JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

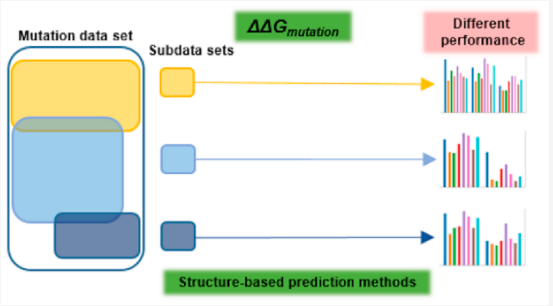
Systematic Investigation of the Data Set Dependency of Protein Stability Predictors

Octav Caldararu, Rukmankesh Mehra, Tom L. Blundell, and Kasper P. Kepp*

Cite This: J. Chem. Inf. Model. 2020, 60, 4772−4784 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Prediction of protein stability changes caused by mutation is of major importance to protein engineering and for understanding protein misfolding diseases and protein evolution. The major limitation to these applications is the fact that different prediction methods vary substantially in terms of performance for specific proteins; i.e., performance is not transferable from one type of mutation or protein to another. In this study, we investigated the performance and transferability of eight widely used methods. We first constructed a new data set composed of 2647 mutations using strict selection criteria for the experimental data and then defined a variety of subdata sets that are unbiased with respect to various aspects such as mutation type, stabilization extent, structure type, and solvent exposure. Benchmarking the methods against these subdata sets enabled us to systematically investigate how data set biases affect predictor performance. In particular, we use a reduced amino acid alphabet to quantify the bias toward mutation type, which we identify as the major bias in current approaches. Our results show that all prediction methods exhibit large biases, stemming not from failures of the models applied but mostly from the selection biases of experimental data used for training or parametrization. Our identification of these biases and the construction of new mutation-type-balanced data should lead to the development of more balanced and transferable prediction methods in the future.



INTRODUCTION

Understanding and predicting how a mutation changes the stability of a protein is a central goal of modern biology, which would enable efficient rational engineering of stable proteins[1−3] and improved understanding of protein stability-related genetic disease risk driven by missense mutations.[4−6] Many steps have been taken in this direction, both experimentally through site-directed mutagenesis studies[7,8] and computationally via the development of diverse algorithms to predict stability changes upon mutation.[9−18]

The change in a protein's free energy of unfolding due to mutation ($\Delta\Delta G$) can be measured experimentally from thermal or chemical denaturation experiments.[19,20] Due to the dependence of the protein stability on the exact interatomic interactions,[21] no computational method can formally calculate the free energy of unfolding, since the structures of the unfolded states and most mutants are unavailable. Thus, $\Delta\Delta G$ is typically interpolated from a single wild-type protein structure. Calculating interaction energies accurately is expensive as it requires both accurate force fields and sampling of conformational space of the altered mutant structure, which is infeasible for proteins with tens of thousands of atoms. Free energy simulations[22,23] are too computationally expensive for extensive use required in the biomedical and protein engineering fields, and even then, performance strongly depends on the quality of the structure, sampling, and force field.

Therefore, computational prediction methods to be used for screening many mutations are necessarily empirical. Some predictors use parametrized energy functions to compute the $\Delta\Delta G$, whereas others apply machine-learning methods. Among methods that use energy functions, some use physics-based force fields,[24−27] and others use statistical potential approaches, where $\Delta\Delta G$ depends on propensities of an amino acid to be in a given environment.[14,28] On the other hand, machine-learning methods[17,18,29] train their model on large data sets and then apply various regression methods to obtain a final $\Delta\Delta G$ prediction. Parameters are obtained from experimental data sets, which are mostly subsets of the largest experimental database available, ProTherm,[30] and a collection of these data sets with annotations can be found on the VariBench website.[31]

Two previous studies[13,32] reported the performance of nine and six prediction methods, respectively, and found that all methods show a poorer performance on data sets that differed from the training sets used by the authors of the various

Received: May 28, 2020
Published: August 10, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

methods. Accordingly, the experimental training data sets have a strong effect on the quality and transferability of the prediction method, i.e., a training data selection bias. Other studies have shown that the performance of some predictors also depends on the quality of the wild-type structure used, with some methods being distinctly more structure-sensitive than others.[33,34] As most predictors do not account for large structural changes upon mutation, structure-breaking mutations are commonly poorly predicted.[35,36] Typically, these involve buried residues, but Khan and Vihinen[32] concluded that predictors perform better on buried mutations, which could also be due to a data selection bias.

Given the data selection bias, no prediction method is perfect, and it is not generally clear which types of mutations and proteins will be better or worse described by a specific method. The most well-known bias is that predictors tend to be more accurate for destabilizing than for stabilizing mutations[37,38] simply because experimental $\Delta\Delta G$ values used for parametrization are heavily skewed toward destabilizing mutations, with the average random mutation destabilizing by typically 1 kcal/mol.[39] Another major bias is a type of amino acid mutation with some types, notably alanine-scan derived data points, being massively over-represented. Quantifying these biases is difficult due to the large number of possible fundamental mutation types (380), many of which are represented by few data if any at all, as discussed below. Instead, studying this bias can (and must) be done via reduced alphabets,[40] as attempted here.

Several recent studies[11,12,38,41] investigated biases in current prediction methods and particularly emphasized the important lack of symmetry of predicted $\Delta\Delta G$ values, i.e., the fact that reverse mutations are not simply producing the sign-inverted result of the direct mutations. The nonsymmetrical nature of some terms in the energy function of the methods has been identified as an issue. Steps have been taken to correct it, but this bias could also be attributed to a data selection bias, which has not been investigated in depth.

Here, we report the performance of eight prediction methods on a novel curated data set, consisting of data both from ProTherm and from other more recent studies. We define subdata sets that are unbiased with respect to various aspects, such as stabilization extent, structure type, solvent exposure, and mutation type, and use them to investigate the effect of data selection biases on predictor performance. In particular, we use a reduced alphabet to quantify the bias toward mutation type, which we identify as the major bias in current approaches. We identify large biases, not due to failures of the actual mathematical models applied but mostly due to the experimental data used for training or parametrization. With the identification of these biases and the proposed mutation-type-balanced data set, we hope to develop more balanced and transferable prediction methods in the future.

METHODS

Data Set Construction. We created a new curated data set of single-point mutations with known $\Delta\Delta G$ values starting from the largest available database, ProTherm,[30] last updated 2013. ProTherm contains very diverse and heterogeneous data, with many different conditions represented, which we expect to produce noise, and with multiple data points for many of the same mutations. Accordingly, our curation was performed with the following criteria: (1) only single-point mutations, (2) known $\Delta\Delta G$ value, (3) known protein structure in the protein data bank, (4) temperature of experiment between 20 and 30 °C, (5) pH of experiment between 5 and 9, (6) only chemical denaturation considered, and (7) no nonstandard amino acids present in the protein. Furthermore, all data points that did not match the sequence were discarded, and all entries with $\Delta\Delta G$ values in kJ/mol were converted to kcal/mol. Thus, we believe that most of the known issues concerning the data in ProTherm[42] have been resolved in our data sets. Duplicate values were averaged if the difference in $\Delta\Delta G$ was less than 0.5 kcal/mol and fully removed from the data set otherwise in order to include as little uncertainty as possible in the final data set.

To this data set, newer data from three other proteins were added, which to a large extent compensate for the missing mutation types in the data above, i.e., the myoglobin and superoxide dismutase (SOD1) mutants from two earlier studies[43,44] and streptococcal protein G mutants from a domain-wide mutagenesis study.[45] Mutations from these data sets were only included if the mutation types were not common (less than five mutations of that type) in the curated ProTherm data.

The sign convention was kept as in the ProTherm database; i.e., negative $\Delta\Delta G$ values indicate destabilizing mutations, whereas positive $\Delta\Delta G$ values indicate stabilizing mutations. The final data set contains 2567 mutations from 106 different proteins and is referred to here as O2567. The full data set and the constructed subdata sets can be found in the Supporting Information, zip file.

Structure Selection. The ProTherm database suggests one structure in the Protein Data Bank (PDB[46]) for each mutation; yet for many proteins, more than one structure has been determined. In order to find the best structures, we performed the following steps: (1) A PDB search was performed based on the Uniprot ID listed in ProTherm, with the option "wild-type only". (2) Only X-ray structures were considered. (3) If possible, an apo structure was selected. (4) If possible, a monomeric structure was selected. (5) If multiple structures fulfilled the conditions above in the same way, the structure with the lowest R_{free} value was selected. (6) In case the R_{free} was not reported for some of the structures, the highest resolution structure was kept. (7) If multiple structures had similar R_{free} or resolution, the real-space Z-difference (RSZD) was calculated for all the residues that underwent mutations in the data set, and the structure with the lowest average RSZD was used. RSZD scores were calculated using EDSTATS[47] from electron density maps calculated in *phenix.maps*.[48]

Of the 106 proteins in the O2567 data set, 70, or about two-thirds, had their PDB codes updated through the above procedure. A full list of the updated PDB codes and their equivalent in ProTherm can be found in the Supporting Information (Table S1).

