Published in final edited form as:

J Phys Chem B. 2024 February 22; 128(7): 1656–1667. doi:10.1021/acs.jpcb.3c08097.

Potts Hamiltonian Models and Molecular Dynamics Free Energy Simulations for Predicting the Impact of Mutations on Protein Kinase Stability

Abhishek Thakur[#],

Center for Biophysics and Computational Biology and Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States

Joan Gizzio[#],

Center for Biophysics and Computational Biology and Department of Chemistry, Temple University, Philadelphia, Pennsylvania 19122, United States

Ronald M. Levy

Center for Biophysics and Computational Biology, Department of Chemistry, and Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States

Abstract

Single-point mutations in kinase proteins can affect their stability and fitness, and computational analysis of these effects can provide insights into the relationships among protein sequence, structure, and function for this enzyme family. To assess the impact of mutations on protein stability, we used a sequence-based Potts Hamiltonian model trained on a kinase family multiple-sequence alignment (MSA) to calculate the statistical energy (fitness) effects of mutations and compared these against relative folding free energies ($\Delta\Delta G$s) calculated from all-atom molecular dynamics free energy perturbation (FEP) simulations in explicit solvent. The fitness effects of

Corresponding Author: Ronald M. Levy – *Center for Biophysics and Computational Biology, Department of Chemistry, and Department of Physics, Temple University, Philadelphia, Pennsylvania 19122, United States*; ronlevy@temple.edu.

[#]Author Contributions

A.T. and J.G. contributed equally to this work.

Notes

The authors declare no competing financial interest.

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.3c08097

Published as part of The Journal of Physical Chemistry B virtual special issue “Gregory A. Voth Festschrift”.

ASSOCIATED CONTENT

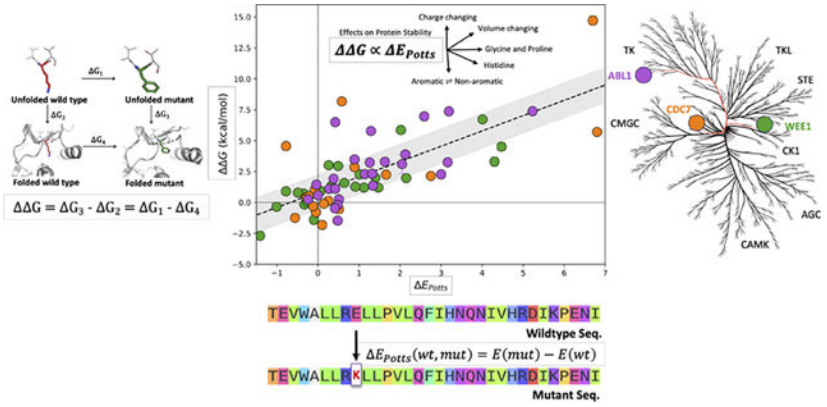
Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.3c08097.

Estimating the stability effects of 65 cancer mutations in WEE1, CDC7, and ABL1 kinases using FEP ($\Delta\Delta G$), Potts energies (ΔE), and MM/GBSA scores; kinome plot to provide visualization of the functional diversity of selected kinases; comparing FEP’s stability predictions to Potts and MM/GBSA via Pearson r and MUE for 23 WEE-1 mutations; comparing FEP’s stability predictions to Potts and MM/GBSA via Pearson r and MUE for 17 CDC7 mutations; comparing FEP’s stability predictions to Potts and MM/GBSA via Pearson r and MUE for 25 ABL1 mutations, analyzing the predictive performance (accuracy and precision) between the Potts model and FEP with experimental results for predicting the effects of mutations on protein stability; comparing different Potts energy calculations to FEP by using threading over crystal structure, FEP starting structure, MD trajectory, and no cutoff; performance summary of the Potts model and MM/GBSA, showing the mutation that leads to large side-chain vdW volume change between wild-type and mutant states, showing improvement in Pearson r between $\Delta\Delta G$ and ΔE after removing outliers from main text Figure 4C, estimating the protein stability effects by FEP and Potts model over 13 mutations that involve transformations between aliphatic and aromatic side chains, demonstrating that sequence-based method implicitly accounts for ensembles of tautomeric/protonation states when mutating to histidine (PDF)

mutations in the Potts model (ΔEs) showed good agreement with experimental thermostability data (Pearson $r = 0.68$), similar to the correlation we observed with $\Delta\Delta G$s predicted from structure-based relative FEP simulations. Recognizing the possible advantages of using Potts models to rapidly estimate protein stability effects of kinase mutations seen in cancer genomics data, we used the Potts statistical energy model to estimate the stability effects of 65 conservative and nonconservative mutations across three distinct kinases (Wee1, Abl1, and Cdc7) with somatic mutations reported in the Genomic Data Commons (GDC) database. The ΔEs of these mutations calculated from the Potts model are consistent with the corresponding $\Delta\Delta G$s from FEP simulations (Pearson ratio of 0.72). The agreement between these methods suggests that the Potts model may be used as a sequence-based tool for high-throughput screening of mutational effects as part of a computational pipeline for predicting the stability effects of mutations. We also demonstrate how the scalability of the fitness-based Potts model calculations permits analyses that are not easily accessed using FEP simulations. To this end, we employed site-saturation mutagenesis in the Potts model in order to investigate the relative stability effects of mutations seen in different cancer evolutionary scenarios. We used this approach to analyze the effects of drug pressure in Abl kinase by contrasting the relative fitness penalties of somatic mutations seen in miscellaneous cancer types with those calculated for mutations associated with cancer drug resistance. We observed that, in contrast to somatic mutations of Abl seen in various tumors that appear to have evolved neutrally, cancer mutations that evolved under drug pressure in Abl-targeted therapies tend to preserve enzyme stability.

Graphical Abstract



INTRODUCTION

A single mutation in the amino acid sequence of a protein can lead to structurally or functionally deleterious or advantageous effects. Deciphering and leveraging the effects of mutations on protein fold stability is challenging due to the vast number of possible mutations and the complexity of the mutational landscape.[1–4] While the full scope of biological consequences for making mutations is difficult to predict without experimental guidance, the effects of single-point mutations on protein thermostability can in principle be probed entirely *in silico.* The impact of mutations on the stability of the protein fold provides insights into the physical and sequence-evolutionary constraints on protein folding

and offers a framework for predicting the biological consequences of mutations. Over the past decade, structure-based computational approaches that have been increasingly utilized to guide and complement experiments include implicit solvent molecular mechanics/Gibbs–Boltzmann surface area (MM/GBSA) and Analytical Generalized Born plus NonPolar (AGBNP) models.[5,6] Free energy perturbation (FEP) is an explicit solvent molecular dynamics-based method, offering a robust approach for assessing the thermodynamic effects of amino acid substitutions on protein stability.[7–9] However, FEP's explicit representation of water and requirement for multiple MD simulations (lambda windows) make it computationally intensive. Nevertheless, the valuable insights gained from FEP simulations contribute to our understanding of protein structure–function relationships and the changes in stability caused by specific mutations.[10] On the other hand, MM/GBSA and AGBNP provide faster, implicit solvent alternatives, although they are more approximate methods.[6,11,5] However, when used in conjunction with FEP as part of a workflow, implicit solvent models can be used as a structure-based tool to more rapidly select mutation targets for further analysis, thus enhancing the scalability of the screening process.[9]

In the past decade, there has been a significant increase in experimental studies that map protein mutations and correlate their effects with biological fitness, providing valuable insights into protein function and structural stability. However, analyzing the effects of mutations over many members of a protein family rather than individual proteins poses challenges; thermodynamically rigorous structure-based methods such as FEP have their limitations and can also be very expensive in terms of resources and time when trying to understand the effects of mutations over many proteins from the same family. In contrast, sequence-based Potts Hamiltonian models can predict the effects of mutations on protein fitness at-scale, allowing large sequence data sets to be rapidly analyzed.

The Potts Hamiltonian is an information theoretic potential function trained on natural sequence covariation in a protein family MSA. Once trained, the statistical energy Hamiltonian is able to predict direct interactions between residues that capture their complex covariation patterns in the MSA.[12] The Hamiltonian is parametrized by pairwise couplings J between pairs of amino acids, which can be interpreted as direct coevolutionary contributions to the total fitness of the protein sequence.[13] The Potts coupling between residues that are "in contact" in the protein tertiary structure can generally be interpreted as a fold stability contribution to fitness due to their side-chain interactions.[13] Potts models are well suited for predicting the effects of mutations on protein stability.[14,15] While the initial determination of the statistical energy parameters of $\sim 10^5$ or more residue pairs for a typical $L \approx 300$ protein family is a computationally intensive task, the Potts Hamiltonian once parametrized can be used to perform mutation stability analysis at-scale for any number of sequences. There are, in principle, up to $20^2 \times \binom{L}{2}$ unique interaction parameters that go into a Potts Hamiltonian for a protein family with (aligned) primary sequence length L and a natural 20-letter amino acid alphabet, but in practice, there are many fewer parameters as only residues that appear at each position in the MSA are fit. These parameters can be inferred from a protein family MSA without analytical approximations using generative methods such as the Mi3-GPU software developed in our lab.[16] FEP simulations, on the other hand, depend on molecular mechanics force fields to directly probe

the relative changes in free energy of two structural ensembles (e.g., folded vs unfolded). In the literature, there is currently no study that shows how well the two methods, structure-based and sequence-based, i.e., FEP and Potts models, are correlated in their prediction of mutational effects on protein stability.

In previous research conducted by our group, we developed a Potts Hamiltonian model for the protein kinase family and used it to analyze relative conformational preferences.[13,17] The kinase Potts model, inferred using Mi3-GPU, has also been shown to reproduce higher-order residue correlations (beyond pairwise) seen in experimental sequence data from the UniProt database.[12,18] In the present work, the fitness effects of mutations, modeled as statistical energy differences between wild-type and mutant sequences in the Potts model (ΔEs), serve as a proxy for the effects of mutations on stability (relative change in folding free-energy, $\Delta\Delta G$). This was done by using the Potts model to predict the effects of single-point mutations on kinase fitness and comparing the results with relative FEP simulations, focusing on a set of somatic mutations in kinases that were identified from tumor samples. We note that the effects of kinase mutations on protein folding and stability make a major contribution to protein fitness; there are other contributions, especially effects on enzymatic activity and substrate recognition, that also contribute to the fitness of the mutant enzyme. As a benchmark analysis, we first compared the relative mutant stability predictions from both methods with experimental thermostability data for three kinases (Abl1,[19] Pim1,[20] and EphA3[21]). We then focused our analysis on a set of 65 somatic mutations obtained from cancer genomic data in the GDC database,[1] which are spread across the catalytic domains of Abl1 and two additional kinases, Wee1 and Cdc7. To evaluate the effects of these mutations on the stability of the active state of the catalytic domain, we calculated the relative change in folding free-energy upon mutation, $\Delta\Delta G$, using the FEP+/REST2 Hamiltonian replica exchange method[23,24] with a tripeptide model of the reference unfolded state (Figure 1), as has been used previously.[9] By assessing the correlation between sequence-based fitness changes for these somatic mutations (Potts ΔEs) and the FEP-derived $\Delta\Delta G$s, we gain insights into the practical utility of Potts models when integrated into computational pipelines to screen the effects of clinically observed mutations.

Computational methods that are able to predict the stability consequences of mutations and are highly scalable, such as the Potts model, allow for comprehensive analyses of sequence data which are not readily accessible by FEP simulations alone. For example, selective pressures underlying the distributions of mutations seen in different cancer evolutionary scenarios can be analyzed in the Potts model by performing *in silico* site-saturation mutagenesis to discern biophysical signatures associated with neutrally evolved versus positively selected variation. We used this approach to analyze the effects of drug pressure on Abl kinase. We observed that, in contrast to the set of somatic mutations of Abl seen in various tumors that tend to have deleterious effects on enzyme stability and appear to have evolved neutrally,[25] mutations that evolved under drug pressure in Abl-targeted therapies[2] tend to preserve enzyme stability.

RESULTS AND DISCUSSION

Effects of Single-Point Mutations on Protein Stability in Kinases.

The Potts model is a sequence-based machine-learning statistical energy function that, once trained, can be used to estimate the effects of mutations on protein fitness.[26] One approach, which we employ here, is to "thread" the Potts couplings of a particular kinase sequence over an experimentally solved and MD-refined structural model of the folded wild-type protein (see Methods) and sum the changes in coupling energies that occur upon mutation in the Potts model:

$$\Delta E_{\text{Potts}}(w,m) = E(m) - E(w) \tag{1}$$

$$\Delta E_{\text{Potts}}(w,m) = \sum_{i<j}^{L} (J^{ij}_{m_i m_j} - J^{ij}_{w_i w_j})\delta[d_{ij}(n) < 6\text{Å}] \tag{2}$$

where amino acids (residues) i and j in structure n contribute to the net change in statistical energy, ΔE_{Potts}, only if their distance in the folded protein structure $d_{ij}(n)$ is less than 6 Å (as measured from their closest approaching side chain heavy atoms, including Cβ atoms). In eq 2, δ is unity when this condition is met and is zero otherwise.

The relationship between the fitness of protein sequences and thermodynamic stability of the protein fold (Figure 1) allows the Potts calculation (eq 1) to be interpreted as a sequence-based analog of $\Delta\Delta G$, a thermodynamic observable defined as the difference in folding free energy between wildtype and mutant amino acid sequences (Figure 1).

$$\Delta\Delta G = \Delta G_{\text{mutant}} - \Delta G_{\text{wildtype}} \tag{3}$$

where the difference in free energy between the folded and unfolded states, ΔG, is defined by the log ratio of their equilibrium probabilities in bulk solution, i.e., $\Delta G = -kT\ln\left(\frac{P(\text{folded})}{P(\text{unfolded})}\right)$. The sign of $\Delta\Delta G$ indicates the direction of the stability change upon mutation where, relative to the unfolded state, $\Delta\Delta G < 0$ is stabilizing to the folded protein and $\Delta\Delta G > 0$ is destabilizing.

To evaluate the ability of our kinase Potts model to capture the stability effects of single-point mutations, we have calculated ΔE_{Potts} for 14 mutations over three different kinases with mutant thermostability data available in the literature (see Methods).[19–21] The Potts-calculated ΔEs have good correspondence with $\Delta\Delta G_{\text{exp}}$ estimated from experimental melting temperature shifts, with a Pearson correlation coefficient of $r = 0.68$ (Figure 2B). It should be noted that prior studies have reported a strong correlation between the Potts Hamiltonian

energy, used for estimating the effects of mutations on protein fitness and fold stability and experimental measurements of fold stability.[14,15,27]

Recent advancements in FEP simulations in terms of speed and accuracy have allowed studies to be performed over large data sets where authors have showcased the efficacy of FEP simulations by comparing predicted $\Delta\Delta G$s with experimental data across a variety of protein families.[9,28] The FEP-calculated $\Delta\Delta G$s in these studies tend to agree well with the experimental benchmarks, within a mean unsigned error (MUE) of ±1 kcal/mol and Pearson correlations ranging from 0.64 to 0.82. Our FEP simulations using these same methods for kinase proteins, performed over the much smaller set of benchmark mutations described above, show a statistically significant correlation with $\Delta\Delta G_{\text{exp}}$ estimated from melting temperature shifts ($r = 0.61$ with P value <0.05) and a MUE of ±1 kcal/mol (Figure 2A). This is similar to our results obtained when using the kinase Potts model (Figure 2B and Figure S5), suggesting a potential for the Potts model to be integrated with structure-based computational methods, i.e., FEP, for practical applications such as interpreting cancer genomics data.

To this end, we have identified a set of 65 somatic mutations curated from tumor samples by the GDC database (see Methods). These mutations are spread across the catalytic domains of three protein kinases (WEE1, CDC7, and ABL1). For each mutation, we calculated ΔE_{Potts} (eq 1) (Figure 2C), which, when compared with $\Delta\Delta G_{\text{FEP}}$ from FEP simulations, exhibit good correlation with a Pearson coefficient of $r = 0.72$ similar to what we observed between $\Delta\Delta G_{\text{FEP}}$ and $\Delta\Delta G_{\text{exp}}$ (Figure 2B). Interestingly, with FEP as a benchmark, the Potts statistical energies significantly outperform an end point implicit solvent molecular mechanics approach (MM/GBSA),[9,29] which displays a weaker correlation with the FEP results ($r = 0.43$) (Figure 2D). Correlation coefficients for the comparison between FEP, Potts, and MM/GBSA can be found in the Supporting Information (Figures S2–S4 and Table S1). This analysis demonstrates the utility of ΔE_{Potts} in predicting the effects of mutations on protein stability, potentially as a prescreening tool for FEP and subsequent experimental characterization. The scale of ΔE_{Potts} in Figure 2C is approximately 1.2 kcal/mol, consistent with the scale observed in a previous study where we used Potts threading and free energy simulations to study the relative conformational stabilities of evolutionarily related kinase sequences.[17,27]

Utility of Potts Models for Predicting the Effects of Structurally Nonconservative Mutations.

Our results from Figure 2B–D suggests that the Potts model is a viable tool for analyzing the effects of large numbers of mutations in a high-throughput manner. This opens up the ability to perform analyses that are not as easily accessible for FEP methods, for example, *in silico* site-saturation mutagenesis. When a mutation of interest is identified, e.g., somatic mutations from cancer genomics data such as those analyzed in Figure 2C, or mutations that arise under chemotherapeutic pressure as discussed in a later section (Figure 4), additional insights of a biological nature can be gleaned when the mutational stability effects are viewed in context with all other mutations that can be made at those same positions, i.e., when the mutational space is “saturated” by systematically mutating each position in a

wild-type protein to all amino acid types at that position observed in the UniProt sequence database. However, the accuracy of this approach relies in part on the ability to predict the effects of nonconservative mutations involving large perturbations to structure and dynamics: for example, large changes in side-chain vdW (van der Waals) volume, charge changes, and mutations involving glycines, prolines, and histidines, which have multiple charge and tautomeric states. In general, challenges may arise when structurally modeling mutations involving glycine residues, known for their increased backbone flexibility, or prolines that must establish covalent bonds with the peptide backbone, resulting in cyclization and introducing further complexities.[9,30] While methods for structural modeling and performing these nonconservative mutations are well established for FEP,[28,31] they can be quite resource-intensive and limited in terms of scalability. In contrast, Potts threading does not require mutant structural models and dynamical trajectories as input (Figure S6) as the changes in fitness-based couplings appear to include this information implicitly (eq 2); the calculation involves “threading” the mutant Potts energy terms onto a single wild-type structure, allowing one to rapidly estimate the effects of mutations at-scale for a large number of sequences in the protein family.

To evaluate the consistency of the Potts’s statistical energy in estimating kinase fold stability for nonconservative mutations involving changes in charge, large side-chain volume changes, glycine/prolines, and histidine, and motivate the site-saturation mutagenesis approach in the subsequent section, we separated our data set of 65 single-point mutations into different “challenge cases” to individually assess the confidence of their predicted relative stabilities from the Potts model in comparison with FEP (Table S2). We observe a Pearson correlation coefficient greater than 0.80 for mutations involving charge changes, considerable changes in side-chain vdW volume[32] compared to the wild type (≥60 $Å^3$), and histidines for which pK_a corrections were taken into account (Figure 3). For mutations that involve significant alterations to the protein backbone, i.e., glycine and proline, we observe a weak correlation between FEP and Potts predictions (Pearson ratio of 0.39) (Table S2 and Figure 3C). Three mutations in particular deviate significantly from the linear fit in Figure 3C and have a significant effect on the correlation coefficient (see Figure S8), all of them involving mutations to or from glycine (G401S and D62G in CDC7 and G436D in ABL1). Among the three glycine mutations, FEP simulations yielded remarkably deleterious $\Delta\Delta G$ values for mutations where glycine is the wild type, i.e., G401S (CDC7) and G436D (ABL1), in contrast to the corresponding values of ΔE_{Potts}, which are small in magnitude. As glycine does not have a side chain, unexpectedly large and positive values of $\Delta\Delta G$ may indicate issues encountered in FEP simulations related to backbone flexibility and prediction of the mutant side chain without a common Cβ atom to model from, resulting in unfavorable conformers or clashes with nearby side chains in the folded state, which occupy space around the wild-type residue.

Histidine side chains exist in a dynamic equilibrium of two neutral tautomers and one fully protonated state with a +1 formal charge. Accurate prediction of the relative fold stability effects of mutating to histidine therefore strongly depends upon the correct modeling of these protonation states in the folded protein, which requires accounting for the entire ensemble of tautomeric and protonation states by running FEP simulation

in all three states (HIE, HID, and HIP). The $\Delta\Delta G$ for mutation involving just one of these states can be subsequently corrected to account for other states in the ensemble by adding a "pK_a correction" term (see eqs 4–6 in Methods). This allows us to account for contributions to $\Delta\Delta G$ resulting from shifts in the pK_a of the histidine side chain between the reference (unfolded) environment and the folded protein environment.[33–35] We find that ΔE_{Potts} is strongly correlated with (pK_a-corrected) $\Delta\Delta G$ calculated from FEP simulations with a Pearson r of 0.90 (Figure 3D), suggesting that the Potts coupling parameters (and fundamentally, the sequence statistics from the MSA it is trained on) have implicitly captured the stability/fitness contributions from the entire ensemble of protonation states that coexist in the folded protein population (see Figure S10 and Table S3 for more details).

Using Potts Saturation Mutagenesis to Probe Selection Pressure on Protein Mutations.

As described above, our Potts model calculations for 65 somatic mutations of the kinases Abl1, Wee1, and Cdc7 seen in the GDC database are consistent with FEP simulations and support the interpretation of ΔE_{Potts} (eqs 1 and 2) as a relative stability contribution to fitness due to mutation. In general, we observe these mutations to have neutral or destabilizing effects on fold stability, where 42 mutations result in $\Delta\Delta G > 0$ and $\Delta E_{\text{Potts}} > 0$ while only one mutation, V508L in Wee1, showed significant stabilizing effects from both methods (see Table S1). An enrichment of destabilizing mutations in tumor cell samples is consistent with a neutral theory of somatic evolution in cancer,[25,36,37] which predicts a relative abundance of functionally deleterious mutations compared with beneficial mutations due to neutral drift. While there are a small number of mutations that drive tumorigenesis and evolve under positive selection pressures,[25,36,38–40] a scenario where subsequent tumor evolution is dominated by neutral selection[25] appears consistent with the pattern of stability effects seen here.

Abl1 is a tyrosine kinase well known for its role in chronic myelogenous leukemia (CML), which is positively driven by aberrant fusion with the BCR gene (breakpoint cluster region) at the N-terminus of the Abl sequence, making the kinase constitutively active. Following a subsequent period of neutral evolution, upon exposure to chemotherapeutic agents such as imatinib (Gleevec), additional point mutant variants of the Abl catalytic domain may become positively selected,[41] resulting in drug resistance. Some of these drug resistance mutations alter drug binding directly,[42] while others resist the inhibitory effects of drugs by increasing the intrinsic fitness of the enzyme, or a combination of both.[2,43] As described below, using the Potts model to analyze a large number of mutations sourced from drug-resistant Abl strains[2] (see Methods), we observe a selective bias in the distribution of Potts ΔEs to preserve the intrinsic fitness of the Abl catalytic domain, whereas the 24 somatic mutations of Abl sampled from miscellaneous non-CML tumors (from the original GDC set) have fitness effects consistent with an outcome expected from neutral selection.

Even though our kinase Potts model is trained on natural, drug-naïve sequences across a variety of species, we can still detect how mutations at these positions affect the intrinsic fitness of human Abl by performing saturation mutagenesis at each position in the sequence where a drug resistance mutation is observed (4D). When analyzing the ΔEs of each drug resistance mutation relative to the average of the site-saturated distribution of ΔEs at those

same positions (Figure 4D), we observe a heavier left-hand tail in the Potts distribution of drug resistance mutations compared with the null model, signifying a selective bias to preserve the fitness and stability of the enzyme. The null model was constructed using the overall distribution of ΔEs acquired from Potts site-saturation mutagenesis at the same positions in the Abl sequence where mutations are observed in the clinical data. For resistance mutations, the null model was rejected using the Kolmogorov–Smirnov (K–S) test with $P \approx 10^{-4}$ (Figure 4A, left), meaning the weight of the left-hand tail for the distribution of (red) drug resistance mutations compared with the null model (black) is highly unlikely to be a random realization of the latter. Consistently, when performing this same analysis over the 24 somatic mutations as a control (Figure 4A, right), the null model is accepted with a P value of 0.47, suggesting that Abl mutations seen in miscellaneous non-CML tumors are likely a product of neutral selection.

