Three Simple Properties Explain Protein Stability Change upon Mutation

Caldararu, Octav; Blundell, Tom L.; Kepp, Kasper Planeta

Published in:
Journal of Chemical Information and Modeling

Link to article, DOI:
10.1021/acs.jcim.1c00201

Publication date:
2021

Document Version
Peer reviewed version

Citation (APA):
Three simple properties explain protein stability change upon mutation

Octav Caldararu¹, Tom L. Blundell², and Kasper P. Kepp¹*

¹ DTU Chemistry, Technical University of Denmark, Building 206, 2800 Kgs. Lyngby, Denmark.
² Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, United Kingdom.

Correspondence: E-mail: kpj@kemi.dtu.dk. Tel: +45 45252409

Abstract: Accurate prediction of protein stability upon mutation enables rational engineering of new proteins and insights into protein evolution and monogenetic diseases caused by single-point amino acid substitutions. Many tools have been developed to this aim, ranging from energy-based models to machine-learning methods that use large amounts of experimental data. However, as the methods become more complex, the interpretation of the chemistry underlying the protein stability effects becomes obscure. It is thus of interest to identify the simplest prediction model that retains complete amino-acid specific interpretation; for a given number of input descriptors, we expect such a model to be almost universal. In this study, we identify such a limiting model, SimBa, a simple multilinear regression model trained on a substitution-type-balanced experimental data set. The model accounts only for the solvent accessibility of the site, volume difference, and polarity difference caused by mutation. Our results show that this very simple and directly applicable model performs comparably to other much more complex, widely used protein stability prediction methods. This suggests that a hard limit of ~1 kcal/mol numerical accuracy and R ~ 0.5 trend accuracy exists and that new features, such as account of unfolded states, water co-localization and amino-acid correlations, are required to improve accuracy to, e.g., ½ kcal/mol.
Introduction

Protein function is, except for intrinsically disordered proteins, inherent to a distinct folded state,\textsuperscript{1–3} which can however unfold upon perturbation into an unstructured state. The ability of a protein to stay folded – its thermodynamic stability – and the forces that govern this stability have been extensively studied since the beginning of protein science.\textsuperscript{4–11} We are especially interested in how protein stability is affected by amino-acid substitution, which occurs repeatedly during natural evolution in a likely trade-off between stability and function\textsuperscript{12–18} and in genetic diseases driven by point mutations.\textsuperscript{19–21} Accurate prediction of protein stability would also aid protein engineering for e.g. industrial purposes, where stability is a common desired trait.\textsuperscript{22–26}

Practically, the difference in the free energy of unfolding between the wild-type protein and the mutant ($\Delta\Delta G$) is measured via thermal or chemical (pH, denaturants) denaturation, producing a mutant unfolding free energy relative to the measured wild type.\textsuperscript{5,27} The stability depends on intermolecular interactions between amino acids, such as hydrogen bonds or hydrophobic effects of the folded state, but also the conformation of the unfolded state, e.g. the extent to which hydrophobic groups are exposed to solvent.\textsuperscript{5,28,29} Although the problem thus requires knowledge of four structures – wild type and mutant folded and unfolded – computation of protein stability can only be done on experimental wild type and sometimes, when available, folded mutant structures. This major limitation may produce a hard wall on accuracy.\textsuperscript{30} In addition, the energy that needs to be calculated is a free energy that includes entropy, which is problematic even for methods that compute the full interaction energies of the protein structure.\textsuperscript{31,32}

Almost all methods used today for fast prediction of $\Delta\Delta G$ utilize either the sequence alone and the nature of the amino acid change, or alternatively also the structure of the wild-type protein, with a few methods also estimating the mutant folded structure. The methods can be split into those using parametrized molecular mechanics force-fields (Eris,\textsuperscript{33} CC/PBSA,\textsuperscript{34} FoldX,\textsuperscript{35} Rosetta\textsuperscript{36}), statistical potentials derived from structural databases, with varying degrees of parametrization (CUPSAT,\textsuperscript{37} SDM,\textsuperscript{38} PoPMuSiC\textsuperscript{39}) and, machine-learning tools that utilize the large amount of experimental data to create predictive models (I-Mutant,\textsuperscript{40} Maestro,\textsuperscript{41} mCSM,\textsuperscript{42} Automute\textsuperscript{25}). Naturally, the machine-learning and knowledge-based methods are faster than energy-based methods and several studies also suggest that they are more accurate.\textsuperscript{43,44}