Prediction Methods Used. Prediction methods were selected based on their ability to model any mutation, to give a quantitative $\Delta\Delta G$ prediction (rather than just qualitatively; destabilizing or stabilizing), and to work at high computational speed; a diverse group of methods was desired to assess the data set dependence broadly. Eight publicly available, widely used predictors were used in this study: FoldX,[27] I-Mutant 3.0,[49] PoPMuSiC 2.1,[28] Maestro,[50] mCSM,[18] SDM,[14] CUPSAT,[51] and Automute 2.0.[52]

To briefly highlight some of the main differences between these methods, FoldX uses an empirical force field to calculate

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

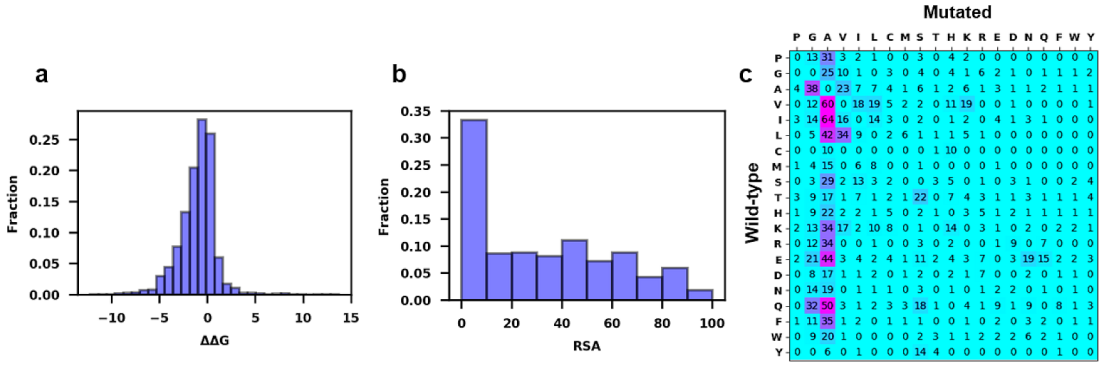


Figure 1. Distribution of (a) $\Delta\Delta G$ values, (b) RSA, and (c) mutation types for the O2567 data set.

the free energy of folding for the wild-type and mutant structures. As the force field is rather sensitive to the structure,[34] a minimization of the wild-type structure was performed before prediction using the FoldX command *RepairPDB*. CUPSAT uses atomic potentials from chemical properties and empirically derived torsion potentials. I-Mutant 3.0 uses support-vector machines that account for amino acid substitution and structural environments. Similarly, Maestro combines support vector machines with a random-forest approach to obtain a consensus free energy. mCSM uses graph-based signatures that encode distance patterns between atoms. Automute 2.0 uses a four-body statistical potential derived through Delaunay tessellation of the whole protein. SDM uses a statistical potential obtained from an environmental-specific substitution table. PoPMuSiC also uses a statistical potential calculated from contact probabilities of amino acids close to the mutated residue.

Unless specified otherwise, all prediction programs were run with default parameters.

Reduced Amino Acid Alphabet. For the analysis of mutation types, we used a reduced amino acid alphabet to minimize the number of mutation types since not all mutation types are covered by the experimentally known data, which produces a bias problem. This alphabet was computed from local structure features of amino acids in approximately 1400 structures in the PDB.[53] While multiple groupings were proposed by the authors, we use here the alphabet consisting of eight groups, with a letter denoting each group in parentheses: hydrophobic—A, L, M (A); aliphatic—I, V (I); aromatic—F, Y, W (F); long polar amino acids—E, Q, K, R (X); short charged/polar amino acids—D, N (N); short polar amino acids—H, S, T, C (S), and two groups consisting of one structure-breaking amino acid each, (G) and (P). We refer to this reduced alphabet as the Etchebest alphabet throughout the paper.

Calculating Global and Local Variables of a Mutated Amino Acid. Relative solvent accessibility was calculated with Naccess,[54,55] using as the default the van der Waals atomic radii. The length of the proteins was computed as the number of amino acids in each PDB file. The secondary structure composition of the proteins was taken from the CATH structural database.[56] A mutation was considered to be volume changing if it changed the volume of the residue by more than 30 $Å^3$ (approximately the volume of a water molecule).

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 a method compared to the experimental data. The mean signed error (MSE) shows the systematic error of a method toward stabilization or destabilization.

Overlap between various data sets used in this work is measured as the number of data points in common. A data point was considered to be the same in multiple data sets if it represented the same wild-type residue and the same mutated residue at the same position in the same protein. However, the PDB code and the $\Delta\Delta G$ value can differ.

RESULTS AND DISCUSSION

Biases in the Total Data Set. The distribution of $\Delta\Delta G$ values in the total data set O2567 is shown in Figure 1a. As expected, there are more destabilizing mutations than stabilizing ones, a trait shared with the full ProTherm database[30] and most other data sets used previously for training prediction programs.[31] $\Delta\Delta G$ values have a median of −0.7 kcal/mol, which is typical for experimental and theoretical data.

The distribution of relative solvent accessibility (RSA) of the mutated residues in the wild-type structures is shown in Figure 1b. The larger percentage of residues with RSA of 0%−10% could indicate a bias in the data set toward more buried residues. However, this obviously depends on the threshold that reasonably defines a "buried" residue. If we consider buried residues only those with RSA of 0%−30% and exposed residues being all residues with greater than 30%, the data set is balanced, with 55% of data points being buried and 45% being exposed. We consider the latter reasonable, as only a few outlying mutations are strongly stabilizing or destabilizing with RSA > 30%, as shown in the Supporting Information, Figure S1.

Next, we investigated the bias of the O2567 data set toward mutation type. The bias in experimental mutation types may have a large impact on the application of the methods to mutation types not covered by the training set. Many experimental studies are alanine scans, wherein residues in a protein are mutated to alanine in order to determine the contribution of each residue to the overall stability of a protein. In contrast, residues such as tryptophan, proline, or cysteine are often not mutated due to the disruptions this might cause in the wild-type protein.

We concluded that evaluating this bias accurately requires a data set with a balanced amount of data points for each type of the 380 possible mutations. The number of mutations of each type in the O2567 data set is shown as a heat map in Figure 1c. There is a strong overrepresentation of certain mutation types,

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

particularly any wild-type residue to alanine. Conversely, many mutation types are poorly represented: 79 out of 380 mutation types are not present at all; 35 of these involve proline or cysteine. We expect this under-representation to be suffered by all data sets used for training methods so far in the literature. An additional 155 mutation types are represented by less than five data points, making their type-specific impact on stability poorly defined.

Comparison of O2567 with Other Data Sets. To be able to rigorously evaluate prediction performance, we investigated more closely how the $\Delta\Delta G$ distribution differs in the O2567 data set compared to other specific data sets used for training and testing prediction programs. The maximum, minimum, mean, median, and standard deviation of $\Delta\Delta G$ values in the training and test data sets of various programs are given in Table 1. The means and medians of all data sets are very similar, whereas the magnitude varies, with O2567 and Automute trainings having the most extreme minimum and maximum $\Delta\Delta G$ values. We expect that this would affect the performance of the prediction methods on the mutations with extreme $\Delta\Delta G$ values.

Table 1. Properties of Distribution of $\Delta\Delta G$ Values in O2567 Data Set Compared to Training and Test Data Sets of Prediction Methods Studied[a]

Data set	Maximum $\Delta\Delta G$	Minimum $\Delta\Delta G$	Mean $\Delta\Delta G$	Median $\Delta\Delta G$	Standard deviation
O2567	13.7	−12.2	−1.0	−0.7	2.0
FoldX training	3.4	−5.4	−1.4	−1.2	1.3
FoldX test	3.0	−7.5	−1.4	−1.1	1.4
I-Mutant training	6.8	−12.0	−1.0	−0.7	1.8
I-Mutant test	12.7	−8.4	−0.8	−0.6	1.8
PoPMuSiC training/ mCSM training/ Maestro training	6.8	−5.0	−1.0	−0.8	1.5
PoPMuSiC test/SDM test/mCSM test/ Maestro test	3.8	−4.9	−0.8	−0.6	1.6
SDM training	6.8	−9.3	−1.1	−0.8	1.7
Automute training	13.7	−12.2	−1.3	−1.1	1.9
Automute test	6.8	−12.0	−1.0	−0.7	1.8

[a]Several programs have used the same data sets for training or testing. All values are in kcal/mol.

Since O2567 is partly a subset of the ProTherm database, it is expected that it will have some overlap with the training sets of the studied prediction methods. Figure 2a shows the number of mutations in common between O2567 and the training and test sets of each predictor. Data for CUPSAT were not available to us and were thus excluded from the analysis. The overlap is less than 50% with all training sets except that of Automute 2. Accordingly, O2567 is a distinct data set with a good balance between similarities and differences relative to other data sets, in particular, considering the data-quality-based selection criteria that we applied. When applying the different methods to the O2567 data set, the trend accuracy, as estimated from the correlation coefficient, correlates fairly linearly with the overlap of O2567 with the method's training set (R = 0.77, Figure 2b). This suggests that the data used in the training of the methods largely control the performance of a given method, as shown before,[32] and raises a concern about overfitting and transferability of these methods in general.

Method Performance for the Full O2567 Data Set. As a first blind test of the eight methods, we studied in detail how they perform for all the mutations of the O2567 data set. We note that among them Maestro could not be applied to structures that contained missing residues, resulting in only two-thirds of the mutations being calculated. This can cause a slight inflation in all the performance metrics of Maestro reported herein.

