JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

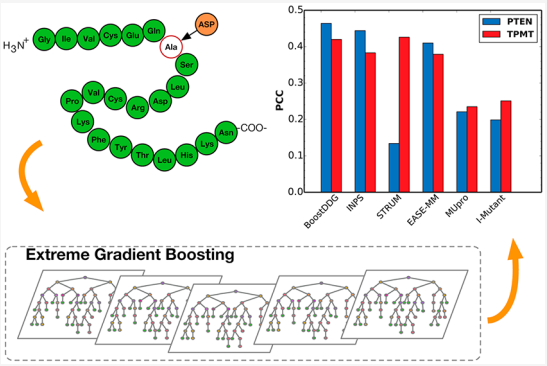
pubs.acs.org/jcim Article

Accurately Predicting Mutation-Caused Stability Changes from Protein Sequences Using Extreme Gradient Boosting

Xuan Lv, Jianwen Chen, Yutong Lu, Zhiguang Chen, Nong Xiao,* and Yuedong Yang*

Cite This: https://dx.doi.org/10.1021/acs.jcim.0c00064 Read Online

ABSTRACT: Accurately predicting the impact of point mutation on protein stability has crucial roles in protein design and engineering. In this study, we proposed a novel method (BoostDDG) to predict stability changes upon point mutations from protein sequences based on the extreme gradient boosting. We extracted features comprehensively from evolutional information and predicted structures and performed feature selection by a strategy of sequential forward selection. The features and parameters were optimized by homologue-based cross-validation to avoid overfitting. Finally, we found that 14 features from six groups led to the highest Pearson correlation coefficient (PCC) of 0.535, which is consistent with the 0.540 on an independent test. Our method was indicated to consistently outperform other sequence-based methods on three precompiled test sets, and 7363 variants on two proteins (PTEN and TPMT). These results highlighted that BoostDDG is a powerful tool for predicting stability changes upon point mutations from protein sequences.



1. INTRODUCTION

Nonsynonymous single nucleotide polymorphisms (SNPs) play fundamental roles in protein evolution mainly through inducing changes in amino acid sequences.[1] Previous studies[2,3] suggest that nonsynonymous SNP mutations can be disease-associated and detrimental to human health by affecting structural stabilities of proteins. Even a single residue substitution, namely, point mutation, may have serious effects on a protein's stability and consequently result in changes in the protein's structure and function. Therefore, accurate prediction of mutation-caused protein stability changes ($\Delta\Delta G$) is essential for guiding directed protein evolution and for understanding the relationship between protein variants and diseases.

Over the years, efforts were devoted to developing methods to evaluate the stability changes upon point mutations. Traditional mutagenesis studies can accurately quantify the thermodynamic effects of mutations via biological experiments. However, these experimental efforts usually take considerable time and cost, especially for large-scale mutations. With the rapid accumulation of experimentally measured data for both proteins and mutants in the ProTherm database,[4] computational methods were greatly improved in recent years. Predictive stability effects of mutations can be obtained from a protein's sequence or tertiary structure information, and the majority of current methods are structure-based, including several energy-based approaches and machine learning methods. Physical energy functions compute $\Delta\Delta G$ by simulating atomic force fields in a protein structure[5] and may not be applied to large-scale data sets due to high computational requirements. Methods using statistical energy functions calculate the conditional probabilities of certain residues or atoms in different structural environments. For instance, SDM[6] characterizes free energy changes with the frequency of environment-specific amino acid substitutions, and CUPSAT[7] utilizes the knowledge-based potentials in a given structure. Empirical potential approaches derive a linear combination of different energy functions, whose weights are fitted to experimental data.[8,9] Machine learning methods can learn complex nonlinear models from sequence or structural features, and various algorithms have been used to generate predictions, such as support vector machine,[10−13] decision tree,[14] artificial Neural network,[15,16] and collaborative filtering model.[8]

One limitation of structure-based approaches is that the tertiary structure of the target protein is often unknown; in fact, less than 0.2% of the proteins in UniProt can retrieve three-dimensional structures in the PDB library.[17] With the development of high-throughput sequencing technologies, the gap between the number of protein sequences and resolved

Received: January 21, 2020
Published: March 23, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

structures is being further widened. In the past, it was generally accepted that the prediction accuracy obtained using sequence alone is lower than using tertiary structure information.[10] Notably, Khan and Vihinen[18] made a systematic analysis of 11 kinds of software, indicating that several sequence-based methods[10,19,20] achieved predictive accuracy comparable to the remaining structural methods. Recent studies have reported that sequence-based predictors[11,12] achieved similar or better performance than structure-based ones. For the above reasons, we devoted our interest to the sequence-based methods in this work.