However, this accuracy relates only to protein types and mutations already in the training sets.\textsuperscript{45} We recently found\textsuperscript{30} that all parametrized methods are biased towards mutations in their training sets,
and it is thus necessary to train with balanced data sets, not just in terms of stabilization and destabilization\textsuperscript{46} but in particular in terms of mutation type (possibly with a reduced amino acid alphabet) which otherwise creates systematic biases, a typical example being excessive reliance on alanine substitutions. In addition, experimental conditions should be selected to be so similar as possible to reduce also the random errors.\textsuperscript{30}

Despite the rapid growth of experimental protein stability data\textsuperscript{47–49} and the development of machine-learning methods, accuracy still struggles with mean absolute errors of 1 kcal/mol and correlation coefficients of 0.5.\textsuperscript{50}\textsuperscript{43,44,50} Indeed, a natural bound to the prediction accuracy of 1 kcal/mol and to the Pearson correlation coefficient of 0.7-0.8 may exist,\textsuperscript{50} depending on noise in experimental data and confounders such as pH, temperature and ion strength, and assumptions in the models, e.g. noise in the crystal structures. The typical destabilization effect of a random mutation is also 1 kcal/mol, making this unsatisfactory.\textsuperscript{51} An accuracy of ½ kcal/mol would thus substantially reduce sign errors (destabilization/stabilization) and improve the rational design of new stable protein mutants or interpretation of natural variants with modest but important stability effects related to disease.\textsuperscript{13,52–54} Considering that current models lack amino acid correlations, water colocalization effects, dynamics/entropy, and structural information of the unfolded states, we consider progress in this direction realistic. Also, more complex models make interpretation of results more cumbersome. For example, some programs use tessellation\textsuperscript{25} or a graph-based description of the protein\textsuperscript{42}, which makes it hard to pinpoint the amino-acid properties that contribute to the observed protein stability.

For these various reasons we want to know how far one can go with simple parametrized models that retain all chemical insight on a residue-specific basis, and how well such a model performs vs. state-of-the-art methods. In this study, we identify such as simple model and show that relative solvent accessibility (RSA), volume difference (\(V_{\text{diff}}\)) and hydrophobicity difference (\(H_{\text{diff}}\)) suffice to describe protein stability upon mutation, trained on a balanced data set, with similar accuracy as much more complex prediction methods. This indicates that all methods are limited by the same hard accuracy wall. We suggest that the main limits defining this wall, beyond the experimental uncertainty in data points itself, are lack of good structural information for the mutant folded and the unfolded states, account of amino-acid correlations, protein dynamics, and water co-localization.
Methods

Data sets
The balanced data set was built by filtering the O2567 data set\textsuperscript{30} to achieve mutation type balance which was identified as the most important source of bias. It contains 1131 mutations (B1131), with each of the 380 mutation types (20 x 19 possible amino acids substitutions) being represented by at most 5 data points. The full O2567 set was also used as a training set for comparison.

Two test sets were used to evaluate the robustness of the model. The first test set (B663) was also balanced by mutation type, starting from the I-Mutant 3.0 training set\textsuperscript{55} after deleting any mutations in common with the O2567 training data set. This balanced test set contained 663 mutations, with each of the 380 mutation types being represented by at most five data points, and it is fully disjoint from both the B1131 and the O2567 training set. The second test set was the well-studied S350 data set\textsuperscript{56}, which was specifically constructed to be disjoint from any training set, and several prediction methods have been tested on it. Both test data sets contain both monomeric and multimeric proteins, with \(~23\%\) data points in either of the datasets belonging to multimeric proteins. We kept both monomeric and multimeric proteins in the test sets, unlike other prediction methods, such as SDM, since all structural information in the models (RSA) is calculated taking into account all chains of a protein, and thus the model should not have any additional difficulties in calculating mutations at interfaces. In order to confirm this, we studied also the monomeric proteins of B663 separately below. The training and test data sets are listed in the Supporting Information excel file (datasets.xlsx).

The sign convention was kept as in the ProTherm database,\textsuperscript{57} i.e. negative $\Delta\Delta G$ values indicate destabilizing mutations, whereas positive $\Delta\Delta G$ values indicate stabilizing mutations. Structural sensitivity of the model was tested on the structural sensitivity dataset described previously\textsuperscript{58}, which includes saturated mutagenesis of 25 distinct proteins with at least 3 structures in the PDB.

Calculation of amino-acid properties
Amino-acid properties were calculated using the deposited PDBs after checking for common errors in the files, e.g. non-matching amino-acid numbering in the training/test data sets and in the PDB files, in which case the PDB was renumbered, or missing residues for the sites of interest, in which case the data point was discarded. No such missing residues were found in either of the data sets used for training and testing. The PDBs were previously curated as described in our earlier work.\textsuperscript{30}
Relative solvent accessibility (RSA) was calculated with Naccess\textsuperscript{59,60}, using default van der Waals atomic radii (provided at the end of the Supporting Information file) and divided by 100 to produce a scale from 0 to 1. RSA is defined as the ratio of the absolute solvent accessibility of the residue divided by that in a Ala-X-Ala tripeptide, where X is the residue in question. RSA calculations were performed on the full PDB structure (including all chains), but without adding hydrogen atoms. A default probe radius of 1.40 Å was used for all calculations.

The volume difference of the mutation was calculated as the difference between the molecular volumes (measured in Å\textsuperscript{3}) of the mutated and wild-type amino acids: \( V_{\text{diff}} = V_{\text{mut}} - V_{\text{wild}} \), i.e. a negative \( V_{\text{diff}} \) corresponds to a large to small mutation, whereas a positive \( V_{\text{diff}} \) corresponds to larger mutant site volume. The final \( V_{\text{diff}} \) used was divided by 100 for rescaling with RSA and \( \Delta H_{\text{diff}} \). Standard volumes of the amino acids were taken from ProtParam.\textsuperscript{61} Similarly, the hydrophobicity change was calculated as the difference between the hydrophobicity of the mutated and wild-type residue: \( \Delta H_{\text{diff}} = H_{\text{mut}} - H_{\text{wild}} \). Multiple hydrophobicity scales from ProtParam were used; the full list can be found in the Supporting Information (Table S1). All results herein are reported only for the Grantham polarity scale.\textsuperscript{62} This is technically not a hydrophobicity scale, but it provides similar chemical information to hydrophobicity. With this scale, a higher H value represents a more polar residue, thus a negative \( \Delta H_{\text{diff}} \) corresponds to a change towards more hydrophobic, whereas a positive \( \Delta H_{\text{diff}} \) corresponds to a decrease in hydrophobicity upon mutation. Both \( V_{\text{diff}} \) and \( \Delta H_{\text{diff}} \) were calculated only from the wild type and mutant residues of the mutated sites, not taking into account other residues in the protein.

In a pilot phase of the project, we studied other descriptors without fitting full models since many descriptors, such as side chain entropy, hydrogen bond tendency, weight, size, number of heteroatoms, hydrophobicity etc. tend to correlate and we settled for the simplest possible model with two descriptors with a very direct interpretation in protein structure context (small/large, polar/hydrophobic). We expect minor improvements would be feasible with additional descriptors but at the expense of added complexity.

**Regression model**

The model was built in Python using the Scikit-learn Linear Regression module.\textsuperscript{63} Two different prediction models are reported, SimBa-NI and SimBa-I. In SimBa-NI, the \( \Delta \Delta G \) is calculated from a least-squares multilinear regression model using as independent variables RSA, \( V_{\text{diff}} \) and \( \Delta H_{\text{diff}} \):

\[
\Delta \Delta G = a + b \cdot \text{RSA} + c \cdot V_{\text{diff}} + d \cdot \Delta H_{\text{diff}}
\]  

(1)
The second model, SimBa-I, considers the second-order interaction between RSA and $V_{\text{diff}}$ and $H_{\text{diff}}$, which is important to protein stability as both hydrophobicity-changing mutations and volume-changing mutations have a different effect when the mutated residue is buried or accessible. Thus, the least-squares multilinear regression becomes:

$$
\Delta \Delta G = a + b \cdot \text{RSA} + c \cdot V_{\text{diff}} + d \cdot H_{\text{diff}} + e \cdot \text{RSA} \cdot V_{\text{diff}} + f \cdot \text{RSA} \cdot H_{\text{diff}}
$$

(2)

Comparison to other prediction programs

We compared the simple model parameterized above to several state-of-the-art models that 1) have the ability to model any mutation, 2) give a quantitative $\Delta \Delta G$ prediction (rather than just qualitatively; destabilizing or stabilizing) and 3) work at high computational speed so as to be competitive with the simple model in this regard. Eight publicly available, widely used predictors were used: FoldX\textsuperscript{35}, I-Mutant 3.0\textsuperscript{40}, PoPMuSiC 2.1\textsuperscript{39}, Maestro\textsuperscript{41}, mCSM\textsuperscript{42}, SDM\textsuperscript{38}, CUPSAT\textsuperscript{37} and Automute 2.0\textsuperscript{64}. CUPSAT, PoPMuSiC 2.1 and mCSM were used as servers, whereas local programs were used for the other methods. All programs were run with default parameters unless specified otherwise. The FoldX prediction was performed after minimizing the input PDB structure with the RepairPDB module. The Automute 2.0 prediction was obtained with REPtree regression.

Statistical measures

The predictors’ performance was evaluated based on three metrics: The Pearson correlation coefficient (R) describes the ability of a method to provide the correct trend in a data set. The mean absolute error (MAE) describes the overall numerical accuracy of the method compared to the experimental data. The mean signed error (MSE) shows the systematic error of a method towards stabilization or destabilization.

Results and Discussion

Balanced training set results in good accuracy for all mutation types

SimBa-NI and SimBa-I were validated by predicting the mutations in the two test sets. For the first mutation-type balanced test set B663, both models presented reasonably high trend accuracy (R ~0.5) and a MAE of 1.19 and 1.11 kcal/mol, respectively (Table 1 and Figure 1A-1B). Moreover, the bias towards destabilizing mutations that many predictors suffer from was low for both models, with
SimBa-I having a mean signed error of only 0.02 kcal/mol. For the S350 test set, both models showed a similar trend accuracy, R=0.46 for SimBa-NI and R=0.5 for SimBa-I (Figure 1C-1D). However, the MAE was lower, 1.07 kcal/mol for both models, whereas the MSE was slightly higher, 0.31 and 0.13 kcal/mol respectively.

In all quality measures reported here and for the two test sets, SimBa-I behaved better than SimBa-NI, confirming what is well-appreciated, that protein stability is affected differently by the volume and hydrophobicity of the wild and mutant amino-acid types depending on whether the residue is buried or not. In our simple regression methods, we can directly evaluate the quantitative significance of this driver of fold stability. We also note that including the second-order interactions does not risk overfitting the predictive model, as all validation was done on independent test sets.

Using the O2567 data set as training set instead of a balanced data set affected little the trend accuracy on B663, as shown in Table S2, with R=0.46 and 0.5. However, the MAE and MSE of the models trained on the full data set were worse, MAE=1.17 and 1.24 kcal/mol and a MSE of 0.31 and 0.25 kcal/mol. These results showcase the strength of using a mutation-type balanced data set for training, as it avoids some bias towards a certain mutation, e.g. to alanine.
Figure 1. Predicted vs experimental ΔΔG values. (A) SimBa-I on the B663 dataset. (B) SimBa-NI on the B663 dataset. (C) SimBa-I on the S350 dataset. (D) SimBa-NI on the S350 dataset.

Table 1. Statistical measures (R, MAE in kcal/mol and MSE in kcal/mol) for the two simple models, SimBa-NI and SimBa-I trained on the balanced B1131 data set against the two test sets.

<table>
<thead>
<tr>
<th></th>
<th>SimBa-NI</th>
<th></th>
<th>SimBa-I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>B663</td>
<td>0.46</td>
<td>1.19</td>
<td>0.12</td>
</tr>
<tr>
<td>S350</td>
<td>0.47</td>
<td>1.07</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Figure 2. Performance of Simba-I. (A) MAE in kcal/mol for the SimBa-I predictions of each mutation type in the B663 data set; (B) Relationship between the absolute error of the prediction of SimBa-I and the absolute value of the experimental ΔΔG.

Performance on mutation types and outliers

The SimBa models constructed in this study use a balanced training set in respect to mutation types, which is meant to reduce the bias other methods have towards certain mutation types, especially alanine mutations. Figure 2a shows the MAE of SimBa-I for each type of mutation found in B663. As can be observed, only a few mutation types have a MAE > 2 kcal/mol and the overall bias is low. In particular mutations to proline, glycine and cysteine are problematic for the SimBa-I model. These mutations often strongly perturb the structure thus it is expected that a model that uses only the wild-type structures would fail in their prediction.

There are several outliers in the SimBa-I prediction (Table S3) that are mutations with very large absolute ΔΔG in experiment. In fact, there is a high correlation between the absolute ΔΔG in experiment and the error of the prediction (R=0.69, Figure 2b). We note that this behaviour is not necessarily very significant for the method, as very large experimental values relative to the full fold stability of the protein could also derive from errors in the experiment.
Table 2. Results for models described here and eight other widely-used protein stability prediction methods for the two test sets B663 and S350: Pearson correlation coefficient (R), mean absolute error (MAE, in kcal/mol), and mean signed error (MSE, in kcal/mol).

<table>
<thead>
<tr>
<th></th>
<th>B663</th>
<th></th>
<th></th>
<th></th>
<th>S350b</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>MAE</td>
<td>MSE</td>
<td>R</td>
<td>MAE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimBa-NI</td>
<td>0.46</td>
<td>1.19</td>
<td>0.12</td>
<td>SimBa-NI</td>
<td>0.47</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>SimBa-I</td>
<td>0.52</td>
<td>1.11</td>
<td>0.02</td>
<td>SimBa-I</td>
<td>0.51</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>FoldX</td>
<td>0.45</td>
<td>1.32</td>
<td>-0.02</td>
<td>FoldX</td>
<td>0.49</td>
<td>2.16</td>
<td></td>
</tr>
<tr>
<td>CUPSAT</td>
<td>0.22</td>
<td>1.62</td>
<td>0.2</td>
<td>CUPSAT</td>
<td>0.37</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>SDM</td>
<td>0.42</td>
<td>1.30</td>
<td>-0.65</td>
<td>SDM</td>
<td>0.61</td>
<td>1.29</td>
<td></td>
</tr>
<tr>
<td>PoPMuSiC</td>
<td>0.62</td>
<td>1.01</td>
<td>-0.12</td>
<td>PoPMuSiC</td>
<td>0.67</td>
<td>1.61</td>
<td></td>
</tr>
<tr>
<td>Maestro</td>
<td>0.49</td>
<td>1.19</td>
<td>-0.19</td>
<td>Maestro</td>
<td>0.70</td>
<td>1.13</td>
<td></td>
</tr>
<tr>
<td>Automute</td>
<td>0.44</td>
<td>1.20</td>
<td>-0.17</td>
<td>Automute</td>
<td>0.46</td>
<td>1.43</td>
<td></td>
</tr>
<tr>
<td>mCSM</td>
<td>0.60</td>
<td>1.11</td>
<td>-0.2</td>
<td>mCSM</td>
<td>0.73</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>I-Mutant 3a</td>
<td>0.70</td>
<td>0.92</td>
<td>-0.12</td>
<td>I-Mutant 3</td>
<td>0.50</td>
<td>1.50</td>
<td></td>
</tr>
</tbody>
</table>

a: The test set is a sub-data set of the I-Mutant training set; b: Data taken from 65

Comparison to other predictors

To compare SimBa-I and SimBa-NI to other structure-based protein stability prediction methods we computed the predictions of eight popular methods on the two test sets tested herein. Statistical measures of the quality of the prediction of all ten methods are shown in Table 2. Scatter plots of ΔΔGexp vs ΔΔGpred for the eight predictors compared via the B663 data set are shown in Figure S1.

Many predictors, especially machine-learning predictors, report a very high correlation (R=0.7-0.8) in their validation, but that does not always translate to blind testing, especially when using a mutation-type balanced data set as test set, as B663.30 Unsurprisingly, I-Mutant 3.0 showed the highest accuracy of all methods in terms of both R and MAE for B663, as all data points in the test set belonged to the training set of I-Mutant, thus a fair comparison with I-Mutant is not possible for this test set. SimBa-I performed the third best in terms of correlation, after PoPMuSiC and mCSM, and the second best in terms of MAE, after PoPMuSiC and tied with mCSM. Furthermore, SimBa-I had the least bias towards stabilization and destabilization, tied for the lowest MSE with FoldX. All other methods except CUPSAT showed a destabilization bias, which is common due to destabilization biases in the experimental data used as training sets. This indicates that the use of a balanced training set quells somewhat the destabilization bias of the methods.
On the popular test set S350, the SimBa models performed worse than current modern machine-learning models, such as mCSM and Maestro, and worse than the newest version of SDM, but better than FoldX and CUPSAT. However, both SimBa-I and SimBa-NI performed almost exactly the same on S350 as on B663, highlighting the robustness of the methods. This is not the case for any other prediction method. PoPMuSiC showed a similar correlation coefficient, but a significantly higher MAE for S350.

The mutations from the two data sets that have the highest absolute errors for each of the eight prediction methods are shown in Table S4. As the SimBa models, all other studied methods displayed large absolute errors for mutations that have a large absolute experimental ΔΔG. Mutations to V13 in proteinase inhibitor SSI (3SSI) and Y97 in ribonuclease A (1RTB) are outliers in all methods studied, including SimBa, indicating that these are particular difficult cases and not necessarily errors in the models. 3SSI is a 2.30-Å resolution structure of a dimer from 1996, and 1RTB is a 2.5-Å resolution crystal structure from 1992.

We note that many prediction methods take into account only one chain per protein and thus may have difficulties predicting mutations at interfaces of multimeric proteins, such as the proteinase inhibitor SSI. However, this is not the case for the SimBa models, as the RSA calculation by Naccess takes into account all chains. In order to confirm this, we evaluated the performance of each method on only monomeric proteins of the B663 data set (Table S5). The results show that our methods are largely unaffected by the multimeric state, as expected. The largest differences are seen for the SimBa models, which perform better on monomeric proteins, most likely because of the B1131 training set, and SDM, which performs worse on monomeric proteins.

All the indicators discussed above relative to the accuracy (trend accuracy, numerical accuracy and systematic error) of the methods. Recently, we proposed that the structural sensitivity of a method is a useful base measure for the precision of a method, with the caveat that a very low structural sensitivity would not be desirable, as it would mean that the method is not really “structure-based”. As SimBa-I and SimBa-NI only use one metric that depends on the structure (RSA), it is expected that the structural sensitivity would be low. Indeed, as shown in Table S6, the structural sensitivity of the two SimBa methods is the lowest (0.09 kcal/mol) except for I-Mutant and generally our simple method is as precise, in terms of producing structure-independent outcomes, as the best machine-learning methods (0.09-0.14 kcal/mol).
Figure 3. Coefficients of each variable in the multilinear regression models. (A) SimBa-NI. (B) SimBa-I. A negative coefficient implies destabilizing and a positive implies stabilizing average effect.

Contributions to protein stability
While it is of fundamental interest by itself that only three properties provide so much of the predictive power, a major practical advantage of a simple model is that the model can be readily applied to understand the driving forces behind the effects of each specific mutation under consideration, which is sometimes very hard with machine-learning methods.

Figure 3 shows the coefficients of each variable used in the SimBa-NI and SimBa-I model to the predict ΔΔG values in the balanced test set. A negative contribution shows a destabilization of the protein due to the mutation, whereas a positive contribution shows a stabilization of the protein due to the mutation. For SimBa-NI (Figure 3a), the optimized intercept shows that the average effect of a mutation was to destabilize a protein by ~1.6 kcal/mol, which is negated if the mutated residue was exposed (positive contribution of RSA). For SimBa-I (Figure 3b), the expected effect of a buried mutation is similar, -1.6 kcal/mol.

A smaller volume of the mutated residue tends to destabilize the protein if the mutation was buried, otherwise it tends to stabilize (positive contribution of V_{diff} and negative contribution of RSA x V_{diff}). Conversely, an increase in hydrophobicity most typically tends to destabilize the protein if the mutation is exposed, otherwise it tends to stabilize (negative contribution of H_{diff} and positive contribution of RSA x H_{diff}), which is quite intuitive, but important to see in a quantitative context.

It is also important to note that the SimBa models capture most of the physics of side chain contributions to the structural stability of the folded state, as the V_{diff} term indirectly includes
information on the change in favourable Van der Waals contacts, whereas the $H_{\text{diff}}$ term is to some extent a proxy of polar (electrostatic and hydrogen-bond) contributions. Thus, we demonstrate that the physical understanding of protein stability can be translated into a simple quantitative model with the similar predictive capacities as advanced models. The models also have the advantage of being easy to use, even by a non-expert. A short set of instructions for the use of the models is included in the Supporting Information.

**Conclusions**

Current models describing the effect of mutation on protein stability are often complex and hard to interpret; at the same time, it is unclear how important their complexity really is. We have built two multilinear regression models, SimBa-I and SimBa-NI, which use only three simple parameters: The relative solvent accessibility, the difference in volume between the wild-type and the mutated residue and their difference in hydrophobicity. Many parametrized models display higher accuracy for mutation types in their training sets, and thus our models were trained using a balanced training set with a maximum of five mutations for each of the 380 mutation types.

The validation of the models on a balanced test set showed that they possess good accuracy, with SimBa-I having $R \sim 0.5$, mean absolute error $\sim 1.1$ kcal/mol and a very low systematic (signed) error of 0.02 kcal/mol. This performance was only slightly worse than that of PoPMuSiC but better than all other much more complex methods that we studied. SimBa-NI, the model that did not consider the solvent exposure context of the other two independent variables performed slightly worse, highlighting the importance of solvent-exposure-weighted chemical properties. SimBa-I also showed similar performance on a popular test set used by many other methods, S350, with $R \sim 0.5$ and a mean absolute error $\sim 1.07$ kcal/mol. Although performance is better for some other methods for this special test set, it confirms that SimBa-I performs consistently well across diverse data sets.

The coefficients of the model suggest an intuitive relationship between protein stability and mutation, notably showing that a hydrophobicity increase in the core of the protein translates to an increase in protein stability, and that the expected default destabilization of a folded protein by a random buried mutation is 1.6 kcal/mol (larger than expected from mutation-imbalanced data sets).

Our results imply that a hard wall of accuracy ($\sim 1$ kcal/mol, $R \sim 0.5$) exists, since more complex models do not add much value. We think this is because all the studied methods miss the (experimentally typically elusive) mutant and unfolded structures, interpolate stability effects only
from the wild-type folded structure, and ignore amino acid correlations and water colocalization effects. We urge future methods to push these features to break the 1 kcal/mol accuracy limit.

In conclusion, our model quantitatively assesses protein stability upon mutation from easy-to-understand amino-acid properties in a very simple least-squares multilinear regression. The strength of the prediction model comes both from its interpretability and from the mutation-type balanced data set used for training and shows that the same accuracy as typically seen by advanced methods is largely achieved already with such simple assumptions.

Acknowledgements
The Danish Council for Independent Research (Grant case: 8022-00041B) is gratefully acknowledged for supporting this work. We acknowledge access to computational resources from the ROBUST Resource for Biomolecular Simulations (Novo Nordisk Foundation, NNF18OC0032608).

Supporting Information
The supporting information pdf file contains hydrophobicity scales studied, quality metrics of the models using them, prediction outliers for SimBa-I, scatter plots for the predictions of other methods tested in this study, instructions for using Simba-I, and van der Waals radii. This information is available free of charge at http://pubs.acs.org

Data and Software Availability
All the data used in this work are uploaded in a separate file called datasets.xlsx. The regression model is fully reproducible from these data via the model parameters (coefficients) described in the "model" sheet in the excel file. Addition information is available from the authors upon request. The additional models tested are described as below:

FoldX: http://foldxsuite.crg.eu/

Maestro: https://pbwww.che.sbg.ac.at/maestro/web/

PoPMuSiC v 2.1: https://soft.dezyme.com/

CUPSAT: http://cupsat.tu-bs.de/
mCSM: http://biosig.unimelb.edu.au/mcsm/

I-Mutant: http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi

SDM: http://marid.bioc.cam.ac.uk/sdm2

Automute: http://binf.gmu.edu/automute/

ORCID
Kasper Planeta Kepp: 0000-0002-6754-7348
Octav Caldararu: 0000-0002-5900-6568
Tom L Blundell: 0000-0002-2708-8992

References


(20) Petukh, M.; Kucukkal, T. G.; Alexov, E. On Human Disease-Causing Amino Acid Variants:


(64) Masso, M.; Vaisman, I. I. AUTO-MUTE 2.0: A Portable Framework with Enhanced

ΔΔG = a + b \cdot RSA + c \cdot V_{\text{diff}} + d \cdot H_{\text{diff}} + e \cdot RSA \cdot V_{\text{diff}} + f \cdot RSA \cdot H_{\text{diff}}