The performance of the prediction methods on the full data set, evaluated by three different metrics (R, MAE, and MSE), is shown in Figure 3. These three metrics measure the trend accuracy, the overall numerical accuracy, and the systematic over- or under-stabilization tendency. We stress that this benchmark does not probe the quality of the methods at all, since as we show below that performance is extremely data set dependent and only performance on balanced test sets are relevant probes of performance. When tested on the full O2567 data set, Maestro showed the highest trend accuracy as measured by R, followed by Automute, PoPMuSiC, mCSM, and I-Mutant 3.0, all with values ranging from 0.43 to 0.53 (Figure 3a). SDM, CUPSAT, and FoldX correlated less well, with R < 0.4. The prediction methods are quite robust in terms of these correlations, as seen when applying artificial data sets constructed by randomly shuffling the experimental $\Delta\Delta G$ values of the O2567 data set (Table S2). However, the correlation was substantially impaired when removing data

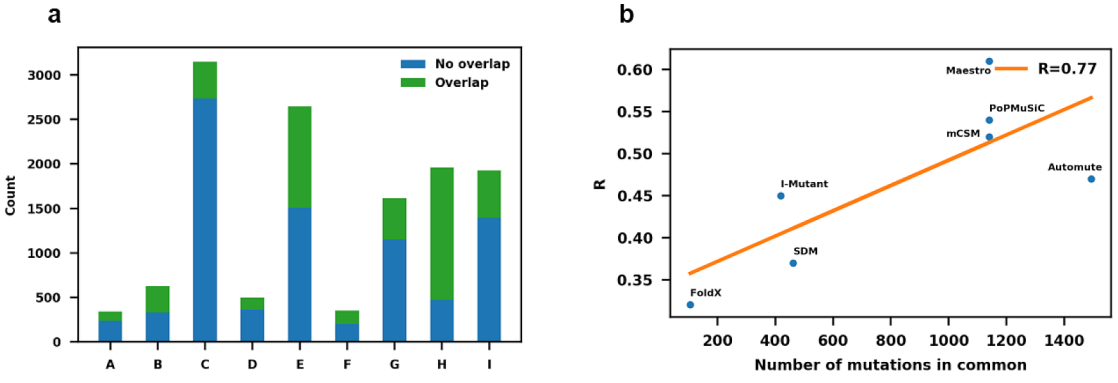


Figure 2. (a) Overlap (number of mutations in common) between the O2567 data set and training and test data sets of methods studied here. Several methods have used the same data sets for training or testing. A, FoldX-training; B, FoldX-test; C, I-Mutant 3.0 training; D, I-Mutant 3.0 test; E, PoPMuSiC training/mCSM training/Maestro training; F, PoPMuSiC test/SDM test/mCSM test/Maestro test; G, SDM training; H, Automute training; I, Automute test. (b) Pearson correlation coefficient for each method tested on the O2567 data set vs training set overlap.

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

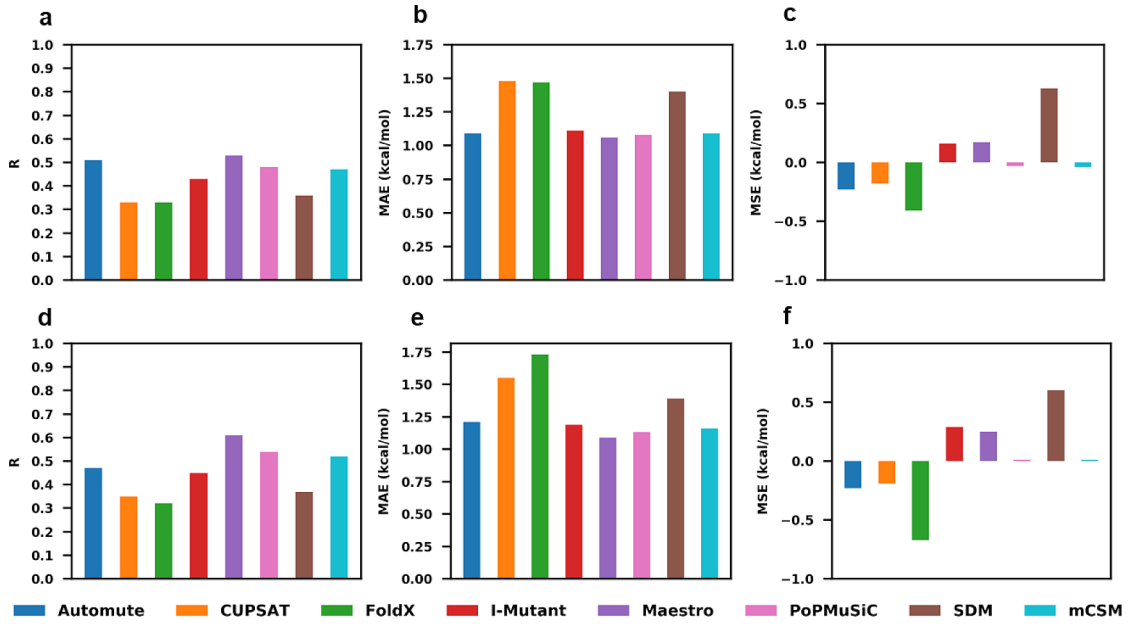


Figure 3. (a) Pearson correlation coefficient, R, (b) mean absolute error (MAE, in kcal/mol), and (c) mean signed error (MSE, in kcal/mol) for the eight prediction methods against the full O2567 data set. (d) R, (e) MAE (kcal/mol), and (f) MSE (kcal/mol) for the eight prediction methods against the balanced data set with a maximum of five mutations of each type.

points included in any training set (Table S3), with the highest trend accuracy (Maestro) being only 0.3. The differences in performance were also reduced correspondingly. This behavior is not surprising and reflects the circularity issue that has been observed when testing other types of machine-learning prediction methods, which results in inflated performance metrics for methods that have the same data points in the training and test data sets.[57] Our analysis allows us to address this important issue by quantifying the training set bias and producing subdata sets that are more representative and relevant, qualities necessary for proper analysis.[58]

The MAE values (Figure 3b) indicate a similar tendency, with Maestro, PoPMuSiC, Automute, mCSM, and I-Mutant 3.0 performing better for this data set than SDM, FoldX, and CUPSAT. This tendency was maintained when data points were removed from any training set (Table S3), although the MAE increased by approximately 0.2 kcal/mol for all methods. The errors of the better performing methods were ∼1.0 kcal/mol, similar to what has been reported previously in independent benchmarks.[13,32,59]

From Figure 3c, PoPMuSiC and mCSM displayed MSE values close to 0 kcal/mol, indicating a balanced description of stabilization effects (no systematic stabilization error). On the other hand, FoldX predicted too many destabilizing mutations (MSE = −0.41 kcal/mol), whereas SDM displayed too much stabilization for the total data set (MSE = 0.63 kcal/mol). However, as seen already from Figure 2b, such performance is largely a matter of overlap with the data points of the test and training sets; i.e., predictive capacity is much smaller for all methods outside their parametrization range, a typical shortcoming of empirical methods. To handle this issue, one needs to either use more sophisticated methods or to train the methods on balanced data sets that are more universal, as discussed below.

Using a Reduced Amino Acid Alphabet to Model Mutations. As already indicated, all data sets used for training and testing methods carry biases toward certain features of their mutations. There are many types of biases, as discussed below, but to obtain a first impression of the biases, we used two approaches. First, we created a balanced data set in which no mutation type (defined as any change from one of the 20 residues to another) was represented by more than five data points; we randomly removed data points from the common mutation types. Due to the random nature of this approach, we created six different balanced data sets. However, the analysis of these data had similar $\Delta\Delta G$ and RSA distributions, with all six data sets exhibiting a mean from −0.94 to −0.83 kcal/mol and 51%−53% buried mutations (Supporting Information, Table S4); we thus report here results for only one of them. Second, in order to account for the many holes in the mutation matrix in Figure 1c (many mutation types are not covered experimentally), we reduced the number of mutation types with a reduced amino acid alphabet, as proposed by Etchebest et al.,[53] discussed in the Methods section.

To understand how mutation bias affects the performance of the methods, we again benchmarked the eight methods on the more balanced subdata set with only five mutations of each type. This benchmark is a more adequate test of the universality (transferability) of the methods. As shown in Figure 3 (bottom panels), all methods show similar metrics as the full data set, with slightly higher correlations (Figure 3a) and slightly lower MAE in general (Figure 3b). Automute and FoldX perform worse in terms of MAE for the balanced data set, whereas the other methods are not much affected. This indicates that Automute and FoldX perform better on more data-wise common mutation types, such as mutations to alanine. This is problematic because these mutations that are common to the data sets are not necessarily common also to nature or protein engineering more broadly. The large bias toward destabilization seen for FoldX (Figure 3c, top) is also substantially aggravated for a more balanced data set (Figure 3c, bottom).

In order for a reduced alphabet to correctly model mutations, all residues within one group should display

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

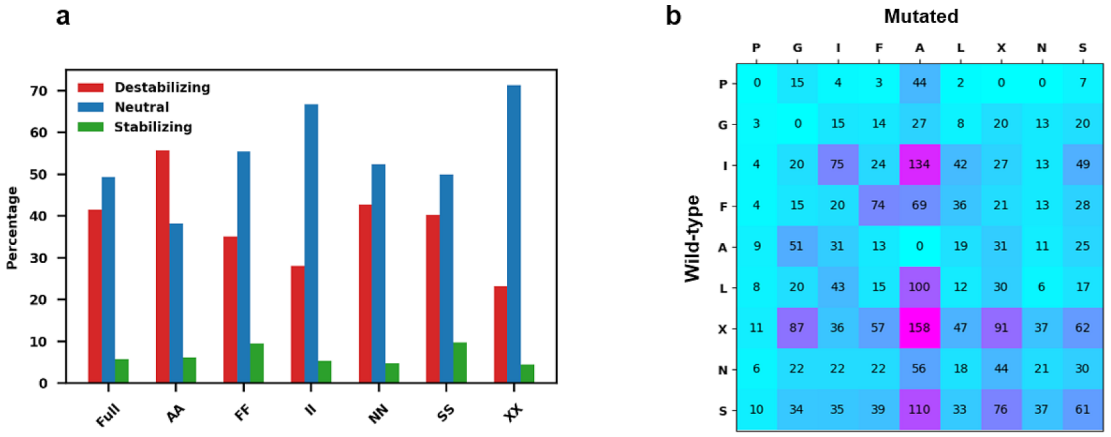


Figure 4. (a) Number of destabilizing, stabilizing, and neutral mutations in each intragroup mutation type for the Etchebest reduced alphabet and for all mutations in the full O2567 data set. (b) Number of mutations per type from the modified Etchebest alphabet: P (P); G (G); I, V (I); F, Y, W (F); A (A); L, M (L); E, Q, K, R (X); D, N (N); H, S, T, C (S).

