with Spruce. Finally, both structures were protonated at pH 7.4 using OEChem.This resulted in a charge of +1 for imatinib at the piperazine ring. Generated structures and scripts for structure preparation are available on github (https://github.com/openkinome/study-abl-resistance).

Perses hybrid topology setup. The hybrid topology, positions, and system for each transformation were generated using Perses 0.10.1 and OpenMM 8.0.0.[46,145,146] The hybrid topology was generated using a single topology approach. The hybrid positions were assembled by copying the positions of all atoms in the WT ("old") topology and then copying the positions of the atoms unique to the mutant ("new") residue (i.e., unique new atoms). Unique new atom positions were generated using the Perses FFAllAngleGeometryEngine, which probabilistically proposes positions for one atom at a time based on valence energies alone. Further details on hybrid topology, positions, and system generation (including definitions of the valence, electrostatic, and steric energy functions) are available in the Perses "RESTCapableHybridTopologyFactory" class. For charge-changing mutations, counterions were added to neutralize the mutant system by selecting water molecules in the WT system that are initially at least 8 Å from the solute and alchemically transforming the WT water molecules into sodium or chloride ions in the mutant system. For example, if the mutation was ALA→ASP, a water molecule in the WT system was transformed into a sodium ion in the mutant system to keep the system endstate neutral. If the mutation was GLU→ALA, a water molecule in the WT system was transformed into a chloride ion in the mutant system. Additional details on the counterion implementation can be found in the Perses "_handle_charge_changes()" function found in "perses.app.relative_point_mutation_setup". To prevent singularities when turning off the nonbonded interactions involving unique old or unique new atoms, a softcore approach was used that involves "lifting" unique old or unique new interaction distances: A padding distance ($w(\lambda)$) was added to the interaction distances involving unique old or unique new atoms so that the atoms could not be on top of each other.[147] w lifting (the maximum value for $w(\lambda)$) was selected to be 0.3 nm. This lifting softcore approach was applied to both the electrostatic and steric interactions, so multi-stage alchemical protocols (e.g., where electrostatics must be turned off before sterics) were not necessary for scaling on or off the electrostatic and steric interactions. Instead, a linear protocol was used for interpolating the valence, nonbonded, and lifting terms. This softcore approach is similar to traditional softcore approaches with the critical difference being with our handling of the Lennard Jones potential. Our approach uses a lifting distance

($w(\lambda)$) that is independent of σ (the distance at which the Lennard Jones potential energy equals zero),[148,149] whereas prior approaches define the lifting distance as a function of σ. In our approach, the lifting distance was defined to be independent of sigma for ease of implementation.

Replica exchange sampling using Perses (RepEx). Alchemical replica exchange (AREX) simulations were performed using Perses 0.10.1 and OpenMMTools 0.21.5 (https://github.com/choderalab/openmmtools). The alchemical protocol was defined with evenly spaced λ values between 0 and 1. Before AREX was performed, the positions were minimized at each of the alchemical states using the OpenMM LocalEnergyMinimizer with an energy tolerance of 10 kJ/mol. Each AREX cycle consisted of running 250 steps (4 femtosecond timestep) with the OpenMM 8.0.0 LangevinMiddleIntegrator at a temperature of 300 K, a collision rate of 1 picosecond^{-1}, and a constraint tolerance of 1e-6.[150–152] All-to-all replica swaps were attempted every cycle.[50] Replica mixing plots were created using OpenMMTools 0.21.5 (https://github.com/choderalab/openmmtools) to extract the mixing statistics from the AREX trajectories. Default settings were used unless otherwise noted. Full details on the AREX implementation are available online (https://github.com/choderalab/perses-b arnase-barstar-paper/blob/main/scripts/04_run_repex/run_repex.py). For each replica, 5000 cycles (i.e., 5 ns) were run, resulting in 30 ns of sampling per phase per mutation. Replicas mixed well for all mutations, indicating good phase space overlap. 10000 cycles were initially run per replica (10 ns/replica), resulting in 240 ns of sampling per phase per neutral mutation and 360 ns of sampling per phase per charge changing mutation. To improve the accuracy of our predicted free energy differences, the sampled alchemical states were bookended with "virtual endstates," which were not sampled during free energy calculation, but for which reliable estimates of the physical endstates could be robustly produced during analysis.

System equilibration and alchemical sampling using Replica Exchange and Nonequilibrium cycling (NEQ). All systems used the AMBER14SB force field for the protein and GAFF-2.11 for the small-molecule.[106,112] Each system was then equilibrated for 3 nanoseconds (ns) where bonds to hydrogen were constrained with CCMA and Hydrogen Mass Repartitioning (HMR) with hydrogen masses set to 4 amu) was applied to allow for a 4 femtosecond (fs) timestep.[153] A nonbonded cutoff of 1.1 nm was used for Lennard-Jones 12-6 interactions. Particle mesh Ewald (PME) was applied for treatment of long-range interactions with a direct-space cutoff of 1 nm, relative error tolerance of 0.0005, and automatic (default) selection of alpha and grid spacing.[122] Alchemical Replica Exchange (RepEx) was performed on Perses using multiple swapping cycles to mix alchemical configurations.[46] Each RepEx cycle consisted of running 250 steps (1 picosecond at a 4 femtosecond timestep) using the Leapfrog Langevin Integrator with a BAOAB splitting at a temperature of 300 K, a collision rate of 1 picosecond^{-1}, and a constraint tolerance of 1e-6.[150,154] All-to-all replica swaps were attempted every cycle. This set of cycles was repeated for every mutation in both complex (inhibitor-bound) and apo (inhibitor-absent) constructs of Abl kinase with both Dasatinib and Imatinib complex structures. Replica mixing plots were created using OpenMMTools 0.21.5 (https://github.com/choderalab/openmmtools) to extract the mixing statistics from the AREX trajectories. Default settings were used unless otherwise noted. Full details on AREX implementation are available on github (https://github.com/choderalab/perses). Non-equilibrium cycles (NEQs) are independently collected on Folding@home where each individual cycle serves as an independent statistical replicate in free energy estimation.[39,91,155] A single cycle consisted of 4 stages each of which lasted 1.5ns. An initial equilibration at λ = 0 was first run, followed by a forward non-equilibrium process which drives λ from 0 to 1 over 100 equally spaced windows, followed by another equilibrium simulation at λ = 1. Lastly, a reverse non-equilibrium process driving λ from 1 to 0 was run over 100 equally spaced windows across another 1.5ns. All cycles were run using a Leapfrog Langevin Integrator with BAOAB splitting at a collision rate of 1 picosecond^{-1} and a constraint tolerance of 1e-8, leading to a total trajectory time of 6ns per cycle. 100 replicate cycles were collected per mutation in both complex (inhibitor-bound) and apo (inhibitor-absent) on Folding@home to obtain forward and reverse work distributions. For both RepEx and NEQ sampling using Perses, free energy differences were calculated using the Multistate Bennett Acceptance Ratio.[156]

ACKNOWLEDGEMENTS AND FUNDING: We want to thank the citizen-scientists of Folding@home for donating their computing resources for the Perses Free Energy estimations (projects 17606–17630). This work used resources from the High-Performance Computing Group at Memorial Sloan Kettering Cancer Center. The authors are grateful to the MSKCC DigITs and HPC team, especially Jamie Cheong, Lohit Valleru, and Monica Chakradeo for their assistance with high-performance computing resources. SS is a Damon Runyon

Quantitative Biology Fellow from the Damon Runyon Cancer Research Foundation (DRQ-14-22) and acknowledges support from a NCI Pathway to Independence Award for Outstanding Early-Stage Postdoctoral Researchers (NCI K99 CA286801). MA was supported by a Research Fellowship of the Alexander von Humboldt Foundation. JDC acknowledges funding from the National Institutes of Health (R35GM152017 and P30CA008748). DS acknowledges financial support from Bayer AG. AMR acknowledges funding from the National Institutes of Health (F30 CA260771). JDC, DS, and AV acknowledge funding from the Stiftung Charité and the BIH Einstein Foundation. MAS acknowledges funding from the National Institutes of Health (R35GM119437).

