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

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

Computational Biomolecular Dynamics Group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

Supporting Information

ABSTRACT: Resistance to small molecule drugs often emerges in cancer cells, viruses, and bacteria as a result of the evolutionary pressure exerted by the therapy. Protein mutations that directly impair drug binding are frequently involved in resistance, and the ability to anticipate these mutations would be beneficial in drug development and clinical practice. Here, we evaluate the ability of three distinct computational methods to predict ligand binding affinity changes upon protein mutation for the cancer target Abl kinase. These structure-based approaches rely on first-principle statistical mechanics, mixed physics- and knowledge-based potentials, and machine learning, and were able to estimate binding affinity changes and identify resistant mutations with remarkable accuracy. We expect that these complementary approaches will enable the routine prediction of resistance-causing mutations in a variety of other target proteins.

INTRODUCTION

Drug resistance is one of the chief challenges to be overcome in the development of robust anticancer and antimicrobial therapies. While resistance can emerge via multiple mechanisms, such as increased drug efflux or activation of alternative signaling pathways, it is often caused by protein mutations directly impacting drug binding.1,2 Anticipating these resistance-causing mutations is of interest for personalized medicine, as it would inform treatment decisions based on the patient’s genotype3,4 and aid the development of combination therapies. It would also benefit drug development by allowing the parallel exploration of inhibitors with different resistance profiles. While large-scale experimentation is feasible,6 it is neither cheap nor convenient, and accurate computer predictions mitigate the need for systematic experimentation.

Protein kinases are among the most exploited drug targets, with 48 inhibitors approved to date in the United States.7 The majority of these inhibitors target tyrosine kinases,6 which play a critical role in the modulation of growth factor signaling.7,8 Hence, tyrosine kinase inhibitors (TKIs) are employed against a number of malignancies, like chronic myelogenous leukemia (CML) and nonsmall-cell lung cancer.7,8 In particular, TKIs targeting the human kinase Abl are the first-line therapy for the treatment of CML.9 However, susceptibility to resistance requires continued development of new-generation inhibitors.8 For instance, in nonsmall-cell lung cancer, acquired resistance inevitably occurs within 1–2 years of starting the therapy.10 In CML, more than 25% of patients switch TKI due to intolerance or resistance,11 the latter being most often caused by mutations in Abl.8 Because kinases display a long tail of rare and uncharacterized mutations,12 the sensitivity of many clinically identified kinase mutants to TKI treatment is often unknown. Thus, rapid and accurate computational approaches could impact clinical decision-making by providing oncologists with a first indication of whether the observed mutation is likely to cause resistance to certain inhibitors.

Here, we show how both physics-based and data-driven computational approaches can be used to accurately estimate the change in affinity of TKIs for the human kinase Abl caused by single-point mutations. To test the different methodologies, we used a data set of 144 Abl:TKI affinity changes (\(\Delta\Delta G\)) across 8 TKIs caused by 31 Abl mutations (Figure 1), which was compiled by Hauser et al.12 using publicly available cell viability IC50 data. The computational approaches tested were based on (i) molecular dynamics (MD) simulations7 with a nonequilibrium free energy calculation protocol;14−16 (ii) Rosetta, a modeling program that uses mixed physics- and knowledge-based potentials;17,18 and (iii) machine learning (ML).19 For completeness, we discuss our results also in light of those obtained by Hauser et al.12 with the OPLS3 force field20 and a MD-based approach similar to ours (Table 1).

RESULTS AND DISCUSSION

The performance of all calculations was assessed by the root-mean-square error (RMSE) between calculated and experimentally measured \(\Delta\Delta G\) values, the Pearson correlation coefficient (\(r\)), and the area under the precision recall curve (AUPRC).21 The latter measures the ability to classify mutations as resistant or susceptible. In particular, precision measures which fraction of mutations classified as resistant are actually resistant, whereas recall measures which fraction of
resistant mutations are correctly classified as resistant. Consistent with previous work, resistant mutations were defined as those causing more than a 10-fold drop in affinity ($\Delta \Delta G_{\text{exp}} > 1.36$ kcal/mol). On this data set, given the fraction of resistant mutations (Figure 1c), a random classifier would return an AUPRC of 0.13.

First, we estimated changes in TKI affinity with an approach based on the first-principles of statistical mechanics. MD simulations were used to sample the isothermal–isobaric ensemble while a specific residue was “alchemically” mutated into another one. From the nonequilibrium work required to transform the residue, it is possible to recover the equilibrium free energy difference and, using a suitable thermodynamic cycle (Figure S1), a rigorous estimate of $\Delta \Delta G$. We tested multiple force fields, and, on this data set, Amber force fields performed better than Charmm (Figure S2, Table S1); halogen bond modeling did not improve the Charmm results (Figure S3). For simplicity, we discuss only the results obtained with the best-performing Amber (A99) and Charmm (C22) force fields (Table 1). The performance of A99 was comparable to that of OP3 (Figure 2), which was obtained by Hauser et al. with the FEP+ program by Schrödinger Inc. and the recently optimized proprietary OPLS force field. C22 performed significantly worse than A99 and OP3 in terms of correlation and classification ability.

Surprisingly, combining force fields in a consensus approach (considering half of the simulations for each parent force field) did not generally improve the results (Figure S4). It is conceivable that, on this data set, the benefits of the additional sampling obtained with one force field outweighs those expected by combining multiple force fields. We also did not observe a dependence of the accuracy of the calculation on

Figure 1. Data set of Abl kinase mutations and associated TKI affinity changes ($\Delta \Delta G$) studied. (a) Structure of human Abl kinase (PDB-ID 1OPJ) with imatinib (light orange) bound. Mutated wild-type residues are shown as violet sticks. (b) Name and chemical structure of the 8 TKIs studied. (c) Distribution of the 144 experimental $\Delta \Delta G$ values. The line at $\Delta \Delta G = 1.36$ kcal/mol separates mutations defined as resistant from susceptible.
the net charge change of the system upon mutation (Figure S5).

While our free energy protocol used a total of 216 nS of MD per ΔΔG estimate,16 OP3 used 720 ns for charge-changing mutations, and 360 ns for charge-conserving ones.12 We thus also tested a more expensive protocol (A99f), which matched exactly the simulation time invested by OP3. A99f improved upon the uncertainty (Figure S6), but not upon the accuracy (Figure 2), of the A99 estimates. Lower uncertainty is expected when more calculation repeats are used as in this case. On the other hand, accuracy is not necessarily proportional to the computation generally resulted in enhanced performance (Figure S7). For instance, combining A99 and R15, we obtained a particularly strong consensus estimate (Figure S8) with RMSE = 0.62 kcal/mol, r = 0.71, and AUPRC = 0.61. This can be attributed to the complementary performance of the two approaches: while the MD based calculations showed better precision performance and a lower number of false positives, Rosetta displayed better recall and a lower number of false negatives (Table S2, Figure 2).

As a purely data-driven approach, we employed extremely randomized regression trees,14 a ML technique that uses ensembles of decision trees similarly to random forest. As input features, we used terms derived from the crystallographic protein–ligand structures (e.g., hydrogen bonds, nonpolar contacts, residue-ligand distance), ligand and residue physical model. The technical details of the ML model construction are provided in the SI Methods. A n

\[ \Delta \Delta G = \text{RMSE} \times \text{AUPRC} \]

For the performance measures, the point estimates from the original samples and their 95% bootstrapped confidence intervals are shown (\( \Delta \Delta G \)). RMSE: root-mean-square error; AUPRC: area under the precision-recall curve.

Data for the MD calculations with the OPLS force field refer to and were taken from Hauser et al.12

Nvidia (Pascal architecture).

Intel Xeon (Broadwell architecture).

Nvidia GeForce GTX 1080 Titan.

Intel Xeon (Broadwell architecture).

Time for charge-changing mutations the simulation time is double.

Table 1. Summary of the Approaches Used, Their Performance, and Computational Cost

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Method</th>
<th>Force field or scoring function</th>
<th>Hardware</th>
<th>Compute hours</th>
<th>RMSE (kcal/mol)</th>
<th>Pearson</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP3</td>
<td>Molecular Dynamics</td>
<td>OPLS3⁷⁶</td>
<td>1 GPU ¹</td>
<td>72²</td>
<td>1.07 ± 0.15</td>
<td>0.49 ± 0.19</td>
<td>0.56 ± 0.12</td>
</tr>
<tr>
<td>C22</td>
<td>Molecular Dynamics</td>
<td>Charmm22* and CGenFF v 3.01</td>
<td>10 CPU ³</td>
<td>59</td>
<td>1.03 ± 0.15</td>
<td>0.24 ± 0.04</td>
<td>0.25 ± 0.12</td>
</tr>
<tr>
<td>A99</td>
<td>Molecular Dynamics</td>
<td>Amber99sb*-ILDN and GAFF v 2.1</td>
<td>10 CPU ³</td>
<td>59</td>
<td>0.91 ± 0.07</td>
<td>0.40 ± 0.10</td>
<td>0.56 ± 0.11</td>
</tr>
<tr>
<td>A99f</td>
<td>Molecular Dynamics</td>
<td>Amber99sb*-ILDN and GAFF v 2.1</td>
<td>10 CPU ³</td>
<td>98²</td>
<td>0.91 ± 0.10</td>
<td>0.42 ± 0.10</td>
<td>0.51 ± 0.23</td>
</tr>
<tr>
<td>R15</td>
<td>Rosetta</td>
<td>REF15</td>
<td>1 CPU ³</td>
<td>32</td>
<td>0.72 ± 0.10</td>
<td>0.67 ± 0.11</td>
<td>0.58 ± 0.12</td>
</tr>
<tr>
<td>R16</td>
<td>Rosetta</td>
<td>βNOV16</td>
<td>1 CPU ³</td>
<td>32</td>
<td>0.71 ± 0.10</td>
<td>0.59 ± 0.13</td>
<td>0.39 ± 0.18</td>
</tr>
<tr>
<td>ML1</td>
<td>Machine Learning</td>
<td>n/a</td>
<td>1 CPU ³</td>
<td>0.02</td>
<td>0.87 ± 0.06</td>
<td>0.12 ± 0.04</td>
<td>0.28 ± 0.39</td>
</tr>
<tr>
<td>ML2</td>
<td>Machine Learning</td>
<td>n/a</td>
<td>1 CPU ³</td>
<td>0.02</td>
<td>0.69 ± 0.05</td>
<td>0.57 ± 0.12</td>
<td>0.49 ± 0.23</td>
</tr>
</tbody>
</table>

²For the performance measures, the point estimates from the original samples and their 95% bootstrapped confidence intervals are shown (\( \Delta \Delta G \)).

³Data for the MD calculations with the OPLS force field refer to and were taken from Hauser et al.12

⁴Nvidia (Pascal architecture).

⁵Intel Xeon E5-2630 v 4.

⁶Nvidia GeForce GTX 1080 Titan.

⁷Intel Xeon (Broadwell architecture).

⁸Time for charge-changing mutations the simulation time is double.

Abbreviation

<table>
<thead>
<tr>
<th>Method</th>
<th>Force field or scoring function</th>
<th>Hardware</th>
<th>Compute hours</th>
<th>RMSE (kcal/mol)</th>
<th>Pearson</th>
<th>AUPRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP3</td>
<td>Molecular Dynamics</td>
<td>OPLS3⁷⁶</td>
<td>1 GPU ¹</td>
<td>72²</td>
<td>1.07 ± 0.15</td>
<td>0.49 ± 0.19</td>
</tr>
<tr>
<td>C22</td>
<td>Molecular Dynamics</td>
<td>Charmm22* and CGenFF v 3.01</td>
<td>10 CPU ³</td>
<td>59</td>
<td>1.03 ± 0.15</td>
<td>0.24 ± 0.04</td>
</tr>
<tr>
<td>A99</td>
<td>Molecular Dynamics</td>
<td>Amber99sb*-ILDN and GAFF v 2.1</td>
<td>10 CPU ³</td>
<td>59</td>
<td>0.91 ± 0.07</td>
<td>0.40 ± 0.10</td>
</tr>
<tr>
<td>A99f</td>
<td>Molecular Dynamics</td>
<td>Amber99sb*-ILDN and GAFF v 2.1</td>
<td>10 CPU ³</td>
<td>98²</td>
<td>0.91 ± 0.10</td>
<td>0.42 ± 0.10</td>
</tr>
<tr>
<td>R15</td>
<td>Rosetta</td>
<td>REF15</td>
<td>1 CPU ³</td>
<td>32</td>
<td>0.72 ± 0.10</td>
<td>0.67 ± 0.11</td>
</tr>
<tr>
<td>R16</td>
<td>Rosetta</td>
<td>βNOV16</td>
<td>1 CPU ³</td>
<td>32</td>
<td>0.83 ± 0.10</td>
<td>0.59 ± 0.13</td>
</tr>
<tr>
<td>ML1</td>
<td>Machine Learning</td>
<td>n/a</td>
<td>1 CPU ³</td>
<td>0.02</td>
<td>0.87 ± 0.06</td>
<td>0.12 ± 0.04</td>
</tr>
<tr>
<td>ML2</td>
<td>Machine Learning</td>
<td>n/a</td>
<td>1 CPU ³</td>
<td>0.02</td>
<td>0.69 ± 0.05</td>
<td>0.57 ± 0.12</td>
</tr>
</tbody>
</table>

²For the performance measures, the point estimates from the original samples and their 95% bootstrapped confidence intervals are shown (\( \Delta \Delta G \)).

³Data for the MD calculations with the OPLS force field refer to and were taken from Hauser et al.12

⁴Nvidia (Pascal architecture).

⁵Intel Xeon E5-2630 v 4.

⁶Nvidia GeForce GTX 1080 Titan.

⁷Intel Xeon (Broadwell architecture).

⁸Time for charge-changing mutations the simulation time is double.
Figure 2. Accuracy of the ΔΔG estimates. (a) Scatter plots of experimental versus calculated ΔΔG values. The identity is shown as a dashed gray line. The four quadrants indicate the location of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) according to the definition of resistant and susceptible mutations used (resistant if ΔΔGexp > 1.36 kcal/mol, susceptible otherwise)12 and an equivalent cutoff (1.36 kcal/mol) for the calculated ΔΔG values. Each ΔΔG estimate is color-coded according to its absolute error with respect to the experimental ΔΔG value; at 300 K, an error of 1.4 kcal/mol corresponds to a ~10-fold error in Kd change, and an error of 2.8 kcal/mol to a ~100-fold error in Kd change. (b) Summary of the performance of the ΔΔG estimates across approaches in terms of RMSE, Pearson correlation, and AUPRC; point estimates from the original samples and 95% bootstrapped confidence intervals are shown (SI Methods). Differences at three levels of significance are reported using labels within the chart: e.g., a "C22**" label above the RMSE mark of OP3 indicates that the RMSE of OP3 is significantly lower (i.e., better agreement with experiment) than that of C22 at α = 0.05.
another ML model based on Gaussian process regression (Figure S9).
Given that ML is expected to perform best when trained on the most relevant data, we asked whether the model performance would improve when trained on tyrosine kinase mutations only. In this case, we effectively assessed the interpolative ability of the model, as it was trained and tested on the TKI data set (ML2). To do this, we used 8-fold nested cross-validation. At each iteration, the model selected a subset of the 128 features via 7-fold cross validation with 7 inhibitors and was then tested on the eighth. This was done 8 times to obtain predictions for all TKIs and thus all entries in the data set. ML2 accurately estimated ΔΔG values and could discriminate well between resistant and susceptible mutations (RMSE = 0.68 kcal/mol, r = 0.57, AUPRC = 0.47).
If the objective is only to identify resistant mutations rather than assessing their impact on ligand binding quantitatively, then AUPRC is the most relevant performance measure to examine. In this regard, MD based calculations with the OPLS3 or Amber99 force fields (OP3, A99, A99f), and Rosetta with the REf15 scoring function (R15) provided the best performance (Figure 3, Figure S10). Furthermore, in this scenario, one is likely willing to accept a larger fraction of false positives if this means recovering more resistant mutations. Thus, a suitable consensus strategy is to select the most positive ΔΔG estimate among those available to maximize the recall (at the expense of precision) at a given threshold (Figure S11). Using this approach with A99 and R15 returns an AUPRC of 0.62 and, at the conventional threshold of ΔΔGcalc > 1.36 kcal/mol, a recall of 0.79 and a precision of 0.48 (Figure 3; Table S2). This means that about half of the mutations classified as resistant are in fact resistant, and 79% of the resistant mutations in the data set have been identified.
From the results presented, it emerges that the latest computational approaches, from physics-based to data-driven ones, are able to predict TKI affinity changes upon Abl mutation and identify resistance-causing mutations. However, the different methods have different strengths and weaknesses. The MD-based calculations achieved a remarkable performance considering that force field parameters are based on fundamental physical properties of organic molecules (e.g., quantum mechanical energies and electrostatic potentials, liquid densities, and enthalpies of vaporization), rather than binding affinities directly. This, together with the rigorous statistical mechanical treatment, might lie at the basis of the demonstrated generalizability of the approach across protein–ligand systems. The downside is the comparably higher computational cost required to obtain precise ΔΔG estimates. On the other hand, ML predictions proved accurate only when the binding affinities for Abl:TKI were used directly for training. With this information, however, a ΔΔG estimate could be generated in a matter of seconds to minutes on a single CPU core, with most of the time spent collecting the necessary structure-based features. Thus, the optimal approach choice in a prospective scenario will depend on the amount of prior information available for the system of interest, the nature of the study (a large-scale scan versus an in-depth study of a few mutations), as well as the computational resources at hand. That being said, we found that Rosetta struck a good balance between accuracy, generalizability, and computational cost. Similar to MD, this approach considers an ensemble of protein–ligand conformations, and its scoring function is based on simple physical and structural properties of organic molecules and biomolecules, so that it does not require binding affinities to the system of interest for training. In addition, approximations such as the use of implicit solvation and a more limited conformational sampling result in a moderate computational cost.

**CONCLUSION**

In summary, we have shown how fundamentally different structure-based approaches are able to predict TKI resistance to similar or better accuracy than reported thus far. Taking advantage of multiple techniques in a consensus fashion resulted in remarkable accuracy and tuning of the classification performance. These results suggest that a diverse set of computational approaches, each with its own strengths and weaknesses, is now available for the reliable estimate of resistance-causing protein mutations. It is our hope that the availability of multiple complementary approaches will enable the routine prediction of resistance-causing mutations across many other protein targets.

**ASSOCIATED CONTENT**

*Supporting Information*

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.9b00590.

Methods, including supporting figures and data tables (PDF)
Data S1: detailed information on the data set and numerical results for all calculations (XLSX)

Data S2: input files pertaining to the molecular-dynamics-based free energy calculations and Rosetta calculations (ZIP)

Data S3: input files and Jupyter notebooks used for the training and testing of the machine learning model. Available at https://doi.org/10.5281/zenodo.3350897

■ AUTHOR INFORMATION

Corresponding Authors
*E-mail: maldegh@gwdg.de.
*E-mail: bgroot@gwdg.de.

ORCID
Matteo Aldeghi: 0000-0003-0019-8806
Vytautas Gapsys: 0000-0002-6761-7780
Bert L. de Groot: 0000-0003-3570-3534

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.A. was supported by a Postdoctoral Research Fellowship of the Alexander von Humboldt Foundation. V.G. was supported by the BioExcel CoE (www.bioexcel.eu), a project funded by the European Union (Contract H2020-EINFRA-2015-1-675728).

■ REFERENCES

(24) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and