Among the machine learning methods used in practice, gradient tree boosting has been shown to give state-of-the-art results in many classification or regression tasks, while rarely being used to predict stability changes upon point mutations. A relatively new ensemble method called extreme gradient boosting (XGBoost) was proposed and shined in many competitions such as KDDCup and Kaggle,[21] outperforming many other known machine learning methods or deep learning techniques both in accuracy and computational efficiency. Recently published studies have also confirmed the successful applications of XGBoost in computational biology,[22−24] including our study to predict genome-wide RNA secondary structure profiles.[25]

In this study, we developed a new sequence-based predictor (BoostDDG) for protein stability change upon point mutation using the XGBoost machine learning model. We collected and curated a reliable training data set from recently published studies and performed feature selection with parallel model optimization by homologue-based cross-validation. Several studies have pointed out that the predictions of most machine learning methods show a bias toward destabilizing mutations since their training sets are dominated by negative $\Delta\Delta G$ values.[26−29] To solve this problem, we trained our model using a balanced learning set by introducing reverse mutations with the opposite sign of corresponding $\Delta\Delta G$ values. The selected features in the final input vector were also evaluated through comprehensive ablation experiments. By testing on different data sets, our method was shown to consistently achieve better accuracies in comparison with other state-of-the-art methods. The Web server and all the collected data of BoostDDG are freely available at http://biomed.nscc-gz.cn/server/BoostDDG/.

2. MATERIALS AND METHODS

2.1. Data Sets. We have collected the training and independent test sets from the recent two studies by Folkman et al.[12] and Cao et al.[16] Folkman's study used 1676 mutations for training, and two independent test sets (S236 and S543 containing 236 and 543 mutations, respectively). Cao's study used S5444 for training and S276 for testing. We merged S1676 and S5444 data sets and removed mutations whose occurring proteins have a sequence identity of ≥25% with any protein in three independent test sets and finally extracted 2815 mutations in 150 proteins for training and cross-validation. Furthermore, we employed the stability changes on coded protein by PTEN and TPMT genes from the Critical Assessment of Genome Interpretation (CAGI) challenge. After removing the variants with unknown amino acid "X", we independently tested a total of 7363 missense mutations for the PTEN (3736 mutations) and TPMT (3627 mutations) proteins. In this challenge, a stability score was deployed to describe the steady-state abundance of missense mutations with 0 signifying unstable, 1 signifying wild-type stability, and >1 signifying better stability than the wild-type. This data set was named the CAGI data set. Table S1 summarizes the details of these data sets.

2.2. Candidate Features. In order to train our model, we extracted several classes of candidate features including evolutionary information, predicted structural features, and physicochemical properties for mutant and wild-type amino acids at target residues in the protein sequences.

Evolutionary Conservation Features. In this study, we employed the position-specific scoring matrix (PSSM) generated from PSI-BLAST2.7.1[30] with an E-value threshold of 0.001 in three iterations to estimate evolutionary conservation of the mutation site. We built three features from the PSSM profile: ΔF_t, ΔP_t, and PE_t. The ΔF_t is computed as

$$\Delta F_t = (F_{t,mt} - F_{t,wt})/100$$

where $F_{t,mt}$ and $F_{t,wt}$ are the occurrence frequency of mutant (mt) and wild-type (wt) amino acids at the mutation site t extracted from the PSSM (the second 20 columns), respectively. In the sites with zeros for all amino acids, the frequencies are extracted from the blosum62 matrix to the wild-type. The ΔP_t is computed as

$$\Delta P_t = P_{t,mt} - P_{t,wt}$$

where $P_{t,mt}$ and $P_{t,wt}$ are extracted from the log-likelihood ratio matrix (the first 20 columns) representing the probability of mutant (mt) and wild-type (wt) amino acids at the mutation site t. The conservation entropy at mutation site t is computed as

$$PE_t = \sum_{j=1}^{20} P_{t,j} \cdot F_{t,j}$$

Predicted Structural Features. Since our method is sequence based, we derived structural features according to structural properties predicted from protein sequences, including the relative solvent accessible surface area (rASA), three probabilities (helix, sheet, coil) of the secondary structure (SS), and the disorder probability. The first two classes of features were obtained from SPIDER 3.0,[31] and the disorder probability was predicted by the SPOT-disorder,[32] both with the default parameters.

Physicochemical Properties. We employed a group of seven physicochemical parameters provided by Meiler et al.[33] including hydrophobicity, isoelectric point, polarizability, steric parameter, volume, helix tendency, and sheet tendency, namely, AAPh7. we adopted ΔAAPh7 = AAPh7$_{mt}$ − AAPh7$_{wt}$ to calculate the difference between the mutant and wild-type amino acids at the mutation site. We also introduced another feature called AAscore inspired by Liu et al.[34] to measure the distinction between wild-type (wt) and mutant (mt) amino acids as

$$\theta(R_{mt}, R_{wt}) = \frac{1}{7}\sum_{j}^{7} (I_j(R_{mt}) - I_j(R_{wt}))^2$$

where $\theta(R_{mt}, R_{wt})$ represents the scores calculated using seven physicochemical properties of amino acids presented in the AAindex database (see Table S2). $I_j(R_{mt})$ and $I_j(R_{wt})$ are the normalized values of mutant and wild-type amino acids in property j, which can be calculated by