DISCLOSURES: CDC, JPB, MA are employees of Bayer AG. CDC and MA are shareholders of Bayer AG. DS is an employee of Nuvisan. VG is an employee of Janssen pharmaceuticals and may own equity. JDC is a current member of the Scientific Advisory Board of OpenEye Scientific Software and Founder and CEO of Achira, and has equity interests in Achira and PICO Therapeutics. As a scientific advisor or invited speaker, he has received consulting or speaking fees from Abbvie, Astex, Boehringer-Ingelheim, Blueprint Medicines, Celgene, Foresite Capital, Foresite Labs, Interline Therapeutics, MPM Capital, OpenEye Scientific, Schrödinger, and Ventus Therapeutics. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, Vir Biotechnology, Bayer, XtalPi, Interline Therapeutics, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open Force Field Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, and the Sloan Kettering Institute. A complete funding history for the Chodera lab can be found at http://choderalab.org/funding.

DISCLAIMER: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

DATA AVAILABILITY AND SUPPLEMENTARY DATA: All input files, data outputs, and relevant structural files for prospective predictions, reported NanoBRET data, analysis notebooks, and mutation counting are available on OSF (https://osf.io/s6ktq/). Input and analysis scripts, structures, and notebooks for the retrospective analysis via Perses are also available in the same OSF repository (https://osf.io/s6ktq/). The full volume of Perses files, containing simulation data, work values, and hybrid topology factories are available freely upon request and require 1.9 TB of storage space. As such, the files are available upon request without question but are not available online due to the lack of a suitable online storage repository. All methods described above are open-source except FEP+, which is not open-source, and Rosetta, which is free to non-commercial users. Source code is available for GROMACS (https://www.gromacs.org/), PMX (https://github.com/deGrootLab/pmx), Perses (https://github.com/choderalab/perses), OpenMM (https://openmm.org/), and OpenMMTools (https://github.com/choderalab/openmmtools).

AUTHOR CONTRIBUTIONS:

Conceptualization: JDC, MAS
Methodology: JDC, MAS
Investigation: CDC, SS, DS, VG, MA, JB, JG, WG, SH
Writing -- Original Draft: SS, JB, JS,
Writing -- Review & Editing: SS, CDC, VG, DS, AV
Funding Acquisition: JDC, MAS, AV
Resources: JDC, MAS, BdG
Supervision: BdG, CDC, JDC, MAS, AV

REFERENCES:

1. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors: A 2024 update. *Pharmacological Research* **200**, 107059 (2024).

2. Chen, A. *et al.* Tumor Genomic Profiling Practices and Perceptions: A Survey of Physicians Participating in the NCI-MATCH Trial. *JCO Precis Oncol* 1207–1216 (2020) doi:10.1200/PO.20.00217.

3. Nussinov, R., Jang, H., Nir, G., Tsai, C.-J. & Cheng, F. Open Structural Data in Precision Medicine. *Annual Review of Biomedical Data Science* **5**, 95–117 (2022).

4. O'Dwyer, P. J. *et al.* The NCI-MATCH trial: lessons for precision oncology. *Nat Med* **29**, 1349–1357 (2023).

5. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).

6. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology* 1–16 (2017) doi:10.1200/PO.17.00011.

7. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* **49**, 170–174 (2017).

8. Prasad, V. Perspective: The precision-oncology illusion. *Nature* **537**, S63–S63 (2016).

9. Flaherty, K. T. *et al.* Molecular Landscape and Actionable Alterations in a Genomically Guided Cancer Clinical Trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH). *JCO* **38**, 3883–3894 (2020).

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

11. Castro, R. L.-R. de *et al.* Lessons learned during the journey of data: from experiment to model for predicting kinase affinity, selectivity, polypharmacology, and resistance. 2024.09.10.612176 Preprint at https://doi.org/10.1101/2024.09.10.612176 (2024).

12. Aldeghi, M., Gapsys, V. & de Groot, B. L. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS Cent. Sci.* **5**, 1468–1474 (2019).

13. Reungwetwattana, T. & Ou, S.-H. I. MET exon 14 deletion (METex14): finally, a frequent-enough actionable oncogenic driver mutation in non-small cell lung cancer to lead MET inhibitors out of "40 years of wilderness" and into a clear path of regulatory approval. *Transl Lung Cancer Res* **4**, 820–824 (2015).

14. Wang, Q. & Cheng, T. Evidences for mutations in the histone modifying gene SETD2 as critical drivers in leukemia development. *Sci. China Life Sci.* **57**, 944–946 (2014).

15. Pecci, F. *et al.* Activating point mutations in the MET kinase domain represent a unique molecular subset of lung cancer and other malignancies targetable with MET inhibitors. *Cancer Discovery* (2024) doi:10.1158/2159-8290.CD-23-1217.

16. Azam, M., Seeliger, M. A., Gray, N. S., Kuriyan, J. & Daley, G. Q. Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nat Struct Mol Biol* **15**, 1109–1118 (2008).

17. An, L. *et al.* Defining the Sensitivity Landscape of 74,389 EGFR Variants to Tyrosine Kinase Inhibitors. 2021.07.18.452818 Preprint at https://doi.org/10.1101/2021.07.18.452818 (2021).

18. Outhwaite, I. R. *et al.* Death by a thousand cuts through kinase inhibitor combinations that maximize selectivity and enable rational multitargeting. *eLife* **12**, e86189 (2023).

19. Luebeck, J. *et al.* Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. *Nature* **616**, 798–805 (2023).

20. Nathanson, D. A. *et al.* Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science* **343**, 72–76 (2014).

21. Patterson, S. E., Statz, C. M., Yin, T. & Mockus, S. M. Abstract 2600: Analysis of drug resistance mechanisms and strategies for overcoming resistance in cancer therapy using a curated clinical knowledgebase. *Cancer Research* **77**, 2600–2600 (2017).

22. Persky, N. S. *et al.* Defining the landscape of ATP-competitive inhibitor resistance residues in protein kinases. *Nat Struct Mol Biol* **27**, 92–104 (2020).

23. Brown, C. A. *et al.* Antagonism between substitutions in β-lactamase explains a path not taken in the evolution of bacterial drug resistance. *Journal of Biological Chemistry* **295**, 7376–7390 (2020).

24. Lyczek, A. *et al.* Mutation in Abl kinase with altered drug-binding kinetics indicates a novel mechanism of imatinib resistance. *Proceedings of the National Academy of Sciences* **118**, e2111451118 (2021).

25. Nolen, B., Taylor, S. & Ghosh, G. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol Cell* **15**, 661–675 (2004).