reasonably similar $\Delta\Delta G$ values relative to the remaining mutation types. This is hard to achieve since $\Delta\Delta G$ values depend substantially on the local context of the mutated site. For a good decomposition, mutations within the same reduced group, e.g., AA or XX, are expected to yield mostly neutral effects on protein stability, with $\Delta\Delta G$ closer to 0 kcal/mol. Figure 4a shows the number of destabilizing, stabilizing, and neutral mutations for each intragroup mutation type. We considered destabilizing mutations as mutations with $\Delta\Delta G < -1$ kcal/mol, stabilizing mutations with $\Delta\Delta G > 1$ kcal/mol, and neutral mutations with -1 kcal/mol $< \Delta\Delta G < 1$ kcal/mol. All intragroup mutations have more neutral mutations, except the AA mutation type, for which most mutations are destabilizing. Furthermore, all intragroup mutations except AA displayed a higher percentage of neutral mutations than the full set of mutations from O2567. Consequently, we separated the AA mutations into AL, LA, AM, LM, MA, and ML mutations to check where the destabilization arises. Here, 71 of 73 destabilizing mutations belonged to the LA or MA mutation types, whereas LM and ML show mostly neutral changes. Thus, we decided to separate the A group into two groups: A (A) and L, M (L).

Our resulting customized reduced alphabet thus consists of nine groups of amino acids and, accordingly, 78 possible mutation types ($9 \times 9 - 3$ for AA, PP, and GG). This reduces the number of mutation types by a factor of ∼5 and increases the number of data points for each mutation type (Figure 4b), making statistical analysis possible. Two mutation types (PX and PN) have no representation in the data set, and six mutation types possess fewer than five data points. All these seven mutation types involve proline, and we thus note that we cannot draw any significant conclusions regarding mutations involving proline; due to the strong structure-disturbing character, such mutations are typically highly destabilizing and potentially denaturing, which is probably why they are so under-represented in the first place.

To properly quantify how each prediction method performs based on mutation type, we computed predictions for all mutation types in the reduced alphabet (Figure 4b). The correlation of all methods per mutation type is shown in Figure 5a. Mutations to glycine displayed good correlation for all prediction methods. Similarly, most intragroup mutations (except LL), which are mostly neutral, showed good correlation for all predictors. Maestro and Automute displayed the best correlation for 19 mutation types each, and PoPMuSiC had the best correlation for 15 mutation types. Although CUPSAT and FoldX showed poorer correlation for the full data set, they have the best correlation for eight and six mutation types, respectively. In particular, FoldX performed well for mutations involving large, charged residues. mCSM exhibited the best correlation for only two mutation types, despite performing well on the full data set. This indicates that mCSM is more transferable across mutation types, which is also confirmed by its box plot (Figure 6). All methods produced negative correlation for certain mutation types. Very interestingly, the trend for the FS mutation type is negative for all methods apart from Automute. I-Mutant 3.0 is the method that displays most negative correlation for mutation types, whereas mCSM has the fewest.

Substantial variations in MAE due to mutation type is apparent from Figure 5b. Mutations involving large aromatic residues or tiny glycine residues show large errors for most prediction methods, especially for FoldX, which has a MAE > 5 kcal/mol for mutations from glycine to other residues. Other mutation types, such as AN or IL, are much less problematic and exhibit errors of less than 1.0 kcal/mol for all prediction methods. Automute showed the smallest absolute errors for 18 mutation types, followed by Maestro for 16 mutation types, and PoPMuSiC for 14. CUPSAT and FoldX show a poorer performance, only displaying lowest MAEs for three and four mutation types, respectively.

The MSE for each prediction method per mutation type (Figure 5c) generally follows the observations from the full data set. FoldX has very few mutation types for which it shows a positive (stabilizing) MSE. All the FoldX outliers in MSE are strongly destabilizing. Similarly, but not to the same extent, CUPSAT and Automute exhibit a systematic destabilization bias. Conversely, SDM shows only positive MSE outliers, although unlike FoldX it also has mutation types for which it has a systematic negative MSE.

Figure 6 summarizes the performance of the eight methods for the mutation-balanced data sets. SDM and CUPSAT show the fewest outliers in correlation per mutation type but a quite low average R. Automute and Maestro display the highest average R of ∼0.5, whereas mCSM and PoPMuSiC are the most transferrable, with the lowest standard deviation in R when excluding outliers. (Figure 6a). The box plots in Figure 6b suggest that Maestro has the fewest outliers per mutation

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

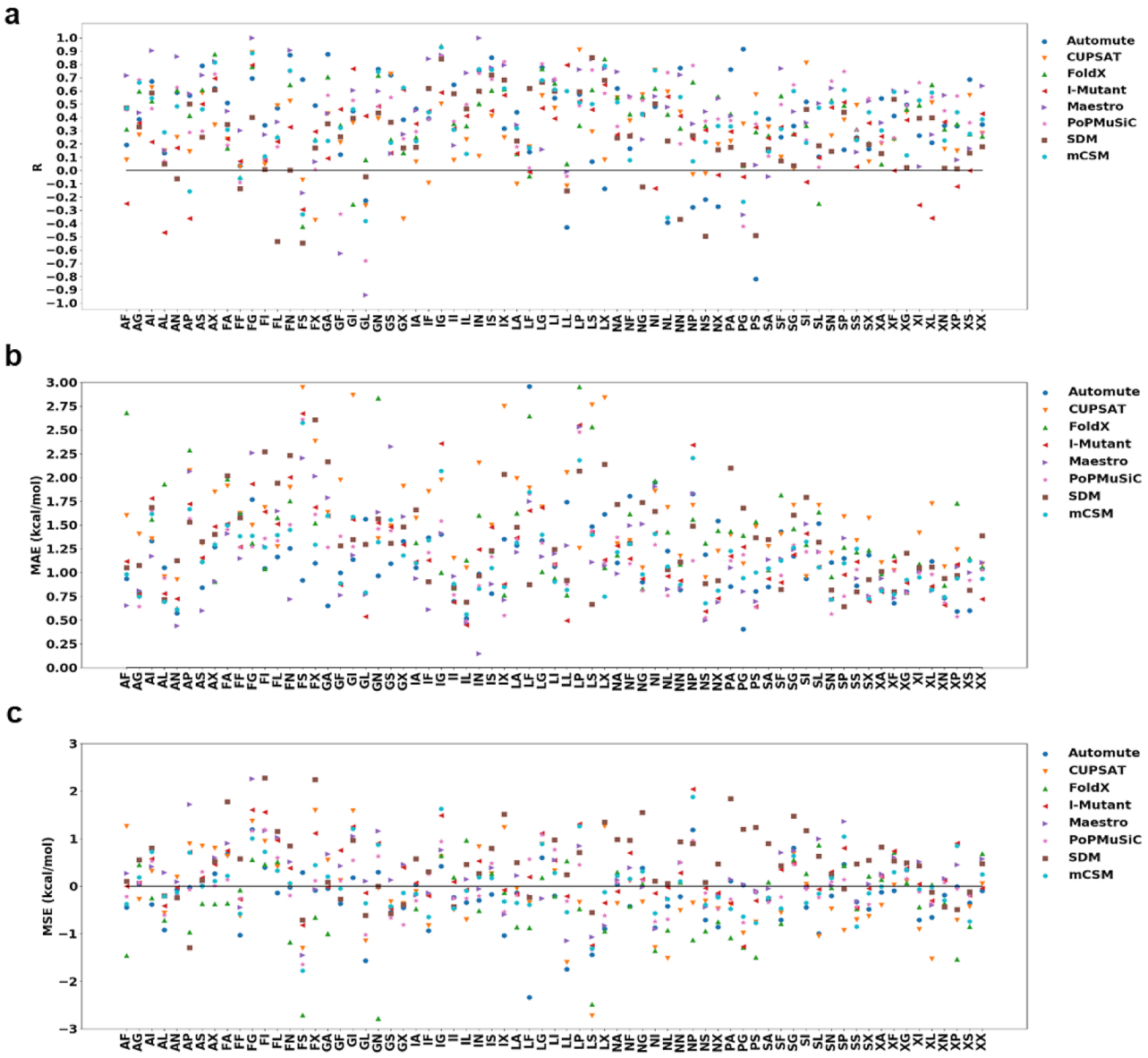


Figure 5. (a) Pearson correlation coefficient, R, (b) MAE (in kcal/mol), and (c) MSE (in kcal/mol) of the prediction methods against subdata sets containing only one mutation type from the modified Etchebest reduced alphabet. Mutation types with fewer than five data points have been excluded.