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$I_j(a_i) = \frac{\tilde{I}_j(a_i) - \frac{1}{20}\sum_{k=1}^{20}\tilde{I}_k(a_k)}{\sqrt{\frac{1}{20}\sum_{k=1}^{20}(\tilde{I}_k(a_k) - \frac{1}{20}\sum_{k=1}^{20}\tilde{I}_k(a_k))^2}}$$

where $\tilde{I}_j(a_i)$ symbolizes the original physicochemical parameters of amino acid a_i in the property j, and $a_k(k = 1,2,3,4\cdots,20)$ stands for the 20 common amino acids.

2.3. XGBoost Technique. BoostDDG was trained using XGBoost, a novel tree-boosting algorithm to develop an ensemble strong learner from multiple weak learners in an additive way; the tree model outputs a weighted sum of the prediction of each learner by readjusting weights of mislabeled samples in each boosting step. The training procedure can be summarized as follows:

(i) For each descriptor, sort the feature values and scan the best splitting point.

(ii) Select the best splitting point that optimizes the objective function.

(iii) Repeat the splitting procedure (the above two steps) until the predetermined maximum tree depth is achieved.

(iv) Assign prediction scores to leaves and prune negative nodes in a bottom-up order.

(v) Repeat the above steps until the predetermined number of trees is reached.

XGBoost has advantages in preventing overfitting by incorporating regularization terms as well as shrinkage and descriptor subsampling techniques. In addition, distributed and parallel computing enables faster model training.

2.4. Training Procedure. To find out the optimal set of features and parameters for our model, we coupled feature selection with a parallel grid search procedure. Evaluation of features and parameters in each step was performed using the 10-fold cross-validation.

Feature Selection. To ensure the effectiveness and non-redundancy of all the input features in our model, we selected the best subset of all candidate features using a bidirectional greedy selection algorithm. Both Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) algorithms were adopted to avoid local optimal results to the greatest extent. SFS (Algorithm 1) begins with an empty set of features F_{sel0} and iteratively searches the best feature f from the remaining feature set F_{left} and adds f to F_{sel0} for a lower root-mean-square error (RMSE). The procedure repeats until F_{left} is empty or RMSE no longer decreases. On the contrary, SBS (Algorithm S1) starts with the full set of features F_{sel0} and removes the worst feature f to yield a lower RMSE. The reverse search process stops when no feature could be discarded. We compared the results generated by the two searching algorithms and selected the one with a higher Pearson's correlation coefficient (PCC) as the final input feature set for our model.

Parallel Grid Search. As the performances are affected by hyperparameters, we have scanned all combinations of parameters by a grid search procedure during the SFS/SBS searches. However, it is considerably time consuming to perform a gird search with the 10-fold cross-validation for each kind of feature set. To speed up the computation, we implemented an MPI-based approach to parallelize the search process through distributing all the parameter combinations among 288 CPU cores on the Tianhe-2 supercomputer, leading to over 200 times faster than the computation on a single CPU core. Table S3 details the tested hyperparameters set up in the parallel grid search process and the final optimal values.

Algorithm 1 Sequential Forward Feature Selection.

Input: The set of all candidate features, F_{list};
Output: Features that have been selected, F_{sel0};

```
function SFS_SELECT(F_list)
    F_left ← F_list, F_sel0 ← ∅, loss0 ← 10
    while F_left is not empty do
        loss_reduction ← [ ]
        for f ∈ F_left do
            F_sel ← F_sel0 ∪ f
            bestParams, loss ← GridSearch(F_sel)
            loss_reduction.insert(loss0 − loss)
        end for
        if max(loss_reduction) < 0 then
            break
        else
            F_sel0 ← F_sel0 ∪ F_left[argmax loss_reduction]
            F_left ← F_left − F_left[argmax loss_reduction]
            loss0 ← loss0 − max(loss_reduction)
        end if
    end while
    return F_sel0
end function
```

Balanced Data Set with Reverse Mutations. Several previous studies have discussed the antisymmetric property of the free energy changes between wild-type and mutant proteins.[26,28,29,35] The $\Delta\Delta G$ of direct mutation should be equal to the $-\Delta\Delta G$ of reverse variation according to thermodynamic reversibility of mutations. However, most of the predictors ignored the self-consistency requirement due to the bias toward destabilizing mutations in their training sets. Here, we trained our model with a balanced data set by introducing reverse mutations. Each reverse data point has the negative value of experimentally measured $\Delta\Delta G$, and the corresponding features were calculated from the mutated protein sequence.