26. Meng, Y., Pond, M. P. & Roux, B. Tyrosine Kinase Activation and Conformational Flexibility: Lessons from Src-Family Tyrosine Kinases. *Accounts of Chemical Research* **50**, 1193–1201 (2017).

27. Xie, T., Saleh, T., Rossi, P. & Kalodimos, C. G. Conformational states dynamically populated by a kinase determine its function. *Science* **370**, (2020).

28. Tong, M. & Seeliger, M. A. Targeting Conformational Plasticity of Protein Kinases. *ACS Chem. Biol.* **10**, 190–200 (2015).

29. Yun, C.-H. *et al.* The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2070 (2008).

30. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol* **26**, 127–132 (2008).

31. Hauser, K. *et al.* Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Commun Biol* **1**, 1–14 (2018).

32. Hayes, T. K. *et al.* Comprehensive mutational scanning of EGFR reveals TKI sensitivities of extracellular domain mutants. *Nat Commun* **15**, 2742 (2024).

33. Hanson, S. M. *et al.* What Makes a Kinase Promiscuous for Inhibitors? *Cell Chemical Biology* **26**, 390-399.e5 (2019).

34. Hari, S. B., Merritt, E. A. & Maly, D. J. Sequence Determinants of a Specific Inactive Protein Kinase Conformation. *Chemistry & Biology* **20**, 806–815 (2013).

35. Morando, M. A. *et al.* Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Scientific reports* **6**, 24439–9 (2016).

36. Gapsys, V. *et al.* Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **11**, 1140–1152 (2020).

37. Novack, D., Zhang, S. & Voelz, V. A. Massively Parallel Free Energy Calculations for in silico Affinity Maturation of Designed Miniproteins. 2024.05.17.594758 Preprint at https://doi.org/10.1101/2024.05.17.594758 (2024).

38. Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* **122**, 5389–5399 (2018).

39. Singh, S. & Hanson, S. Running and analyzing massively parallel molecular simulations. Preprint at https://doi.org/10.26434/chemrxiv-2024-vwpww (2024).

40. Singh, S., Sun, X., Blumer, K. & Bowman, G. R. Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. *eLife* **7**, (2018).

41. Knoverek, C. R. *et al.* Opening of a cryptic pocket in β-lactamase increases penicillinase activity. *PNAS* **118**, (2021).

42. Keddy, C. *et al.* Resistance profile and structural modeling of next-generation ROS1 tyrosine kinase inhibitors. *Mol Cancer Ther* **21**, 336–346 (2022).

43. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucl Acids Res* **44**, D1045–D1053 (2016).

44. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2012).

45. Khalak, Y., Tresadern, G., Hahn, D. F., de Groot, B. L. & Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J. Chem. Theory Comput.* **18**, 6259–6270 (2022).

46. Zhang, I. *et al.* Identifying and Overcoming the Sampling Challenges in Relative Binding Free Energy Calculations of a Model Protein:Protein Complex. *J. Chem. Theory Comput.* **19**, 4863–4882 (2023).

47. Mey, A. S. J. S. *et al.* Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living Journal of Computational Molecular Science* **2**, 18378–18378 (2020).

48. Meng, Y., Sabri Dashti, D. & Roitberg, A. E. Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations. *J. Chem. Theory Comput.* **7**, 2721–2727 (2011).

49. Gallicchio, E., Levy, R. M. & Parashar, M. Asynchronous replica exchange for molecular simulations. *Journal of Computational Chemistry* **29**, 788–794 (2008).

50. Chodera, J. D. & Shirts, M. R. Replica exchange and expanded ensemble simulations as Gibbs sampling: Simple improvements for enhanced mixing. *The Journal of Chemical Physics* **135**, 194110 (2011).

51. Meng, Y., Sabri Dashti, D. & Roitberg, A. E. Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations. *J. Chem. Theory Comput.* **7**, 2721–2727 (2011).

52. Shirts, M. R., Bair, E., Hooker, G. & Pande, V. S. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Phys. Rev. Lett.* **91**, 140601 (2003).

53. Crooks, G. E. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *Journal of Statistical Physics* **90**, 1481–1487 (1998).

54. Dellago, C. & Hummer, G. Computing Equilibrium Free Energies Using Non-Equilibrium Molecular Dynamics. *Entropy* **16**, 41–61 (2014).

55. 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.* n/a-n/a (2016) doi:10.1002/anie.201510054.

56. Aldeghi, M., de Groot, B. L. & Gapsys, V. Accurate Calculation of Free Energy Changes upon Amino Acid

Mutation. in *Computational Methods in Protein Evolution* (ed. Sikosek, T.) 19–47 (Springer, New York, NY, 2019). doi:10.1007/978-1-4939-8736-8_2.

57. Gozgit, J. M., Schrock, A., Chen, T.-H., Clackson, T. & Rivera, V. M. Comprehensive Analysis Of The In Vitro Potency Of Ponatinib, and All Other Approved BCR-ABL Tyrosine Kinase Inhibitors (TKIs), Against a Panel Of Single and Compound BCR-ABL Mutants. *Blood* **122**, 3992 (2013).

58. Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* **358**, eaan4368 (2017).

59. O'Hare, T., Eide, C. A. & Deininger, M. W. N. Bcr-Abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia. *Blood* **110**, 2242–2249 (2007).

60. Soverini, S. *et al.* Contribution of ABL Kinase Domain Mutations to Imatinib Resistance in Different Subsets of Philadelphia-Positive Patients: By the GIMEMA Working Party on Chronic Myeloid Leukemia. *Clinical Cancer Research* **12**, 7374–7379 (2006).

61. Landrum, G. A. & Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).

62. Hanson, S. M., Ekins, S. & Chodera, J. D. Modeling error in experimental assays using the bootstrap principle: understanding discrepancies between assays using different dispensing technologies. *Journal of Computer-Aided Molecular Design* **29**, 1073–1086 (2015).

63. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

64. van Staveren, W. C. G. *et al.* Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1795**, 92–103 (2009).

65. Yang, X. *et al.* Development of Cell Permeable NanoBRET Probes for the Measurement of PLK1 Target Engagement in Live Cells. 2023.02.25.529946 Preprint at https://doi.org/10.1101/2023.02.25.529946 (2023).

66. Single tracer-based protocol for broad-spectrum kinase profiling in live cells with NanoBRET. https://star-protocols.cell.com/protocols/1010.

67. Ong, L. L. *et al.* A High-Throughput BRET Cellular Target Engagement Assay Links Biochemical to Cellular Activity for Bruton's Tyrosine Kinase. *SLAS Discovery* **25**, 176–185 (2020).

68. Robers, M. B. *et al.* Target engagement and drug residence time can be observed in living cells with BRET. *Nat Commun* **6**, 10091 (2015).

69. Tokarski, J. S. *et al.* The structure of Dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. *Cancer Res* **66**, 5790–5797 (2006).

70. Azam, M., Latek, R. R. & Daley, G. Q. Mechanisms of Autoinhibition and STI-571/Imatinib Resistance Revealed by Mutagenesis of BCR-ABL. *Cell* **112**, 831–843 (2003).

71. Nagar, B. *et al.* Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. *Cell* **112**, 859–871 (2003).

72. Gapsys, V., Michielssens, S., Seeliger, D. & de Groot, B. L. pmx: Automated protein structure and topology generation for alchemical perturbations. *J Comput Chem* **36**, 348–354 (2015).

