Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations

Kevin Hauser, Christopher Negron, Steven K. Albanese, Soumya Ray, Thomas Steinbrecher, Robert Abel, John D. Chodera & Lingle Wang

The therapeutic effect of targeted kinase inhibitors can be significantly reduced by intrinsic or acquired resistance mutations that modulate the affinity of the drug for the kinase. In cancer, the majority of missense mutations are rare, making it difficult to predict their impact on inhibitor affinity. We examine the potential for alchemical free-energy calculations to predict how kinase mutations modulate inhibitor affinities to Abl, a major target in chronic myelogenous leukemia (CML). These calculations have useful accuracy in predicting resistance for eight FDA-approved kinase inhibitors across 144 clinically identified point mutations, with a root mean square error in binding free-energy changes of $1.10^{\pm 0.3}$ kcal mol$^{-1}$ (95% confidence interval) and correctly classifying mutations as resistant or susceptible with $88.93\%$ accuracy. This benchmark establishes the potential for physical modeling to collaboratively support the assessment and anticipation of patient mutations to affect drug potency in clinical applications.
Targeted kinase inhibitors are a major therapeutic class in the treatment of cancer. A total of 38 selective small-molecule kinase inhibitors have now been approved by the FDA, including 34 approved to treat cancer, and perhaps 50% of all current drugs in development target kinases. Despite the success of selective inhibitors, the emergence of drug resistance remains a challenge in the treatment of cancer and has motivated the development of second- and then third-generation inhibitors aimed at overcoming recurrent resistance mutations.

While a number of drug resistance mechanisms have been identified in cancer (e.g., induction of splice variants, or alleviation of feedback), inherent or acquired missense mutations in the kinase domain of the target of therapy are a major form of resistance to tyrosine kinase inhibitors (TKI) 15,19. Oncology is entering a new era with major cancer centers now deep sequencing tumors to reveal genetic alterations that may render subclonal populations susceptible or resistant to targeted inhibitors, but the use of this information in precision medicine has lagged behind. It would be of enormous value in clinical practice if an oncologist could reliably ascertain whether these mutations render the target of therapy resistant or susceptible to available inhibitors; such tools would facilitate the enrollment of patients in mechanism-based basket trials, help prioritize candidate compounds for clinical trials, and aid the development of next-generation inhibitors.

While some cancer missense mutations are highly recurrent and have been characterized clinically or biochemically, a long tail of rare mutations collectively accounts for the majority of clinically observed missense mutations (Fig. 1a), leaving clinicians and researchers without knowledge of whether these uncharacterized mutations might lead to resistance. While rules-based and machine learning schemes are still being assessed in oncology contexts, work in predicting drug response to microbial resistance has shown that rare mutations present a significant challenge to approaches that seek to predict resistance to therapy. Clinical cancer mutations may impact drug response through a variety of mechanisms by altering kinase activity, ATP affinity, substrate specificities, and the ability to participate in regulatory interactions, compounding the difficulties associated with limited datasets that machine learning approaches face. In parallel with computational approaches, high-throughput experimental techniques such as MITE-Seq have been developed to assess the impact of point mutations on drug response. However, the complexity of defining selection schemes that reliably correlate with in vivo drug effectiveness and long turn-around times might limit their ability to rapidly and reliably impact clinical decision-making.

Physics-based approaches could be complementary to machine-learning and experimental techniques in predicting changes in TKI affinity due to mutations with few or no prior clinical observations. Alchemical free-energy methods permit receptor-ligand binding energies to be computed rigorously, including all relevant entropic and enthalpic contributions. Encouragingly, kinase-inhibitor binding affinities have been predicted using alchemical free-energy methods with mean unsigned errors of 1.0 kcal mol$^{-1}$ for CDK2, JNK1, p38, and Tyk2.

Recently, one study has hinted at the potential utility of alchemical free-energy calculations in oncology by predicting the impact of a single clinical mutation on the binding free energies of the TKIs dasatinib and R4L534. Here, we ask whether physical modeling techniques may be useful in predicting whether clinically identified kinase mutations lead to drug resistance or drug sensitivity. We perform state-of-the-art relative alchemical free-energy calculations using FEP$^+$, recently demonstrated to achieve sufficiently good accuracy to drive the design of small-molecule inhibitors for a broad range of...
targets during lead optimization, to calculate the effect of point mutation on the binding free energy between the inhibitor and the kinase receptor (Fig. 1b, c). We compare this approach against a fast but approximate physical modeling method implemented in Prime (an MM-GBSA approach) in which an implicit solvent model is used to assess the change in minimized interaction energy of the ligand with the mutant and wild-type kinase. We consider whether these methods can predict a ten-fold reduction in inhibitor affinity (corresponding to a binding free-energy change of 1.36 kcal mol$^{-1}$) to assess baseline utility. As a benchmark, we compile a set of reliable inhibitor ΔpIC50 data for 144 clinically identified mutants of the human kinase Abl, an important oncology target dysregulated in cancers like chronic myelogenous leukemia (CML), for which six FDA-approved TKIs are available. While ΔpIC50 can approximate a dissociation constant $K_D$, other processes contributing to changes in cell viability might affect IC50 in ways that are not accounted for by a traditional binding experiment, motivating a quantitative comparison between ΔpIC50 and $K_D$. The results of this benchmark demonstrate the potential for FEP+ to predict the impact that mutations in Abl kinase have on drug binding, and a classification accuracy of 88.93% (for all statistical metrics reported in this paper, the 95% confidence intervals (CI) is shown in the form of $(\hat{x}_{\text{upper}}, \hat{x}_{\text{lower}})$, an RMSE of 1.07$^{+0.26}_{-0.89}$ kcal mol$^{-1}$, and an MUE of 0.70$^{+0.92}_{-0.67}$ kcal mol$^{-1}$ was achieved.

Results

A benchmark of ΔpIC50s for predicting mutational resistance. To construct a benchmark evaluation dataset, we compiled a total of 144 ΔpIC50 measurements of Abl:TKI affinities, summarized in Table 1 while ensuring all measurements for an individual TKI were reported in the same study from experiments run under identical conditions. 131 ΔpIC50 measurements were available across the six TKIs with available co-crystal structures with wild-type Abl—26 for axitinib and 21 for bosutinib, dasatinib, imatinib, nilotinib, and ponatinib. 13 ΔpIC50 measurements were available for the two TKIs for which docking was necessary to generate Abl:TKI structures—7 for erlotinib and 6 for gefitinib. For added diversity, this set includes TKIs for which Abl is not the primary target—axitinib, erlotinib, and gefitinib. All mutations in this benchmark dataset have been clinically observed (Supplementary Table 1). Due to the change in bond topology required by mutations involving proline, which is not currently supported by the FEP+ technology for protein residue mutations, the three mutations H396P (axitinib, gefitinib, erlotinib) were excluded from our assessment. As single-point mutations were highly represented in the Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) study analyzed in Fig. 1a, we excluded double mutations from this work. However, the impact of mutations from multiple sites can potentially be modeled by sequentially mutating each site and this will be addressed in future work.