Homologue-Based 10-Fold Cross-Validation. In general cross-validations, the training set can be separated based on mutations (completely random), residues (mutations on the same residue put together), proteins (mutations on the same protein put together), or homologues (mutations on homologous proteins put together). Since mutation, residue, and protein-based cross-validations may share protein information with the training samples, these schemes are likely to cause an overestimate of the performance, as also indicated in the previous study.[12] Relatively, *homologue-based* evaluation is the strictest assessment since all mutations occurring on the same protein or homologous proteins are always grouped in the same fold to avoid overfitting for a specific group of proteins. We employed the *homologue-based* scheme on the S2815 data set to conduct our 10-fold cross-validation combined with feature selection and parameter optimization. We strictly separated the proteins from 111 homologous clusters in S2815 into 10 folds to ensure that no two folds share proteins with sequence similarity $\geq 25\%$. The S2815 data set was augmented by introducing hypothetical reverse variations, and each pair of mutations (direct and reverse) was included in the same fold to avoid overfitting. Moreover, we repeated each cross-validation procedure 10 times with 10 regenerated 10-folds and averaged the results to gain an unbiased evaluation.

Considering that our method is evolutionary based and it is time consuming to generate PSSM profiles from multiple sequence alignment, we also developed another method DeepDDG-single, which is single-sequence based and more

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

computationally efficient by removing evolutionary conservation information and replacing the predicted secondary structure features from SPIDER 3.0[31] with those generated by SPIDER3-single[36] that also made predictions from only sequences without using evolutionary information.

3. RESULTS

3.1. Feature Selection. We applied a bidirectional greedy feature selection algorithm on the S2815 data set with 10-fold cross-validation. The final combination of predictive features for BoostDDG was composed of the following features: ΔAAPh7 (the changes of hydrophobicity, volume, polarizability, isoelectric, helix tendency, sheet tendency, and steric), rASA, ΔF_t, AAscore, SS (helix probability, sheet probability, and coil probability), and disorder probability. All these features were real-valued and constituted a 14-dimensional input vector. Table 1 (left) displays the performances of XGBoost-based models using additive feature groups in each iteration of the SFS algorithm. The SFS selected six feature groups containing totally 14 features. The trained model achieved a PCC of 0.535, slightly higher than the 0.530 by using all features. The robustness of our trained model with selected features further proved essentially same results on the independent test of S543. The PCC of 0.540 is consistently higher than 0.531 by using all features. The results suggest that our feature selection procedure is helpful to improve prediction performance by removing redundancy and irrelevant features.

Table 1. PCC between Experimental ΔΔG and ΔΔG Predicted by Using SFS-Selected Feature Groups or by Removing Each Feature Group from the Final Model

Feature groups[a]	CV[c]	Ind. test[c]	Feature groups[b]	CV[c]	Ind. test[c]
ΔAAPh7	0.385	0.332	BoostDDG	0.535	0.540
+rASA	0.461	0.440	−ΔAAPh7	0.325	0.451
+ΔF_t	0.510	0.508	−rASA	0.502	0.511
+SS	0.524	0.526	−AAscore	0.529	0.537
+AAscore	0.531	0.538	−ΔF_t	0.478	0.471
+disorder	**0.535**	**0.540**	−disorder	0.523	0.544
all features	0.530	0.531	−SS	0.520	0.517
SVM[d]	0.515	0.505	SVM[e]	0.519	0.511

[a]Additive feature groups selected in SFS algorithm. [b]Removed feature group from the final model. [c]CV, 10-fold cross-validation; Ind. test, independent test on S543. [d]Performance of SVM model using all features. [e]Performance of SVM model using selected features.

Table 2. Comparison of BoostDDG and Related Methods on Three Benchmark Data Sets[a]

	S236 test set		S543 test set		S276 test set	
Methods	PCC	RMSE	PCC	RMSE	PCC	MAE
I Mutant[b]	0.443	1.18	0.323	1.37	0.391	1.08
I Mutant[c]	0.521	1.07	0.356	1.34	0.453	0.91
MUpro	0.362	1.20	0.332	1.32	0.190	1.06
Rosetta	0.270	1.88	0.380	3.58	0.339	5.25
FoldX	0.277	1.70	0.405	1.87	0.300	2.13
DFIRE	0.535	1.18	0.450	1.44	0.230	1.25
PoPMuSiC	0.570	1.05	0.533	1.21	0.443	0.91
EASE-MM	0.589	1.03	0.530	1.22	0.402	0.91
STRUM	–	–	–	–	0.447	0.88
mCSM	–	–	–	–	0.467	0.90
INPS[d]	–	–	–	–	0.474	0.89
SDM	–	–	–	–	0.483	1.02
BoostDDG-single	0.511	1.17	0.431	1.39	0.355	1.02
BoostDDG	**0.600**	**1.01**	**0.540**	**1.21**	**0.514**	**0.78**

[a]Predictions of S236 and S543 were calculated from the data provided in Folkman's study.[12] Predictions of S276 were collected from Cao's study.[16]. [b]Sequence-based I Mutant. [c]Structure-based I Mutant. [d]Computed by ourselves.