73. Engelberger, F., Zakary, J. D. & Künze, G. Guiding protein design choices by per-residue energy breakdown analysis with an interactive web application. *Frontiers in Molecular Biosciences* **10**, 1178035 (2023).

74. Valanciute, A. *et al.* Accurate protein stability predictions from homology models. *Computational and*

Structural Biotechnology Journal **21**, 66–73 (2023).

75. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. *eLife* **12**, e82593 (2023).

76. Høie, M. H., Cagiada, M., Beck Frederiksen, A. H., Stein, A. & Lindorff-Larsen, K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Reports* **38**, 110207 (2022).

77. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach Learn* **63**, 3–42 (2006).

78. Shan, Y. *et al.* How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **133**, 9181–9183 (2011).

79. Shan, Y. *et al.* A conserved protonation-dependent switch controls drug binding in the Abl kinase. *Proceedings of the National Academy of Sciences* **106**, 139–144 (2009).

80. Ayaz, P. *et al.* Structural mechanism of a drug-binding process involving a large conformational change of the protein target. *Nat Commun* **14**, 1885 (2023).

81. Meng, Y. *et al.* Predicting the Conformational Variability of Abl Tyrosine Kinase using Molecular Dynamics Simulations and Markov State Models. *J. Chem. Theory Comput.* **14**, 2721–2732 (2018).

82. Meng, Y., Lin, Y. & Roux, B. Computational Study of the “DFG-Flip” Conformational Transition in c-Abl and c-Src Tyrosine Kinases. *J Phys Chem B* **119**, 1443–1456 (2015).

83. Aiba, H., Nakamura, T., Mitani, H. & Mori, H. Mutations that alter the allosteric nature of cAMP receptor protein of Escherichia coli. *The EMBO Journal* **4**, 3329–3332 (1985).

84. Brosey, C. A. *et al.* Defining NADH-Driven Allostery Regulating Apoptosis-Inducing Factor. *Structure* (2016) doi:10.1016/j.str.2016.09.012.

85. Horn, J. R. & Shoichet, B. K. Allosteric inhibition through core disruption. *Journal of Molecular Biology* **336**, 1283–1291 (2004).

86. Cortina, G. A. & Kasson, P. M. Predicting allostery and microbial drug resistance with molecular simulations. *Current opinion in structural biology* **52**, 80–86 (2018).

87. Pemovska, T. *et al.* Axitinib effectively inhibits BCR-ABL1(T315I) with a distinct binding conformation. *Nature* **519**, 102–105 (2015).

88. Chan, W. W. *et al.* Conformational Control Inhibition of the BCR-ABL1 Tyrosine Kinase, Including the Gatekeeper T315I Mutant, by the Switch-Control Inhibitor DCC-2036. *Cancer Cell* **19**, 556–568 (2011).

89. Ford, M. C. & Babaoglu, K. Examining the Feasibility of Using Free Energy Perturbation (FEP+) in Predicting Protein Stability. *Journal of Chemical Information and Modeling* **57**, 1276–1285 (2017).

90. Cui, G. Affinity Predictions with FEP+: A Different Perspective on Performance and Utility. (2016).

91. Voelz, V. A., Pande, V. S. & Bowman, G. R. Folding@home: Achievements from over 20 years of citizen science herald the exascale era. *Biophysical Journal* (2023) doi:10.1016/j.bpj.2023.03.028.

92. Perner, F. *et al.* MEN1 mutations mediate clinical resistance to menin inhibition. *Nature* 1–7 (2023) doi:10.1038/s41586-023-05755-9.

93. Boby, M. L. *et al.* Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science* **382**, eabo7201 (2023).

94. Grossfield, A. *et al.* Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular

Simulations [Article v1.0]. *Living J Comput Mol Sci* **1**, 5067 (2018).

95. Grossfield, A. & Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annu Rep Comput Chem* **5**, 23–48 (2009).

96. Yang, M., Jiang, X. & Jiang, N. Protonation state and free energy calculation of HIV-1 protease–inhibitor complex based on electrostatic polarisation effect. *Molecular Physics* **112**, 1659–1669 (2014).

97. Hernández González, J. E. & de Araujo, A. S. Alchemical Calculation of Relative Free Energies for Charge-Changing Mutations at Protein–Protein Interfaces Considering Fixed and Variable Protonation States. *J. Chem. Inf. Model.* **63**, 6807–6822 (2023).

98. Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Commun Chem* **6**, 1–12 (2023).

99. Lorenz, S., Deng, P., Hantschel, O., Superti-Furga, G. & Kuriyan, J. Crystal structure of an SH2–kinase construct of c-Abl and effect of the SH2 domain on kinase activity. *Biochemical Journal* **468**, 283–291 (2015).

100. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).

101. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Research* **33**, W382 (2005).

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

103. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **27**, 112–128 (2018).

104. 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.* **7**, 2284–2295 (2011).

105. 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.* **7**, 525–537 (2011).

106. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).

107. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *The Journal of Physical Chemistry B* **113**, 9004–9015 (2009).

108. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

109. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

110. 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* **118**, 6438–6446 (2014).

111. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983).

112. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a

general amber force field. *Journal of Computational Chemistry* **25**, 1157–1174 (2004).

113. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).

114. Besler, B. H., Merz Jr., K. M. & Kollman, P. A. Atomic charges derived from semiempirical methods. *Journal of Computational Chemistry* **11**, 431–439 (1990).

115. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **5**, 129–145 (1984).

116. Kolář, M. & Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **8**, 1325–1333 (2012).

117. 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. *Journal of Chemical Theory and Computation* **11**, 2743 (2015).

118. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *Journal of computational chemistry* **26**, 1701–1718 (2005).

119. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

120. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).

121. 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. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).

122. Essmann, U. *et al.* A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577–8593 (1995).

123. Páll, S. & Hess, B. A flexible algorithm for calculating pair interactions on SIMD architectures. *Computer Physics Communications* **184**, 2641–2650 (2013).

124. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 116–122 (2008).

125. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics* **52**, 7182–7190 (1981).

126. Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22**, 245–268 (1976).

127. Crooks, G. E. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E* **61**, 2361–2366 (2000).

128. Levinson, N. M. & Boxer, S. G. Structural and Spectroscopic Analysis of the Kinase Inhibitor Bosutinib and an Isomer of Bosutinib Binding to the Abl Tyrosine Kinase Domain. *PLOS ONE* **7**, e29828 (2012).

129. Weisberg, E. *et al.* Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer Cell* **7**, 129–141 (2005).

130. Zhou, T. *et al.* Structural Mechanism of the Pan-BCR-ABL Inhibitor Ponatinib (AP24534): Lessons for Overcoming Kinase Inhibitor Resistance. *Chemical Biology & Drug Design* **77**, 1–11 (2011).

131. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

132. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research* **43**, W443 (2015).

133. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455–461 (2010).

134. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

135. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

136. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

137. team, T. pandas development. pandas-dev/pandas: Pandas. Zenodo https://doi.org/10.5281/zenodo.13819579 (2024).

138. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

139. Waskom, M. seaborn: statistical data visualization. *JOSS* **6**, 3021 (2021).

140. Robers, M. B. *et al.* Quantifying Target Occupancy of Small Molecules Within Living Cells. *Annu Rev Biochem* **89**, 557–581 (2020).

141. Yung-Chi, C. & Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology* **22**, 3099–3108 (1973).

142. Cowan-Jacob, S. W. *et al.* Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Crystallogr D Biol Crystallogr* **63**, 80–93 (2007).

143. van Linden, O. P. J., Kooistra, A. J., Leurs, R., de Esch, I. J. P. & de Graaf, C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space. *J. Med. Chem.* **57**, 249–277 (2014).

144. Kanev, G. K., de Graaf, C., Westerman, B. A., de Esch, I. J. P. & Kooistra, A. J. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Research* **49**, D562–D569 (2021).

145. Rufa, D. A. *et al. Towards Chemical Accuracy for Alchemical Free Energy Calculations with Hybrid Physics-Based Machine Learning / Molecular Mechanics Potentials*. 2020.07.29.227959 https://www.biorxiv.org/content/10.1101/2020.07.29.227959v1 (2020) doi:10.1101/2020.07.29.227959.

146. Eastman, P. *et al.* OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *J. Phys. Chem. B* **128**, 109–116 (2024).

147. Pomès, R., Eisenmesser, E., Post, C. B. & Roux, B. Calculating excess chemical potentials using dynamic simulations in the fourth dimension. *The Journal of Chemical Physics* **111**, 3387–3395 (1999).

148. Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R. & van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters* **222**, 529–539 (1994).

149. Lee, T.-S. *et al.* Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore

Potentials. *J. Chem. Theory Comput.* **16**, 5512–5525 (2020).

150. Fass, J. *et al.* Quantifying Configuration-Sampling Error in Langevin Simulations of Complex Molecular Systems. *Entropy* **20**, 318 (2018).

151. Leimkuhler, B. & Matthews, C. Robust and efficient configurational molecular sampling via Langevin dynamics. *The Journal of Chemical Physics* **138**, 174102 (2013).

152. Leimkuhler, B. & Matthews, C. Efficient molecular dynamics using geodesic integration and solvent–solute splitting. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **472**, 20160138 (2016).

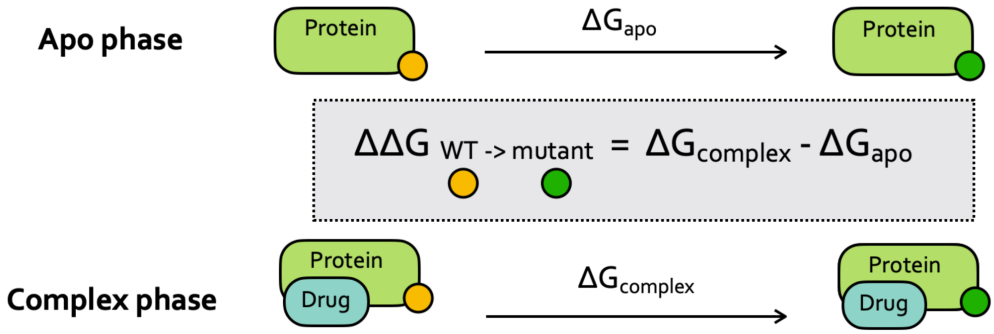
153. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation* **11**, 1864–1874 (2015).

154. Bussi, G. & Parrinello, M. Accurate sampling using Langevin dynamics. *Phys. Rev. E* **75**, 056707 (2007).

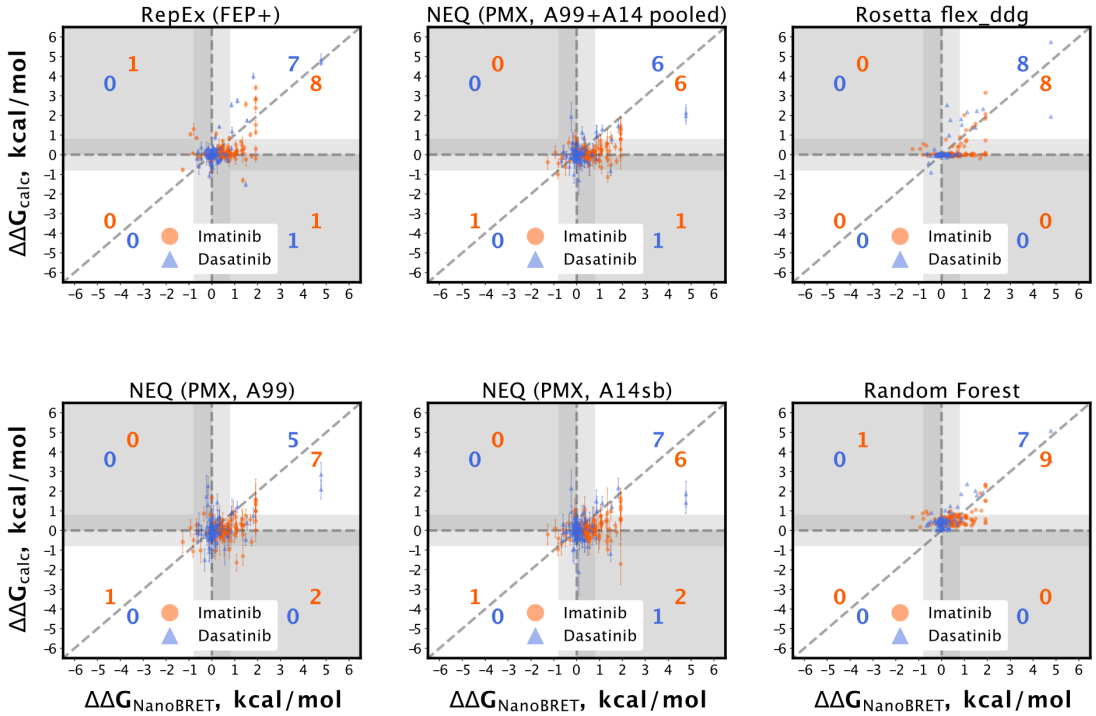
155. Shirts, M. & Pande, V. S. Screen Savers of the World Unite! *Science* **290**, 1903–1904 (2000).

156. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys* **129**, 124105 (2008).

SUPPLEMENTARY FIGURES:



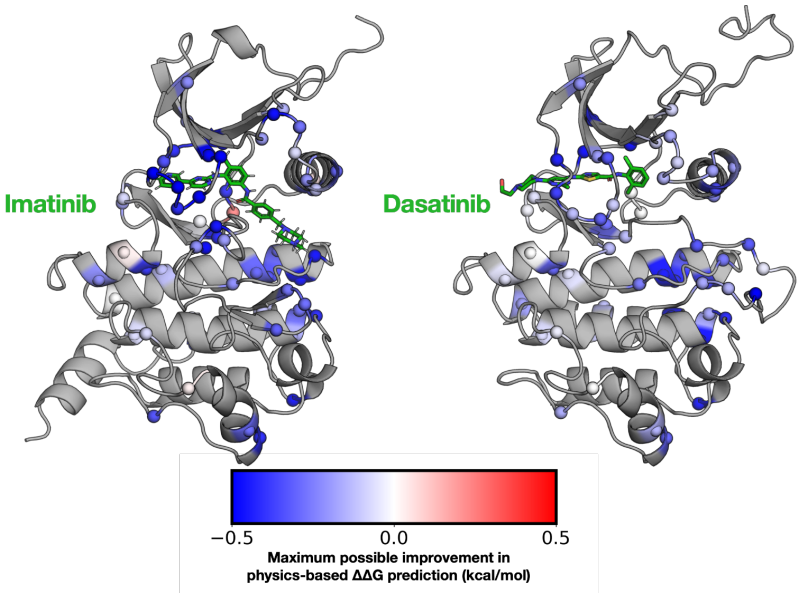
Supp. Figure 1. Representative thermodynamic cycle that highlights the set of transformations used to predict the ΔΔG of mutations upon drug binding in a protein-ligand system. Transformations are computed by computing the energetic cost of mutating a residue from wild type (yellow) to mutant (dark green), in both apo phase (top row) and in complex with drug binding (bottom row). From these calculations, ΔΔG can be computed by subtracting the apo phase (ΔG_{apo}) from the complex phase ($\Delta G_{complex}$).