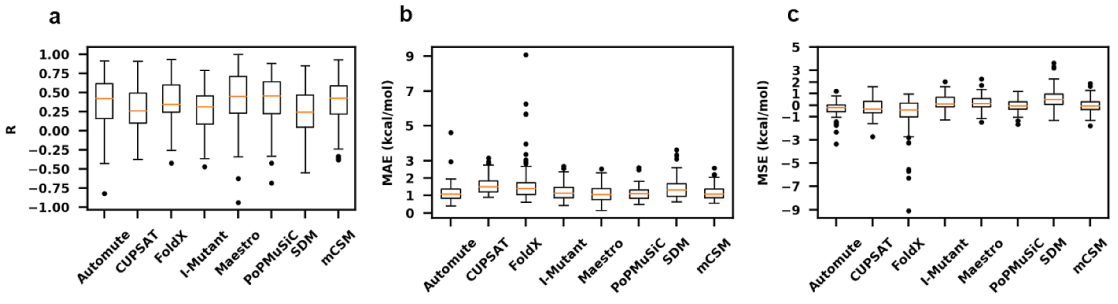


Figure 6. Box plots of (a) Pearson correlation coefficients, (b) mean absolute errors (kcal/mol), and (c) mean signed errors (kcal/mol) for the eight studied methods on mutation-type-balanced subdata sets containing only one mutation type from the modified Etchebest reduced alphabet.

type, based on MAE. PoPMuSiC and mCSM display the lowest standard deviation in MAE when excluding outliers, with PoPMuSiC only having two mutation types with MAE > 2.0 kcal/mol. Automute exhibits good MAE values, in general, but displays almost 5.0 kcal/mol error for one mutation type (LF). It is apparent from Figure 6c that PoPMuSiC, mCSM, I-Mutant 3.0, and Maestro all show good MSE for all mutation types, with low variances and an average centered around 0 kcal/mol. Automute also has a low standard deviation among mutation types when excluding outliers. On the other hand,

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

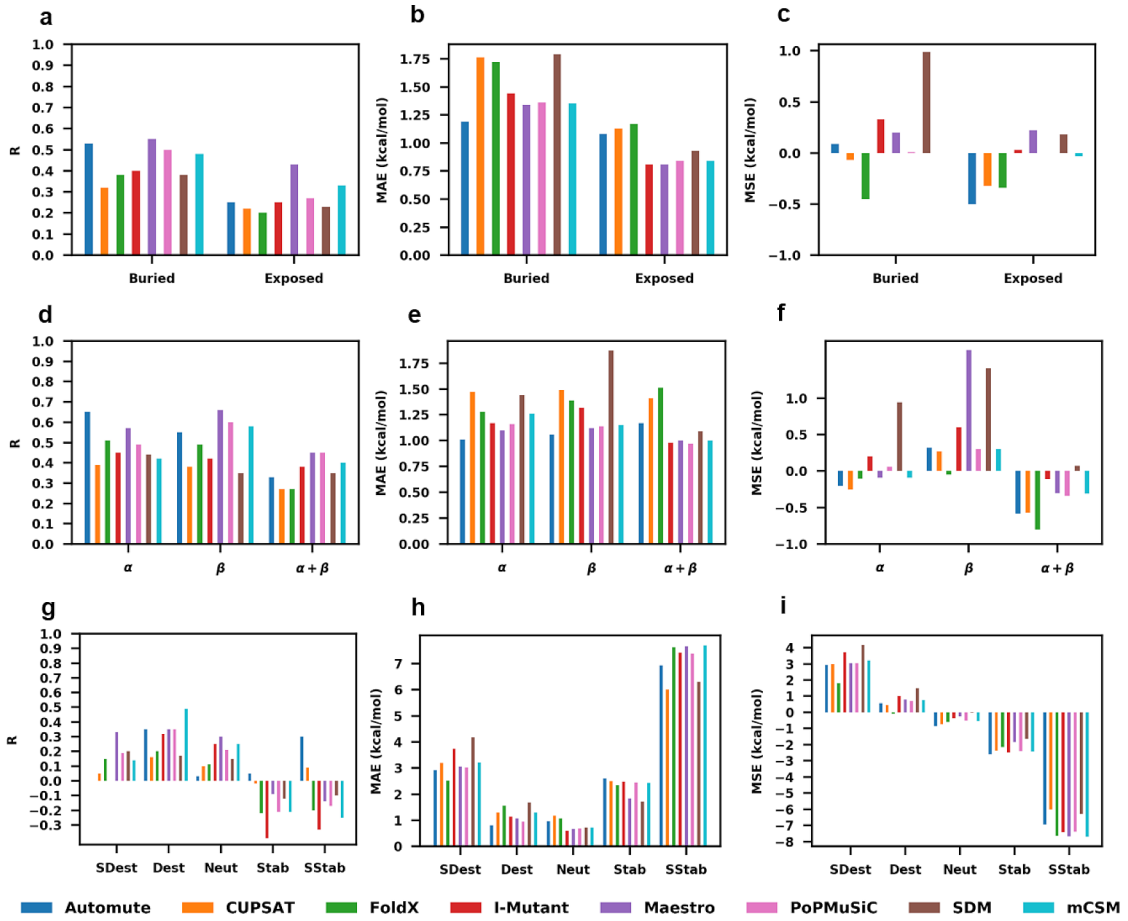


Figure 7. Performance of methods based on mutation properties. (a) Pearson correlation coefficient, R, (b) mean absolute error (MAE, in kcal/mol), and (c) mean signed error (MSE, in kcal/mol) for the eight prediction methods against the subdata sets containing only buried (RSA < 30%) or only exposed (RSA > 30%) mutations. (d) R, (e) MAE (kcal/mol), and (f) MSE (kcal/mol) for the eight prediction methods against the data sets containing mutations only in α proteins, only in β proteins, or only in α+β proteins, according to CATH classification. (g) R, (h) MAE (kcal/mol), and (i) MSE (kcal/mol) for the eight prediction methods against data sets containing mutations with a certain experimental $\Delta\Delta G$: strongly destabilizing ($\Delta\Delta G \leq -4.0$ kcal/mol), destabilizing ($-4.0 < \Delta\Delta G \leq -1.0$ kcal/mol), neutral ($-1.0 < \Delta\Delta G < 1.0$ kcal/mol), stabilizing ($1.0 \leq \Delta\Delta G < 4.0$ kcal/mol), and strongly stabilizing ($\Delta\Delta G \geq 4.0$ kcal/mol).

FoldX has the most outliers, with MSE for some mutation types up to −9 kcal/mol, almost three times as large as any other method.

These results indicate large variations in the behavior of each prediction method depending on the mutation type, of direct relevance to the transferability of the methods.

Performance of Methods vs Properties of Mutations. Above, we have emphasized the mutation type based on a reduced alphabet, which is the logical way to classify amino acid mutations into groups in order to obtain a type-balanced data set. However, other variables apart from mutation type can be of interest. To understand these other properties, we split the original O2567 data set into several subdata sets that share specific properties, such as the length, secondary structure, solvent exposure, or amino acid volume change upon mutation.

The subdata sets considered were (1) buried mutations with RSA ≤ 30%, (2) exposed mutations with RSA > 30%, (3) all mutations with a α-helix structure, according to CATH classification, (4) all mutations with a β-sheet structure, (5) all mutations with an $\alpha + \beta$ structure, (6) all mutations without a fold class, (7) long mutations in proteins with greater than 150 residues, (8) short mutations in proteins with less than 150 residues, (9) large-to-small (L2S) mutations where the wild-type residue has a larger volume than the mutated residue, (10) same-to-same (S2S) mutations where the wild-type residue has approximately the same volume as the mutated residue, (11) small-to-large (S2L) mutations where the wild-type residue has a smaller volume than the mutated residue, (12) monomeric mutations in monomeric structures according to the PDB structure used for protein stability prediction, (13) oligomeric mutations in oligomeric structures according to the PDB structure used for protein stability prediction, (14) mutations in structures without ligands (apo), and (15) mutations in structures with ligands (holo).

Furthermore, we split the data set depending on the magnitude of $\Delta\Delta G$ into strongly destabilizing ($\Delta\Delta G \leq -4.0$ kcal/mol), destabilizing ($-4.0 < \Delta\Delta G \leq -1.0$ kcal/mol), neutral ($-1.0 < \Delta\Delta G < 1.0$ kcal/mol), stabilizing ($1.0 \leq \Delta\Delta G < 4.0$ kcal/mol), and strongly stabilizing ($\Delta\Delta G \geq 4.0$ kcal/mol). A full description of the subdata sets, including number of mutations and distribution of $\Delta\Delta G$ values in each subdata set, is given in the Supporting Information (Table S5).

The performance of each prediction method on these subdata sets is given in Figure 7 and Table S6. Buried mutations showed a correlation comparable to the full data set, R = 0.32−0.55, whereas exposed mutations displayed weaker correlations, R = 0.20−0.43 (Figure 7a), in accordance to a

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

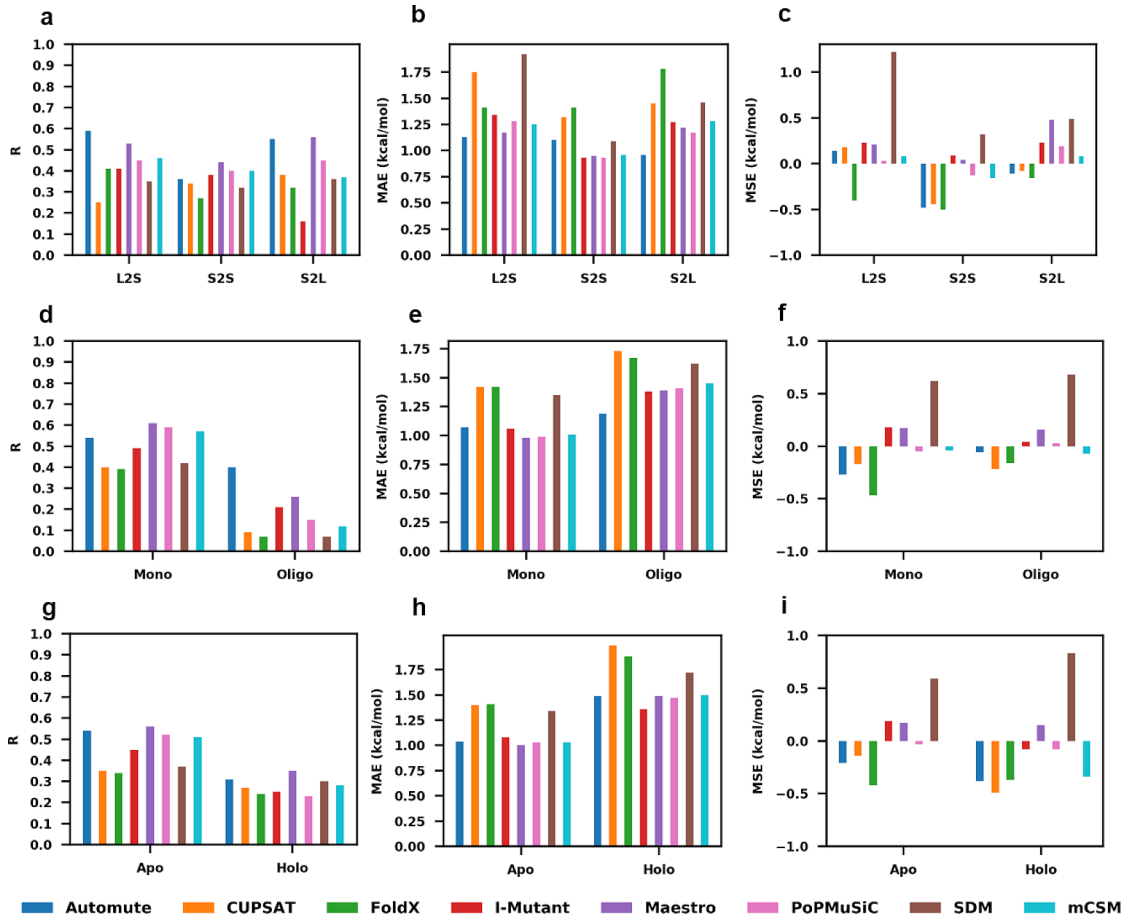


Figure 8. Performance of methods based on mutation properties: (a) Pearson correlation coefficient, R, (b) mean absolute error (MAE, in kcal/mol), and (c) mean signed error (MSE, in kcal/mol) for the eight prediction methods against subdata sets containing only mutations from a large residue to a smaller one (L2S), a small residue to a larger one (S2L), or keeping the volume of the residues constant (S2S). (d) R, (e) MAE (kcal/mol), and (f) MSE (kcal/mol) for the eight prediction methods against the data sets containing mutations only in monomeric proteins or only in oligomeric proteins, according to their PDB structure. (g) R, (h) MAE (kcal/mol), and (i) MSE (kcal/mol) for the eight methods against data sets containing mutations only in proteins without ligands (apo) or with ligands (holo), according to their PDB structures.

previous study.[13] Conversely, the MAE of exposed mutations (MAE = 0.81−1.17 kcal/mol) was lower compared to the buried mutations (MAE = 1.19−1.79 kcal/mol), most probably due to the smaller magnitude of exposed mutations, which have a lower effect on the overall stability of the protein. No prediction method stood out as performing better on buried or exposed mutations than on the full data set. The long and short subdata sets gave similar trends as the buried and exposed subdata sets, probably because the long proteins tend to have a larger percentage of buried residues (Table S7).

A notable variation in predictor performance was observed among the structure-based subdata sets (Figure 6b); FoldX and Automute showed better correlation, 0.51 and 0.65, for α proteins compared to the full O2567 data set, 0.33 and 0.51, respectively. mCSM performed less well on α proteins than on the full data set in terms of correlation, R = 0.42, whereas SDM performed better, with R = 0.44. Maestro and PoPMuSiC show both good correlation, 0.66 and 0.6, and a good MAE, 1.12 and 1.14 kcal/mol, for β-proteins. All methods exhibited poorer correlation but a better MAE for the $\alpha + \beta$ subdata set. This apparently derives from the fact that the $\alpha + \beta$ proteins in this data set are shorter than the α or β ones, leading to more exposed mutations. Moreover, the different behavior of the methods on the structure-balanced subdata sets also stems from each training data set composition. For example, the training set of Maestro and PoPMuSiC contains many β structures (Table S7), which is also reflected in their performance. The subdata set without a secondary structure is statistically inconclusive as it only contains 17 data points.

The subdata sets split according to the $\Delta\Delta G$ values of mutations show the most striking differences in the prediction power of all methods (Figure 7c). As the magnitude of $\Delta\Delta G$ values is naturally lower in each of these subdata sets, the correlation of all methods is also lower than for the full data set. All prediction methods except FoldX and CUPSAT show MAE < 1.0 kcal/mol for neutral mutations. Furthermore, similar MAE can be observed for the destabilizing subdata set as for the full data set. Interestingly, the MSE is always positive for the destabilizing subdata set (except for FoldX) and always negative for the neutral subdata set. This suggests that all methods are built with a predictive bias toward mutations with $\Delta\Delta G \sim -1$ kcal/mol. Studying the stabilizing, strongly stabilizing, and strongly destabilizing subdata sets reveals that all the methods completely fail for unusual mutations. For example, SDM, which displayed a MSE of 0.63 kcal/mol for the full data set, has a MSE of −1.65 kcal/mol for the stabilizing subdata set.

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

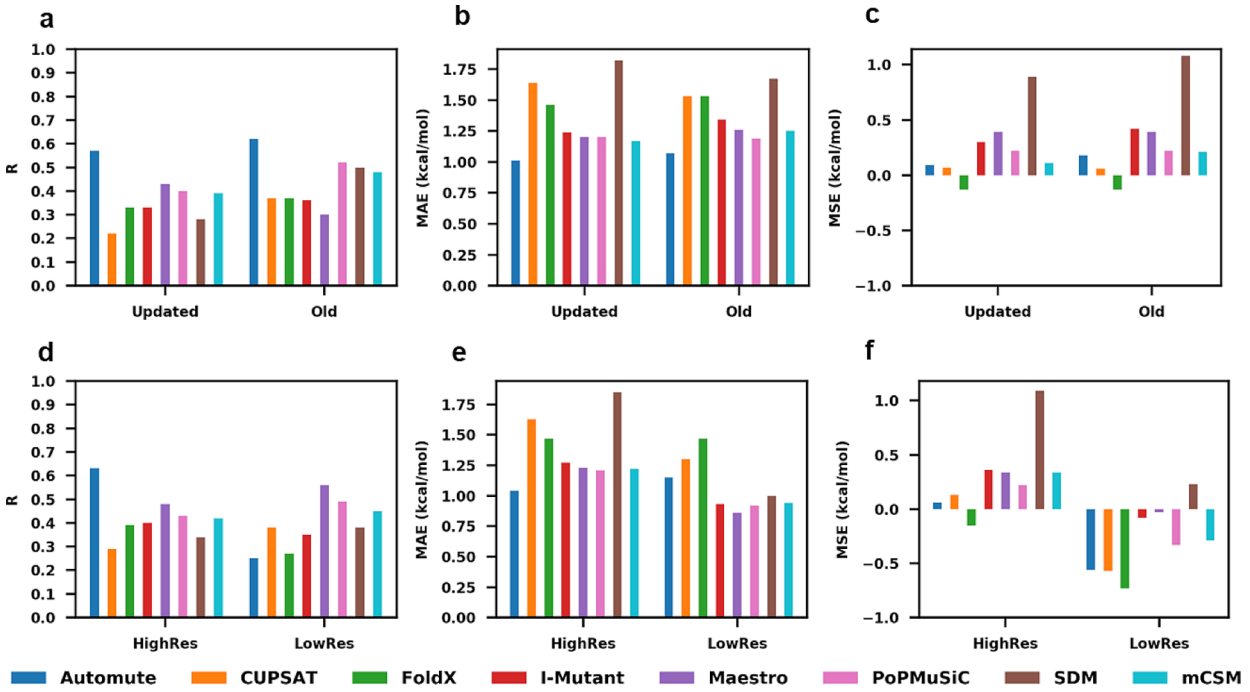


Figure 9. (a) Pearson correlation coefficient, R, (b) mean absolute error (MAE, in kcal/mol), and (c) mean signed error (MSE, in kcal/mol) for the eight prediction methods with predictions started from the ProTherm PDB structure (old) or from an updated PDB structure, (d) R, (e) MAE (kcal/mol), and (f) MSE (kcal/mol) for the eight prediction methods against the data sets containing mutations only in high-resolution crystal structures (<2.0 Å) or only in low-resolution crystal structures (≥2.0 Å).

Method performance also strongly depended on the relative volume of residues involved in the mutation (Figure 8a). In general, the correlation was lower for all methods predicting the same-to-same mutations, R = 0.27−0.44, but displayed lower MAE, 0.93−1.41 kcal/mol. This can again be explained by the small effect of same-to-same mutations on the overall stability of the proteins. Interestingly, not all mutations performed worse on predicting small-to-large mutations than on predicting large-to-small mutations. This behavior would be expected as small-to-large mutations should change more drastically the structure of the protein due to clashes with other residues, which prediction algorithms cannot generally model. In particular, CUPSAT displayed a higher correlation (R = 0.38) and a lower MAE (1.45 kcal/mol) for small-to-large mutations than for large-to-small mutations (R = 0.25, MAE = 1.75 kcal/mol). Conversely, I-Mutant 3.0 behaved as expected, with a much lower ability to predict the trend of small-to-large mutations, with R = 0.16 and MAE = 1.27 kcal/mol. The performance of the methods on these subdata sets is not driven by the composition of the training data sets, as all training data sets have the same composition of large-to-small, same-to-same, and small-to-large mutations as the O2567 data set, with only around 15% small-to-large mutations (Table S7).

It is well known that oligomerization and ligand-binding contribute to the overall stability of a protein; yet ideally, the sample used to determine the mutant stability should reflect the crystal structure composition. Oligomer states in real cells and even in experimental samples can be very heterogeneous and complex, and thus, we define here oligomer state as that extracted from the PDB. Of the studied methods, only PoPMuSiC and Maestro take oligomerization into account; mCSM has a version specific for this purpose, which was not tested (mCSMppi).[60] Predictors that calculate interactions through an empirical or statistical potential add the contribution of the ligand but probably feature worse parametrization for ligands.

Our results show that all prediction methods perform much better on monomeric structures, both in terms of correlation, with R = 0.39−0.61 for the monomeric subdata set and only 0.07−0.26 for the oligomeric subdata set, and in terms of absolute error, with MAE of 0.98−1.42 kcal/mol for the monomeric subdata set and 1.19−1.73 kcal/mol for the oligomeric subdata set (Figure 8b). This indicates that oligomerization plays a very important role in protein stability, and this should be taken into account. Similarly, all prediction methods show a higher correlation and a lower MAE for the apo proteins (R = 0.34−0.56 and MAE = 1.00−1.41 kcal/mol) than for the holo proteins (R = 0.23−0.35 and MAE = 1.36−1.99 kcal/mol), as shown in Figure 8c. The differences in performance are not as major as for the oligomerization subdata sets, although ligands do affect protein stability and therefore prediction.

Interestingly, the conclusions drawn here also hold when using subdata sets that contain no data points used in any training set (Table S3). The only property of the mutation which was affected was the structure of the protein (i.e., α, β or $\alpha + \beta$), with the bias in the methods clearly originating from the training set composition. For other properties, however, the differences in performance stem also from the prediction model used.

Structure Sensitivity. Finally, the question of protein structure input is often overlooked, and since there are several structures available for many of the proteins, it is a matter of interest whether methods display sensitivity and possibly overfitting to choice of structure used for the calculation. In order to test this sensitivity, we calculated the stability effects on two different sets of structures. The first set of crystal structures represents the PDB codes annotated by the

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

ProTherm database, whereas the second set uses updated PDB codes as described in the Methods section.

Figure 9a shows the side-by-side performance of the eight prediction methods using the updated crystal structures and original PDB codes. Proteins that did not have their PDB code updated were not included in this calculation. Perhaps surprisingly, all methods exhibited a better performance when the original PDB codes from ProTherm were used in terms of correlation, with R = 0.30−0.62 compared to the updated PDB codes (R = 0.22−0.57). MAE was similar for both calculations starting from the updated PDB codes (1.01−1.64 kcal/mol) and from the original PDB codes (1.07−1.53 kcal/mol). In many cases, the differences are rather small, with CUPSAT being the most structure-sensitive method, as also shown before.[44] The fact that the updated structures impair prediction can again be attributed to data selection bias, this time in structure use, as the methods were mostly trained using the original structures from ProTherm.

Moreover, we investigated how each method performs on high- and low-quality structures. To this end, we split the O2567 data set into mutations in high-resolution structures (<2.0 Å) and low-resolution structures (≥2.0 Å). It would be expected that methods that take into account local interactions, such as SDM or FoldX, will perform better on high-resolution structures, whereas methods that use a simple description of the protein, e.g., mCSM or Automute, would not exhibit a difference in performance depending on the quality of the structure. Figure 9b also shows the performance of the eight prediction methods on these two subdata sets. Interestingly, only three prediction methods, FoldX, I-Mutant 3.0, and Automute, showed better correlation for mutations in the high-resolution data set, with Automute having the biggest difference, R = 0.63 for the high-resolution subdata set and R = 0.25 for the low-resolution subdata set. Furthermore, only Automute had a lower MAE for the high-resolution data set, 1.04 kcal/mol, compared to 1.15 kcal/mol for the low-resolution subdata set.

These unexpected results confirmed the finding from the use of old and updated structures that most methods carry a bias toward older, less accurate crystal structures used for training. However, excluding any structures that were also used in the training sets of any method heightened the difference in performance on high- and low-resolution structures, with methods showing very low correlations of R = −0.01−0.35 and MAE > 1.5 kcal/mol for the high-resolution data (Table S3). Thus, this bias does not derive from the training data sets of experimental $\Delta\Delta G$, but it could arise from the way that the energy functions were constructed, taking into account structural information from the entire PDB, which consists of >60% structures with ≥2.0. On the positive side, one could argue that most current prediction methods do not need a high-quality structure, making it possible to use homology models for prediction when the structure is not known.

CONCLUSIONS

In this study, we studied how the performance of common protein stability predictors depends on biases in data sets used for training the methods and, as a consequence, the transferability of the methods. All methods displayed a correct trend in their predictions of our full data set, with correlations ranging between 0.3 and 0.5 and errors of the order of 1.0 kcal/mol, similar to what has been previously reported in other independent studies but generally worse than in the original papers.[13,32] The performance of the methods was correlated to how similar our benchmark data set was to data sets used for training the methods, clearly highlighting data set bias for most methods.

To move forward on the transferability problem of protein stability prediction, we then systematically investigated how the performance of each predictor varies with mutation type. For this purpose, we created a mutation-type-balanced subdata set, with a maximum of five data points for each mutation type. All predictors exhibited a similar accuracy for this data set as for the full data set, with FoldX and Automute performing worse. However, many mutation types are not covered in such a data set, making this benchmark of transferability very mild. Instead, we grouped mutation types using a reduced alphabet that presumably captures most of the chemical properties and stability tendencies of the mutations and evaluated the performance of all predictors on each mutation type. From this stricter test, we identified major bias, with predictions exhibiting excellent correlation for certain mutation types and negative correlation for others. In general, mutations involving glycine or aromatic residues were more problematic than, for example, mutations involving small hydrophobic amino acids, again due to overrepresented data points (data set bias). Importantly, we show that this bias could be quantified by using a mutation-type-balanced subdata set, which we expect can help efforts toward training transferable prediction methods.

We also studied other characteristics of mutations by splitting the full data sets into several subdata sets, unbiased by solvent exposure, stabilization extent, and volume of the involved residues or global aspects of the protein structure. All methods performed well on neutral and destabilizing mutation but were unable to predict strongly stabilizing or destabilizing mutations and mildly stabilizing mutations. This is quite disappointing, especially for protein engineering, where we often want to design mutant proteins which are more stable than the wild type. We also observe a difference in predictor performance depending on the solvent exposure of the mutated residue, with solvent exposed mutations exhibiting a poorer correlation than more buried residues. As prediction methods do not explicitly take solvent into account, this is not unexpected, and solvent contribution to protein stability should be studied more closely to remedy this situation.

Interestingly, the protein structure also affects the performance of the prediction methods. All predictors give poorer results if the protein is in an oligomeric state or has a ligand bound, which is not surprising. Thus, care should be taken to have a starting structure that is in a monomeric, apo state. More concerning, the use of newer and better-quality structures than annotated by ProTherm and typically used for training tends to impair performance in many cases. Strikingly, all predictors except Automute perform better on lower resolution structures (>2.0 Å) than on higher resolution structures (<2.0 Å).

Our analysis of these biases has outlined the capabilities of each method when applied to specific types of mutations, which should be of value when judging the significance of applied predictions. We also expect that our study and the proposed data sets will enable the improvement of prediction methods which are currently strongly biased toward over-represented mutation types in their training sets, both by constructing a training data set as free of bias as possible and

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784

by modifying the energy function in order to improve the prediction for the most biased mutation types.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00591.

PDB codes used and additional information on data set performance, sensitivity, and analysis (PDF)

O2567 data set (XLSX)

All constructed subdata sets (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Kasper P. Kepp − *DTU Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark;* orcid.org/0000-0002-6754-7348; Phone: + + 45 45252409; Email: kpj@kemi.dtu.dk

Authors

Octav Caldararu − *DTU Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*

Rukmankesh Mehra − *DTU Chemistry, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*

Tom L. Blundell − *Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, United Kingdom*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00591

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The Danish Council for Independent Research (Grant 8022-00041B) is gratefully acknowledged for supporting this work.

■ REFERENCES

(1) Huang, P.-S.; Boyken, S. E.; Baker, D. The Coming of Age of de Novo Protein Design. *Nature* **2016**, *537* (7620), 320.

(2) Street, A. G.; Mayo, S. L. Computational Protein Design. *Structure* **1999**, *7* (5), R105−R109.

(3) Yeung, N.; Lin, Y.-W.; Gao, Y.-G.; Zhao, X.; Russell, B. S.; Lei, L.; Miner, K. D.; Robinson, H.; Lu, Y. Rational Design of a Structural and Functional Nitric Oxide Reductase. *Nature* **2009**, *462* (7276), 1079−1082.

(4) Kucukkal, T. G.; Petukh, M.; Li, L.; Alexov, E. Structural and Physico-Chemical Effects of Disease and Non-Disease NsSNPs on Proteins. *Curr. Opin. Struct. Biol.* **2015**, *32*, 18−24.

(5) Petukh, M.; Kucukkal, T. G.; Alexov, E. On Human Disease-Causing Amino Acid Variants: Statistical Study of Sequence and Structural Patterns. *Hum. Mutat.* **2015**, *36* (5), 524−534.

(6) Yue, P.; Li, Z.; Moult, J. Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. *J. Mol. Biol.* **2005**, *353* (2), 459−473.

(7) Hughson, F. M.; Barrick, D.; Baldwin, R. L. Probing the Stability of a Partly Folded Apomyoglobin Intermediate by Site-Directed Mutagenesis. *Biochemistry* **1991**, *30* (17), 4113−4118.

(8) Lee, C.; Levitt, M. Accurate Prediction of the Stability and Activity Effects of Site-Directed Mutagenesis on a Protein Core. *Nature* **1991**, *352* (6334), 448−451.

(9) Topham, C. M.; Srinivasan, N.; Blundell, T. L. Prediction of the Stability of Protein Mutants Based on Structural Environment-Dependent Amino Acid Substitution and Propensity Tables. *Protein Eng., Des. Sel.* **1997**, *10* (1), 7−21.

(10) Kulshreshtha, S.; Chaudhary, V.; Goswami, G. K.; Mathur, N. Computational Approaches for Predicting Mutant Protein Stability. *J. Comput.-Aided Mol. Des.* **2016**, *30* (5), 401−412.

(11) Montanucci, L.; Savojardo, C.; Martelli, P. L.; Casadio, R.; Fariselli, P. On the Biases in Predictions of Protein Stability Changes upon Variations: The INPS Test Case. *Bioinformatics* **2019**, *35* (14), 2525−2527.

(12) Pucci, F.; Bernaerts, K. V.; Kwasigroch, J. M.; Rooman, M. Quantification of Biases in Predictions of Protein Stability Changes upon Mutations. *Bioinformatics* **2018**, *34* (21), 3659−3665.

(13) Potapov, V.; Cohen, M.; Schreiber, G. Assessing Computational Methods for Predicting Protein Stability upon Mutation: Good on Average but Not in the Details. *Protein Eng., Des. Sel.* **2009**, *22* (9), 553−560.

(14) Worth, C. L.; Preissner, R.; Blundell, T. L. SDM–a Server for Predicting Effects of Mutations on Protein Stability and Malfunction. *Nucleic Acids Res.* **2011**, *39*, W215−W222.

(15) Gilis, D.; Rooman, M. PoPMuSiC, an Algorithm for Predicting Protein Mutant Stability Changes: Application to Prion Proteins. *Protein Eng., Des. Sel.* **2000**, *13* (12), 849−856.

(16) Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M. Fast and Accurate Predictions of Protein Stability Changes upon Mutations Using Statistical Potentials and Neural Networks: PoPMuSiC-2.0. *Bioinformatics* **2009**, *25* (19), 2537−2543.

(17) Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* **2005**, *33*, W306−W310.

(18) Pires, D. E. V.; Ascher, D. B.; Blundell, T. L. MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* **2014**, *30* (3), 335−342.

(19) Dill, K. A.; Shortle, D. Denatured States of Proteins. *Annu. Rev. Biochem.* **1991**, *60*, 795−825.

(20) Robertson, A. D.; Murphy, K. P. Protein Structure and the Energetics of Protein Stability. *Chem. Rev.* **1997**, *97* (5), 1251−1268.

(21) Lazaridis, T.; Karplus, M. *Biophys. Chem.* **2002**, *100*, 367−395.

(22) Ford, M. C.; Babaoglu, K. Examining the Feasibility of Using Free Energy Perturbation (FEP+) in Predicting Protein Stability. *J. Chem. Inf. Model.* **2017**, *57* (6), 1276−1285.

(23) 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.

(24) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability. *Proteins: Struct., Funct., Genet.* **2011**, *79* (3), 830−838.

(25) Benedix, A.; Becker, C. M.; de Groot, B. L.; Caflisch, A.; Böckmann, R. A. Predicting Free Energy Changes Using Structural Ensembles. *Nat. Methods* **2009**, *6*, 3−4.

(26) Pokala, N.; Handel, T. M. Energy Functions for Protein Design: Adjustment with Protein-Protein Complex Affinities, Models for the Unfolded State, and Negative Design of Solubility and Specificity. *J. Mol. Biol.* **2005**, *347* (1), 203−227.

(27) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33*, W382−W388.

(28) Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; Rooman, M. PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality. *BMC Bioinf.* **2011**, *12* (1), 151.

(29) Masso, M.; Vaisman, I. I. Accurate Prediction of Stability Changes in Protein Mutants by Combining Machine Learning with Structure Based Computational Mutagenesis. *Bioinformatics* **2008**, *24* (18), 2002−2009.

(30) Bava, K. A.; Gromiha, M. M.; Uedaira, H.; Kitajima, K.; Sarai, A. ProTherm, Version 4.0: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.* **2004**, *32*, D120−D121.

(31) Sasidharan Nair, P.; Vihinen, M. VariBench: A Benchmark Database for Variations. *Hum. Mutat.* **2013**, *34* (1), 42−49.
(32) Khan, S.; Vihinen, M. Performance of Protein Stability Predictors. *Hum. Mutat.* **2010**, *31* (6), 675−684.
(33) Christensen, N. J.; Kepp, K. P. Stability Mechanisms of Laccase Isoforms Using a Modified FoldX Protocol Applicable to Widely Different Proteins. *J. Chem. Theory Comput.* **2013**, *9* (7), 3210−3223.
(34) Christensen, N. J.; Kepp, K. P. Accurate Stabilities of Laccase Mutants Predicted with a Modified FoldX Protocol. *J. Chem. Inf. Model.* **2012**, *52* (11), 3028−3042.
(35) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J. Mol. Biol.* **2002**, *320* (2), 369−387.
(36) Thiltgen, G.; Goldstein, R. a. Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. *PLoS One* **2012**, *7* (10), e46084.
(37) Fariselli, P.; Martelli, P. L.; Savojardo, C.; Casadio, R. INPS: Predicting the Impact of Non-Synonymous Variations on Protein Stability from Sequence. *Bioinformatics* **2015**, *31* (17), 2816−2821.
(38) Pucci, F.; Bernaerts, K.; Teheux, F.; Gilis, D.; Rooman, M. Symmetry Principles in Optimization Problems : An Application to Protein Stability Prediction. *IFAC-PapersOnLine* **2015**, *48* (1), 458−463.
(39) Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D. S. The Stability Effects of Protein Mutations Appear to Be Universally Distributed. *J. Mol. Biol.* **2007**, *369* (5), 1318−1332.
(40) Melo, F.; Marti-Renom, M. A. Accuracy of Sequence Alignment and Fold Assessment Using Reduced Amino Acid Alphabets. *Proteins: Struct., Funct., Genet.* **2006**, *63* (4), 986−995.
(41) Usmanova, D. R.; Bogatyreva, N. S.; Bernad, J. A.; Eremina, A. A.; Gorshkova, A. A.; Kanevskiy, G. M.; Lonishin, L. R.; Meister, A. V.; Yakupova, A. G.; Kondrashov, F. A.; Ivankov, D. N. Self-Consistency Test Reveals Systematic Bias in Programs for Prediction Change of Stability upon Mutation. *Bioinformatics* **2018**, *34* (21), 3653−3658.
(42) Yang, Y.; Urolagin, S.; Niroula, A.; Ding, X.; Shen, B.; Vihinen, M. Pon-Tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. *Int. J. Mol. Sci.* **2018**, *19* (4), 1009.
(43) Kepp, K. P. Towards a "Golden Standard" for Computing Globin Stability: Stability and Structure Sensitivity of Myoglobin Mutants. *Biochim. Biophys. Acta, Proteins Proteomics* **2015**, *1854*, 1239−1248.
(44) Kepp, K. P. Computing Stability Effects of Mutations in Human Superoxide Dismutase 1. *J. Phys. Chem. B* **2014**, *118* (7), 1799−1812.
(45) Nisthal, A.; Wang, C. Y.; Ary, M. L.; Mayo, S. L. Protein Stability Engineering Insights Revealed by Domain-Wide Comprehensive Mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (33), 16367−16377.
(46) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Wessig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(47) Tickle, I. J. Statistical Quality Indicators for Electron-Density Maps. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 454−467.
(48) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX : A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (2), 213−221.
(49) Capriotti, E.; Fariselli, P.; Rossi, I.; Casadio, R. A Three-State Prediction of Single Point Mutations on Protein Stability Changes. *BMC Bioinf.* **2008**, *9*, S6.
(50) Laimer, J.; Hofer, H.; Fritz, M.; Wegenkittl, S.; Lackner, P. MAESTRO - Multi Agent Stability Prediction upon Point Mutations. *BMC Bioinf.* **2015**, *16* (1), 116.
(51) Gromiha, M. M. Prediction of Protein Stability upon Point Mutations. *Biochem. Soc. Trans.* **2007**, *35* (6), 1569−1573.
(52) Masso, M.; Vaisman, I. I. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. *Adv. Bioinf.* **2014**, *2014*, 278385.
(53) Etchebest, C.; Benros, C.; Bornot, A.; Camproux, A. C.; De Brevern, A. G. A Reduced Amino Acid Alphabet for Understanding and Designing Protein Adaptation to Mutation. *Eur. Biophys. J.* **2007**, *36* (8), 1059−1069.
(54) Hubbard, S.; Thornton, J. *NACCESS*; Department of Biochemistry Molecular Biology, University College London: London, 1993.
(55) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55* (3), 379−400.
(56) Knudsen, M.; Wiuf, C. The CATH Database. *Hum. Genomics* **2010**, *4* (3), 207−212.
(57) Grimm, D. G.; Azencott, C. A.; Aicheler, F.; Gieraths, U.; Macarthur, D. G.; Samocha, K. E.; Cooper, D. N.; Stenson, P. D.; Daly, M. J.; Smoller, J. W.; Duncan, L. E.; Borgwardt, K. M. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Hum. Mutat.* **2015**, *36* (5), 513−523.
(58) Vihinen, M. How to Evaluate Performance of Prediction Methods? Measures and Their Interpretation in Variation Effect Analysis. *BMC Genomics* **2012**, *13*, S2.
(59) Montanucci, L.; Martelli, P. L.; Ben-Tal, N.; Fariselli, P. A Natural Upper Bound to the Accuracy of Predicting Protein Stability Changes upon Mutations. *Bioinformatics* **2019**, *35* (9), 1513−1517.
(60) Ascher, D. B.; Jubb, H. C.; Pires, D. E. V; Ochi, T.; Higueruelo, A.; Blundell, T. L. Protein-Protein Interactions: Structures and Druggability. In *Multifaceted Roles of Crystallography in Modern Drug Discovery*; Springer, 2015; pp 141−163.

https://dx.doi.org/10.1021/acs.jcim.0c00591
J. Chem. Inf. Model. 2020, 60, 4772−4784