Table 3. Independent Test Results for PTEN and TPMT Data Sets in Terms of PCC and RMSE between Expected Stability Scores and Linear-Fitted Predictions

	PTEN		TPMT		TPMT and PTEN	
Methods	PCC	RMSE	PCC	RMSE	PCC	RMSE
MUpro1.1	0.221	0.320	0.235	0.349	0.229	0.336
I Mutant2.0	0.199	0.322	0.251	0.348	0.228	0.336
EASE-MM	0.410	0.299	0.379	0.332	0.391	0.317
STRUM	0.134	0.325	0.426	0.325	0.320	0.327
INPS	0.444	0.294	0.383	0.332	0.418	0.313
BoostDDG	**0.464**	**0.290**	**0.420**	**0.323**	**0.440**	**0.310**

For comparison, we also trained SVM regression models implemented in the scikit-learn[37] python library using all features and selected features based on the S2815 data set. Both SVM models achieved PCCs lower than those generated by BoostDDG whether in cross-validation (0.515 and 0.519) or the independent test (0.505 and 0.511), indicating that XGBoost is better for this task. We also tested the SBS algorithm to remove redundant features from the whole feature set, but no feature was filtered out.

3.2. Assessment of Feature Importance. To get further insights into the factors that contribute to the mutation-induced protein stability changes, we employed ablation experiments by removing each feature group from the final model and re-evaluating the performance based on the cross-validation and test set (Table 1, right). Both the PCCs of cross-validation and the independent test decrease the most when removing ΔAAPh7. This is understandable since ΔAAPh7 is a seven-dimensional feature vector that is larger than other feature groups. Among the remaining features, the exclusion of ΔF_t caused the largest drop of PCC, confirming that evolutionary conservation is an essential factor in the prediction of stability changes. The predicted solvent accessibility has a significant effect on our model as well, which is consistent with the results by Dehghanpoor et al.[38] and Cao et al.,[16] likely because residues with smaller accessible surface areas (buried) are more vulnerable to destabilizing substitutions. Different from the positive contribution of disorder probability in the cross-validation, the removal of disorder probability led to a slight increase in PCC from 0.540 to 0.544 in the test set, likely due to the difference of data sets, so we still kept the disorder probability in our final model.

3.3. Method Comparison on Three Benchmark Data Sets. We adopted three different benchmark test sets, S236 together with S543 from Folkman et al.[12] and S276 from Cao et al.[16] to verify if BoostDDG generalizes well. None of the three test sets was used for training or optimizing our method. For comparing with other predictors, we only tested the experimentally measured ΔΔG values (direct mutations). We compared our method with 11 state-of-the-art methods, including four sequence-based methods (I Mutant,[19] MUpro,[10] INPS,[11] and EASE-MM[12]) and seven structure-