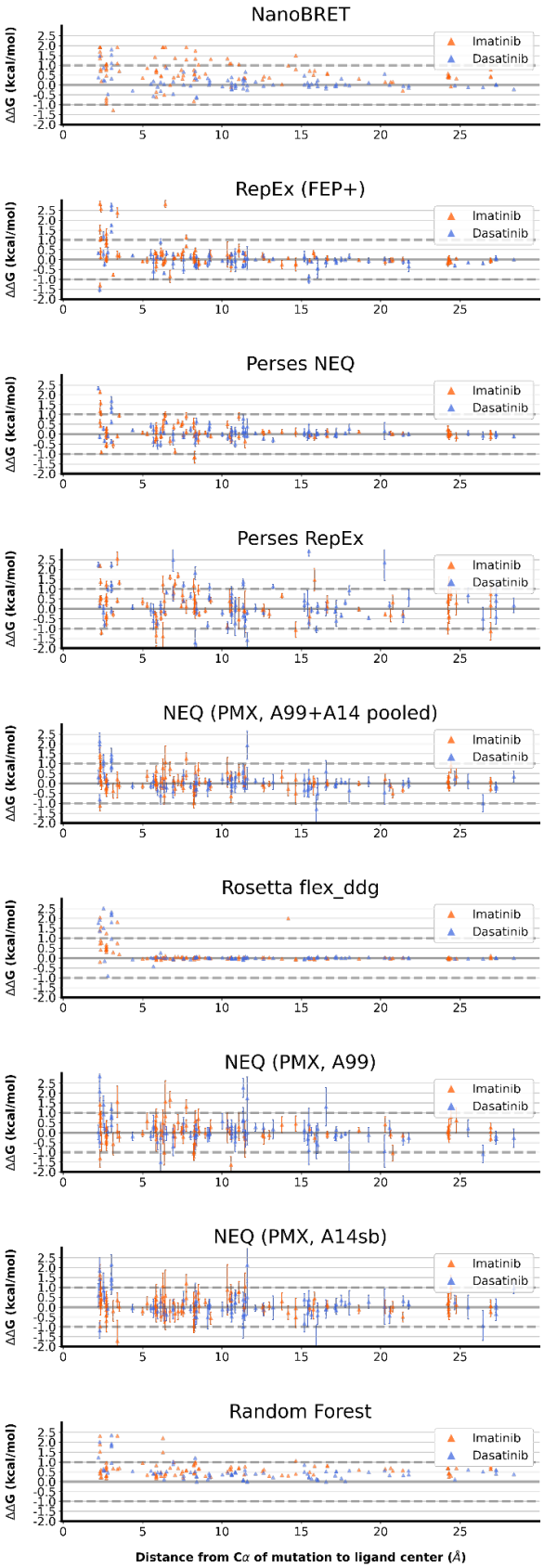
Supp. Figure 2. Truth tables for each method demonstrate the capacity of computational methods to act as classifiers. The number of "true positives" of ΔΔG predictions for a variety of methods are shown (top right and bottom left integers). A prediction is considered a true positive if its has the same sign as the experimental NanoBRET predictions, and both have |ΔΔG| > 1 kcal/mol (top right and bottom left quadrants). True positives for resistant (top right integers) and sensitizing mutations (bottom left integers) are shown for both imatinib (orange) and dasatinib (blue). Experimental ΔΔG measurements with magnitude below 1 kcal/mol are not labeled as "true positives." Prospective methods shown are Replica Exchange using FEP+ (top left), Nonequilibrium switching using PMX using Amber99 force field (bottom left), the Amber14sb force field (bottom middle), and the resultant prediction taken from pooling the work values from both force fields (top middle). Rosetta's flex_ddg (top right) and a random forest model trained on prior data (bottom right) are also shown.

Supp. Table 1. Summary statistics from AUPRC curves highlighting the similarity in performance for each method. The 95% confidence intervals are calculated based on bootstrapping 1000 repeats with replacement

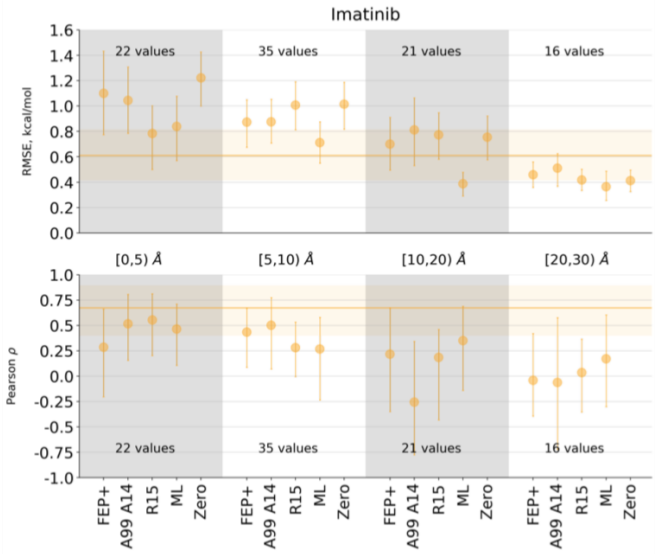
Method	Pooled AUPRC	95% Conf. Interval	Pooled accuracy	Distance from imatinib baseline at 1 kcal/mol	Distance from dasatinib baseline at 1 kcal/mol
FEP+	0.6	0.45–0.73	0.82	0.27	0.68
PMX A99/A14	0.55	0.38–0.71	0.82	0.52	0.57
PMX A99	0.47	0.31–0.65	0.81	0.66	0.63
PMX A14	0.52	0.37–0.68	0.82	0.36	0.46
Rosetta 15	0.6	0.44–0.77	0.85	0.41	0.54
Random forest	0.58	0.44–0.71	0.79	0.19	0.44



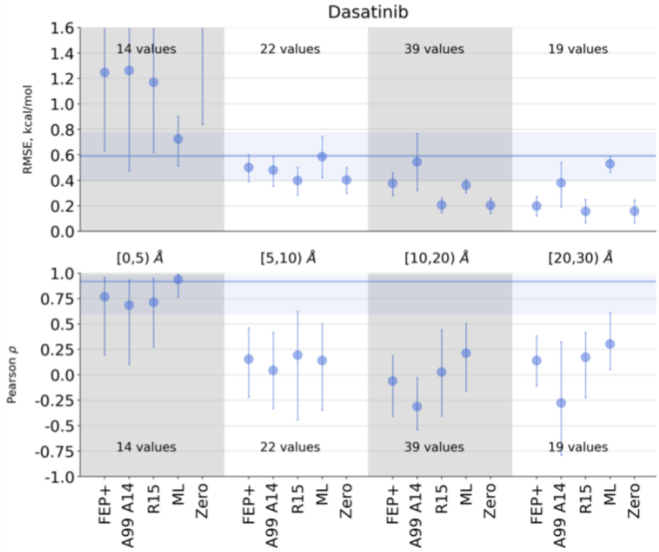
Supp. Figure 3. Maximum possible improvement in ΔΔG prediction by an alchemical simulation method relative to a non-alchemical method. The maximum possible improvement the is computed by taking the least accurate non-alchemical ΔΔG prediction and subtracting it from the most accurate ΔΔG prediction, where a negative score indicates the best possible improvement for any predicted ΔΔG value. These values are mapped onto the structure of Abl kinase (PDB: 1OPJ) for each mutation (spheres), and colored to indicate the degree to which alchemical methods are able to improve upon individual non-alchemical predictions (color scale, below). Mappings are made for both imatinib predictions (left) and dasatinib (right).