Experimental ΔpIC50 measurements for wild-type and mutant Abl were converted to ΔΔG in order to make direct comparisons between physics-based models and experiment. However, computation of experimental uncertainties were required to understand the degree to which differences between predictions and experimental data were significant. Since experimental error estimates for measured IC50s were not available for the data in Table 1, we compared data to other sources that have published IC50s for the same mutations in the presence of the same TKIs (Fig. 2a–c). Cross-comparison of 97 experimentally measured ΔΔGs derived from cell viability assay IC50 data led to an estimate of experimental variability of 0.32$^{+0.26}_{-0.28}$ kcal mol$^{-1}$ root mean square error (RMSE) that described the expected repeatability of the measurements. Because multiple factors influence the IC50 aside from direct effects on the binding affinity we also compared ΔΔGs derived from ΔpIC50s with those derived from binding affinity measurements (Δ$G_f$) for which data for a set of 27 mutations was available (Fig. 2d). The larger computed RMSE of 0.81$^{+0.04}_{-0.03}$ kcal mol$^{-1}$ represents an estimate of the lower bound of the RMSE to the IC50-derived ΔΔGs that we might hope to achieve with FEP+ or Prime, which were performed using non-phosphorylated models, when comparing sample statistics directly. Comparing 31 mutations for which phosphorylated and non-phosphorylated Δ$G_f$s were available, we found a strong correlation between the ΔΔGs derived from those data ($r = 0.94$, Supplementary Figure 1).

Most mutations do not significantly reduce TKI potency. The majority of mutations do not lead to resistance by our 10-fold affinity loss threshold: 86.3% of the co-crystal set ($n = 113$) and 86.8% of the total set ($n = 125$). Resistance mutations, which are likely to result in a failure of therapy, constitute 13.7% of the co-crystal set ($n = 18$) and 13.2% of the total set of mutations ($n = 19$). The ΔpIC50s for all 144 mutations are summarized in Supplementary Tables 2–7. Two mutations exceeded the dynamic range of the assays (IC50 > 10,000 nM); as these two mutations

### Table 1 Public ΔpIC50 datasets for 144 Abl kinase mutations and eight TKIs with corresponding wild-type co-crystal structures used in this study

| TKI       | N_mut | R | S | PDB   | ΔG<sub>max</sub>−ΔG<sub>min</sub> | Source | ΔG<sub>WT</sub> |
|-----------|-------|---|---|-------|---------------------------------|--------|---------------|
| Axitinib  | 26    | 0 | 26 | 4wa9  | 2.05                            | 52     | –8.35         |
| Bosutinib | 21    | 4 | 17 | 3ue4  | 2.79                            | 79     | –9.81         |
| Dasatinib | 21    | 5 | 16 | 4xey  | 5.08                            | 79     | –11.94        |
| Imatinib  | 21    | 5 | 16 | 1opj  | 2.16                            | 79     | –9.19         |
| Nilotinib | 21    | 4 | 17 | 3cs9  | 3.88                            | 17     | –10.74        |
| Ponatinib | 21    | 0 | 21 | 3oxz  | 1.00                            | 79     | –11.70        |
| Subtotal  | 131   | 18 | 113|       |                                  |        |               |
| Erlotinib | 7     | 1 | 6  | Dock to 3ue4 | 1.73                        | 82     | –9.77         |
| Gefitinib | 6     | 0 | 6  | Dock to 3ue4 | 1.79                        | 82     | –8.84         |
| Total     | 144   | 19 | 125|       |                                  |        |               |

N<sub>mut</sub>: Total number of mutants for which ΔpIC50 data was available
Number of Resistant, Susceptible mutants using 10-fold affinity change threshold
PDB Source PDB ID, or Dock to 3ue4, which used 3ue4 as the receptor for Glide-SP docking inhibitors without co-crystal structure
ΔG<sub>WT</sub>: Binding free energy of inhibitor to wild-type Abl, as estimated from IC50 data.
clearly raise resistance, we excluded them from quantitative analysis (RMSE and MUE) but included them in truth table analyses and classification metrics (accuracy, specificity, and sensitivity).

FEP+ predicts affinity changes for clinical Abl mutants. Figure 1b depicts the thermodynamic cycle that illustrates how we used relative free-energy calculations to compute the change in ligand binding free energy in response to the introduction of a single mutation in the kinase (Fig. 1c). From prior experience with relative alchemical free-energy calculations for ligand design, good initial receptor-ligand geometry was critical to obtaining accurate and reliable free-energy predictions, so we first focused on the 131 mutations in Abl kinase across six TKIs for which wild-type Abl:TKI co-crystal structures were available. Figure 3 summarizes the performance of predicted binding free-energy changes (ΔΔG) for all 131 mutants in this set for both a fast MM-GBSA physics-based method that only captures interaction energies for a single structure (Prime) and rigorous alchemical free-energy calculations (FEP+). Scatter plots compare experimental and predicted free-energy changes (ΔΔG) and characterize the ability of these two techniques to predict experimental measurements. Statistical uncertainty in the predictions and experiment-to-experiment variability in the experimental values are shown as ellipse height and widths, respectively. The value for experimental variability was 0.32 kcal mol⁻¹, which was the standard error computed from the cross-comparison in Fig. 2. For FEP+, the uncertainty was taken to be the standard error of the average from three independent runs for a particular mutation, while Prime results are deterministic and are not contaminated by statistical uncertainty.

To better assess whether discrepancies between experimental and computed ΔΔGs simply arise for known forcefield limitations or might indicate more significant effects, we incorporated an additional error model in which the forcefield error was taken to be a random error $\sigma_{PF} = 0.9$ kcal mol⁻¹, a value established from previous benchmarks on small molecules absent conformational sampling or protonation state issues. Thin error bars in Fig. 2 represent the overall estimated error due to both this forcefield error and experimental variability or statistical uncertainty.

To assess overall quantitative accuracy, we computed both RMSE—which is rather sensitive to outliers, and mean unsigned error (MUE). For Prime, the MUE was 1.16 ± 0.25 kcal mol⁻¹ and the RMSE was 1.72 ± 0.40 kcal mol⁻¹. FEP+, the alchemical free-energy approach, achieved a significantly higher level of quantitative accuracy with an MUE of 0.82 ± 0.06 kcal mol⁻¹ and an RMSE of 1.11 ± 0.30 kcal mol⁻¹. Notably, alchemical free-energy calculations come substantially closer than MMGBSA approach to the minimum achievable RMSE of 0.81 ± 0.04 kcal mol⁻¹ (due to experimental error; Fig. 2) for this dataset.

FEP+ accurately classifies affinity changes for Abl mutants. While quantitative accuracy (MUE, RMSE) is a principle metric of model performance, an application of potential interest is the ability to classify mutations as raising resistance to a specific TKI. To characterize the accuracy with which Prime and FEP+ classified mutations in a manner that might be therapeutically relevant, we classified mutations by their experimental impact on the binding affinity as susceptible (affinity for mutant is diminished by no more than 10-fold, ΔΔG ≤ 1.36 kcal mol⁻¹) or as resistant (affinity for mutant is diminished by at least 10-fold, ΔΔG > 1.36 kcal mol⁻¹). Summary statistics of experimental and computational predictions of these classes are shown in Fig. 2 (bottom) as truth tables (also known as confusion matrices).