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

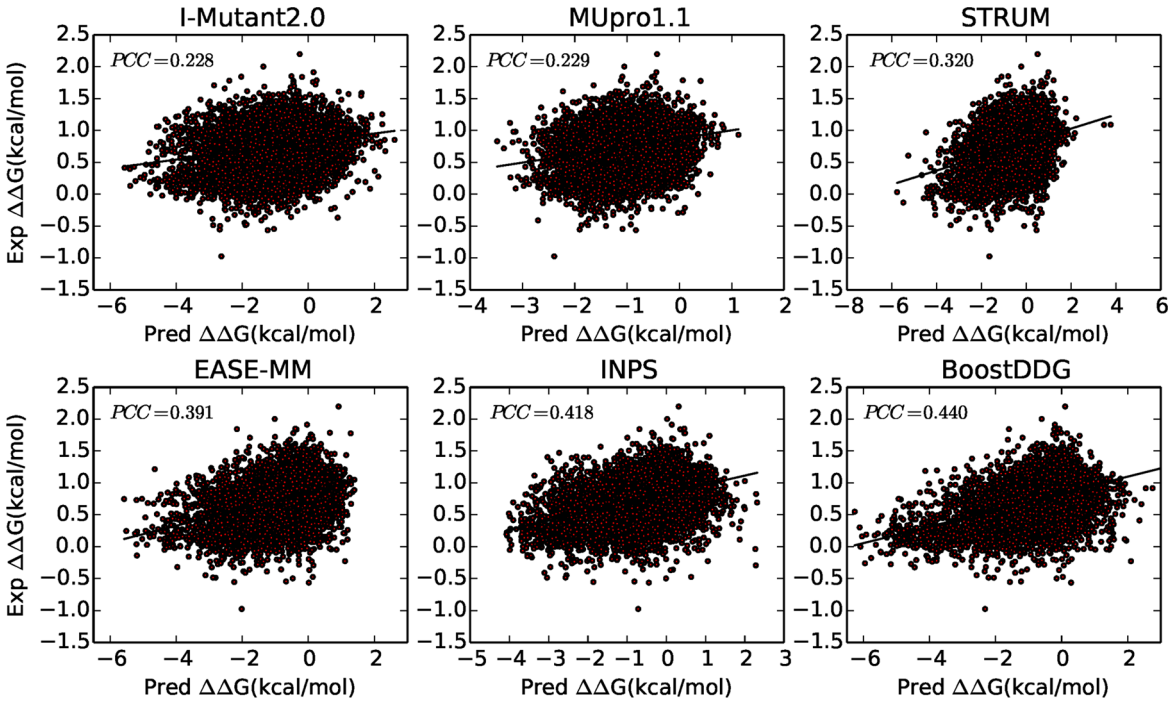


Figure 1. Comparison between expected stability scores and scores predicted with six sequence-based methods for the CAGI data set.

based energy functions (Rosetta,[39] FoldX,[40] DFIRE,[41] PoPMuSiC,[42] STRUM,[17] mCSM,[43] SDM[6]). The structure-based version of I Mutant was also included. We tested a few methods (STRUM, mCSM, INPS, and SDM) only on the S276 set since their training sets contain many proteins sequentially similar to S543 and S236 sets.

The performance was assessed in terms of PCC and RMSE. As S276 does not have reported RMSE in the previous study,[16] we used the reported mean absolute error (MAE) instead on this data set for a consistent comparison. Table 2 outlines the results obtained with each of the above-mentioned methods on three test sets. Remarkably, BoostDDG produced the highest PCC and lowest RMSE (or MAE) for all three data sets, outperforming all sequence-based and structure-based methods tested in previous studies. BoostDDG achieved a higher PCC in the S236 but a lower PCC in the S276 than the one in S543. We noticed that the PCC in the S236 is even higher than the PCC (= 0.54) in our training set S2815. This is likely because S2815 and S276 data sets contain $\Delta\Delta G$ values obtained under various experimental conditions, while the S236 and S543 data sets only contain the $\Delta\Delta G$ measurements under the standard pH and temperature. Thus, S236 is less challenging as also seen by a typically higher PCC in S236 and a lower PCC in S276 by other methods as well. Figures S1−S3 show the predicted versus experimentally measured $\Delta\Delta G$ on three test sets, respectively.

We also tested our single-sequence-based method BoostDDG-single on these data sets. Despite the gap with evolutionary-based methods, it still surpasses some popular sequence-based predictors (I Mutant[19] and MUpro[10]). In fact, the availability of evolutionary information can be limited since most proteins have few known homologous sequences.[36] Additionally, the generation of evolutionary sequence profile is time consuming even for a short sequence. Therefore, the single-sequence-based method is able to constitute a trade-off between accuracy and computational efficiency.

3.4. Independent Testing of CAGI Data Set. We further tested our method on two proteins, PTEN (phosphatase and TEnsin homologue) and TPMT (thiopurine S-methyl transferase), from the CAGI challenge that was provided by Fowler and Fields.[44] Both proteins have pairwise identities of less than 25% with proteins in our training set. We used the full-length sequence of the two proteins to compare BoostDDG with five other leading sequence-based approaches (MUpro1.1,[10] I Mutant2.0,[19] EASE-MM,[12] STRUM,[17] and INPS[11]). STRUM is a structure-based method but can be applied to protein sequences with predicted 3D structures. Pejaver et al. analyzed the performances of several computational predictors on the PTEN and TPMT protein variants and found that predictors ranked different by metrics.[45] For consistent comparisons, we calculated the PCC and RMSE between stability scores and the predictions of the six methods.

Table 3 outlines the PCC between stability scores and the prediction data of the six methods. BoostDDG produced a more balanced and accurate prediction for the two proteins with a PCC of 0.420−0.464. The second best method is INPS, achieving a PCC of 0.44 in total, which is significantly lower than our method according to the Fisher test (P = 0.05). Although STRUM yielded a comparable correlation (0.426) for the TPMT data, the accuracy for PTEN data (0.134) is much lower than other methods probably because the accuracy of the simulated 3D structure decreases when the sequence length is too long (>400). The independent test further demonstrates that our method is able to give an improved prediction when experimental structures are not available. Figure 1 illustrates the expected stability score versus predicted $\Delta\Delta G$.

4. DISCUSSION

In this work, we have developed a new sequence-based method, BoostDDG, to predict the stability effects of point mutations based on XGBoost. We carefully selected features

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

from initial input with a bidirectional greedy algorithm to generate the best nonredundant feature set. To select features in a fair and efficient way, we performed parallel model optimization in each step of feature selection with the strictest homologue-based 10-fold cross-validation, in which no two folds shared homologous proteins.