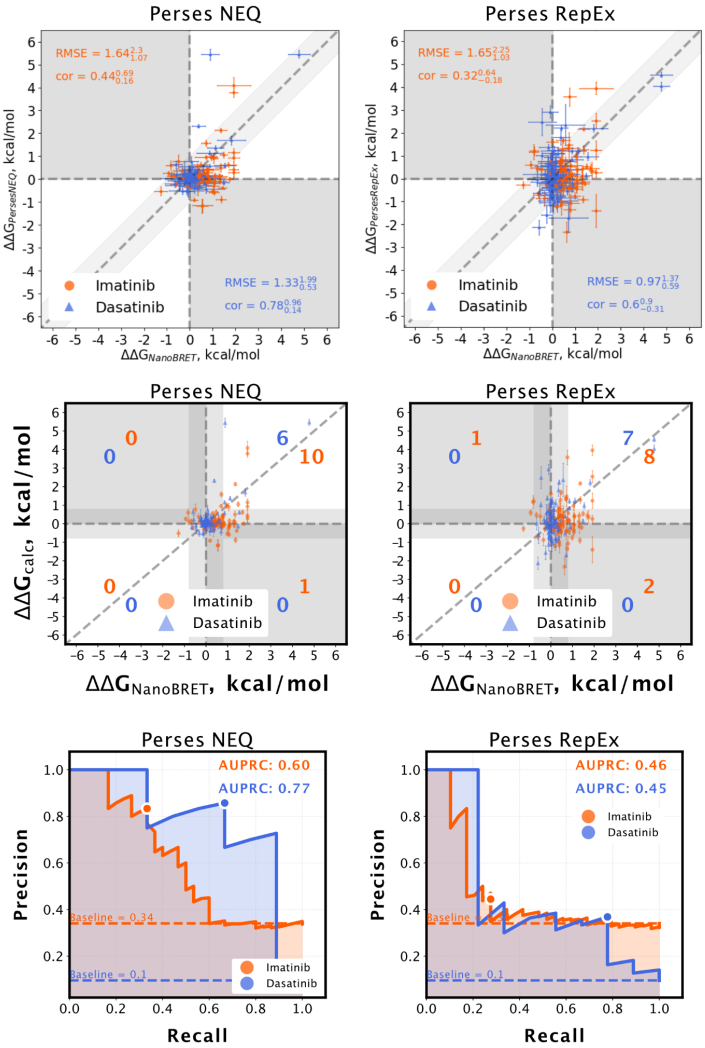
Supp. Figure 4. ΔΔG estimations from both experiment and computation. Computed ΔΔG for NanoBRET and each computational method is plotted for imatinib (orange) and dasatinib (blue) as a function of the distance from the residue's Cα carbon to the center of mass of the ligand in the crystal structure.



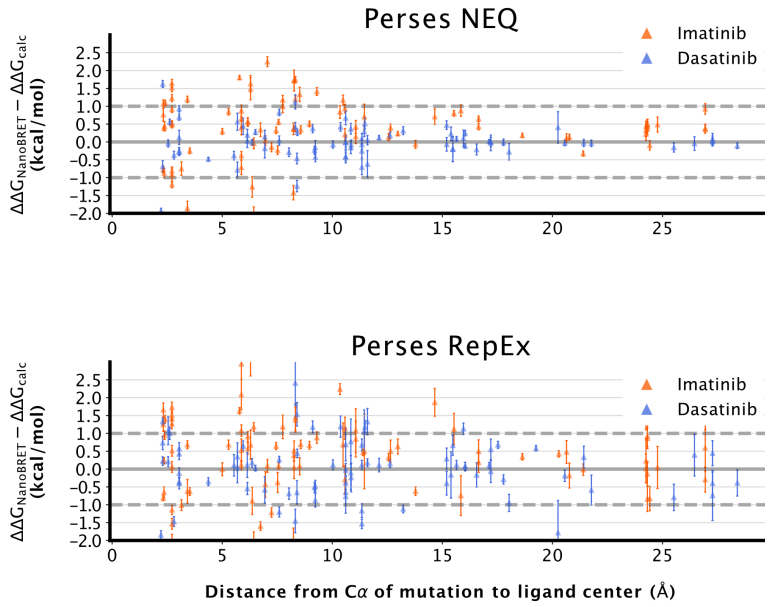
Supp. Figure 5. **Estimation accuracy and summary statistics is dependent on the degree of distance from the active site.** Calculation accuracy grouped in ranges of the distance between the mutated residue and the imatinib. “Zero” denotes a prediction where every mutation is predicted to be neutral in impact on inhibitor binding (ΔΔG = 0 kcal/mol). Horizontal lines mark RMSE and correlation values between two experimental measurements.



Supp. Figure 6: Calculation accuracy grouped in ranges of the distance between the mutated residue and the inhibitor. “Zero” denotes a prediction where every mutation is predicted to be neutral in impact on inhibitor binding (ΔΔG = 0 kcal/mol). Horizontal lines mark RMSE and correlation values between two experimental measurements (Fig. 2).



Supp. Figure 7. Open-source tools like Perses can estimate ΔΔG and act as a classifier while allowing a user to robustly tune parameters and investigate simulation details. Scatterplots (top row) showing predictions made using Perses using both NEQ (left column) and Replica Exchange (RepEx, right column), including both Root Mean Square Error (RMSE) and Pearson correlation (labeled “cor”). These can be converted into truth tables (middle row) that generate Precision-Recall curves (bottom row) for both sampling methods on Perses are also shown for both Imatinib (orange) and dasatinib (blue).



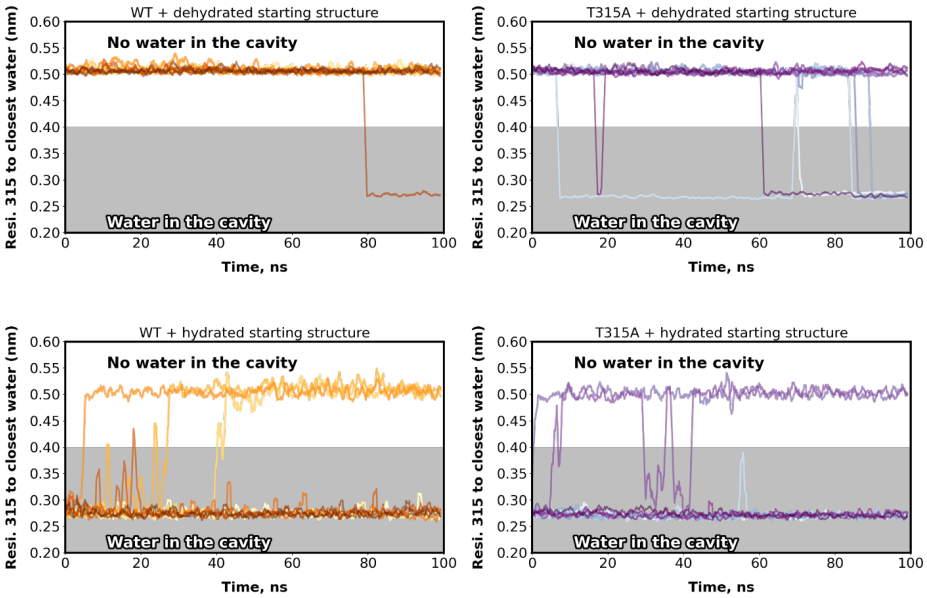
Supp. Figure 8. Perses-based alchemical simulations ΔΔG estimations are also able to predict the impact of distal mutations on imatinib and dasatinib binding. The deviation from predicted ΔΔG to experiment is plotted for imatinib (orange) and dasatinib (blue) as a function of the distance from the residue's Cα carbon to the center of mass of the ligand in the crystal structure. Values are shown for predictions made with Perses using Nonequilibrium Cycling (top) and replica exchange (bottom).

A.

$\Delta\Delta G_{comp}$= 1.30 ± 0.2 kcal/mol
$\Delta\Delta G_{expt}$= 1.40 kcal/mol
Error: 0.10 kcal/mol

B.

$\Delta\Delta G_{comp}$= -1.0 ± 0.3 kcal/mol
$\Delta\Delta G_{expt}$= 1.40 kcal/mol
Error: 2.40 kcal/mol



Supp. Figure 9. Trajectories started with the crystal water removed both increase alchemical ΔΔG prediction and decreased water binding. A. The distance from residue 315 to the closest water molecule in equilibrium MD trajectories (right) started from a crystal structure that removed the closest water molecule to residue 315, rendering the residue in a "dehydrated" state. Trajectories were collected for both wild-type (middle) and T315A constructs (right) Alchemical free energy calculations estimating ΔΔG using this structure yield accurate predictions (left). **B.** Distance from residue 315 to the closest water molecule in equilibrium trajectories starting from the same crystal structure with all water molecules present (middle and right) show increased proximity and potential interaction between residue 315 and water molecule in both WT (middle) and T315A constructs (right). However, alchemical ΔΔG predictions made using this starting structure are far less accurate (left).

Supp. Table 2. Protonation state of imatinib alters ΔΔG predictions for several simulated mutations. Predictions were considered significantly changed when the error in predicted ΔΔG compared to the simulations with positively charged imatinib changed by 0.6 kcal/mol or more. Reported values of ΔΔG calculations are the mean of at least 3 independent repeats. Each cell is colored by how much the prediction deviates from the NanoBRET (bottom row of table).

Mutation	Ligand and net charge	NanoBRET (kcal/mol)	FEP+ (kcal/mol)	a99	a14	PersesNEQ	PersesRepEx
L298F	imatinib+0	**-1.27**	-0.77	-0.58	-0.2	-0.55	0.28
L298F	imatinib+1					-0.79	-2.91
L298F	Imatinib (bulk)					-0.77	-2.86
T315A	imatinib+0	**1.36**	-1.31	-1.31	-0.97	2.01	2.48
T315A	imatinib+1					2.36	2.45
T315A	Imatinib (bulk)					2.30	2.45
		Colorscale for difference between computational prediction vs. NanoBRET measurement: (kcal/mol)	< 0.5	0.5 - 1.0	1.0 - 1.5	> 1.5	

Supp. Table 3. Dasatinib alters ΔΔG predictions for several simulated mutations. Predictions were considered significantly changed when the error in predicted ΔΔG compared to the simulations with positively charged imatinib changed by 0.6 kcal/mol or more. Reported values of ΔΔG calculations are the mean of at least 3 independent repeats. Each cell is colored by how much the prediction deviates from the NanoBRET (bottom row of table).

Mutation	Ligand	NanoBRET (kcal/mol)	FEP+ (kcal/mol)	A99 (kcal/mol)	A14 (kcal/mol)	PersesNEQ (kcal/mol)	PersesRepEx (kcal/mol)
L298F	dasatinib	**-0.31**	0.07	-1.5	0.27	-0.47	0.60
T315A	dasatinib	**1.48**	-1.51	-0.46	-1.19	2.39	2.24
		Colorscale for difference between computational prediction vs. NanoBRET measurement:	< 0.5	0.5 - 1.0	1.0 - 1.5	> 1.5	

Supp Table 4. A neutral protonation state of imatinib alters FEP+ predictions for several simulated mutations. Reported values of FEP+ calculations are the mean of at least 3 independent repeats.

		FEP+ prediction: ΔΔG ± SD (kcal/mol)		
Mutation:	**Experiment value (kcal/mol)**	**Prediction for imatinib^{+1}**	**Prediction for imatinib^{+0}**	**Bulk Imatinib prediction**
Y353H	1.367	0.077 ± 0.059	0.880 ± 0.220	0.113 ± 0.070
F359I	1.068	0.033 ± 0.284	0.733 ± 0.317	0.066 ± 0.286
N368S	0.002	-0.853 ± 0.522	-0.180 ± 1.146	-0.825 ± 0.553
E282G	1.284	0.673 ± 0.090	0.007 ± 0.352	0.582 ± 0.107
E282K	1.920	1.130 ± 0.198	0.173 ± 0.345	0.966 ± 0.208
V289F	0.710	0.430 ± 0.270	2.02 ± 0.435	0.475 ± 0.281
E292V	0.388	0.080 ± 0.227	-0.540 ± 0.220	-0.001 ± 0.225
M351K	1.920	1.920 ± 0.303	0.440 ± 0.422	1.546 ± 0.311
E355G	0.539	-0.017 ± 0.330	-0.740 ± 0.169	-0.120 ± 0.315

Supp Table 5. Alternative side chain protonation states do not improve FEP+ predictions. Predictions were considered significantly changed when the error in predicted DDG compared to the simulations with the default side chain protonation state changed by 0.6 kcal/mol or more. Protonation state of imatinib was +1. Reported values of FEP+ calculations are the mean of at least 3 independent repeats.

		Experimental	**Default protonation state**		**Alternative protonation state**	
Mutation	**Ligand**	**ΔΔG [kcal/mol]**	**Mutation**	**ΔΔG [kcal/mol] ± SD**	**Mutation**	**ΔΔG [kcal/mol] ± SD**
Y253H	imatinib	1.920	TYR253HID	1.517 ± 0.429	TYR253HIE	0.877 ± 0.452
E282K	imatinib	1.920	GLU282LYS	1.130 ± 0.198	GLU282LYN	0.260 ± 0.110
M351K	imatinib	1.920	MET351LYS	1.920 ± 0.303	MET351LYN	0.017 ± 0.442
L248R	dasatinib	0.883	LEU248ARG	1.184 ± 0.468	LEU248ARN	2.230 ± 0.184
G250E	dasatinib	0.187	GLY250GLU	0.028 ± 0.416	GLY250GLH	-0.877 ± 0.685