CONCLUSIONS

This study uses a sequence-based kinase family Potts Hamiltonian model to predict the effects of single-point mutations in protein kinases. Its performance was compared with experimentally derived $\Delta\Delta G$s taken from the literature for a set of 14 single-point mutations, showing a good correlation (Pearson r of 0.68). Leveraging all-atom molecular dynamics FEP simulations as a computational benchmark, we calculated the relative folding free energies of point mutants, $\Delta\Delta Gs$, and compared these with fitness-based ΔEs calculated from the Potts model. The comparison of $\Delta\Delta G$ versus ΔE for 65 somatic mutations observed from cancer genomics data[1] of three kinases (Wee1, Cdc7, and Abl1) suggests that the sequence-based Potts model can be utilized as a prescreening tool for much more compute intensive FEP simulations and subsequent labor-intensive laboratory experiments. We note that the Pearson correlation observed when comparing Potts ΔEs with FEP-calculated $\Delta\Delta Gs$ ($r = 0.72$) is similar to that observed when comparing Potts ΔEs with experimental benchmark data ($r = 0.68$).

This study also begins to explore the application of the Potts model for probing the effects of cancer mutations on protein stability; our results from Figure 4 suggest that the analysis of Potts ΔEs, when performed over cancer genomics data, can discern whether kinase mutations observed in a particular disease environment have arisen under positive versus neutral selection. This analysis involved site-saturation mutagenesis in the Potts model, requiring nonconservative amino acid changes to be made to the protein sequence. To substantiate this approach, we first paid special attention to specific categories of amino acid substitutions among the 65 somatic mutations, which represent well-known “challenge cases” for computational modeling (Figure 3) and thoroughly evaluated the consistency between Potts ΔEs and FEP-derived $\Delta\Delta Gs$ for these nonconservative substitutions involving changes in charge, large side-chain vdW volume changes, and mutations involving proline, glycine, and histidine. Our results for these mutations are largely consistent (Figure 3), adding confidence to our broader observation of the stability trends for somatic mutations in Wee1, Cdc7, and Abl1; the results from Figure 2C show an enrichment of deleterious ($\Delta\Delta G > 0$) and neutral mutations ($\Delta\Delta G \approx 0$), consistent with the characteristics of passenger mutations expected under the theory of neutral drift, which is thought to explain much of the sequence variation observed in cancer data.[25,37] In contrast, for the case where

chemotherapeutic pressures have led to the evolution of drug resistance in Abl-targeted therapies, our analysis with the Potts model suggests that drug resistance mutations are biased to preserve the stability of the enzyme (Figure 4)

Overall, the results discussed in this study highlight the potential of the Potts Hamiltonian model to be used as a tool for predicting the effects of mutations on protein stability and function generally, even for nonconservative mutations; these results suggest the Potts model may also be useful as a sequence covariation-based tool to analyze some features of the dynamics involved in cancer development. The Potts model can rapidly assess the effects of single-point mutations across a broad spectrum of proteins within the same family without the necessity of constructing accurate structural models of the mutant proteins. In terms of net-computational cost, when used for pre-FEP screening or as a high-throughput alternative to FEP simulations, the initial Potts model inference of the Hamiltonian parameters is the most significant bottleneck.

METHODS

FEP Calculations.

FEP represents a physics-based molecular dynamics simulation approach applied to understand how changes in amino acid residues affect the stability of a protein. This involves transforming the wild-type residue into the mutated one using alchemical techniques, in both folded and unfolded protein states, using a thermodynamic cycle as described in Figure 1. The simulations yield estimates of the free energy difference ($\Delta\Delta G$) between folded and unfolded states in both the wild type and mutant as described in eq 1 and signify the mutation-induced shift in protein thermostability. In this study, FEP+ software from the Schrödinger Suite 2021−4 was employed to compute the latter described thermostability effects on three kinases (WEE1, CDC7, and ABL1). The initial wild-type complexes were prepared using the Protein Preparation Wizard tool.

Crystal structures of the catalytic domain for Weel, Cdc7, and Abl1 (PDB ID: 1X8B,[44] 6YA7,[45] and 2V7A,[46] respectively) provided the starting Cartesian coordinates for folded-state FEP simulations. The initial apo wild-type structure was generated by deleting any cocrystallized inhibitors followed by default protocol for protein preparation in maestro. However, the Abl1 crystal structure gatekeeper residue Thr315 was mutated to Ile, so it was mutated back to the wild type and the rotamer state of the Ile side chain was compared with the wild-type Abl1 crystal structure (PDB ID: 2HZ4[47]). To overcome the artifacts in the protein due to deleting the ligands and mutating back to the wild type, all three apo wild-type systems were subjected to equilibration over 50 ns of classical molecular dynamic simulations using the Desmond module. The capped tripeptide with native conformation and sequence (where the mutation site is in the center of the peptide) extracted from the crystal structure was used as a model for the unfolded state.

The topology files for all wild-type and mutant systems were generated using the OPLS4 force field.[48] Each system was solvated using a SPC cubic water box of dimensions 10 Å from the edge of protein in each direction. To simulate the system at physiological conditions, counterions were first added to neutralize the overall charge followed by

additional Na+ and Cl− ions incorporated randomly in the simulation box to achieve a concentration of 0.15 M NaCl. The mutated residue was included in the Hamiltonian replica exchange with solute tempering (REST) region.[49] Followed by the solvation steps, all systems were subjected to a series of short minimizations and equilibrations using the Desmond default protocol. The alchemical λ windows, which connect the wild-type and mutant states, were established at a default count of 16 for mutations that do not involve a charge change, whereas mutations that involved charge changes or had difficulty in convergence were extended up to 32 lambdas. Within each λ window, production molecular dynamics (MD) simulations were conducted for 10–20 ns under the NPT ensemble. To maintain overall neutrality for mutations that result in a net change in charge, the coal chemical water approach was employed.[50] For mutations involving prolines, a set of 16–24 core-hopping λ windows were utilized. In this approach, the CG-CD bond within the pyrrolidine ring was substituted with a softcore bond, enabling the bond to be broken to accommodate noncyclic side-chain mutations.[9] This same protocol and parameters were applied to both the wild-type and mutant unfolded tripeptide systems.

The protonation equilibrium of the residue fragment in bulk solution is formally described by the microscopic acid dissociation constant of each titratable site, which can, in principle, be determined through quantum calculations. Absolute determination of pK_a for protein residue side chains in the folded protein environment is significantly more challenging, both experimentally and computationally.[51] However, if a reference pK_a is known, e.g., solvated residue in the unfolded state, then shifts in the protonation equilibrium from the model reference state to the folded protein environment can be determined classically using MD and FEP, which reduces to a relatively straightforward calculation under the assumption of independent pK_as. By expressing the pK_a shift in terms of relative folding free energies via FEP where the different tautomeric and charged states are held fixed and simulated independently, assuming the pK_as of nearby titratable residues are uncoupled, the contribution of the entire ensemble of histidine protonation and tautomeric states can be determined using a "pK_a correction" formula that takes three different $\Delta\Delta G$ values calculated via FEP simulations, which alchemically protonate/deprotonate each of the three titratable sites on the histidine side chain in succession.[33–35]

The "true" value of $\Delta\Delta G_f^{w \to m}$ for mutating a nontitratable wild-type residue "w" to mutant histidine "m", which rigorously accounts for the relative populations of all protonation states, can be calculated by taking the estimated value $\Delta\Delta G_f^{w \to m_\epsilon}$, which comes from an FEP simulation of the mutation to a single fixed tautomer ϵ, and adding a "correction" term $\Delta\Delta G_f^{m_\epsilon \to m}$.

$$\Delta\Delta G_f^{w \to m} = \Delta\Delta G_f^{w \to m_\epsilon} + \Delta\Delta G_f^{m_\epsilon \to m} \tag{4}$$

The additional term is sometimes called a "pK_a correction" because it derives from the ability to calculate shifts in pK_a that occur between the reference environment and the folded protein environment. From this, we obtain information about the population shifts

between tautomeric and protonation states between the two protein environments, and thus the folding free-energy cost for restricting the ensemble to ϵ, $\Delta\Delta G_{\mathrm{f}}^{m_\epsilon \to m}$, can be determined with the following formula[51] and included as a correction to the calculated $\Delta\Delta G_{\mathrm{f}}^{w \to m_\epsilon}$.

$$\begin{aligned}\Delta\Delta G_{\mathrm{f}}^{m_\epsilon \to m} &\equiv -RT\ln\left(\frac{P(\epsilon \mid \text{folded})}{P(\epsilon \mid \text{unfolded})}\right)\\ &= -RT\ln\left(\frac{\left\{\exp\left[-\frac{1}{RT}\Delta\Delta G_{\mathrm{f}}^{m_\epsilon \to m_{\epsilon+}}\right]\left(10^{(\mathrm{p}K_{\mathrm{a}}^{\epsilon} - \mathrm{pH})}\right) + \exp\left[-\frac{1}{RT}\Delta\Delta G_{\mathrm{f}}^{m_\epsilon \to m_\delta}\right]\left(10^{(\mathrm{p}K_{\mathrm{a}}^{\epsilon} - \mathrm{p}K_{\mathrm{a}}^{\delta})}\right)\right\}\left(\exp\left[-\frac{1}{RT}\Delta\Delta G_{\mathrm{f}}^{m_\delta \to m_{\delta+}}\right]10^{(\mathrm{p}K_{\mathrm{a}}^{\delta} - \mathrm{pH})} + 1\right) + 1}{10^{(\mathrm{p}K_{\mathrm{a}}^{\epsilon} - \mathrm{pH})} + 10^{(\mathrm{p}K_{\mathrm{a}}^{\epsilon} - \mathrm{p}K_{\mathrm{a}}^{\delta})}\left(10^{(\mathrm{p}K_{\mathrm{a}}^{\delta} - \mathrm{pH})} + 1\right) + 1}\right)\end{aligned} \tag{5}$$

where the microscopic acid dissociation constants associated with the formation of ϵ and δ tautomers and resulting pK_{a}s in the fully solvated or state were adopted from ref 33 and assumed to correspond to the ionization equilibrium in the unfolded state:

$$\mathrm{His}_{\epsilon}^{+} \overset{K_{\mathrm{a}}^{\epsilon}}{\longleftrightarrow} \mathrm{His}_{\epsilon} + \mathrm{H}^{+}$$

$$\mathrm{His}_{\delta}^{+} \overset{\mathrm{K}_{\mathrm{a}}^{\delta}}{\leftrightarrow} \mathrm{His}_{\delta} + \mathrm{H}^{+}$$

For the case where the titratable residue is the wild type rather than the mutant, the correction term $\Delta\Delta G_{\mathrm{f}}^{w_\epsilon \to w}$ is calculated the same way but should instead be subtracted from the FEP-calculated $\Delta\Delta G_{\mathrm{f}}^{w_\epsilon \to m}$:

$$\Delta\Delta G_{\mathrm{f}}^{w \to m} = \Delta\Delta G_{\mathrm{f}}^{w_\epsilon \to m} - \Delta\Delta G_{\mathrm{f}}^{w_\epsilon \to w} \tag{6}$$

Potts Model Calculations.

The kinase family Potts Hamiltonian model was constructed from an MSA of Hanks-type “eukaryotic” protein kinase catalytic domains, as described previously,[12,17] which has a total depth of ~10^5 sequences and 259 columns (sequence “length” $L = 259$). The Hamiltonian inference was performed as described previously, using Mi3-GPU: a program that utilizes massively parallelized MCMC (Markov chain Monte Carlo) simulations on GPUs to generate ensembles of sequences from a given Potts Hamiltonian, starting from an initial “guess”, and iteratively perturbing the coupling parameters until subsequent MCMC runs can generate ensembles with residue–residue correlations that agree with the training MSA. For more details, we refer the reader to previous publications.[16,26]

When considering the fitness of a mutant sequence relative to its wild-type form, it is standard practice to calculate the quantity ΔE as a function of the wild-type and mutant sequence,

$$\Delta E(w,m) = \sum_{i<j}^{L} (J^{ij}_{m_i m_j} - J^{ij}_{w_i w_j}) + \sum_{i}^{L} (h^{i}_{m_i} - h^{i}_{w_i}) \tag{7}$$

where J and h represent the Hamiltonian residue-pair couplings and position-specific field terms, respectively. ΔE is the relative statistical energy of a wild-type sequence "w" with the mutant form "m", which relates to the relative probability of observing the wild type in the MSA compared with the mutant via the Boltzmann factor $\frac{P(m)}{P(w)} = e^{-\Delta E}$. The Boltzmann factor is a measure of the relative fitness of the mutant to the wild-type fold, of which the fold stability of the mutant sequence compared with the wild type is often the largest contribution,[27] which is the quantity of interest in this study.

Consistent with our previous work where we used the kinase Potts model to determine conformational propensities of folded kinases,[13,17] the stability of a single free-energy basin (e.g., the active conformation) can be probed relative to the unfolded state by threading the Potts couplings of a given sequence (e.g., the wild type) over pairs of residues observed in contact in structure n;

$$E(w,n) = \sum_{i<j}^{L} J^{ij}_{w_i w_j} \delta[d_{ij}(n) < 6\text{Å}] \tag{8}$$

where $\delta = 1$ if the distance between the closest approaching side-chain heavy atoms of residues i and j (including Cβ atoms) is within 6 Å and $\delta = 0$ otherwise. For glycine residues, the side chain is defined as the Cα atom. The mutant sequence m can be threaded over the same structure, and the difference in threaded energy with respect to the wild type is interpreted as a statistical energy analog of the relative folding free energy calculated from FEP simulations using structure n.

$$\begin{aligned}\Delta E(w,m,n) &= E(m,n) - E(w,n) \\ &= \sum_{i<j}^{L} (J^{ij}_{m_i m_j} - J^{ij}_{w_i w_j}) \delta[d_{ij}(n) \\ &< 6\text{Å}]\end{aligned} \tag{9}$$

We find that the Pearson correlation coefficient of the structure-independent calculation (eq 7) vs $\Delta\Delta G$ calculated from FEP is similar to the correlation observed when threading the zero-gauge couplings (eq 9) over the MD-refined starting structures used for FEP simulations (Supporting Information). Interestingly, for specific examples where histidine is involved, the threading calculation has better correspondence with FEP-calculated $\Delta\Delta G$s particularly when the effects of the protonation equilibrium are rigorously accounted for (eqs 4–6). This suggests that the Potts zero-gauge couplings capture the relevant interactions

formed between the mutant side chain and nearby residues when histidine is in its different protonation states.

MM/GBSA Calculation for Estimating the Protein Stability Effects.

This study employed a standard protocol to assess the impact of protein mutations on stability using the MM/GBSA method implemented in BioLuminate (Schrödinger Suite 2021–4).[11] The protocol systematically involves mutating each residue to a reference amino acid and analyzing the resultant change in protein stability, as illustrated in eq 3. This method utilizes the OPLS4 force field[48] in conjunction with the implicit solvent model (VSGB)[52] to assess the impact of mutations on protein stability. It solely focuses on side-chain rotamers by keeping the protein backbone and neighboring side chains fixed and eliminating the need for extensive MD simulations sampling, thereby significantly enhancing computational efficiency. Prior to residue scanning calculations, crystallographic water molecules were removed from the input structures to ensure compatibility with the VSGB solvation model.

Conversion of Experimental Melting Temperatures to $\Delta\Delta G$.

Protein melting temperatures (T_m) are a common experimental measurement of protein fold stability, and the effect of mutations on stability is reflected by T_m shifts between wild type and mutant (ΔT_m). However, ΔT_m is difficult to compare directly with $\Delta\Delta G$s calculated from FEP simulations (kcal/mol) without further treatment. As reviewed in a recent study,[53] the Gibbs–Helmholtz equation relates the free-energy of folding with melting temperature:

$$\Delta G = -\Delta H_m\left(1-\frac{T}{T_m}\right)+\Delta C_p\left(T_m - T + T\ln\left(\frac{T}{T_m}\right)\right) \tag{10}$$

In practice, ΔH_m and ΔC_p are unknown values; however, based on early protein folding experiments, Razban[53] suggests the following approximation:

$$\Delta G \approx -N[0.698 + 0.014(T_m - 333)]\frac{T_m - T}{T}\frac{\text{kcal}}{\text{mol}} \tag{11}$$

where N is the number of residues in the protein construct used for experimental determination of T_m and T is set to 300 K.

Data Set of Somatic Mutations in Cancer.

Somatically occurring mutations located on the Abl1, Wee1, and CDC7 kinase domains were identified from cancer genomic data using the National Cancer Institute GDC (Genomic Data Commons) database.[1] Only mutations satisfying the following criteria were retained for further analysis: (1) mutations must involve nonsynonymous protein coding changes, (2) mutations must be located on the catalytic domain sequence, and (3) mutations do not target conserved residues in the DFG (“Asp-Phe-Gly”) or catalytic loop motifs.

Overall, we collected 24 such mutations for Abl1, 23 for Wee1, and 16 for CDC7, which are plotted in Figure 2C (data available in Table S1).

Data Set of Drug Resistance Mutations in Abl1.

For our comparison in Figure 4 of somatic mutations in cancer (presumably under neutral selection) with cancer drug resistance mutations (presumably under positive and/or purifying selection), we used the set of 93 Abl1 drug resistance mutations studied by Lyczek et al.,[19] originally sourced from the COSMIC database.[22] The "somatic mutations in miscellaneous cancers" data set refers to the 24 somatic mutations of Abl described above, initially identified from GDC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grants, numbers R35 GM132090 and R01 AI177849, and by NIH Computer Equipment Grant (OD020095). We thank Dr. Allan Haldane at Temple University for many helpful discussions related to the kinase Potts model and for providing the Python code used to make the plots in Figure 4,,C,D.

REFERENCES

(1). Grossman RL; Heath AP; Ferretti V; Varmus HE; Lowy DR; Kibbe WA; Straudt LM Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Med. 2016, 375 (12), 1109–1112. [PubMed: 27653561]

(2). Lyczek A; Berger BT; Rangwala AM; Paung YT; Tom J; Philipose H; Guo J; Albanese SK; Robers MB; Knapp S; et al. Mutation in Abl kinase with altered drug-binding kinetics indicates a novel mechanism of imatinib resistance. Proc. Natl. Acad. Sci. U.S.A. 2021, 118 (46), No. e2111451118. [PubMed: 34750265]

(3). Hopf TA; Ingraham JB; Poelwijk FJ; Schärfe CP; Springer M; Sander C; Marks DS Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 2017, 35 (2), 128–135. [PubMed: 28092658]

(4). Høie MH; Cagiada M; Frederiksen AHB; Stein A; Lindorff-Larsen K Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. Cell Rep. 2022, 38 (2), No. 110207. [PubMed: 35021073]

(5). Deng N; Flynn WF; Xia J; Vijayan RSK; Zhang B; He P; Mentes A; Gallicchio E; Levy RM Large scale free energy calculations for blind predictions of protein–ligand binding: the D3R Grand Challenge 2015. Journal of Computer-Aided Molecular Design 2016, 30 (9), 743–751. [PubMed: 27562018]

(6). Xia J; Flynn W; Gallicchio E; Uplinger K; Armstrong JD; Forli S; Olson AJ; Levy RM Massive-Scale Binding Free Energy Simulations of HIV Integrase Complexes Using Asynchronous Replica Exchange Framework Implemented on the IBM WCG Distributed Network. J. Chem. Inf. Model. 2019, 59 (4), 1382–1397. [PubMed: 30758197]

(7). Steinbrecher T; Abel R; Clark A; Friesner R Free Energy Perturbation Calculations of the Thermodynamics of Protein Side-Chain Mutations. J. Mol. Biol. 2017, 429 (7), 923–929. [PubMed: 28279701]

(8). Pucci F; Bernaerts KV; Kwasigroch JM; Rooman M Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics 2018, 34 (21), 3659–3665. [PubMed: 29718106]

(9). Duan J; Lupyan D; Wang L Improving the Accuracy of Protein Thermostability Predictions for Single Point Mutations. Biophys. J. 2020, 119 (1), 115–127. [PubMed: 32533939]

(10). Sergeeva AP; Katsamba PS; Liao J; Sampson JM; Bahna F; Mannepalli S; Morano NC; Shapiro L; Friesner RA; Honig B Free Energy Perturbation Calculations of Mutation Effects on SARS-CoV-2 RBD::ACE2 Binding Affinity. J. Mol. Biol. 2023, 435 (15), No. 168187. [PubMed: 37355034]

(11). Beard H; Cholleti A; Pearlman D; Sherman W; Loving KA Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. PLoS One 2013, 8 (12), No. e82849. [PubMed: 24340062]

(12). McGee F; Hauri S; Novinger Q; Vucetic S; Levy RM; Carnevale V; Haldane A The generative capacity of probabilistic protein sequence models. Nat. Commun. 2021, 12 (1), 6302. [PubMed: 34728624]

(13). Haldane A; Flynn WF; He P; Vijayan RSK; Levy RM Structural propensities of kinase family proteins from a Potts model of residue co-variation. Protein Sci. 2016, 25 (8), 1378–1384. [PubMed: 27241634]

(14). Hopf TA; Ingraham JB; Poelwijk FJ; Springer M; Sander C; Marks DS Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv preprint arXiv:1510.04612 2015.

(15). Lapedes A; Giraud B; Jarzynski C Using sequence alignments to predict protein structure and stability with high accuracy. arXiv preprint arXiv:1207.2484 2012.

(16). Haldane A; Levy RM Mi3-GPU: MCMC-based inverse Ising inference on GPUs for protein covariation analysis. Comput. Phys. Commun. 2021, 260, No. 107312. [PubMed: 33716309]

(17). Gizzio J; Thakur A; Haldane A; Levy RM Evolutionary divergence in the conformational landscapes of tyrosine vs serine/threonine kinases. eLife 2022, 11, No. e83368. [PubMed: 36562610]

(18). Haldane A; Flynn WF; He P; Levy RM Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs. Biophys. J. 2018, 114 (1), 21–31. [PubMed: 29320688]

(19). Ayaz P; Lyczek A; Paung Y; Mingione VR; Iacob RE; de Waal PW; Engen JR; Seeliger MA; Shan Y; Shaw DE Structural mechanism of a drug-binding process involving a large conformational change of the protein target. Nat. Commun. 2023, 14 (1), 1885. [PubMed: 37019905]

(20). Lori C; Lantella A; Pasquo A; Alexander LT; Knapp S; Chiaraluce R; Consalvi V Effect of single amino acid substitution observed in cancer on Pim-1 kinase thermodynamic stability and structure. PloS one 2013, 8 (6), No. e64824. [PubMed: 23755147]

(21). Mohanty S; Oruganty K; Kwon A; Byrne DP; Ferries S; Ruan Z; Hanold LE; Katiyar S; Kennedy EJ; Eyers PA; et al. Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. PLOS Genetics 2016, 12 (2), No. e1005885. [PubMed: 26925779]

(22). Forbes SA; Bindal N; Bamford S; Cole C; Kok CY; Beare D; Jia M; Shepherd R; Leung K; Menzies A; et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011, 39, D945–D950. [PubMed: 20952405]

(23). Wang L; Friesner RA; Berne BJ Replica exchange with solute scaling: A more efficient version of replica exchange with solute tempering (REST2). J. Phys. Chem. B 2011, 115 (30), 9431–9438. [PubMed: 21714551]

(24). Wang L; Berne B; Friesner R On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. Proc. Natl. Acad. Sci. U. S. A. 2012, 109 (6), 1937. [PubMed: 22308365]

(25). Cannataro VL; Townsend JP Neutral Theory and the Somatic Evolution of Cancer. Mol. Biol. Evol. 2018, 35 (6), 1308–1315. [PubMed: 29684198]

(26). Levy RM; Haldane A; Flynn WF Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. Curr. Opin. Struct. Biol. 2017, 43, 55–62. [PubMed: 27870991]

(27). Morcos F; Schafer NP; Cheng RR; Onuchic JN; Wolynes PG Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc. Natl. Acad. Sci. U. S. A. 2014, 111 (34), 12408–12413. [PubMed: 25114242]

(28). Steinbrecher T; Zhu C; Wang L; Abel R; Negron C; Pearlman D; Feyfant E; Duan J; Sherman W 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. 2017, 429 (7), 948–963. [PubMed: 27964946]

(29). Negron C; Pearlman DA; del Angel G Predicting mutations deleterious to function in beta-lactamase TEM1 using MM-GBSA. PloS one 2019, 14 (3), No. e0214015. [PubMed: 30889230]

(30). Hayes RL; Brooks CL 3rd A strategy for proline and glycine mutations to proteins with alchemical free energy calculations. Journal of computational chemistry 2021, 42 (15), 1088–1094. [PubMed: 33844328]

(31). Clark AJ; Negron C; Hauser K; Sun M; Wang L; Abel R; Friesner RA Relative Binding Affinity Prediction of Charge-Changing Sequence Mutations with FEP in Protein–Protein Interfaces. J. Mol. Biol. 2019, 431 (7), 1481–1493. [PubMed: 30776430]

(32). Miller S; Janin J; Lesk AM; Chothia C Interior and surface of monomeric proteins. J. Mol. Biol. 1987, 196 (3), 641–656. [PubMed: 3681970]

(33). Coskun D; Chen W; Clark AJ; Lu C; Harder ED; Wang L; Friesner RA; Miller EB Reliable and Accurate Prediction of Single-Residue pKa Values through Free Energy Perturbation Calculations. J. Chem. Theory Comput. 2022, 18 (12), 7193–7204. [PubMed: 36384001]

(34). Uranga J; Mikulskis P; Genheden S; Ryde U Can the protonation state of histidine residues be determined from molecular dynamics simulations? Computational and Theoretical Chemistry 2012, 1000, 75–84.

(35). Wilson CJ; Karttunen M; de Groot BL; Gapsys V Accurately predicting protein pKa values using non-equilibrium alchemy. J. Chem. Theory Comput 2023, 19 (21), 7833–7845. [PubMed: 37820376]

(36). Bányai L; Trexler M; Kerekes K; Csuka O; Patthy L Use of signals of positive and negative selection to distinguish cancer genes and passenger genes. eLife 2021, 10, No. e59629. [PubMed: 33427197]

(37). Martincorena I; Raine KM; Gerstung M; Dawson KJ; Haase K; Van Loo P; Davies H; Stratton MR; Campbell PJ Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 2017, 171 (5), 1029–1041.e21. [PubMed: 29056346]

(38). Vogelstein B; Papadopoulos N; Velculescu VE; Zhou S; Diaz LA; Kinzler KW Cancer Genome Landscapes. Science 2013, 339 (6127), 1546–1558. [PubMed: 23539594]

(39). Saito Y; Koya J; Araki M; Kogure Y; Shingaki S; Tabata M; McClure MB; Yoshifuji K; Matsumoto S; Isaka Y; et al. Landscape and function of multiple mutations within individual oncogenes. Nature 2020, 582 (7810), 95–99. [PubMed: 32494066]

(40). Zehir A; Benayed R; Shah RH; Syed A; Middha S; Kim HR; Srinivasan P; Gao J; Chakravarty D; Devlin SM; et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature Medicine 2017, 23 (6), 703–713.

(41). Friedman R Drug resistance in cancer: molecular evolution and compensatory proliferation. Oncotarget 2016, 7 (11), 11746. [PubMed: 26909596]

(42). Zuccotto F; Ardini E; Casale E; Angiolini M Through the “gatekeeper door”: Exploiting the active kinase conformation. J. Med. Chem. 2010, 53 (7), 2681–2694. [PubMed: 20000735]

(43). Hoemberger M; Pitsawong W; Kern D Cumulative mechanism of several major imatinib-resistant mutations in Abl kinase. Proc. Natl. Acad. Sci. U.S.A. 2020, 117 (32), 19221–19227. [PubMed: 32719139]

(44). Squire CJ; Dickson JM; Ivanovic I; Baker EN Structure and Inhibition of the Human Cell Cycle Checkpoint Kinase, Wee1A Kinase: An AtypicalTyrosine Kinase with a Key Role in CDK1 Regulation. Structure 2005, 13 (4), 541–550. [PubMed: 15837193]

(45). Dick SD; Federico S; Hughes SM; Pye VE; O’Reilly N; Cherepanov P Structural Basis for the Activation and Target Site Specificity of CDC7 Kinase. Structure 2020, 28 (8), 954–962.e4. [PubMed: 32521228]

(46). Modugno M; Casale E; Soncini C; Rosettani P; Colombo R; Lupi R; Rusconi L; Fancelli D; Carpinelli P; Cameron AD; et al. Crystal Structure of the T315I Abl Mutant in Complex with the Aurora Kinases Inhibitor PHA-739358. Cancer Res. 2007, 67 (17), 7987–7990. [PubMed: 17804707]

(47). Cowan-Jacob SW; Fendrich G; Floersheimer A; Furet P; Liebetanz J; Rummel G; Rheinberger P; Centeleghe M; Fabbro D; Manley PW Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. Acta Crystallographica Section D 2007, 63 (1), 80–93.

(48). Lu C; Wu C; Ghoreishi D; Chen W; Wang L; Damm W; Ross GA; Dahlgren MK; Russell E; Von Bargen CD; et al. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. J. Chem. Theory Comput. 2021, 17 (7), 4291–4300. [PubMed: 34096718]

(49). Wang L; Deng Y; Knight JL; Wu Y; Kim B; Sherman W; Shelley JC; Lin T; Abel R Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. J. Chem. Theory Comput. 2013, 9 (2), 1282–1293. [PubMed: 26588769]

(50). Chen W; Deng Y; Russell E; Wu Y; Abel R; Wang L Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges. J. Chem. Theory Comput. 2018, 14 (12), 6346–6358. [PubMed: 30375870]

(51). de Oliveira C; Yu HS; Chen W; Abel R; Wang L Rigorous Free Energy Perturbation Approach to Estimating Relative Binding Affinities between Ligands with Multiple Protonation and tautomeric States. J. Chem. Theory Comput. 2019, 15 (1), 424–435. [PubMed: 30537823]

(52). Li J; Abel R; Zhu K; Cao Y; Zhao S; Friesner RA The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. Proteins 2011, 79 (10), 2794–2812. [PubMed: 21905107]

(53). Razban RM Protein Melting Temperature Cannot Fully Assess Whether Protein Folding Free Energy Underlies the Universal Abundance-Evolutionary Rate Correlation Seen in Proteins. Mol. Biol. Evol. 2019, 36 (9), 1955–1963. [PubMed: 31093676]

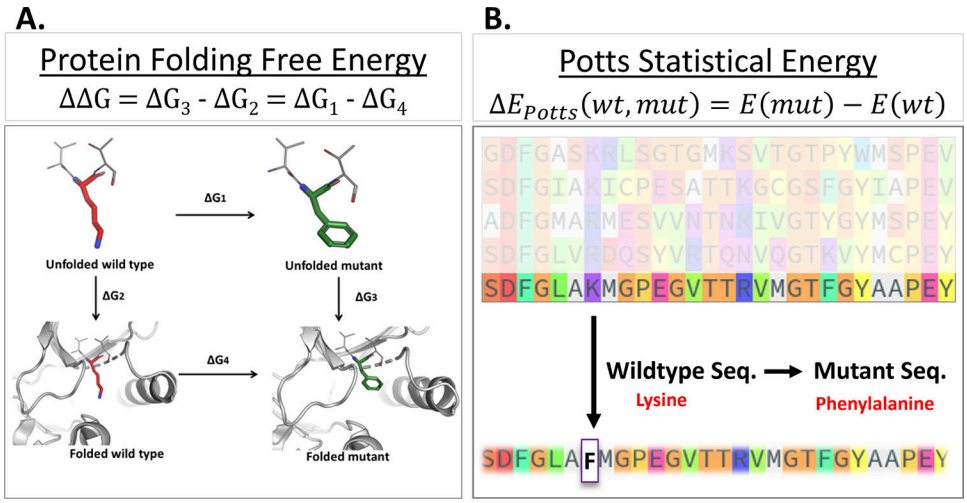


Figure 1.

Two different methods that have been applied to understand the effect of a mutation on protein stability, where (A) represents the thermodynamic cycle to estimate the change in free energy between wild type and mutant associated with protein folding, where vertical and horizontal paths represent the physical and alchemical paths, respectively, and (B) illustrates the sequence-based Potts model derived solely from sequence covariation observed in multiple-sequence alignments.

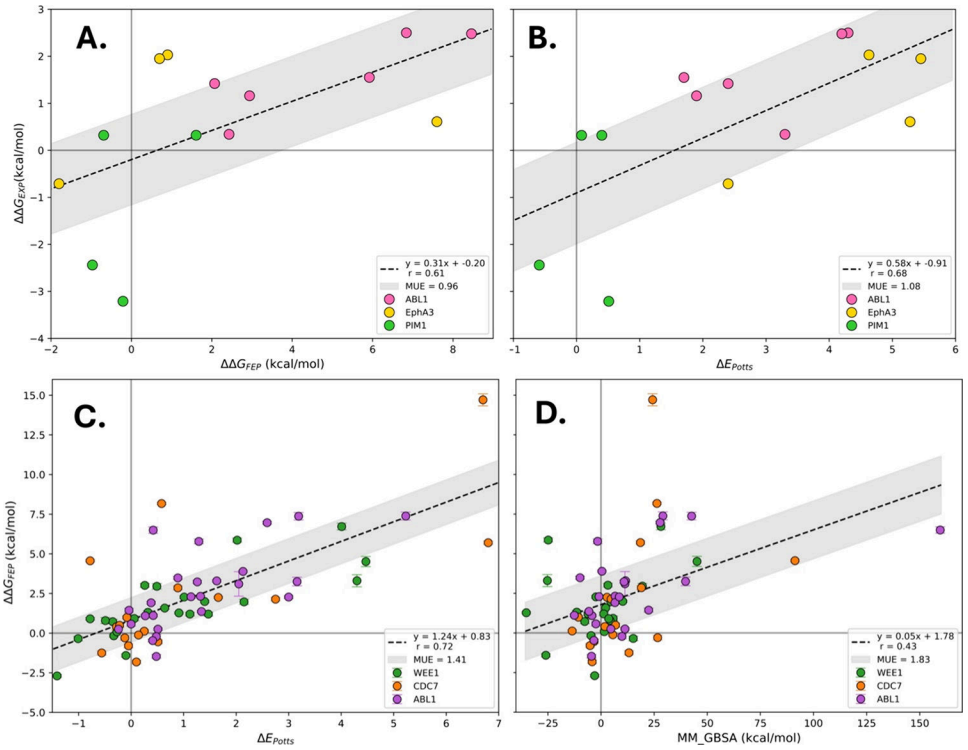


Figure 2.

Analyzing the consistency between different computational methods for predicting the relative stability effects of mutations in kinases: (A) $\Delta\Delta G_{\text{FEP}}$ calculated from FEP simulations compared with a benchmark, $\Delta\Delta G_{\text{exp}}$ (kcal/mol) derived from experimental thermostability data. (B) Potts statistical energy penalty of mutations (ΔE_{Potts}) compared with $\Delta\Delta G_{\text{exp}}$. (C) Comparing the two computational predictors, $\Delta\Delta G_{\text{FEP}}$ and ΔE_{Potts}, over a larger set of kinase mutations from the GDC database, where the slope of 1.24 indicates that 1 Potts ΔE corresponds to approximately 1.24 kcal/mol of $\Delta\Delta G$. (D) $\Delta\Delta G_{\text{FEP}}$ compared instead with stability predictions from an implicit solvent end point method, MM/GBSA (kcal/mol). The MUE between $\Delta\Delta G$ values plotted on the vertical axis and values from the horizontal axis projected onto the regression line is represented by the gray region.

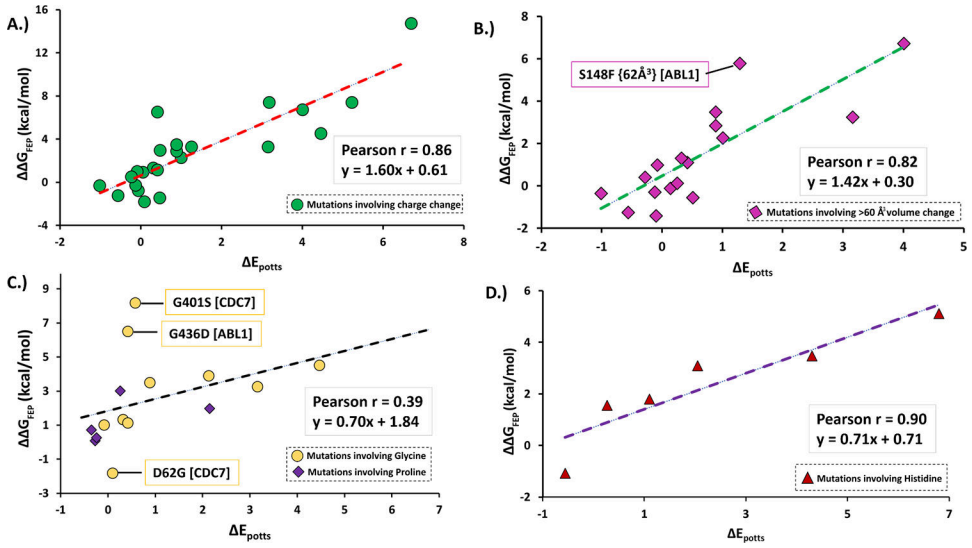
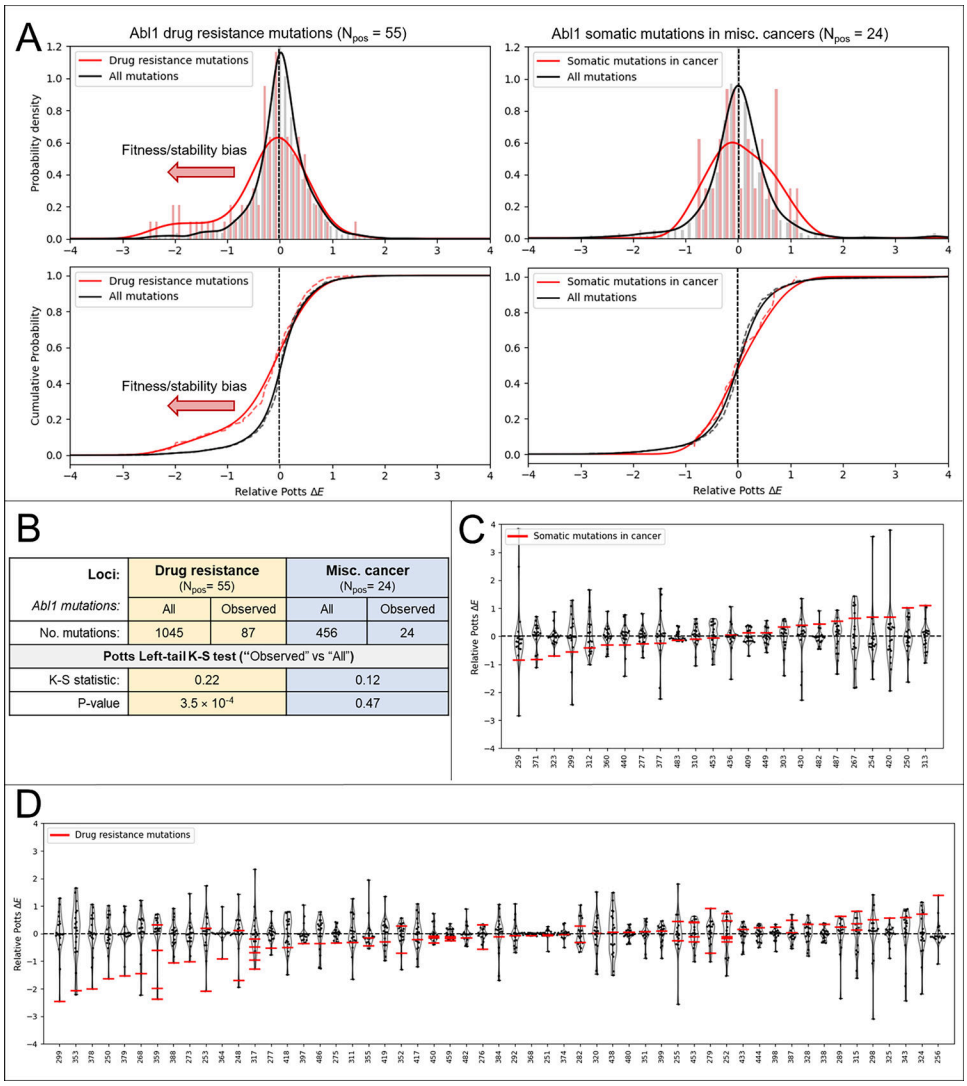


Figure 3.

Plots showing the correlation between protein stability predicted from FEP and the Potts model for the mutations from Figure 2C, separated into different "challenge cases" commonly recognized in computational modeling. The data set was categorized into four groups: (A) 23 mutations involving charge changes, (B) 15 mutations that involve large changes in side-chain vdW volume between wild type and mutant (>60 $Å^3$), (C) 15 mutations involving proline and glycine, and (D) 6 mutations involving histidine, for which we apply a "pK_a correction" to account for changes in the physicochemical environment of titratable sites between the folded and unfolded states (see Methods for details).



Loci:	Drug resistance (N_{pos}= 55)		Misc. cancer (N_{pos}= 24)	
Abl1 mutations:	All	Observed	All	Observed
No. mutations:	1045	87	456	24
Potts Left-tail K-S test ("Observed" vs "All")				
K-S statistic:	0.22		0.12	
P-value	3.5×10^{-4}		0.47	

Figure 4.

Distribution of relative Potts Δ*E*s for Abl1 per position (loci) where a mutation is observed in (A) imatinib-resistant CML (left) or miscellaneous non-CML tumors (right). Relative Potts Δ*E* for each mutation was calculated by subtracting the average Δ*E* from the site-saturated distribution at that position. Normalized histograms and kernel density estimates (KDE) of the probability density are shown in the top plots, and the corresponding cumulative probability distributions are plotted below as dotted and solid curves, respectively. (B) Summary statistics and result of the K–S (Kolmogorov– Smirnov) test. (C, D) Violin plots of the Δ*E* distributions at each position in the Abl sequence where mutations are observed, where the observed mutations are plotted as horizontal red bars, and the overall distribution/"violin" was plotted using KDE with a bandwidth of 0.3.