We employed three different independent test sets from Folkman et al.[12] and Cao et al.[16] and made a comparison with other methods tested in their works. The results showed that BoostDDG produced consistently the best predictions measured in both PCC and RMSE. Furthermore, we independently tested our method with the PTEN and TPMT data sets from the CAGI challenge and found that BoostDDG achieved a balanced correlation of 0.42−0.46 that is superior to those generated by other sequence-based methods. This can be largely attributed to our strict homologue-based validation scheme and the robust regularization performance of XGBoost. Since a majority of the previous methods were trained and evaluated using a residue-based scheme or a protein-based scheme with a high sequence identity cutoff, their performance may be overestimated due to the correlation between available data sets. XGBoost is a regularized scalable extension of traditional tree-boosting algorithms which is much less susceptible to overfitting by tuning the corresponding parameters. Following the previous studies,[11,16] we included the thermodynamic reversibility of variations to enhance the symmetry of our model. We found the inclusion slightly increased the PCC on the CV (0.530 to 0.535) relative to the one without the inclusion and led to small and divergent changes on four test sets (0.64−0.60 on S236, 0.538−0.54 on S543, 0.49−0.51 on S276, and 0.45−0.44 on the CAGI data sets). Although far from extensive, the independent test performance highlights that as a sequence-based method BoostDDG is a promising alternative to existing popular predictors when experimental structures are not available.

Typically, the stability change $\Delta\Delta G$ is sensitive to experimental conditions such as pH and temperature. However, these environmental parameters were not encoded into input features considering that the number of measurements of each mutation at different environmental conditions is too small to provide generalized predictions. Despite the favorable independent test performance that our method has achieved, the correlation values became lower on the larger test sets, indicating that further improvement should be considered. Given that our method only focuses on the features extracted from the mutation site, we aim to design more useful features to capture the interaction changes between neighboring residues caused by mutations in our future work.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00064.

Table S1: Summary of data sets used. Table S2: Seven physicochemical properties of amino acids used for AAscore. Table S3: Hyperparameters set up in the parallel grid search process and the optimal values used for the final model. Algorithm S1: Sequential Backward Selection algorithm. Figure S1: Experimentally measured $\Delta\Delta G$ from the S236 data set versus $\Delta\Delta G$ predicted with the four sequence-based methods (BoostDDG, I Mutant2.0, MUpro1.1, and EASE-MM) and four structure-based methods (I Mutant2.0, FoldX, DFIRE, and PoPMuSiC2.1). Figure S2: Experimentally measured $\Delta\Delta G$ from the S543 data set versus $\Delta\Delta G$ predicted with the four sequence-based methods (BoostDDG, I Mutant2.0, MUpro1.1, and EASE-MM) and four structure-based methods (I Mutant2.0, FoldX, DFIRE, and PoPMuSiC2.1). Figure S3: Experimentally measured $\Delta\Delta G$ from the S276 data set versus $\Delta\Delta G$ obtained from BoostDDG. (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Yuedong Yang − *School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510275, China; Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Ministry of Education, Guangzhou, Guangdong 510275, China;* orcid.org/0000-0002-6782-2813; Email: yangyd25@mail.sysu.edu.cn

Nong Xiao − *State Key Laboratory of High-Performance Computing, School of Computer Science, National University of Defense Technology, Changsha, Hunan 410073, China; School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510275, China*; Email: xiaon6@mail.sysu.edu.cn

Authors

Xuan Lv − *State Key Laboratory of High-Performance Computing, School of Computer Science, National University of Defense Technology, Changsha, Hunan 410073, China;* orcid.org/0000-0001-8821-0026

Jianwen Chen − *School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510275, China;* orcid.org/0000-0001-7999-2070

Yutong Lu − *School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510275, China*

Zhiguang Chen − *School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510275, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00064

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2018YFC0910500), National Natural Science Foundation of China (U1611261, 61772566, and 81801132), Guangdong Frontier & Key Tech Innovation Program (2018B010109006, 2019B020228001), and Introducing Innovative and Entrepreneurial Teams (2016ZT06D211). We appreciate the National Supercomputer Center in Guangzhou for providing us with computing resources.

■ ABBREVIATIONS

PCC, Pearson correlation coefficient; XGBoost, extreme gradient boosting; SS, secondary structure; rASA, relative solvent accessible surface area; SNPs, single nucleotide polymorphisms; PSSM, position-specific scoring matrix; SFS, Sequential Forward Selection; SBS, Sequential Backward Selection; CAGI, critical assessment of genome interpretation

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

REFERENCES

(1) Tennessen, J. A.; Bigham, A. W.; O'Connor, T. D.; Fu, W.; Kenny, E. E.; Gravel, S.; McGee, S.; Do, R.; Liu, X.; Jun, G.; et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **2012**, *337*, 64−69.

(2) Tokuriki, N.; Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **2009**, *19*, 596−604.

(3) Yates, C. M.; Sternberg, M. J. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J. Mol. Biol.* **2013**, *425*, 3949−63.

(4) Kumar, M. D.; Bava, K. A.; Gromiha, M. M.; Prabakaran, P.; Kitajima, K.; Uedaira, H.; Sarai, A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* **2006**, *34*, D204−6.