The simple minimum-energy scoring method Prime correctly classified 9 of the 18 resistance mutations in the dataset while merely 85 of the 113 susceptible mutations were correctly classified (28 false positives). In comparison, the alchemical free-energy method FEP+, which includes entropic and enthalpic contributions as well as explicit representation of solvent, correctly classified 9 of the 18 resistance mutations while a vast majority, 105, of the susceptible mutations were correctly classified (merely 8 false positives). Prime achieved a classification accuracy of 0.720 ± 0.064, while FEP+ achieved an accuracy that is significantly higher (both in a statistical sense and in overall magnitude), achieving an accuracy of 0.870 ± 0.06. Sensitivity (also called true positive rate) and specificity (true negative rate) are also informative statistics in assessing the performance of a binary classification scheme. For Prime, the sensitivity was 0.50 ± 0.23, while the specificity was 0.75 ± 0.08. To put this in perspective, a CML
patient bearing a resistance mutation in the kinase domain of Abl has an equal chance of Prime correctly predicting this mutation would be resistant to one of the TKIs considered here, while if the mutation was susceptible, the chance of correct prediction would be ~75%. By contrast, the classification specificity of FEP+ was substantially better. For FEP+, the sensitivity was 0.50 while the specificity was 0.90. There is a very high probability that FEP+ will correctly predict that one of the eight TKIs studied here will remain effective for a patient bearing a susceptible mutation.

How reliant are classification results on choice of cutoff? Previous work by O’Hare et al. utilized TKI-specific thresholds for dasatinib, imatinib, and nilotinib, which were ~2 kcal mol⁻¹. Supplementary Figure 2 shows that when our classification threshold was increased to a 20-fold change in binding (1.77 kcal mol⁻¹), FEP+ correctly classified 8 of the 13 resistant mutations and with a threshold of 100-fold change in binding (2.72 kcal mol⁻¹), FEP+ correctly classified the only two resistant mutations (T315I/dasatinib and T315I,nilotinib). With the extant multilayered and multinodal decision-making algorithms used by experienced oncologists to manage their patients’ treatment, or by medicinal chemists to propose candidate compounds for clinical trials, the resistant or susceptible cutoffs could be selected with more nuance than the simple 10-fold affinity threshold we consider here. With a larger affinity change cutoff, for example, the accuracy with which physical models predict resistance mutations increases beyond 90% (Supplementary Figure 2). For the alchemical
Bayesian analysis can estimate the true error. The statistical metrics—MUE, RMSE, accuracy, specificity, and sensitivity—discussed above are based on analysis of the apparent performance of the observed modeling results compared with the

|      | Prime | FEP+ |
|------|-------|------|
| **Axitinib** | ![Image](https://via.placeholder.com/150) | ![Image](https://via.placeholder.com/150) |
| S | R | S | R |
| **Imatinib** | ![Image](https://via.placeholder.com/150) | ![Image](https://via.placeholder.com/150) |
| **Dasatinib** | ![Image](https://via.placeholder.com/150) | ![Image](https://via.placeholder.com/150) |
| **Nilotinib** | ![Image](https://via.placeholder.com/150) | ![Image](https://via.placeholder.com/150) |
| **Ponatinib** | ![Image](https://via.placeholder.com/150) | ![Image](https://via.placeholder.com/150) |

Bayesian analysis can estimate the true error. The statistical metrics—MUE, RMSE, accuracy, specificity, and sensitivity—discussed above are based on analysis of the apparent performance of the observed modeling results compared with the
observed experimental data via sample statistics. However, this analysis considers a limited number of mutants, and both measurements and computed values are contaminated with experimental or statistical error. To obtain an estimate of the intrinsic performance of our physical modeling approaches, accounting for known properties of the experimental variability and statistical uncertainties, we used a hierarchical Bayesian model to infer posterior predictive distributions from which expectations and 95% predictive intervals could be obtained. The results of this analysis are presented in Fig. 3 (central tables).

FEP+ is significantly better than Prime at predicting the impact of mutations on TKI binding affinities, as the apparent performance as well as the intrinsic performance were well-separated outside their 95% CI in nearly all metrics. Applying the Bayesian model, the MUE and RMSE for FEP+ was $0.70_{0.32}$ and $0.91_{0.25}$ kcal mol$^{-1}$, respectively ($N = 129$). For the classification metrics accuracy, specificity, and sensitivity, the model yields $0.89_{0.03}$, $0.91_{0.04}$, and $0.69_{0.06}$, respectively ($N = 131$). The intrinsic RMSE and MUE of Prime were $1.76_{0.51}$ and $1.40_{0.24}$ kcal mol$^{-1}$ ($N = 129$), respectively, and the classification accuracy, specificity, and sensitivity was $0.73_{0.17}$, $0.74_{0.16}$, and $0.57_{0.18}$, respectively ($N = 131$). The intrinsic MUE of Prime obtained by this analysis is larger than the observed MUE reflecting the non-Gaussian, fat-tailed error distributions of Prime results.

How transferable is FEP+ across the six TKIs? The impact of point mutations on drug binding are not equally well predicted for the six TKIs. Figure 4 expands the results in Fig. 3 on a TKI-by-TKI basis to dissect the particular mutations in the presence of a specific TKI. Prime and FEP+ correctly predicted that most mutations in this dataset ($N = 26$) do not raise resistance to axitinib, though FEP+ predicted 4 false positives compared with 3 false positives by Prime. The MUE and RMSE of FEP+ was excellent for this inhibitor, $0.70_{0.32}$ and $0.91_{0.25}$ kcal mol$^{-1}$, respectively. While the classification results for bosutinib ($N = 21$) were equally well predicted by Prime as by FEP+, FEP+ was still able to achieve superior, but not significant, predictive performance for the quantitative metrics MUE and RMSE, which were $0.96_{0.42}$ and $1.41_{0.32}$ kcal mol$^{-1}$, respectively (FEP+) and $1.13_{0.33}$ and $1.80_{0.63}$ kcal mol$^{-1}$, respectively (Prime). For dasatinib, FEP+ achieved an MUE and RMSE of $0.76_{0.23}$ and $1.07_{0.57}$ kcal mol$^{-1}$, respectively, whereas the results were, as expected, less quantitatively predictive for Prime ($N = 20$). The results for imatinib were similar to those of dasatinib above, where the MUE and RMSE for FEP+ were $0.82_{0.15}$ and $1.09_{0.42}$ kcal mol$^{-1}$, respectively ($N = 20$). Nilotinib, a derivative of imatinib, led to nearly identical quantitative performance results for FEP+ with an MUE and RMSE of $0.82_{0.15}$ and $1.06_{0.39}$ kcal mol$^{-1}$, respectively ($N = 21$). Similar to axitinib, ponatinib presented an interesting case because there were no mutations in this dataset that raised resistance to it. Despite the wide dynamic range in the computed values of $\Delta$DG for other inhibitors, FEP+ correctly predicted a narrow range of $\Delta$DGs for this drug. This is reflected in the MUE and RMSE of $0.87_{0.16}$ and $1.09_{0.46}$ kcal mol$^{-1}$, respectively, which are in-line with the MUEs and RMSEs for the other TKIs.

Understanding the origin of mispredictions. Resistance mutations that are mispredicted as susceptible are particularly critical because they might mislead the clinician or drug designer into believing the inhibitor will remain effective against the target. Which resistance mutations did FEP+ mispredict as susceptible? Nine mutations were classified by FEP+ to be susceptible when experimentally measured $\Delta$IC$\text{_{50}}$ data indicate the mutations should have increased resistance according to our 10-fold affinity cutoff for resistance. Notably, the 95% CI for five of these mutations included the 1.36 kcal mol$^{-1}$ threshold, indicating these misclassifications are not statistically significant when the experimental error and statistical uncertainty in FEP+ are accounted for: bosutinib/L248R ($\Delta$DG$\text{_{FEP+}} = 1.32_{0.70}$ kcal mol$^{-1}$), imatinib/E255K ($\Delta$DG$\text{_{FEP+}} = 0.43_{0.20}$ kcal mol$^{-1}$), imatinib/Y253F ($\Delta$DG$\text{_{FEP+}} = 0.95_{0.44}$ kcal mol$^{-1}$), and nilotinib/Y253F ($\Delta$DG$\text{_{FEP+}} = 0.89_{0.09}$ kcal mol$^{-1}$). The bosutinib/V299L mutation was also not significant because the experimental $\Delta$DG, $1.70_{0.88}$ kcal mol$^{-1}$, included the $1.36$ kcal mol$^{-1}$ cutoff; the value of $\Delta$DG predicted by FEP+ for this mutation was $0.91_{0.12}$ kcal mol$^{-1}$, the upper bound of the predicted value was within the $0.66$ kcal mol$^{-1}$ of the lower bound of the experimental value.

Four mutations, however, were misclassified to a degree that is statistically significant: dasatinib/T315A, bosutinib/T315I, imatinib/E255V, and nilotinib/E255V. For dasatinib/T315A, although the T315A mutations for bosutinib, imatinib, nilotinib, and ponatinib were correctly classified as susceptible, the predicted free-energy changes for these four TKIs were consistently more negative than the corresponding experimental measurements, like dasatinib/T315A, indicating there might be a generic driving force contributing to the errors in T315A mutations for these five TKIs. Abl is known to be able to adopt many different conformations (including DFG-in and DFG-out), and it is very likely that the T315A mutation induces conformational changes in the apo protein, the inadequate sampling of which may have led to the errors for the T315A mutation. By comparison, the T315I mutations for axitinib, bosutinib, imatinib, nilotinib, and ponatinib were all accurately predicted with the exception of bosutinib/T315I being the only misprediction, suggesting an issue specific to bosutinib. The interactions between the 2,4-dichloro-5- methoxyphenyl ring in bosutinib and the positively charged amine of the catalytic Lys271 may not be accurately captured by the fixed-charge OPLS force field, possibly leading to the misprediction for bosutinib/T315I mutation.

Insufficient sampling might also belie the imatinib/E255V and nilotinib/E255V mispredictions because they reside in the highly flexible P-loop. Since E255V was a charge change mutation, we utilized a workflow that included a transmutable explicit ion (see Methods). The distribution of these ions in the simulation box around the solute might not have converged to their equilibrium state on the relatively short timescale of our simulations (5 ns),
and the insufficient sampling of ion distributions coupled with P-loop motions might lead to misprediction of these two mutations.

How strongly is accuracy affected for docked TKIs? To assess the potential for utilizing physics-based approaches in the absence of a high-resolution experimental structure, we generated models of Abl bound to two TKIs—erlotinib and gefitinib—for which co-crystal structures with wild-type kinase are not currently available. In Fig. 5, we show the Ablerotinib and Abl:gefitinib complexes that were generated using a docking approach (Glide-SP, see Methods). These two structures were aligned against the co-crystal structures of EGFR:erlotinib and EGFR:gefitinib (Glide) to highlight the structural similarities between the binding pockets of Abl and EGFR and the TKI binding mode in Abl versus EGFR. As an additional test of the sensitivity of FEP+ to system preparation, a second set of Ablerotinib and Abl:gefitinib complexes was generated in which crystallographic water coordinates were transferred to the docked inhibitor structures (see Methods).

Alchemical free-energy simulations were performed on 13 mutations between the two complexes; 7 mutations for erlotinib and 6 mutations for gefitinib. The quantitative accuracy of FEP+ in predicting the value of \( \Delta G \) was excellent—MUE and RMSE of 0.58(33) kcal mol\(^{-1}\) and 0.80(99) kcal mol\(^{-1}\), respectively, if crystal waters are omitted, and 0.50(26) kcal mol\(^{-1}\) and 0.69(35) kcal mol\(^{-1}\) if crystal waters were restored after docking. Encouragingly, these results indicate that our initial models of Abl bound to erlotinib and gefitinib were reliable because the accuracy and dependability of our FEP+ calculations were not sensitive to crystallographic waters. Our secondary concern was the accuracy with which the approach classified mutations as resistant or susceptible.

While the results presented in (Fig. 5) indicate that FEP+ is capable of achieving good quantitative accuracy when a co-crystal structure is unavailable, it is important to understand why a mutation was predicted to be susceptible but was determined experimentally to be resistant. F317I was the one mutation that increased resistance to erlotinib (or gefitinib) because it destabilized binding by more than 1.36 kcal mol\(^{-1}\)—1.35(67) kcal mol\(^{-1}\) (gefitinib) and 1.58(90) kcal mol\(^{-1}\) (erlotinib), but the magnitude of the experimental uncertainty means we are unable to confidently discern whether this mutation induces more than 10-fold resistance to either TKI. Therefore, the one misclassification by FEP+ in Fig. 5 is not statistically significant and the classification metrics presented there underestimate the nominal performance of this alchemical free-energy method.

Discussion
The results presented in this work are summarized in Table 2. The performance metrics summarized in Table 2 indicates that the set of 131 mutations for the six TKIs in which co-crystal structures were available is on par with the complete set (144 mutations), which included results based on Abl:TKI complexes generated from docking models. The performance results for the 13 mutations for the two TKIs (erlotinib and gefitinib) in which co-crystal structures were unavailable exhibited good quantitative accuracy (MUE and RMSE) and good classification power.

Overall (N = 144), the MM-GBSA approach Prime classified mutations with good accuracy (0.73(80) and specificity (0.76(84) while the alchemical approach FEP+ was a significant improvement in classification accuracy (0.88(93) and specificity (0.94(98). The quantitative accuracy with which Prime was able to predict the experimentally measured change in Abl:TKI binding (N = 142) characterized by RMSE and MUE was 1.70(40) and 1.14(35) kcal mol\(^{-1}\), respectively. In stark contrast, the quantitative accuracy of FEP+ was statistically superior to Prime with an RMSE and an MUE of 1.07(26) and 0.79(62) kcal mol\(^{-1}\), respectively.
Table 2 Summary of FEP+ and Prime statistics in predicting mutational resistance or sensitivity to FDA-approved TKIs

| Dataset | Method | N_{quant} | MUE (kcal mol$^{-1}$) | RMSE (kcal mol$^{-1}$) | N_{class} | Accuracy | Specificity | Sensitivity |
|---------|--------|-----------|------------------------|------------------------|-----------|----------|------------|------------|
| all     | FEP+   | 142       | 0.790 ± 0.03           | 1.077 ± 0.04           | 144       | 0.880 ± 0.03 | 0.940 ± 0.05 | 0.420 ± 0.08 |
| all     | Prime  | 142       | 1.14 ± 0.03            | 1.70 ± 0.06            | 144       | 0.72 ± 0.08  | 0.76 ± 0.08  | 0.53 ± 0.08  |
| xtal    | FEP+   | 129       | 0.822 ± 0.03           | 1.11 ± 0.08            | 131       | 0.87 ± 0.03  | 0.93 ± 0.08  | 0.50 ± 0.08  |
| xtal    | Prime  | 129       | 1.16 ± 0.02            | 1.72 ± 0.02            | 131       | 0.72 ± 0.05  | 0.75 ± 0.05  | 0.50 ± 0.05  |
| axitinib| FEP+   | 26        | 0.70 ± 0.02            | 0.91 ± 0.04            | 26        | 0.85 ± 0.02  | 0.85 ± 0.02  | NA         |
| axitinib| Prime  | 26        | 1.05 ± 0.03            | 1.85 ± 0.06            | 26        | 0.88 ± 0.03  | 0.88 ± 0.03  | NA         |
| bosutinib| FEP+  | 21        | 0.96 ± 0.03            | 1.41 ± 0.02            | 21        | 0.76 ± 0.07  | 0.88 ± 0.10  | 0.25 ± 0.01  |
| bosutinib| Prime | 21        | 1.12 ± 0.01            | 1.80 ± 0.02            | 21        | 0.81 ± 0.07  | 0.82 ± 0.07  | 0.72 ± 0.07  |
| dasatinib| FEP+  | 20        | 0.76 ± 0.02            | 1.07 ± 0.05            | 20        | 0.90 ± 0.04  | 0.94 ± 0.07  | 0.80 ± 0.07  |
| dasatinib| Prime | 20        | 1.05 ± 0.04            | 1.48 ± 0.04            | 21        | 0.86 ± 0.05  | 0.88 ± 0.08  | 0.80 ± 0.08  |
| imatinib| FEP+   | 20        | 0.82 ± 0.03            | 1.09 ± 0.09            | 20        | 0.81 ± 0.06  | 1.00 ± 0.00  | 0.40 ± 0.00  |
| imatinib| Prime  | 20        | 1.32 ± 0.01            | 1.69 ± 0.05            | 21        | 0.43 ± 0.07  | 0.50 ± 0.25  | 0.20 ± 0.04  |
| nilotinib| FEP+  | 21        | 0.82 ± 0.03            | 1.06 ± 0.09            | 21        | 0.86 ± 0.07  | 0.94 ± 0.08  | 0.50 ± 0.08  |
| nilotinib| Prime | 21        | 1.50 ± 0.02            | 1.86 ± 0.07            | 21        | 0.48 ± 0.07  | 0.53 ± 0.75  | 0.25 ± 0.09  |
| ponatinib| FEP+  | 21        | 0.87 ± 0.04            | 1.09 ± 0.04            | 21        | 1.00 ± 0.06  | 1.00 ± 0.00  | NA         |
| ponatinib| Prime | 21        | 0.94 ± 0.03            | 1.57 ± 0.09            | 21        | 0.81 ± 0.07  | 0.81 ± 0.07  | NA         |
| Glide   | FEP+   | 13        | 0.50 ± 0.02            | 0.69 ± 0.15            | 13        | 0.92 ± 0.07  | 1.00 ± 0.00  | 0.00 ± 0.00  |
| Glide   | Prime  | 13        | 0.91 ± 0.09            | 1.45 ± 0.04            | 13        | 0.80 ± 0.07  | 0.83 ± 0.58  | 1.00 ± 0.00  |

Accuracy, specificity, and sensitivity were computed to assess two-class prediction performance: resistant (ΔΔG > 1.36 kcal mol$^{-1}$) or susceptible (ΔΔG ≤ 1.36 kcal mol$^{-1}$). 95% CIs (sub/superscripts) were estimated from 1000 bootstrap replicates. The sensitivity for axitinib and ponatinib is NA, because there is no resistant mutation for these two drugs.

FEP+ was found better than Prime. Likewise, a clearer picture of the true classification accuracy, specificity, and sensitivity of FEP+ was found—0.90±0.03, 0.92±0.05, and 0.68±0.00 respectively. The high accuracy of FEP+ is very encouraging, and the accuracy can be further improved with more accurate modeling of a number of physical chemical effects not currently considered by the method. While highly optimized, the fixed-charged OPLS37 force field can be further improved by explicit consideration of polarizability effects43, as hinted by some small-scale benchmarks44. These features could be especially important for bosutinib, whose 2,4-dichloro-5-methoxyphenyl ring is adjacent to the positively charged amine of the catalytic Lys271. Many simulation programs also utilize a long-range isotropic analytical dispersion correction intended to correct for the truncation of dispersion interactions at finite cutoff, which can induce an error in protein–ligand binding free energies that depends on the number of ligand heavy atoms being modified.44 Recently, efficient Lennard–Jones PME methods45,46 and perturbation schemes44 have been developed that can eliminate the errors associated with this truncation. While the currently employed methodology for alchemical transformations involving a change in system charge reduces artifacts that depend on the simulation box size and periodic boundary conditions, the explicit ions that were included in these simulations may not have sufficiently converged to their equilibrium distributions in these relatively short simulations. Kinases and their inhibitors are known to possess multiple titratable sites with either intrinsic or effective pK_a's near physiological pH, while the simulations here treat protonation states and proton tautomers fixed throughout the bound and unbound states; the accuracy of the model can be further improved with the protonation states or tautomers shift upon binding or mutation considered47,48. Similarly, some systems display significant salt concentration dependence19, while the simulations for some systems reported here did not rigorously mimic all aspects of the experimental conditions of the cell viability assays. While we have shown that predicting the direct impact of mutations on the binding affinity of ATP-competitive TKIs for a single kinase conformation has useful predictive capacity, many additional physical effects that can contribute to cell viability are not currently captured by examining only the predicted change in inhibitor binding affinity. For example, kinase missense mutations can also shift the populations of kinase conformations (which may affect ATP and inhibitor affinities differentially), modulate ATP affinity, modulate affinity for protein substrate, or modulate the ability of the kinase to be regulated or bounded by scaffolding proteins. While many of these effects are in principle tractable by physical modeling in general it is valuable to examine...
Clinical observations and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms).

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.

A simple threshold of 10-fold TKI affinity change is a crude metric for classifying resistance or susceptibility due to the myriad biological factors that contribute to the efficacy of a drug in a person. In addition to affecting the binding affinity of inhibitors, missense mutations can also cause drug resistance through other physical mechanisms including induction of splice variants or alleviation of feedback. While the current study only focused on the effect of mutation on drug binding affinity, resistance from these other physical mechanisms could be similarly computed using physical modeling. For example, some mutations are known to activate the kinase by increasing affinity to ATP, which could be computed using free-energy methods like FEP.

In this communication, we hypothesized that FEP+, a fully automated relative-chemical-free-energy workflow, had reached the point where it can accurately and reliably predict how clinically observed mutations in Abl kinase alter the binding affinity to ATP, which could be computed using free-energy methods now had mispredictions and outliers to identify whether any of these cases are likely to induce resistance (as observed by ΔpIC50 shifts) by one of these alternative mechanisms.
Force Field Builder. The two ligands that contained a fragment with a torsion parameter not covered by OPLS were axitinib and bosutinib; Force Field Builder was used with standard parameters. SPC water was used for all water. For mutations that change the net change of the system, counterions were included to neutralize the system with additional Na\(^+\) and Cl\(^-\) ions added to achieve 0.15 M excess to mimic the solution conditions of the experimental assay.

Prime (MM-GBSA). Prime was used to predict the geometry of mutant side chains and to calculate relative changes in free energy using MM-GBSA single-point estimates. VSBG\(^{36}\) was used as the implicit solvent model to calculate the solvation free energy for the four states (complex/wild-type, complex/mutant, apo-protein/wild-type, and apo-protein/mutant) and ΔΔΔG calculated using the thermodynamic cycle depicted in Fig. 1b. Unlike FEP (see below), which simulates the horizontal legs of the thermodynamic cycle, MM-GBSA models the vertical legs by computing the interaction energy between the ligand and protein in both wild-type and mutant states, subtracting these to obtain the ΔΔΔG of mutation on the binding free energy.

Alchemical free-energy perturbation calculations using FEP+. Alchemical free-energy calculations were performed using the FEP+ tool in the Schrödinger suite version 2016-4, which offers a fully automated workflow requiring only an input structure (wild-type complex) and specification of the desired mutation. The default protocol was used throughout: It assigns protein and ligand force parameters (as above), generates a dual-topology alchemical system for transitions between wild-type and mutant states, and performs 12 ps MD simulations with 1 fs time step using Langevin thermostat and barostat at 10 K and constant pressure of 1 atmosphere, using the same restraints; (iii) 12 ps MD simulations with 1 fs time step using Langevin thermostat and barostat at 300 K and constant pressure of 1 atmosphere. System equilibration was automated. It followed the default 5-stage Desmond protocol: (i) 100 ps with 1 fs time steps of Brownian dynamics with positional restraints of solute heavy atoms to their initial geometry using a restraint force constant of 50 kcal mol\(^{-1}\) Å\(^{-2}\); this Brownian dynamics integrator corresponds to a Langevin integrator in the limit when τ\(_{\text{r}}\) → 0, modified to stabilize equilibration of starting configurations with high potential energies; particle and positional restraints were clipped so that particle displacements were limited to 0.1 Å, in any direction. (ii) 12 ps MD simulations with 1 fs time step using Langevin thermostat at 10 K with constant volume, using the same restraints; (iii) 12 ps MD simulations with 1 fs time step using Langevin thermostat and barostat\(^{38}\) at 10 K and constant pressure of 1 atmosphere, using the same restraints; (iv) 12 ps MD simulations with 1 fs time step using Langevin thermostat and barostat at 300 K and constant pressure of 1 atmosphere, using the same restraints; (v) a final unrestrained equilibration MD simulation of 240 ps with 2 fs time step using Langevin thermostat and barostat at 300 K and constant pressure of 1 atmosphere. Electrostatic interactions were computed with particle-mesh Ewald (PME)\(^{45}\) and a 9 Å cutoff distance was used for van der Waals interactions. The production MD simulation was performed in the NPT ensemble using the mtk method\(^{57}\) with integration time steps of 4, 4, and 8 fs, respectively, for the bonded, near, and far interactions following the RESPMA method\(^{48}\) through hydrogen mass repartitioning. Production FEP+ calculations used Hamiltonian replica exchange with solute tempering (REST)\(^{70}\), with automated definition of the REST region. Dynamics were performed with constant pressure of 1 atmosphere and constant temperature of 300 K for 5 ns in which exchanges between windows was attempted every 1.2 ps. The system closure could not be used to reduce statistical errors via path redundancy,\(^{50}\) we instead performed mutual free-energy calculations in triplicate by initializing dynamics with different random seeds. The relative free energies for each mutation in each independent run were calculated using BAR\(^{1,72}\). The reported ΔΔΔG was computed as the mean of the computed ΔΔΔG from three independent simulations. Triplicate simulations were performed in parallel using four NVIDIA Pascal Architecture GPUs per alchemical free-energy simulation (12 GPUs in total), requiring ~6 h in total to compute ΔΔΔG.

Obtaining ΔΔG from ΔΔΔpIC\(_{50}\) benchmark set data. Reference relative free energies were obtained from three publicly available sources of ΔΔΔpIC\(_{50}\) data (Table 1). Under the assumption of Michaelis–Menten binding kinetics (pseudo-first-order, but relative free energies are likely consistent), the inhibitor is competitive with ATP (eq: ic50). This assumption has been successfully used to estimate relative free energies\(^{34,35}\) using the relationship between IC\(_{50}\) and competitive inhibitor affinity K\(_{i}\).

If the Michaelis constant for ATP (K\(_{M}\)) is much smaller than the initial ATP concentration S, the relation in eq:ic50 will tend towards the equality IC\(_{50}\) = K\(_{i}\). The relative change in binding free energy of Abl/TKI binding due to protein mutation is simply:

\[
\Delta \Delta G = R \ln \frac{IC_{50,\text{WT}}}{IC_{50,\text{mut}}}
\]

where IC\(_{50,\text{WT}}\) is the IC\(_{50}\) value for the wild-type protein and IC\(_{50,\text{mut}}\) is the IC\(_{50}\) value for the mutant protein. R is the ideal gas constant and T is taken to be room temperature (300 K).

As alluded to above, relating ΔΔΔpIC\(_{50}\) to ΔΔΔG assumes that the Michaelis constant for ATP is much larger than the initial concentration of ATP, and that the experimentally observed ΔΔΔpIC\(_{50}\) change is solely from changes in kinase:TKI binding affinity. In practice, not all of these assumptions may hold. For example, the experimentally observed ΔΔΔpIC\(_{50}\) might depend on the metabolism of drugs, and for drugs with different mechanisms of action than directly binding to the kinase binding pocket (e.g., binding to the transition structures of kinases), target gene amplification, up/downregulation of positive/negative-feedback effectors, diminished synergism of pro-apoptotic machinery, decoupling of the target from cell survival circuits\(^{56}\), their inhibition ability might not correlate well with binding affinity. However, the comparison between ΔΔΔpIC\(_{50}\) and ΔΔΔG is presented in this manuscript; this comparison highlights the assumptions we used to relate ΔΔΔpIC\(_{50}\) to ΔΔΔG are reasonable for the dataset we studied.

Quantitative accuracy metrics. MUE was calculated by taking the average absolute difference between predicted and experimental estimates of ΔΔΔG. RMSE was calculated by taking the square root of the average squared difference between predicted and experimental estimates of ΔΔΔG. MUE depends linearly on errors for the most severe errors. For example, if the guide line were fixed at 1 threshold representing 10-fold change in affinity, accuracy was calculated as the fraction of all predictions that were correctly classified as sensitive, neutral, or resistant. Sensitivity and specificity were calculated using a binary classification of resistant (ΔΔΔG > 1.36 kcal mol\(^{-1}\)) or susceptible (ΔΔΔG ≤ 1.36 kcal mol\(^{-1}\)). Sensitivity was calculated as the fraction of correctly predicted non-resistant mutations out of all truly susceptible mutations S. Specificity was calculated as the fraction of correctly predicted resistant mutations out of all truly resistant mutations R. The number of susceptible mutations was 113 for axitinib, bosutinib, dasatinib, imatinib, nilotinib and ponatinib, and 12 for erlotinib and gefitinib; the number of resistant mutations R was 18 for axitinib, bosutinib, dasatinib, imatinib, nilotinib, and ponatinib, and 1 for erlotinib and gefitinib.

Consensus model. First, Prime and FEP+ (n = 142) were scaled by minimizing their RMSE to experiment by optimizing slope using linear regression. The resulting (minimum) RMSE was used in a subsequent step to combine the scaled FEP+ and scaled Prime free energies with inverse-variance weighted averaging.

ROC. A ROC curve was generated by computing the true positive rate (sensitivity) and the true negative rate (specificity) when the classification cutoff differencing resistant from sensitizing mutations is changed for (only) the predicted values of ΔΔΔG. Cutoffs were taken by choosing the minimum and maximum value of ΔΔΔG for a dataset (Prime or FEP+), and iteratively computing specificity and sensitivity in steps of 0.001 kcal mol\(^{-1}\), which by this definition will be in the range [0,1]. Experimental positives and negatives were classified with the 1.36 kcal mol\(^{-1}\) cutoff. ROC-AUC was computed using the trapezoidal rule.

Estimating uncertainties of physical-modeling results. 95% symmetric CI (95%) for all performance metrics were calculated using bootstrap by resampling all datasets with replacement, with 1000 resampling events. Confidence intervals were estimated for all performance metrics and reported as ±1.96\(\sqrt{n}\) where \(n\) is the mean statistic calculated from the complete dataset (e.g., RMSE), and \(\bar{x}_{\text{low}}\) and \(\bar{x}_{\text{high}}\) are the values of the statistic at the 2.5th and 97.5th percentiles of the value-sorted list of the bootstrap samples. Uncertainty for ΔΔΔG was computed by the standard
deviation between three independent runs (using different random seeds to set initial velocities), where the 95% CI was [ΔΔG ⊳ 1.96 × σounce, ΔΔG ⊳ 19.66 × σounce] kcal mol⁻¹. It is used in plots for FEP+ and experiment; for Prime.

Bayesian hierarchical model to estimate intrinsic error. We used Bayesian inference to estimate the true underlying prediction error of Prime and FEP+ by making use of known properties of the experimental variability (characterized in Fig. 2) and statistical uncertainty estimates generated by our calculations under weak assumptions about the character of the error.

We presume the true free-energy differences of mutation i, ΔΔGtrue, comes from a normal background distribution of unknown mean and variance, 

\[ ΔΔG_{true} \sim (\mu_{true}, \sigma_{true}^2) \]

where there are M mutations in our dataset. We assign weak priors to the mean and variance

\[ \mu_{true} \sim U(-6, +6) \]
\[ \sigma_{true} \sim 1 \]

where we limit σ > 0.

We presume the true computational predictions (absent statistical error) differ from the (unknown) true free-energy difference of mutation ΔΔGtrue normally distributed errors with zero bias but standard deviation equal to the RMSE for either Prime or FEP+, the quantity we are focused on estimating:

\[ ΔΔG_{true} \sim (ΔΔG_{true}, \sigma_{true}^2) \]

In the case of Prime, since the computation is deterministic, we actually calculate ΔΔGtrue, for each mutant. For FEP+, however, the computed free-energy changes are corrupted by statistical error, which we also presume to be normally distributed with standard deviation σounce.

\[ ΔΔG_{true} \sim (ΔΔG_{calc}, \sigma_{calc}^2) \]

where ΔΔG_{calc} is the free energy computed for mutant i by FEP+, and \( \sigma_{calc} \) is the corresponding statistical error estimate.

The experimental data we observe is also corrupted by error, which we presume to be normally distributed with standard deviation \( \sigma_{exp} \).

\[ ΔΔG_{exp} \sim (ΔΔG_{calc}, \sigma_{exp}^2) \]

Here, we used an estimate of \( K_{eq} \) and \( C_{eq} \)-derived ΔΔG variation derived from the empirical RMSE of 0.81 kcal mol⁻¹, where we took \( \sigma_{exp} \) \approx 0.81 \( \sqrt{2} \) \( \approx 0.57 \) kcal mol⁻¹ to ensure the difference between two random measurements of the same mutant would have an empirical RMSE of 0.81 kcal mol⁻¹.

Under the assumption that the true ΔΔG is normally distributed and the calculated value differs from the true value via a normal error model, it can easily be shown that the MUE is related to the RMSE via

\[ \text{MUE} = \frac{1}{\sqrt{2\pi\sigma_{true}^2}} \int_{-\infty}^{\infty} f(x_{true}) f(x_{true}) \text{d}x_{true} \]

\[ = \frac{1}{\sqrt{2\pi\sigma_{true}^2}} \int_{-\infty}^{\infty} e^{-\frac{(x_{true} - x_{calc})^2}{2\sigma_{true}^2}} \text{d}x_{true} \]

\[ = \sqrt{\frac{1}{2\pi}} \text{RMSE} \]

The model was implemented using PyMC293 observable quantities were set to their computed or experimental values, and 500 samples drawn from the posterior (after discarding an initial 500 samples to burn-in) using the default NUTS sampler. Expectations and posterior predictive intervals were computed from the marginal distributions obtained from the resulting traces.

Code availability. Scripts used for statistics analysis (including the Bayesian inference model) can be found at the following URL: https://github.com/kehauer/Predicting-resistance-of-clinical-Abm-mutations-to-targeted-kinase-inhibitors-using-FEP.

Data availability. All relevant data are publicly available: compiled experimental datasets, input files for Prime and FEP+, and computational results that support our findings can be found at GitHub by following the URL: https://github.com/kehauer/Predicting-resistance-of-clinical-Abm-mutations-to-targeted-kinase-inhibitors-using-FEP.

Received: 8 January 2018 Accepted: 15 May 2018 Published online: 13 June 2018

References

1. Roskoski, R. Jr. USFDA approved protein kinase inhibitors. http://www.brinm.org/PIK/PKIs.htm (2017).
2. Santos, R. et al. A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. 16, 19–34 (2016).
3. Shah, N. P. et al. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. Cancer Cell 2, 117–125 (2002).
4. Buczek, M., Escudier, B., Bartnik, E., Szczyluk, C. & Czarnecka, A. Resistance to tyrosine kinase inhibitors in clear cell renal cell carcinoma: From the patient’s bed to molecular mechanisms. Biochim. Biophys. Acta Rev. Cancer 1845, 31–41 (2014).
5. Huang, L. & Fu, L. Mechanisms of resistance to EGFR tyrosine kinase inhibitors. Acta Pharmacol. Sin. 8, 390–401 (2015).
6. Meyer, S. C. & Levine, R. L. Molecular pathways: molecular basis for sensitivity and resistance to JAK kinase inhibitors. Clin. Cancer Res. 20, 2051–2059 (2014).
7. Davare, M. A. et al. Structural insight into selectivity and resistance profiles of ROS1 tyrosine kinase inhibitors. Proc. Natl Acad. Sci. USA 112, E5381–E5390 (2015).
8. Van Allen, E. M. et al. The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. Cancer Discov. 4, 94–109 (2014).
9. Rani, S. et al. Neurenomed U: a candidate biomarker and therapeutic target to predict and overcome resistance to HER-tyrosine kinase inhibitors. Cancer Res. 74, 3821–3833 (2014).
10. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. Nat. Rev. Cancer 13, 714–726 (2013).
11. Weisberg, E., Manley, P. W., Cowan-Jacob, S. W., Hochhaus, A. & Griffin, J. D. Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukemia. Nat. Rev. Cancer 7, 345–356 (2007).
12. Lu, Y. X., Cai, Q. & Ding, K. Recent developments in the third generation inhibitors of Bcr-Abi for overriding T315I mutation. Curr. Med. Chem. 18, 2146–2157 (2011).
13. Juchum, M., Günther, M. & Laufer, S. A. Fighting cancer drug resistance: opportunities and challenges for mutation-specific EGFR inhibitors. Drug Resist. Updat. 20, 12–28 (2015).
14. Song, Z., Wang, M. & Zhang, A. Alectinib: a novel second generation anaplastic lymphoma kinase (ALK) inhibitor for overcoming clinically-acquired resistance. Acta Pharmacol. Sin. 35, 34–37 (2015).
15. Neel, D. S. & Bivona, T. G. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. Npj Precis. Oncol. 1, 1–6 (2017).
16. Gruber, F., Jhoirth-Hansen, H., Mikkola, I., Stenke, L. & TA, J. A novel BCR-ABL splice isoform is associated with the L248V mutation in CML patients with acquired resistance to imatinib. Leukemia 20, 2057–60 (2006).
17. Chandraratnapy, S. et al. AKT inhibition relieves feedback suppression of receptor tyrosine kinase expression and activity. Cancer Cell 19, 58–71 (2011).
18. Knight, Z. A., Lin, H. & Shokat, K. M. Targeting the kinase crosstalk between polypharmacology. Nat. Rev. Cancer 10, 130 (2010).
19. Housman, G. et al. Drug resistance in cancer: an overview. Nature 509, 1769–1792 (2014).
20. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature 526, 703–713 (2017).
21. Redig, A. J. & Janne, P. A. Basket trials and the evolution of clinical trial design in an era of genomice medicine. J. Clin. Oncol. 33, 973–977 (2015).
22. Hyman, D. M., Taylor, B. S. & Baselga, J. Implementing genome-driven oncology. Cell 168, 584–599 (2017).
23. Pesesky, M. W. et al. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data. Front. Microbiol. 7, 1887 (2016).
24. Melnikov, A., Rogov, P., Wang, L., Gneke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 42, e112–e112 (2014).
25. Chodera, J. D. et al. Alchemical free energy methods for drug discovery: progress and challenges. Curr. Opin. Struct. Biol. 21, 150–160 (2011).
26. Wang, L. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol. J. Med. Chem. 57, 2695–2703 (2015).

27. Abel, R. et al. Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. Curr. Opin. Struct. Biol. 43, 38–44 (2017).

28. Aldeghi, M., Heftet, A., Bodkin, M. J., Knapp, S. & Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. Chem. Sci. 7, 207–218 (2016).

29. Cappel, D. et al. Relative binding free energy calculations applied to protein homology models. J. Chem. Inf. Model. 56, 2388–2400 (2016).

30. Clark, A. J. et al. Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of HIV-1. J. Mol. Biol. 429, 930–947 (2017).

31. Steinbrecher, T. et al. Predicting the effect of amino acid single-point mutations on protein stability-large-scale validation of md-based relative free energy calculations. J. Mol. Biol. 429, 948–963 (2017).

32. Ford, M. C. & Babaoglu, K. Examining the feasibility of using free energy perturbation (FEP+) in predicting protein stability. J. Chem. Inf. Model. 57, 1276–1285 (2017).

33. Zou, J., Song, R., Simmerling, C. & Raleigh, D. Experimental and computational analysis of protein stabilization by Glys-to- d-Ala substitution: a convocation of native state and unfolded state effects. J. Am. Chem. Soc. 138, 15682–15689 (2016).

34. Mehandal, J., Tiwary, P. & Berne, B. J. How a kinase inhibitor withstands gatekeeper residue mutations. J. Am. Chem. Soc. 138, 4608–4615 (2016).

35. Lovering, F. et al. Imidazotriazines: spleen tyrosine kinase (Syk) inhibitors identified by free-energy perturbation (FEP). ChemMedChem 11, 217–233 (2016).

36. Rapp, C., Kalyanaraman, C., Schifflammer, A., Schoenbrun, E. L. & Jacobson, M. P. A molecular mechanics approach to modeling protein–ligand interactions: relative binding affinities in congeric series. J. Chem. Inf. Model. 51, 2082–2089 (2011).

37. 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).

38. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. J. Comput. Chem. 34, 2135–2145 (2013).

39. Maier, J. A. et al. fli4SB: improving the accuracy of protein side chain and backbone parameters from fli99SB. J. Chem. Theory Comput. 11, 3696–3713 (2015).

40. O’Hare, T. Combined Abl inhibitor therapy for minimizing drug resistance in chronic myeloid leukemia: Src/Abl inhibitors are compatible with imatinib. Clin. Cancer Res. 11, 6987–6993 (2005).

41. Shan, Y. et al. A conserved protonation-dependent switch controls drug resistancelarge-scale validation of md-based relative free energy calculations employing lattice-sum methods: an accurate correction scheme for electrostatic finite-size effects. J. Chem. Phys. 139, 184105 (2013).

42. Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: The langevin piston method. J. Chem. Phys. 103, 4631–4621 (1995).

43. Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. J. Chem. Phys. 101, 4177–4189 (1994).

44. Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. J. Chem. Phys. 97, 1990–2001 (1992).

45. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time scale molecular dynamics through hydrogen mass repartitioning. J. Chem. Theory Comput. 11, 1864–1874 (2015).

46. Wang, L., Berne, B. J. & Friesner, R. A. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. Proc. Natl. Acad. Sci. USA 109, 6200–6205 (2012).

47. Price, D. J. & Jorgensen, W. L. Computational binding studies of human gp60-src SH2 domain with a series of nonpeptide, phosphophenyl-containing ligands. Bioorg. Med. Chem. Lett. 10, 2067–2070 (2000).

48. Luccarelli, J., Michel, J., Tirado-Rives, J. & Jorgensen, W. L. Effects of water placement on predictions of binding affinities for p38 MAP kinase inhibitors. J. Chem. Theory Comput. 6, 3850–3856 (2010).

49. Michel, J., Verdonk, M. L. & Essex, J. W. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? J. Med. Chem. 49, 7427–7439 (2006).

50. Barouch-Bentor, R. & Sauer, K. Mechanisms of drug resistance in kinases. Expert Opin. Invest. Drugs 20, 153–208 (2011).

51. McDermott, U. et al. Discrimination of genotype-related sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. Proc. Natl. Acad. Sci. USA 104, 19936–19941 (2007).

52. Salvatier, J., Wiecki, T. V. & Fong, E. Probabilistic programming in python using pymc3. PeerJ Comput. Sci. 2, e55 (2016).

53. Schrock, A., Chen, T.-H., Clarkson, T. & Rivera, V. M. Comprehensive analysis 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).

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

55. Soverini, S. et al. Contribution of abl kinase domain mutations to imatinib resistance in different subsets of Philadelphia-positive patients: By the gimea
working party on chronic myeloid leukemia. Clin. Cancer Res. 12, 7374–7379 (2006).

82. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1046–1051 (2011).

Acknowledgements
We thank Daniel Robinson (Schrödinger), Sonya M. Hanson (MSKCC), and Gregory A. Ross (MSKCC) for helpful discussions. J.D.C. acknowledges support from NIH National Cancer Institute Cancer Center Core Grant P30 CA008748; J.D.C. and S.K.A. acknowledge support from the Sloan Kettering Institute, Cycle for Survival, and NIH grant R01 GM121505. K.H. acknowledges help from Wei Chen (Schrödinger) and Anthony Clark (Schrödinger) for instructions on running mutations changing the net charge of the system, and Simon Gao (Schrödinger) for assistance in computational resources.

Author contributions
K.H., J.D.C., C.N., R.A., and L.W. designed the research; K.H., S.A., T.S., and L.W. identified experimental datasets; K.H. and L.W. performed the simulations; K.H., C.N., S.K.A., S.R., T.S., R.A., J.D.C., and L.W. analyzed the data; K.H., J.D.C., S.K.A., and L.W. wrote the paper.

Additional information
Supplementary information accompanies this paper at https://doi.org/10.1038/s42003-018-0075-x.

Competing interests: J.D.C. is a member of the Scientific Advisory Board for Schrödinger Inc. S.R. is a former employee of Schrödinger Inc.; and K.H., C.N., R.A., T.S., and L.W. are employees of Schrödinger Inc.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018