(5) 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.

(6) Pandurangan, A. P.; Ochoa-Montano, B.; Ascher, D. B.; Blundell, T. L. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* **2017**, *45*, W229−W235.

(7) Parthiban, V.; Gromiha, M. M.; Schomburg, D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **2006**, *34*, W239−W242.

(8) Wainreb, G.; Wolf, L.; Ashkenazy, H.; Dehouck, Y.; Ben-Tal, N. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics* **2011**, *27*, 3286−3292.

(9) 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.

(10) Cheng, J.; Randall, A.; Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1125−1132.

(11) 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*, 2816−2821.

(12) Folkman, L.; Stantic, B.; Sattar, A.; Zhou, Y. EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.* **2016**, *428*, 1394−1405.

(13) Capriotti, E.; Fariselli, P.; Calabrese, R.; Casadio, R. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* **2005**, *21*, ii54−ii58.

(14) Huang, L.-T.; Gromiha, M. M.; Ho, S.-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* **2007**, *23*, 1292−1293.

(15) Capriotti, E.; Fariselli, P.; Casadio, R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **2004**, *20*, i63−i68.

(16) Cao, H.; Wang, J.; He, L.; Qi, Y.; Zhang, J. Z. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1508−1514.

(17) Quan, L.; Lv, Q.; Zhang, Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* **2016**, *32*, 2936−2946.

(18) Khan, S.; Vihinen, M. Performance of protein stability predictors. *Hum. Mutat.* **2010**, *31*, 675−684.

(19) 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.

(20) Dosztányi, Z.; Fiser, A.; Simon, I. Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.* **1997**, *272*, 597−612.

(21) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, August 2016; pp 785−794.

(22) Mendik, P.; Dobronyi, L.; Hári, F.; Kerepesi, C.; Maia-Moço, L.; Buszlai, D.; Csermely, P.; Veres, D. V. Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. *Nucleic Acids Res.* **2019**, *47*, D495−D505.

(23) Zou, L. S.; Erdos, M. R.; Taylor, D. L.; Chines, P. S.; Varshney, A.; Parker, S. C.; Collins, F. S.; Didion, J. P. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics* **2018**, *19*, 390.

(24) Chen, X.; Huang, L.; Xie, D.; Zhao, Q. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* **2018**, *9*, 3.

(25) Ke, Y.; Rao, J.; Zhao, H.; Lu, Y.; Xiao, N.; Yang, Y. Accurate Prediction of Genome-wide RNA Secondary Structure Profile Based On Extreme Gradient Boosting. *bioRxiv*, April 16, **2019**. DOI: 10.1101/610782.

(26) Pucci, F.; Bernaerts, K. V.; Kwasigroch, J. M.; Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **2018**, *34*, 3659−3665.

(27) Usmanova, D. R; Bogatyreva, N. S; Arino Bernad, J.; 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*, 3653−3658.

(28) Fang, J. A Critical Review of Five Machine Learning-Based Algorithms for Predicting Protein Stability Changes upon Mutation. *Briefings Bioinf.* **2019**, *na*, na DOI: 10.1093/bib/bbz071.

(29) Thiltgen, G.; Goldstein, R. A. Assessing predictors of changes in protein stability upon mutation using self-consistency. *PLoS One* **2012**, *7*, e46084.

(30) Schäffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V.; Altschul, S. F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research* **2001**, *29*, 2994−3005.

(31) Heffernan, R.; Dehzangi, A.; Lyons, J.; Paliwal, K.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y.; Yang, Y. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics* **2016**, *32*, 843−849.

(32) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2016**, *33*, 685−692.

(33) Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model.* **2001**, *7*, 360−369.

(34) Liu, B.; Wang, S.; Wang, X. DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Sci. Rep.* **2015**, *5*, 15479.

(35) 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.

(36) Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.* **2018**, *39*, 2210−2216.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(38) Dehghanpoor, R.; Ricks, E.; Hursh, K.; Gunderson, S.; Farhoodi, R.; Haspel, N.; Hutchinson, B.; Jagodzinski, F. Predicting the effect of single and multiple mutations on protein structural stability. *Molecules* **2018**, *23*, 251.

(39) 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*, 830−838.

(40) 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*, 369−387.

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(41) Zhou, H.; Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein science* **2002**, *11*, 2714−2726.
(42) 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*, 2537−2543.
(43) Pires, D. E.; Ascher, D. B.; Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014**, *30*, 335−342.
(44) Fowler, D. M.; Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **2014**, *11*, 801.
(45) Pejaver, V.; Babbi, G.; Casadio, R.; Folkman, L.; Katsonis, P.; Kundu, K.; Lichtarge, O.; Martelli, P. L.; Miller, M.; Moult, J.; et al. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. *Hum. Mutat.* **2019**, *40*, 1495−1506.

https://dx.doi.org/10.1021/acs.jcim.0c00064
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX