EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models

Lukas Folkman[a,b], Bela Stantic[a], Abdul Sattar[a,b], Yaoqi Zhou[a,c,*]

[a] *Institute for Integrated and Intelligent Systems, Griffith University, 170 Kessels Road, Brisbane, Queensland 4111, Australia*
[b] *Queensland Research Laboratory, NICTA – National ICT Australia, 70-72 Bowen Street, Spring Hill, Queensland 4000, Australia*
[c] *Institute for Glycomics, Griffith University, Parklands Drive, Southport, Queensland 4222, Australia*

Abstract

Protein engineering and characterisation of non-synonymous single nucleotide variants (SNVs) require accurate prediction of protein stability changes ($\Delta\Delta G_u$) induced by single amino acid substitutions. Here, we have developed a new prediction method called *Evolutionary, Amino acid, and Structural Encodings with Multiple Models* (EASE-MM), which comprises five specialised support vector machine (SVM) models and makes the final prediction from a consensus of two models selected based on the predicted secondary structure and accessible surface area of the mutated residue. The new method is applicable to single-domain monomeric proteins and can predict $\Delta\Delta G_u$ with a protein sequence and mutation as the only inputs. EASE-MM yielded a Pearson correlation coefficient of 0.53-0.59 in 10-fold cross-validation and independent testing and was able to outperform other sequence-based methods. When compared to structure-based energy functions, EASE-MM achieved a comparable or better performance. The application to a large dataset of human germline non-synonymous SNVs showed that the disease-causing variants tend to be associated with larger magnitudes of $\Delta\Delta G_u$ predicted with EASE-MM. The EASE-MM web-server is available at http://sparks-lab.org/server/ease.

Keywords: missense mutation, amino acid substitution, non-synonymous SNV, free energy change, machine learning

1. Introduction

Accurate prediction of stability changes ($\Delta\Delta G_u$) induced by protein mutations is essential for successful protein engineering and characterisation of non-synonymous single nucleotide variants (SNVs)

*Corresponding author: Yaoqi Zhou, yaoqi.zhou@griffith.edu.au, +61-7-555-28228
Email addresses: lukas.folkman@griffithuni.edu.au (Lukas Folkman), b.stantic@griffith.edu.au (Bela Stantic), a.sattar@griffith.edu.au (Abdul Sattar), yaoqi.zhou@griffith.edu.au (Yaoqi Zhou)

[1]. The most reliable methods for estimating $\Delta\Delta G_u$ employ the three-dimensional structure of a target protein and calculate the free energy difference before and after the mutation with an energy function [2, 3, 4, 5, 6, 7]. The main disadvantage of this approach is that it cannot be used when the three-dimensional structure of the target protein is not available. Therefore, a number of machine learning methods emerged that can predict stability changes knowing the protein sequence only [8, 9, 10, 11]. However, when independently evaluated, the performance of these methods is limited [12], especially for mutations in previously unseen non-homologous proteins [13]. Moreover, the accuracy varies for different types of mutations depending on the secondary structure (SS) or accessible surface area (ASA) of the mutation site [13].

Recently, we have reported that combining feature-based multiple models for the *two-state classification* of stability changes as stabilising or destabilising improves prediction accuracy and achieves more balanced results for different types of mutations [14]. In this work, we designed feature-based multiple models for the *real-value prediction* of $\Delta\Delta G_u$ and developed a publicly available web-server. Our method is called *Evolutionary, Amino acid, and Structural Encodings with Multiple Models* (EASE-MM). We compared the prediction performance of EASE-MM with related work in an extensive independent validation, in which we employed only proteins with $< 25\%$ sequence identity to the dataset used for the design and training of our method. EASE-MM yielded improvements in both cross-validation and independent testing compared to other sequence-based methods. Moreover, EASE-MM, a sequence-based method, achieved a performance comparable to or better than structure-based energy functions. We applied our method to a large dataset of human germline non-synonymous SNVs and found that the disease-causing SNVs tend to be associated with larger magnitudes of $\Delta\Delta G_u$ predicted with EASE-MM. The significance of this finding is that EASE-MM, being a sequence-based method, can be applied to most single-domain monomeric proteins encoded in the human or other genomes.

2. Results

We have built EASE-MM, which comprises five specialised models to predict $\Delta\Delta G_u$ of mutations in residues located in different SS elements (helix, sheet, or coil) and with different levels of ASA (exposed or buried with a 25% threshold). The final prediction is the average of $\Delta\Delta G_u$ predicted with two models, one selected based on the predicted SS and the other based on the predicted ASA of the mutation site. We used a dataset of 1676 mutations (S1676) to design our method and estimate its performance using 10-fold cross-validation. Next, we employed two independent datasets of 543 and 236 mutations (S543 and S236, respectively) to confirm the robust performance of our method. Both datasets had a sequence

identity $< 25\%$ to the dataset used for the design and training of our method. Finally, we studied the relationship between disease-causing germline SNVs and $\Delta\Delta G_u$ predicted with EASE-MM.

2.1. Individual features

First, using the S1676 dataset, we examined the correlation between $\Delta\Delta G_u$ and each individual feature from a diverse set of 19 features encoding evolutionary conservation, amino acid parameters, and predicted structural properties of the mutation site (see Materials and Methods). The feature derived from the amino acid parameter bulkiness as the difference in the bulkiness of the mutant and wild-type amino acids (Δ bulkiness) yielded the highest Pearson correlation coefficient (r) of 0.35. The best feature derived from evolutionary conservation was the difference of the position-specific scoring matrix (PSSM) probabilities of the mutant and wild-type amino acids (Δ PSSM) with an r of 0.27. The feature relative ASA (rASA) of the mutation site had the strongest correlation (r of 0.27) from all predicted structural properties. Supplementary Table S1 lists all individual features and their correlation coefficients for the S1676 dataset.

To illustrate the significance of combining features of different types, we evaluated every possible combination of two features. We used the support vector machine (SVM) [15] algorithm with a *linear* kernel function and 10-fold cross-validation to predict $\Delta\Delta G_u$ for the S1676 dataset. The top nine feature pairs were different amino acid parameters combined with either rASA or Δ PSSM with an r in the range of 0.42–0.46. Namely, the best two combinations were Δ bulkiness+rASA and Δ hydrophobicity+Δ PSSM. Supplementary Figure S1 shows $\Delta\Delta G_u$ as a function of Δ bulkiness and rASA. The figure shows that the introduction of bulkier (relative to wild-type) amino acids in the protein core (low rASA) has a tendency to destabilise the protein structure.

2.2. Feature-based multiple models

Each of the five models employed by EASE-MM is specialised for a different type of mutations. To build these specialised models, we partitioned the training data (S1676) according to SS and ASA predicted from the protein sequence with the SPIDER method [16]. Then, a greedy sequential forward floating selection (SFFS) [17] algorithm was used to select the most relevant features for each data partition from the set of 19 features encoding evolutionary conservation, amino acid parameters, and predicted structural properties (see Materials and Methods). Figure 1 depicts how EASE-MM predicts $\Delta\Delta G_u$. First, SS and ASA of the mutation site are predicted from the protein sequence. Then, one model is selected based on the predicted SS and one based on the predicted ASA of the mutation site. The final prediction is calculated as the average of $\Delta\Delta G_u$ predicted with the two selected models.

Supplementary Table S2 lists the selected features for each EASE-MM model, ranked by their contributions to the prediction performance. At least one of the two best performing amino acid parameters (Supplementary Table S1), Δ bulkiness and Δ hydrophobicity, was represented in each of the five models. Every model contained at least one predicted structural feature, the best structural feature, rASA, was included in all but one model. Also, at least one feature encoding evolutionary conservation was represented in all but one model. This highlights that combining features of different types is beneficial for predicting $\Delta\Delta G_u$. At the same time, each model comprised specific features not represented in other models, supporting our hypothesis of building multiple specialised models for different types of mutations. For instance, the amino acid attribute Δ compressibility was selected for the *helix* and *sheet* models but not for the *coil* model. Regarding the two ASA-based models, features Δ isoelectric point and Δ polarisability were selected for the *buried* but not for the *exposed* model. To demonstrate that the five models of EASE-MM are indeed specialised, we permuted the labels of the five data partitions and observed a statistically significant decrease in prediction performance of 11–43% (Williams' test, $p < 0.01$, Supplementary Table S3).

2.3. Cross-validation performance

We conducted the first validation of our method using *unseen-protein* 10-fold cross-validation on the S1676 dataset (1676 mutations in 70 proteins). The *unseen-protein* 10-fold cross-validation ensures that any cluster of similar proteins is contained within a single cross-validation fold to prevent over-estimation of the prediction results (see Materials and Methods). We replicated the cross-validation 100 times with randomly re-generated folds and averaged the results. We compared the prediction performance of EASE-MM with our previous work, EASE-AA [13], which employs a single model trained using two evolutionary, three predicted structural, and seven amino-acid-based features. In the 10-fold cross-validation, EASE-MM yielded a Pearson correlation coefficient (r) of 0.56, which constitutes a relative improvement of 8% compared to EASE-AA with an r of 0.52 (Table 1). We performed Williams' test for comparing correlation coefficients [18] to confirm that the improvement was statistically significant with $p \ll 0.01$. Supplementary Figure S2 shows the experimentally measured $\Delta\Delta G_u$ as a function of $\Delta\Delta G_u$ predicted with EASE-AA and EASE-MM for the S1676 dataset.

Using multiple models in this work was motivated by the fact that our previous work, EASE-AA, yielded an unbalanced prediction performance for different types of mutations. Figure 2 shows the Pearson correlation coefficient (r) for different types of mutations based on SS and ASA. In this analysis, SS and ASA were calculated using DSSP [19] from experimentally determined structures. Regarding different SS types, EASE-MM was able to improve EASE-AA's performance for mutations located in α-helices

with a relative improvement of 12% from r of 0.51 to 0.57 (Williams' test, $p \ll 0.01$). For β-sheets and coils, the improvements were not significant (r of 0.56 and 0.59, $p = 0.123$; r of 0.40 and 0.43, $p = 0.166$, respectively). Regarding buried (rASA $\leq 25\%$) and exposed (rASA $> 25\%$) mutation sites, both methods yielded a considerably lower correlation for the exposed residues. While EASE-AA yielded an r of only 0.31, EASE-MM achieved an r of 0.42, which represents a relative improvement of 35% (Williams' test, $p \ll 0.01$).

We also divided mutations based on the type of the wild-type and mutant amino acids (denoted as 'wild-type→mutant', Figure 2). We considered mutations to alanine and mutations to any other amino acid type, mutations from small to large amino acids and *vice versa*, mutations from hydrophobic to hydrophilic amino acids and *vice versa*, and mutations from charged to polar amino acids and *vice versa*. EASE-MM achieved a statistically significant improvement for mutations to amino acid types other than alanine (Williams' test, $p \ll 0.01$) with a relative r improvement of 10% (r of 0.54) compared to EASE-AA (r of 0.49). This is interesting because many available datasets have a strong bias towards mutations to alanine. Particularly for S1676, 23% of mutations were 'any→alanine', whereas the second most common mutation, 'any→valine', represented only 7%. This result shows that EASE-MM is not biased (in fact, it has less bias relative to EASE-AA) towards 'any→alanine' mutations despite their abundance in the training dataset. Regarding the other amino acid categories, EASE-MM achieved statistically significant improvements for 'large→small' (Williams' test, $p \ll 0.01$) and 'charged→polar' mutations ($p = 0.003$).

2.4. Independent test performance

We employed two different independent test sets, which were not used in any way for neither feature selection nor model optimisation, to verify if EASE-MM performs robustly. The S543 dataset comprised 543 mutations in 55 proteins with a sequence identity $< 25\%$ to the S1676 training set. The second dataset (S236) contained 236 mutations in 23 proteins with a sequence identity $< 25\%$ to S1676. The two test sets were disjoint (sequence identity $< 25\%$). We compared the performance of EASE-MM with three sequence-based methods (I-Mutant2.0 [8], MUpro [9], and EASE-AA [13]), four structure-based energy functions (Rosetta [5], FoldX [4], DFIRE [7], and PoPMuSiC [3]), and the structure-based version of I-Mutant2.0 [8]. Table 1 summarises the results in terms of r and root mean square error (RMSE). On the larger dataset (S543), EASE-MM yielded an r of 0.53 and was able to outperform all sequence-based as well as structure-based methods except for PoPMuSiC (Williams' test, $p < 0.01$). The relative improvements were of 66% (sequence-based I-Mutant2.0), 61% (MUpro), 47% (structure-based I-Mutant2.0), 39% (Rosetta), 29% (FoldX), 18% (DFIRE), and 10% (EASE-AA). Figure 3 shows the experimentally measured $\Delta\Delta G_u$ as a function of $\Delta\Delta G_u$ predicted with the nine compared methods for the

S543 dataset. On the smaller dataset (S236), EASE-MM yielded an r of 0.59 and statistically significant improvements ($p < 0.01$) compared to sequence-based I-Mutant2.0 and MUpro as well as structure-based Rosetta and FoldX. When compared to sequence-based EASE-AA (r of 0.53), the improvement was not significant ($p = 0.025$). None of the nine compared methods was able to outperform EASE-MM. Supplementary Figure S3 shows the experimentally measured $\Delta\Delta G_u$ as a function of predicted $\Delta\Delta G_u$ for the S236 datasets.

To see if the performance of the structure-based energy functions was affected by the quality of the experimentally determined structures, we also evaluated the performance of all methods on subsets of S543 and S236 comprising only mutations in proteins with high-resolution (≤ 3 Å) crystal structures. We refer to these datasets as S405 with 405 mutations in 44 proteins and S157 with 157 mutations in 16 proteins. While the performance of all methods changed marginally, no method was able to outperform EASE-MM at significance level $\alpha = 0.01$ (Table 1).

Figure 4 shows the prediction performance (r) of the three best-performing methods (EASE-AA, PoPMuSiC, and EASE-MM) on the S543 dataset for different types of mutations based on the SS and ASA of the mutation site and for different wild-type and mutant amino acid types. As it was expected, the most challenging mutations were located in coils, mutations to other amino acids than alanine, and mutations introducing a larger amino acid. Supplementary Figure S4 depicts the same comparison for all nine compared methods for both S543 and S236 datasets.

EASE-MM estimates the SS and ASA of the mutation site using SPIDER [16] from the protein sequence. To see if EASE-MM could make better $\Delta\Delta G_u$ predictions using experimentally determined structures, we substituted SPIDER predictions with the true SS and ASA assignments calculated with DSSP [19] from three-dimensional structures from the Protein Data Bank [20]. There was no statistical difference in the prediction performance (r) of $\Delta\Delta G_u$ for the two methods (Supplementary Table S4, Williams' test, $p = 0.973$ and 0.446 for S543 and S236, respectively). As a control, we used randomly generated SS and ASA assignments and observed a significant decrease in the prediction performance from an r of 0.53 to 0.36 for S543 ($p \ll 0.01$) and from 0.59 to 0.31 for S236 ($p = 0.002$). These results demonstrate that SPIDER can reliably predict SS and ASA, and that their accurate prediction is important for reliable prediction of $\Delta\Delta G_u$ by EASE-MM.

2.5. Stability changes of disease-causing mutations

Finally, we investigated if the predicted stability changes could be used as an indication of disease-causing mutations. We compiled a dataset of 10,511 disease-causing non-synonymous SNVs in 2201 proteins from ClinVar [21] and 278,760 putatively neutral non-synonymous SNVs in 20,096 proteins

from the 1000 Genomes Project [22]. Figure 5 plots the distributions of the absolute value of $\Delta\Delta G_u$ predicted with EASE-MM for the 10,511 disease-causing and 50,910 putatively neutral mutations with allele frequency (AF) $\geq 1\%$. The two distributions were significantly different (Wilcoxon's test, $p \ll 0.01$). As expected, the figure shows that while putatively neutral mutations are characterised by small magnitudes of $\Delta\Delta G_u$, disease-causing mutations tend to have a larger absolute effect on the protein stability according to EASE-MM predictions. We employed the *absolute value* of $\Delta\Delta G_u$ because it has been shown that both protein destabilisation [1, 23] as well as stabilisation, via loss of flexibility [24, 25], can lead to a loss of protein function. That is, it is the magnitude rather than the direction of $\Delta\Delta G_u$ that might be indicative. The idea of using the absolute value of $\Delta\Delta G_u$ is further supported by a large-scale analysis of Casadio *et al.* [26]. There, the authors showed that the correlation of disease probability with the absolute value of experimentally measured $\Delta\Delta G_u$ was stronger than with $\Delta\Delta G_u$ of only the destabilising or stabilising mutations.

We also performed a binary classification of disease-causing and neutral SNVs using predicted $\Delta\Delta G_u$. Based on the receiver operating characteristic (ROC) curve analysis, EASE-MM yielded the area under the curve (AUC) of 0.69. We chose a prediction threshold of $0.8\,\mathrm{kcal\,mol^{-1}}$ based on the point where the distributions of disease-causing and neutral SNVs intersect in Figure 5. This threshold resulted in a Matthews correlation coefficient (MCC) of 0.25 with a sensitivity (correctly predicted disease-causing SNVs) of 49% and specificity (correctly predicted neutral SNVs) of 80%, binary accuracy of 75%, precision (positive predictive value) of 33%, and negative predictive value of 88%. Thus, the statistically different distributions of predicted $\Delta\Delta G_u$ can provide some discriminative power between disease-causing and neutral SNVs. This means that $\Delta\Delta G_u$ predicted with EASE-MM can be useful in combination with other predictive features for improving classification of disease-causing SNVs.

Next, we investigated the relationship between the AF and $\Delta\Delta G_u$. The AF should generally reflect the fitness of the allele with respect to its intended biological function [27]. Hence, we hypothesised that protein stability could be one of the factors affecting the AF. Figure 6 shows the absolute value of $\Delta\Delta G_u$ predicted with EASE-MM as a function of *binned* $\log_{10}$ AF for 278,760 non-synonymous SNVs from the 1000 Genomes Project [22]. There, AFs were grouped into bins so that each bin contained at least 500 SNVs. However, some bins were larger due to many SNVs with the same AF value. As expected, there was a strong negative correlation (r of -0.85) between the AF bins and the average of the absolute value of $\Delta\Delta G_u$. That is, highly populated alleles tend to have smaller changes in the protein stability.

3. Discussion

We have developed a machine learning method named EASE-MM, which can predict the change in the protein stability upon a single amino acid substitution based on sequence information, without the experimentally determined three-dimensional structure. EASE-MM employs multiple SVM models specifically designed for different types of mutations based on the SS and ASA of the mutation site. Each SVM model combines a different set of features encoding evolutionary conservation, amino acid parameters, and predicted structural properties. The new method yielded a robust performance with an r of 0.56 in 10-fold cross-validation and 0.53–0.59 in independent testing. EASE-MM outperformed other sequence-based methods and achieved a comparable or better performance than structure-based energy functions (Table 1).

To avoid over-training in a particular group of proteins, we coupled feature selection with the *unseen-protein* cross-validation procedure, in which we ensured that no two folds shared sequences with a pairwise identity $\geq 25\%$. A similar approach was used for our previous work [13, 14] as well as for the prediction of disease-causing genetic variants [28, 29, 30]. Furthermore, we employed two different independent test sets to confirm that our method was not over-trained. These test sets had $< 25\%$ sequence identity to our training data to prevent over-estimation of prediction results [31].

Our work reveals several predictive features that are important for protein stability. For instance, sequence conservation is an expected feature because it has been long known to be a strong indicator of functionally and structurally important residues. Indeed, we found a positive correlation between $\Delta\Delta G_u$ and Δ PSSM of 0.27 ($p \ll 0.01$). Here, $\Delta\,\mathrm{PSSM} = \mathrm{PSSM}_{mt} - \mathrm{PSSM}_{wt}$ and thus, the positive correlation means that an introduction of an uncommon amino acid type in a conserved position results in the destabilisation of the protein. Different features, however, contribute differently to different models. As shown in Supplementary Table S2, the most important features are rASA and changes in helix tendency for the *helix* model, sequence conservation and changes in amino acid volume for the *sheet* model, changes in hydrophobicity and flexibility for the *coil* model, changes in isoelectric point and bulkiness for the *buried* model, and changes in amino acid volume and predicted helix probability for the *exposed* model. These features were selected automatically using a feature selection algorithm to maximise the performance on the training dataset. Sometimes seemingly similar features were selected for distinct models. However, their interpretation might vary. For instance, on the one hand, the feature changes in helix tendency (Δ helix tendency) in the *helix* model suggests whether the preference to form helix is different for the mutant and wild-type amino acid types. On the other hand, the feature helix probability in the *exposed* model expresses the likelihood of the mutation site adopting a helical structure in the wild-type protein.

One disadvantage of using machine learning models is that the selected features are not always easy to interpret. For example, also the *sheet* model contains the feature Δ helix tendency but its contribution in this model is not very significant.

We found that the new method not only outperformed our previous work, EASE-AA [13], but yielded a more balanced performance for different types of mutations based on the ASA of the mutation site (Fig. 2). That is, EASE-MM yielded improvements in correlation with experimentally measured $\Delta\Delta G_u$ for the exposed residues (rASA $> 25\%$), which were more challenging to predict for both methods, while retaining the same performance as EASE-AA for the buried residues. Since EASE-AA uses a single model trained with the same types of features as implemented in EASE-MM, it means that the improved predictions can be mainly attributed to employing specialised models for different types of mutations.

While the new method can be applied universally to any amino acid sequence of a monomeric protein, it was trained and tested only using structured, soluble proteins. In fact, this is a common limitation to most prediction methods being a direct consequence of the available experimental $\Delta\Delta G_u$ data lacking membrane or intrinsically disordered proteins. Furthermore and particularly for EASE-MM, another limitation is that structural properties (SS and ASA) can be reliably predicted only for single-domain proteins. For the prediction of stability changes of protein-protein complexes, several structure-based methods are available [4, 32].

Finally, we applied EASE-MM to a large dataset of human disease-causing and putatively neutral germline non-synonymous SNVs. We found that the distributions of predicted $\Delta\Delta G_u$ are significantly different (Fig. 5). Moreover, there was a strong negative correlation between the binned AF and average of the absolute value of predicted $\Delta\Delta G_u$ (Fig. 6). Our results show that highly populated alleles tend to be associated with smaller changes in protein stability. As pointed out elsewhere [33, 34], the value of $\Delta\Delta G_u$ alone is not sufficient to provide reliable classification of SNVs because protein stability is only one of many disease-causing factors. Nevertheless, our results indicate that EASE-MM can be combined with other predictive features for improved characterisation of disease-causing mutations. The significance of this finding is that EASE-MM, being a sequence-based method, can be applied to most single-domain monomeric proteins encoded in the human or other genomes.

4. Materials and Methods

4.1. Datasets

We used several different datasets to design, validate, and independently test our method. Table 2 provides a summary of these datasets. We used the S1676 (1676 mutations in 70 proteins) and S236

(236 mutations in 23 proteins) datasets compiled in Folkman *et al.* [14] from ProTherm [35] (version February 2013). ProTherm defines a stability change as the difference in the unfolding free energy: $\Delta\Delta G_u[\mathrm{kcal\,mol^{-1}}] = \Delta G_u(mutant) - \Delta G_u(wild\text{-}type)$. Thus, destabilising mutations yield $\Delta\Delta G_u < 0$. We verified all records in ProTherm and corrected incorrect entries according to the original publications. Next, we removed all duplicate entries of the same amino acid substitutions (*e.g.*, different concentrations of chemicals). If several measurements of the same mutation under the same experimental conditions were present, we averaged $\Delta\Delta G_u$. If several measurements of the same mutation under different experimental conditions were present, we kept only the measurement closest to the physiological pH 7. The S1676 dataset was used to design our method and optimise all parameters. S1676 and S236 are mutually independent and do not share proteins with $\geq 25\%$ sequence identity. Moreover, S236 is also independent ($< 25\%$ sequence identity) to the two datasets which were used to build I-Mutant2.0 [8] and MUpro [9]. Therefore, we used S236 for independent testing and comparison with related work. Another independent test set, S543 (543 mutations in 55 proteins), was compiled as a subset of the 2648 mutations from Dehouck *et al.* [3]. S543 has $< 25\%$ sequence identity to both S1676 and S236. Supplementary Figure S5 shows the distributions of the experimentally measured $\Delta\Delta G_u$ for the three datasets. Supplementary Figures S6, S7, and S8 show the frequencies and EASE-MM's prediction errors of all wild-type and mutant amino acid types for S1676, S543, and S236, respectively.

To see if the comparison with structure-based energy functions (Rosetta [5], FoldX [4], DFIRE [7], and PoPMuSiC [3]) was affected by structures determined with nuclear magnetic resonance (NMR), we also compiled the S157 (a subset of S236 comprising 157 mutations in 16 proteins) and S405 (a subset of S543 comprising 405 mutations in 44 proteins) datasets of high-resolution (≤ 3 Å) crystal structures from the Protein Data Bank [20].

Finally, to study the relationship between the predicted $\Delta\Delta G_u$ and human germline non-synonymous SNVs, we compiled a dataset of 10,511 disease-causing (2201 proteins) and 278,760 putatively neutral (20,096 proteins) SNVs from ClinVar [21] and 1000 Genomes Project [22], respectively. For the distribution analysis, we considered the subset comprising 50,910 putatively neutral SNVs (14,113 proteins) with AF $\geq 1\%$.

4.2. Predictive features

We employed three types of predictive features in our method: evolutionary conservation, amino acid parameters, and predicted structural properties. To estimate evolutionary conservation of the mutation site, we used three iterations of PSI-BLAST [36] on the NCBI non-redundant database with an *e*-value threshold of 10^{-3}. From the PSSM generated with PSI-BLAST, we extracted the probability of the wild-

type (PSSM_{wt}) and mutant (PSSM_{mt}) amino acids at the mutation site. We implemented two different features, PSSM_{wt} and $\Delta\,\mathrm{PSSM} = \mathrm{PSSM}_{mt} - \mathrm{PSSM}_{wt}$. We also included a feature encoding the overall conservation of the mutation site as property entropy (PE) with respect to six sets grouping amino acids based on their chemical properties as aliphatic (A, V, L, I, M, C), aromatic (F, W, Y, H), polar (S, T, N, Q), positive (K, R), negative (D, E), and special (G, P) [37]. The property entropy was calculated from a multiple sequence alignment of the 30 most similar sequences from the NCBI non-redundant database ranked by e-value with PSI-BLAST (using 100, 500, or all sequences resulted in a lower correlation with $\Delta\Delta G_u$, S1676 dataset). We used the implementation from Capra and Singh [38] to calculate the property entropy based on the following equation:

$$\mathrm{PE}(msa_i) = 1 - \left(- \frac{\sum_{g \in G} p(msa_i, g) \times \log p(msa_i, g)}{\log |msa_i|} \right),$$

$$p(msa_i, g) = \sum_{aa \in g} p(msa_i, aa),$$

where msa_i is the i-th column of the multiple sequence alignment msa, G is the set of the defined property groups, and $p(msa_i, g)$ is the probability of the property group g at msa_i, which is equal to the sum of probabilities of the amino acid types (aa) belonging to g.

Different amino acid parameters have been used for the prediction of stability changes [39, 40, 41, 42]. We adopted a total of 11 amino acid parameters: hydrophobicity, volume, polarisability, isoelectric point, helix tendency, sheet tendency, and a steric parameter (graph shape index) from Meiler *et al.* [43]; compressibility, bulkiness, and equilibrium constant with reference to the ionisation property of COOH group from Gromiha *et al.* [44]; and flexibility from Vihinen *et al.* [45]. For each amino acid parameter (AAP), we calculated $\Delta\,\mathrm{AAP} = \mathrm{AAP}_{mt} - \mathrm{AAP}_{wt}$, where AAP_{mt} and AAP_{wt} denote the value of the given AAP for the mutant and wild-type amino acids, respectively. Supplementary Table S5 provides the values of the 11 parameters for the 20 common amino acids.

Finally, we considered five structural features predicted from the protein sequence: rASA, helix, sheet, coil, and disorder probabilities. The rASA and SS probabilities were predicted using SPIDER [16]. The disorder probability was calculated using SPINE-D [46].

4.3. Feature selection and multiple models

To build the five models employed by EASE-MM, we partitioned the S1676 training dataset according to SS (helix, sheet, and coil) and ASA (buried or exposed with a 25% threshold). SS and ASA were predicted from the protein sequence with SPIDER [16]. Supplementary Figure S5 shows the distributions of the experimentally measured $\Delta\Delta G_u$ for the different data partitions.

A unique set of features was identified for each of the five SVM models using the SFFS algorithm [17]. SFFS starts with an empty set of features S_0 and iteratively searches for a better set of features in two steps. First, the best feature f is selected as the one for which $S_i = S_{i-1} \cup \{f\}$ yields the lowest RMSE. Second, features f^* for which $S_i - \{f^*\}$ yields a lower RMSE than S_{i-1} are iteratively removed. Thus, the number of features in S is not monotonously increasing because the search is 'floating' up and down. Supplementary Table S2 lists the selected features for each EASE-MM model.

4.4. Training and evaluation

We employed the S1676 dataset to design our method, perform feature selection, and optimise all parameters using the *unseen-protein* 10-fold cross-validation. The *unseen-protein* cross-validation is used to avoid over-fitting on specific proteins by splitting the dataset into cross-validation folds so that all mutations of a cluster of similar proteins ($\geq 25\%$ sequence identity) are always contained within a single fold. These clusters were identified with Blastclust [47]. To devise a robust estimate of the prediction performance, we replicated the cross-validation procedure 100 times with randomly re-generated folds and averaged the results.

We implemented EASE-MM with ϵ-SVR (support vector regression) and radial basis function (RBF) kernel using the LibSVM [48] library. For ϵ-SVR, we optimised three parameters (C, γ, and ϵ) using a grid search in the range of $C \in \{2^{-1}, 2^0, \ldots, 2^6\}$, $\gamma \in \{2^{-8}, 2^{-7}, \ldots, 2^0\}$, and $\epsilon \in \{2^{-8}, 2^{-7}, \ldots, 2^{-1}\}$. Then, each model was trained on S1676 and tested independently using S543 and S236 to confirm that our approach did not result in over-fitting. Importantly, S543 and S236 did not share similar sequences ($\geq 25\%$ sequence identity) with S1676 and were not used during the design, feature selection, or parameter optimisation in any way. Furthermore, S543 and S236 were disjoint with $< 25\%$ sequence identity. The performance of EASE-MM was assessed in terms of Pearson correlation coefficient (r) and root mean square error (RMSE):

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}},$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum (x_i - y_i)^2}.$$

Finally, we re-optimised the ϵ-SVR parameters and trained the final EASE-MM models, which are deployed on our web-server, using a joint S1676+S236 dataset in order to maximise the size of the training data.

Acknowledgements

Helpful discussions with Yuedong Yang and Jaroslav Bendl are gratefully acknowledged. This work was supported in part by National Health and Medical Research Council (1059775 and 1083450) of Australia and Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme (project number LE150100161) to Y.Z. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, and the Australian Research Council through the ICT Centre of Excellence program. The authors also gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster 'Gowonda' to complete this research. This works was implemented using the GNU Parallel program [49].

Appendix A. Supplementary Data

Supplementary data to this article can be found online.

Abbreviations used:

1000G, 1000 Genomes Project; Å, Angstrom; AAP, amino acid parameter; AF, allele frequency; ASA, accessible surface area; AUC, area under the curve; $\Delta\Delta G_u$, stability change; MCC, Matthews correlation coefficient; NMR, nuclear magnetic resonance; PSSM, position-specific scoring matrix; r, Pearson correlation coefficient; rASA, relative accessible surface area; RBF, radial basis function; RMSE, root mean square error; ROC, receiver operating characteristic; *seq*, sequence-based; SFFS sequential forward floating selection; SNV, single nucleotide variant; SS, secondary structure; SVM, support vector machine; SVR, support vector regression; *str*, structure-based; Δ X, a difference in the property X between the mutant and wild-type amino acids.

References

[1] P. Yue, Z. Li, J. Moult, Loss of protein structure stability as a major causative factor in monogenic disease, Journal of Molecular Biology 353 (2) (2005) 459–473.

[2] A. Benedix, C. Becker, B. de Groot, A. Caflisch, R. Bockmann, Predicting free energy changes using structural ensembles, Nature Methods 6 (1) (2009) 3–4.

[3] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, M. Rooman, Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0, Bioinformatics 25 (19) (2009) 2537.

[4] R. Guerois, J. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations, Journal of Molecular Biology 320 (2) (2002) 369–387.

[5] E. Kellogg, A. Leaver-Fay, D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability, Proteins: Structure, Function, and Bioinformatics 79 (2011) 830–838.

[6] S. Yin, F. Ding, N. Dokholyan, Eris: An automated estimator of protein stability, Nature Methods 4 (6) (2007) 466–467.

[7] H. Zhou, Y. Zhou, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, Protein Science 11 (11) (2002) 2714–2726.

[8] E. Capriotti, P. Fariselli, R. Casadio, I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure, Nucleic Acids Research 33 (Suppl 2) (2005) W306–W310.

[9] J. Cheng, A. Randall, P. Baldi, Prediction of protein stability changes for single-site mutations using support vector machines, Proteins: Structure, Function, and Bioinformatics 62 (4) (2006) 1125–1132.

[10] L. Folkman, B. Stantic, A. Sattar, Sequence-only evolutionary and predicted structural features for the prediction of stability changes in protein mutants, BMC Bioinformatics 14 (Suppl 2) (2013) S6.

[11] L. Huang, M. Gromiha, S. Ho, iPTREE-STAB: Interpretable decision tree based method for predicting protein stability changes upon mutations, Bioinformatics 23 (10) (2007) 1292.

[12] S. Khan, M. Vihinen, Performance of protein stability predictors, Human Mutation 1 (1) (2010) 675.

[13] L. Folkman, B. Stantic, A. Sattar, Towards sequence-based prediction of mutation-induced stability changes in unseen non-homologous proteins, BMC Genomics 15 (Suppl 1) (2014) S4.

[14] L. Folkman, B. Stantic, A. Sattar, Feature-based multiple models improve classification of mutation-induced stability changes, BMC Genomics 15 (Suppl 4) (2014) S6.

[15] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[16] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific Reports 5 (2015) 11476.

[17] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (11) (1994) 1119–1125.

[18] E. J. Williams, The comparison of regression variables, Journal of the Royal Statistical Society. Series B (Methodological) (1959) 396–399.

[19] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.

[20] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The protein data bank, Nucleic Acids Research 28 (1) (2000) 235–242.

[21] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, D. R. Maglott, Clinvar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Research 42 (D1) (2014) D980–D985.

[22] The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature 467 (7319) (2010) 1061–1073.

[23] Z. Wang, J. Moult, SNPs, protein structure, and disease, Human Mutation 17 (4) (2001) 263–270.

[24] B. M. Beadle, B. K. Shoichet, Structural bases of stability—function tradeoffs in enzymes, Journal of Molecular Biology 321 (2) (2002) 285–296.

[25] P. Zavodszky, J. Kardos, A. Svingor, G. A. Petsko, Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins, Proceedings of the National Academy of Sciences 95 (13) (1998) 7406–7411.

[26] R. Casadio, M. Vassura, S. Tiwari, P. Fariselli, P. Luigi Martelli, Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome, Human Mutation 32 (10) (2011) 1161–1170.

[27] G. T. Marth, F. Yu, A. R. Indap, K. Garimella, S. Gravel, W. F. Leong, C. Tyler-Smith, M. Bainbridge, T. Blackwell, X. Zheng-Bradley, et al., The functional spectrum of low-frequency coding variation, Genome Biology 12 (9) (2011) R84.

[28] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations, Nature Methods 7 (4) (2010) 248–249.

[29] L. Folkman, Y. Yang, Z. Li, B. Stantic, A. Sattar, M. Mort, D. N. Cooper, Y. Liu, Y. Zhou, DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels, Bioinformatics 31 (10) (2015) 1599–1606.

[30] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney, P. Radivojac, Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (21) (2009) 2744–2750.

[31] D. G. Grimm, C.-A. Azencott, F. Aicheler, U. Gieraths, D. G. MacArthur, K. E. Samocha, D. N. Cooper, P. D. Stenson, M. J. Daly, J. W. Smoller, L. E. Duncan, K. M. Borgwardt, The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity, Human Mutation 36 (5) (2015) 513–523.

[32] Y. Dehouck, J. M. Kwasigroch, M. Rooman, D. Gilis, BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations, Nucleic Acids Research 41 (W1) (2013) W333–W339.

[33] T. G. Kucukkal, M. Petukh, L. Li, E. Alexov, Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins, Current Opinion in Structural Biology 32 (2015) 18–24.

[34] M. Petukh, T. G. Kucukkal, E. Alexov, On human disease-causing amino acid variants: statistical study of sequence and structural patterns, Human Mutation 36 (5) (2015) 524–534.

[35] M. Kumar, K. Bava, M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, A. Sarai, ProTherm and ProNIT: Thermodynamic databases for proteins and protein–nucleic acid interactions, Nucleic Acids Research 34 (Suppl 1) (2006) D204.

[36] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research 25 (17) (1997) 3389.

[37] L. A. Mirny, E. I. Shakhnovich, Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, Journal of Molecular Biology 291 (1) (1999) 177–196.

[38] J. A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics 23 (15) (2007) 1875–1882.

[39] L. Huang, K. Saraboji, S. Ho, S. Hwang, M. Ponnuswamy, M. Gromiha, Prediction of protein mutant stability using classification and regression tool, Biophysical Chemistry 125 (2–3) (2007) 462–470.

[40] S. Kang, G. Chen, G. Xiao, Robust prediction of mutation-induced protein stability change by property encoding of amino acids, Protein Engineering Design and Selection 22 (2) (2009) 75.

[41] B. Shen, J. Bai, M. Vihinen, Physicochemical feature-based classification of amino acid mutations, Protein Engineering Design and Selection 21 (1) (2008) 37–44.

[42] S. Teng, A. Srivastava, L. Wang, Sequence feature-based prediction of protein stability changes upon amino acid substitutions, BMC Genomics 11 (Suppl 2) (2010) S5.

[43] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, Molecular modeling annual 7 (9) (2001) 360–369.

[44] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, Journal of Protein Chemistry 18 (5) (1999) 565–578.

[45] M. Vihinen, E. Torkkila, P. Riikonen, Accuracy of protein flexibility predictions, Proteins: Structure, Function, and Bioinformatics 19 (2) (1994) 141–149.

[46] T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, Y. Zhou, SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method, Journal of Biomolecular Structure and Dynamics 29 (4) (2012) 799–813.

[47] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215 (3) (1990) 403–410.

[48] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (3) (2011) 27:1–27:27.

[49] O. Tange, GNU Parallel – the command-line power tool, ;login: The USENIX Magazine 36 (1) (2011) 42–47.

Table 1: Comparison of EASE-MM and related work in terms of correlation and error between the predicted and experimentally measured stability changes ($\Delta\Delta G_u$)

	All proteins				High-resolution crystal (≤ 3 Å) structures			
Method	**Dataset**	r^a	p^a	**RMSE**[a]	**Dataset**	r^a	p^a	**RMSE**[a]
EASE-AA	S1676[b]	0.52	4.2×10^{-5}	1.56	—	—	—	—
EASE-MM		0.56	—	1.52		—	—	—
I-Mutant2.0 *seq*[c]	S543[d]	0.32	4.8×10^{-11}	1.37	S405[e]	0.38	1.3×10^{-4}	1.34
MUpro 1.1		0.33	6.9×10^{-8}	1.32		0.37	4.6×10^{-4}	1.30
I-Mutant2.0 *str*[c]		0.36	1.8×10^{-8}	1.34		0.37	1.1×10^{-4}	1.34
Rosetta 3.5		0.38[h]	3.8×10^{-5}	3.58[hj]		0.35[h]	1.6×10^{-4}	4.00[hj]
FoldX 3		0.41	1.1×10^{-3}	1.87		0.42	0.027	1.92
DFIRE		0.45	8.4×10^{-4}	1.44		0.46	0.048	1.49
EASE-AA		0.48	2.0×10^{-3}	1.25		0.48	0.081	1.25
PoPMuSiC 2.1		0.53	0.909	1.21		0.49	0.404	1.27
EASE-MM		0.53	—	1.22		0.51	—	1.25
I-Mutant2.0 *seq*[c]	S236[f]	0.44	1.0×10^{-3}	1.18	S157[g]	0.43	0.032	1.08
MUpro 1.1		0.36	2.7×10^{-5}	1.20		0.29	3.9×10^{-4}	1.14
I-Mutant2.0 *str*[c]		0.52	0.105	1.07		0.47	0.160	0.94
Rosetta 3.5		0.27[i]	1.2×10^{-7}	1.88[ij]		0.34[i]	3.9×10^{-3}	1.94[ij]
FoldX 3		0.28	3.3×10^{-6}	1.70		0.34	9.6×10^{-3}	1.80
DFIRE		0.54	0.162	1.18		0.52	0.563	1.03
EASE-AA		0.53	0.025	1.10		0.51	0.172	1.04
PoPMuSiC 2.1		0.57	0.630	1.05		0.58	0.640	0.98
EASE-MM		0.59	—	1.03		0.55	—	0.97

[a] r, Pearson correlation coefficient; p, probability that the correlation coefficients (r) of the given method and EASE-MM are different due to random chance (Williams' test for comparing correlation coefficients); RMSE, root mean square error [kcal mol^{-1}].

[b] S1676 was used for feature selection and 10-fold cross-validation; the sequence identity of any two proteins from two different cross-validation folds was $< 25\%$.

[c] *seq*, sequence-based; *str*, structure-based.

[d] S543 is an independent test set, not used for feature selection nor model optimisation, compiled from a subset of the dataset from Dehouck *et al.* [3] with a sequence identity $< 25\%$ to S1676, S236, and S157.

[e] S405 is a subset of S543 containing only mutations in high-resolution (≤ 3 Å) crystal structures.

[f] S236 is an independent test set, not used for feature selection nor model optimisation, with a sequence identity $< 25\%$ to S1676, S543, and S405.

[g] S157 is a subset of S236 containing only mutations in high-resolution (≤ 3 Å) crystal structures.

[h] Three mutations were removed due to atomic clashes ($E_{rep} > 7$).

[i] Four mutations were removed due to atomic clashes ($E_{rep} > 7$).

[j] $\Delta\Delta G_u$ predicted with Rosetta is in 'Rosetta energy units', no scaling was performed.

Table 2: Datasets used to design, validate, and independently test EASE-MM

	Mutation count						Protein count						
Dataset	***all***	**H**[a]	**S**[a]	**C**[a]	**B**[a]	**E**[a]	***all***	**H**[a]	**S**[a]	**C**[a]	**B**[a]	**E**[a]	**Source**
S1676[b]	1676	615	438	623	744	932	70	51	41	56	54	59	Folkman *et al.* [14]
S543[c]	543	155	224	164	292	251	55	32	33	39	44	42	Dehouck *et al.* [3]
S405[d]	405	87	195	123	231	174	44	25	27	31	34	35	subset of S543
S236[e]	236	81	62	93	109	127	23	12	12	17	18	16	Folkman *et al.* [14]
S157[f]	157	33	58	66	61	96	16	7	8	12	12	11	subset of S236
human	10,511 (disease) + 278,760[g] (neutral)						2201 (disease) + 20,096[g] (neutral)						ClinVar [21], 1000G [22]

[a] H, helix; S, sheet; C, coil; B, buried; E, exposed; the five SS and ASA partitions were predicted with SPIDER [16].

[b] training set and 10-fold cross-validation, the sequence identity of any two proteins from two different folds was $< 25\%$.

[c] test set; S543 is independent to S1676, S236, and S157 with a sequence identity $< 25\%$.

[d] test set; S405 is a subset of S543 containing only mutations in high-resolution (≤ 3 Å) crystal structures.

[e] test set; S236 is independent to S1676, S543, and S405 with a sequence identity $< 25\%$.

[f] test set; S157 is a subset of S236 containing only mutations in high-resolution (≤ 3 Å) crystal structures.

[g] For the distribution analysis, a subset of 50,910 mutations (14,113 proteins) with allele frequency (AF) $\geq 1\%$ was used.

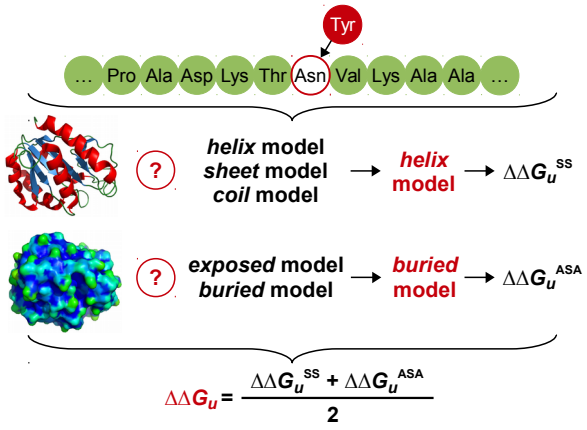

Figure 1: EASE-MM calculates the stability change ($\Delta\Delta G_u$) as the average of $\Delta\Delta G_u$ predicted with two distinct models chosen based on the *predicted* secondary structure and accessible surface area of the mutation site.

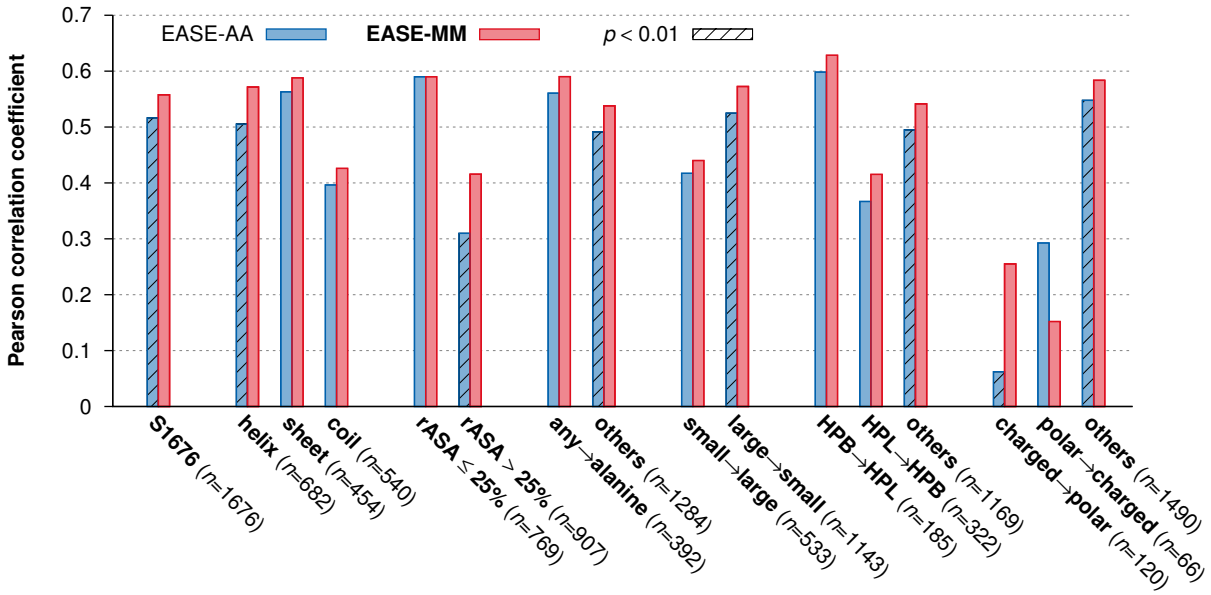

Figure 2: Pearson correlation coefficient (r) as the performance of single-model EASE-AA [13] and multiple-models EASE-MM for different types of mutations from the S1676 dataset. The striped bars show results which are statistically different from EASE-MM (Williams' test, $p < 0.01$). The secondary structure elements (helix, sheet, coil) and relative accessible surface area (rASA) of the mutation site were calculated with DSSP [19]. We also divided mutations based on the type of the wild-type and mutant amino acids (denoted as 'wild-type→mutant'). Small and large amino acids were defined based on the non-hydrogen atom counts. Amino acids were grouped based on their side-chains as hydrophobic (HPB): A, V, I, L, M, F, Y, W; polar: S, T, N, Q; charged: D, E, K, R, H; and hydrophilic (HPL): polar + charged.

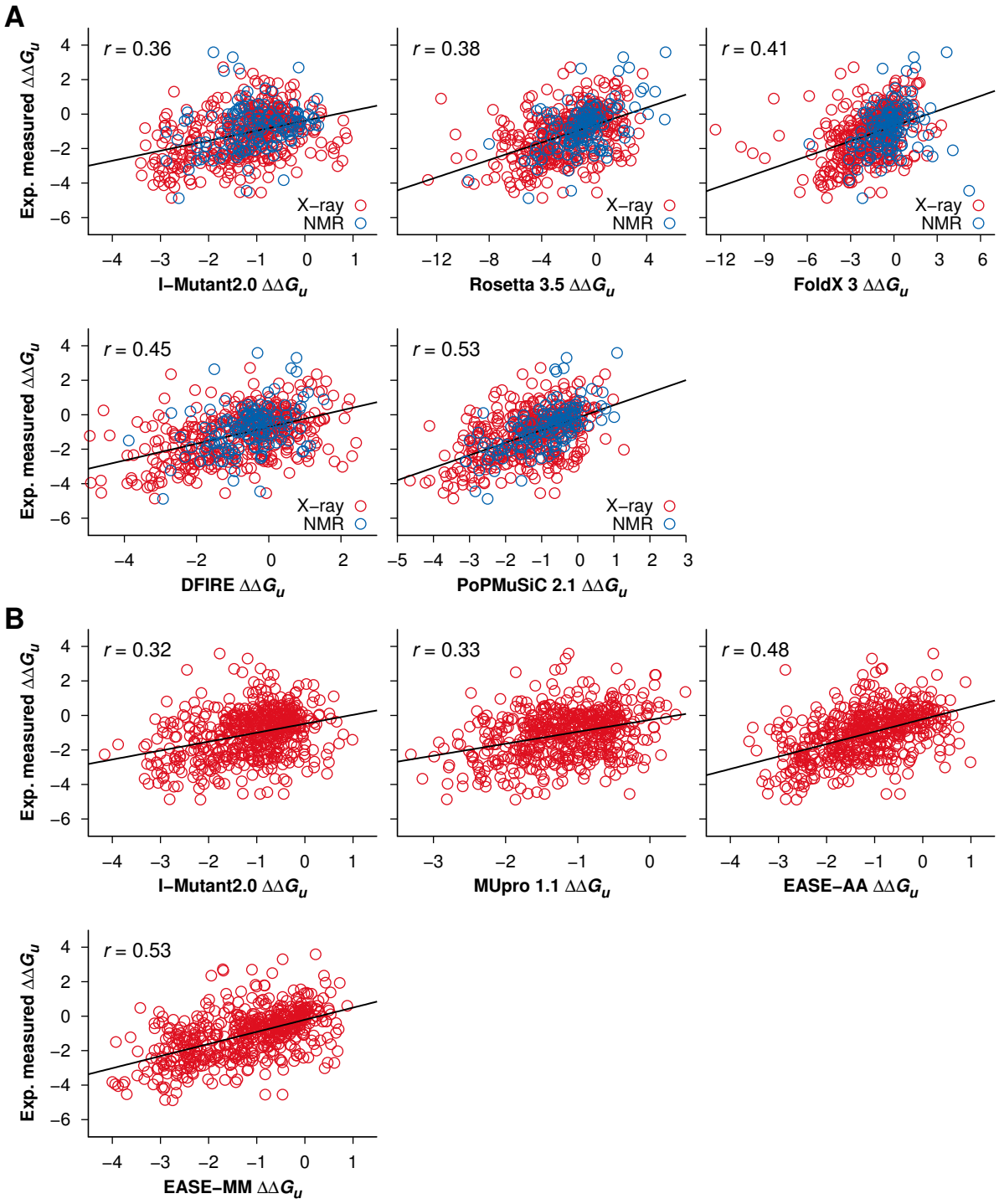

Figure 3: Experimentally measured stability changes ($\Delta\Delta G_u$) from the S543 dataset as a function of $\Delta\Delta G_u$ predicted with the five *structure-based* methods **(A)** and four *sequence-based* methods **(B)** including EASE-MM. Three predictions which caused atomic clashes during structure optimisation with Rosetta ($E_{rep} > 7$) were removed from the Rosetta plot. For the structure-based methods **(A)**, X-ray denotes predictions for proteins with high-resolution (≤ 3 Å) crystal structures (405 mutations), and NMR denotes predictions for protein structures determined with nuclear magnetic resonance (138 mutations). The black lines are the linear regression fits.

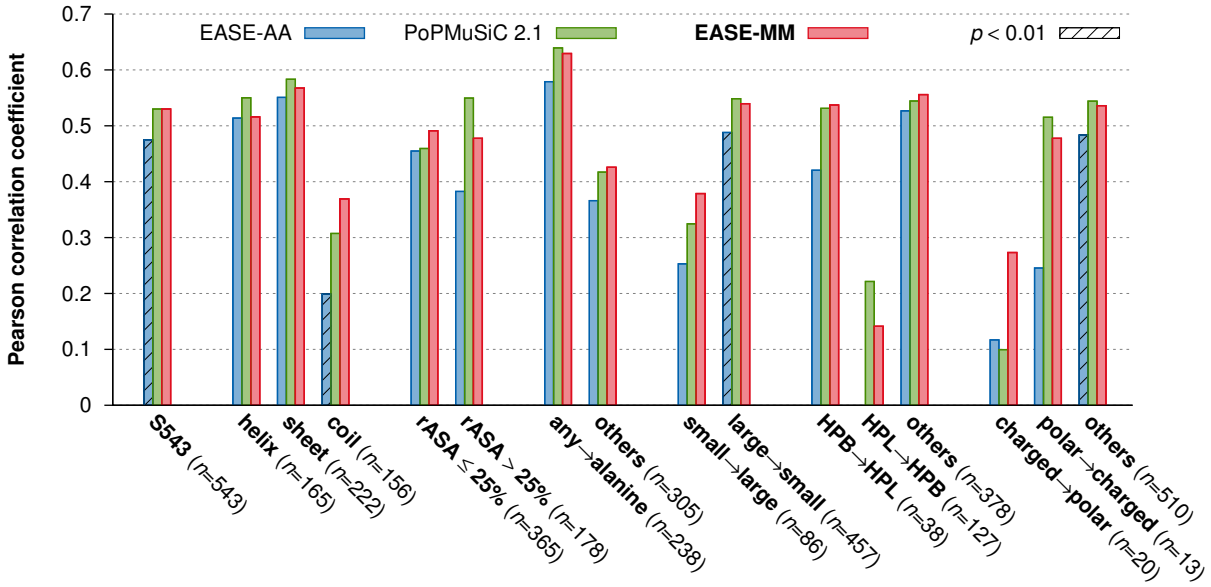

Figure 4: Pearson correlation coefficient (r) as the performance of EASE-AA, PoPMuSiC, and EASE-MM for different types of mutations from the S543 dataset. The striped bars show results which are statistically different from EASE-MM (Williams' test, $p < 0.01$). Some methods yielded a negative correlation, which is shown here as a missing bar. The secondary structure elements (helix, sheet, coil) and relative accessible surface area (rASA) of the mutation site were calculated with DSSP [19]. We also divided mutations based on the type of the wild-type and mutant amino acids (denoted as 'wild-type→mutant'). Small and large amino acids were defined based on the non-hydrogen atom counts. Amino acids were grouped based on their side-chains as hydrophobic (HPB): A, V, I, L, M, F, Y, W; polar: S, T, N, Q; charged: D, E, K, R, H; and hydrophilic (HPL): polar + charged.

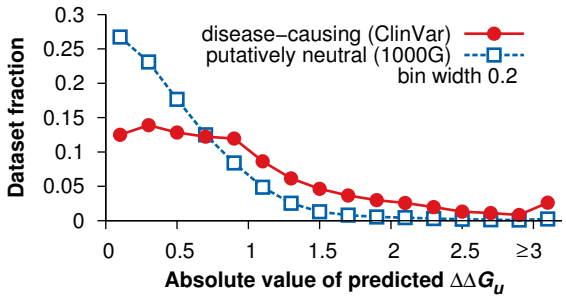

Figure 5: Distributions of the absolute value of stability changes ($\Delta\Delta G_u$) predicted with EASE-MM for 10,511 disease-causing non-synonymous single nucleotide variants (SNVs) from ClinVar [21] and 50,910 putatively neutral non-synonymous SNVs from the 1000 Genomes Project (1000G) [22] with allele frequency (AF) $\geq 1\%$. The figure shows that neutral mutations prevail for small stability changes, while disease-causing mutations are characterised by more significant changes in the protein stability.

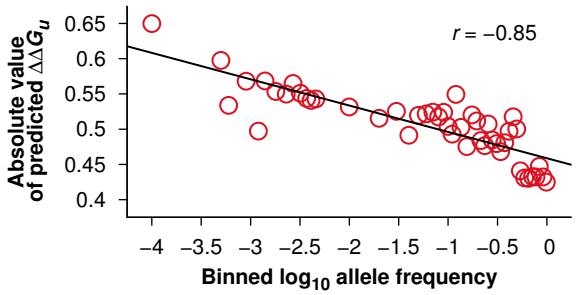

Figure 6: Average of the absolute value of stability changes ($\Delta\Delta G_u$) predicted with EASE-MM as a function of binned $\log_{10}$ allele frequency (AF) for 278,760 non-synonymous single nucleotide variants (SNVs) from the 1000 Genomes Project [22], the average being calculated for bins containing at least 500 SNVs (some bins were larger due to many mutations with the same AF value). The black line is the linear regression fit. The figure shows a strong negative correlation demonstrating that highly populated alleles tend to have smaller changes in the protein stability.

Supplementary Data

EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models

Lukas Folkman, Bela Stantic, Abdul Sattar, Yaoqi Zhou*

Supplementary Figures

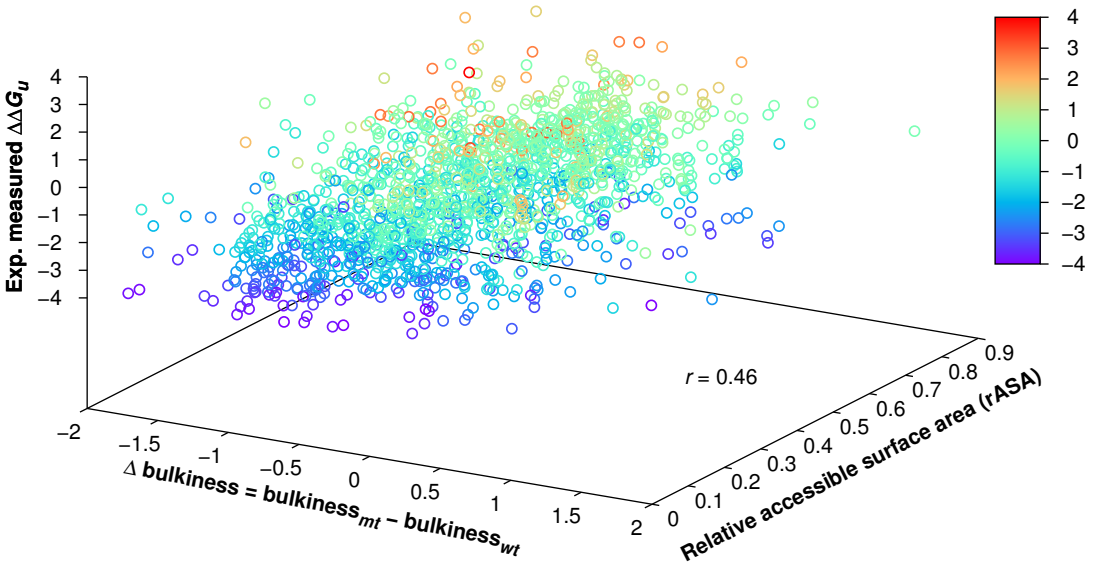

Figure S1: Experimentally measured stability changes ($\Delta\Delta G_u$) as a function of the amino acid parameter Δ bulkiness and predicted structural property relative accessible surface area (rASA) for the S1676 dataset. Δ bulkiness denotes the difference of the bulkiness of the mutant (bulkiness$_{mt}$) and wild-type (bulkiness$_{wt}$) amino acids. $\Delta\Delta G_u$ predicted based on Δ bulkiness and rASA with a *linear* support vector machine (SVM) model yielded a Pearson correlation coefficient (r) of 0.46. The figure shows that the introduction of a bulkier (relative to wild-type) amino acid in the protein core (low rASA) has a tendency to destabilise the protein structure.

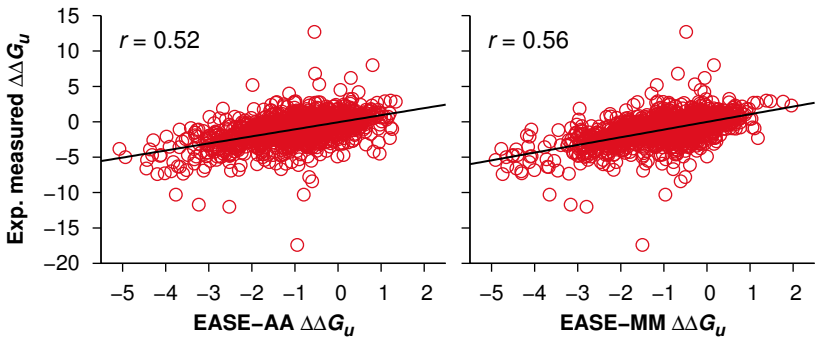

Figure S2: Experimentally measured stability changes ($\Delta\Delta G_u$) as a function of $\Delta\Delta G_u$ predicted with EASE-AA and EASE-MM for the S1676 dataset. The black lines are the linear regression fits.

*Corresponding author

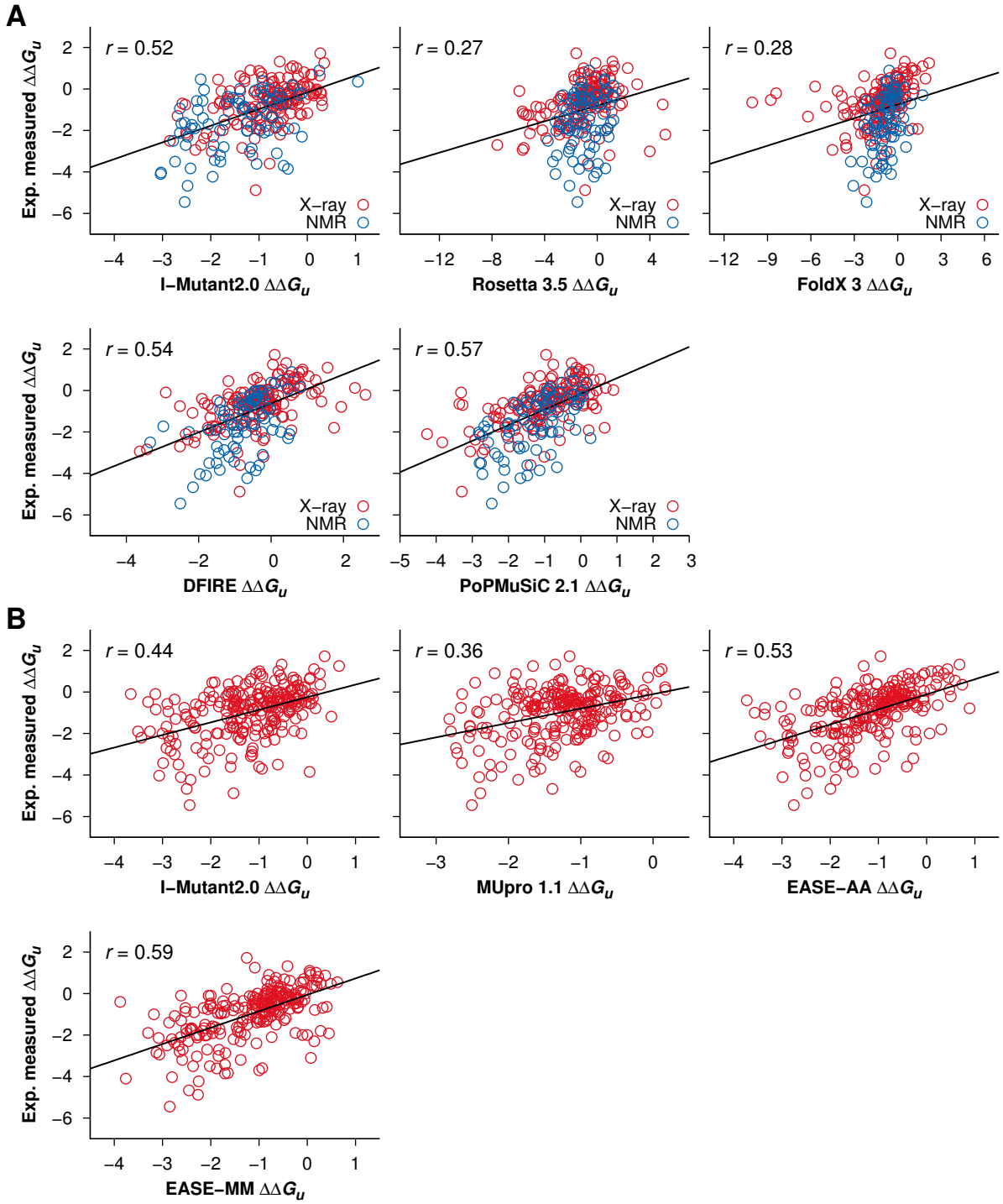

Figure S3: Experimentally measured stability changes ($\Delta\Delta G_u$) from the S236 dataset as a function of $\Delta\Delta G_u$ predicted with the five *structure-based* methods **(A)** and four *sequence-based* methods **(B)** including EASE-MM. Four predictions which caused atomic clashes during structure optimisation with Rosetta ($E_{rep} > 7$) were removed from the Rosetta plot. For the structure-based methods **(A)**, X-ray denotes predictions for proteins with high-resolution (≤ 3 Å) crystal structures (157 mutations), and NMR denotes predictions for protein structures determined with nuclear magnetic resonance (79 mutations). The black lines are the linear regression fits.

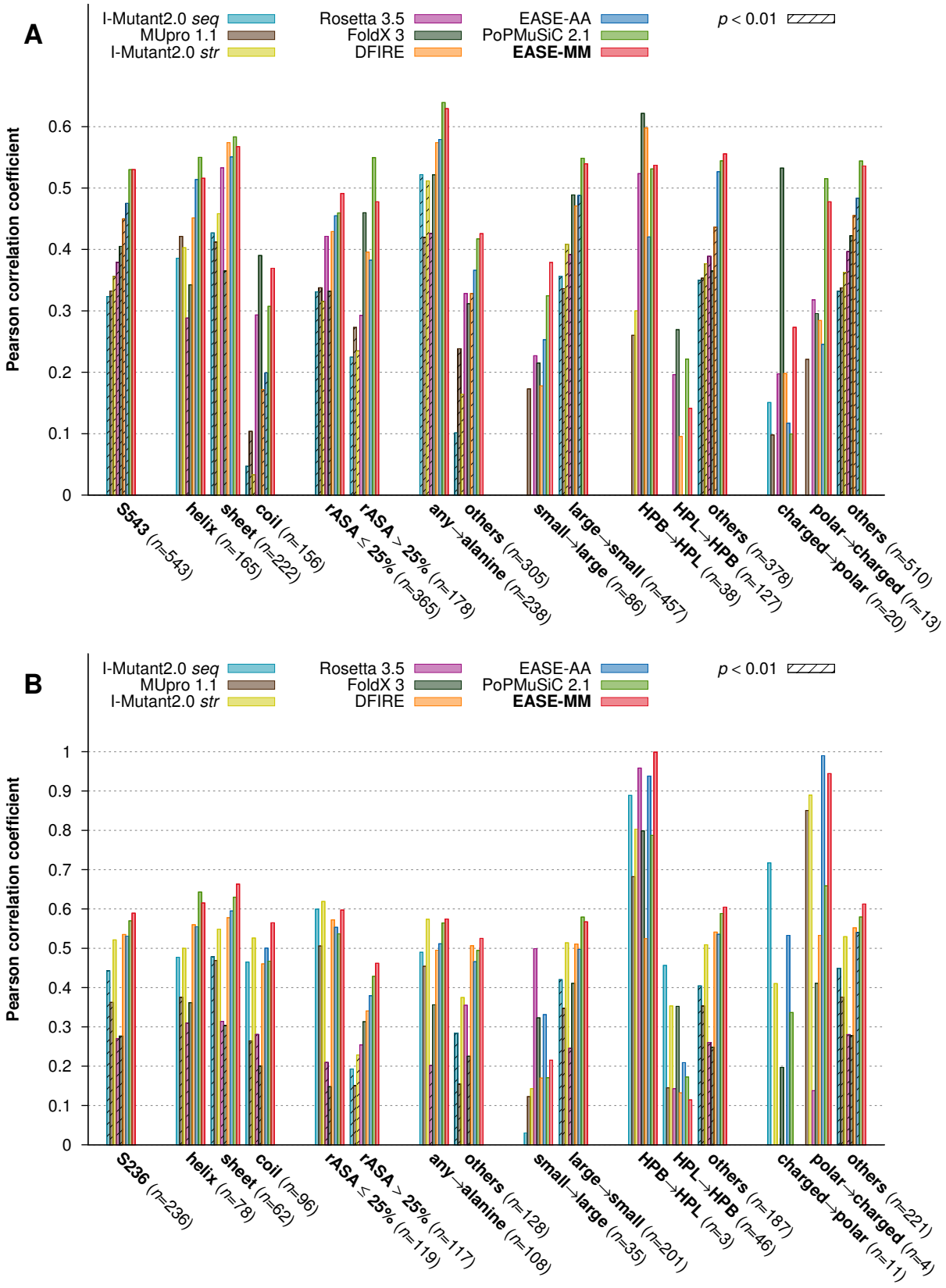

Figure S4: Pearson correlation coefficient (r) as the performance of EASE-MM and the eight compared methods for different types of mutations from the S543 **(A)** and S236 **(B)** datasets. The striped bars show results which are statistically different from EASE-MM (Williams' test, $p < 0.01$). Some methods yielded a negative correlation, which is shown here as a missing bar. The secondary structure elements (helix, sheet, coil) and relative accessible surface area (rASA) of the mutation site were calculated with DSSP [1]. We also divided mutations based on the type of the wild-type and mutant amino acids (denoted as 'wild-type→mutant'). Small and large amino acids were defined based on the non-hydrogen atom counts. Amino acids were grouped based on their side-chains as hydrophobic (HPB): A, V, I, L, M, F, Y, W; polar: S, T, N, Q; charged: D, E, K, R, H; and hydrophilic (HPL): polar + charged.

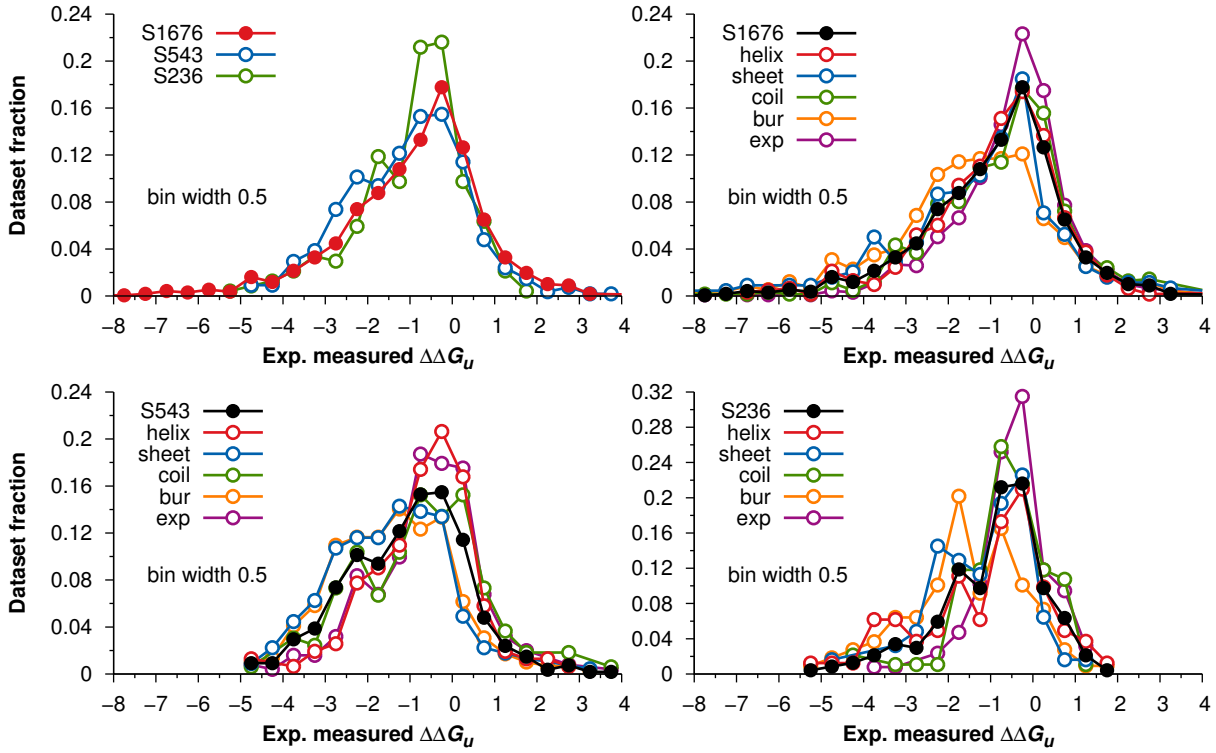

Figure S5: The distributions of the experimentally measured stability changes ($\Delta\Delta G_u$) for the three different datasets and for the five data partitions of each dataset. The five data partitions were created based on SPIDER [2] predictions.

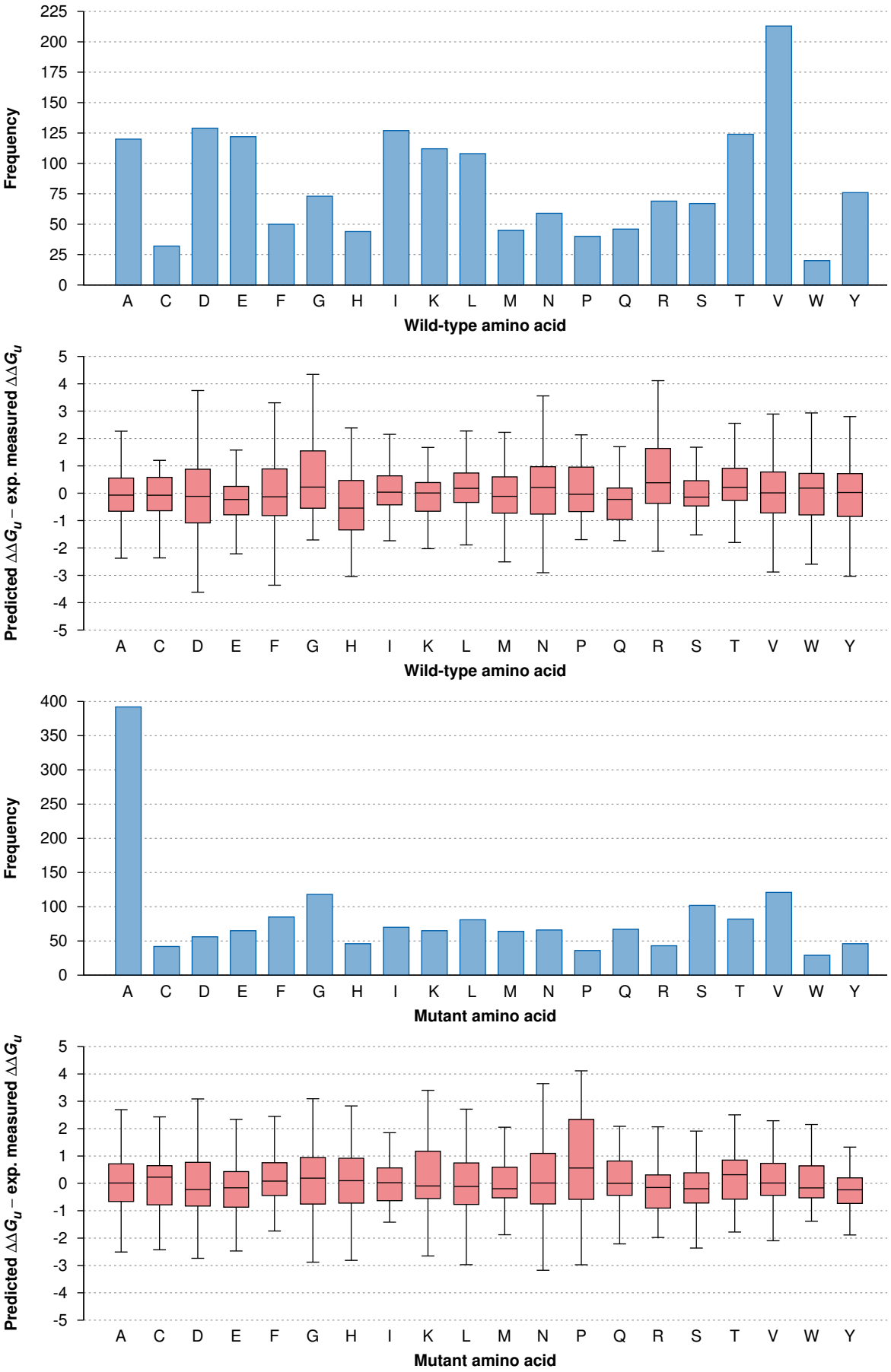

Figure S6: Frequencies and EASE-MM's prediction errors for different wild-type and mutant amino acid types from the S1676 dataset.

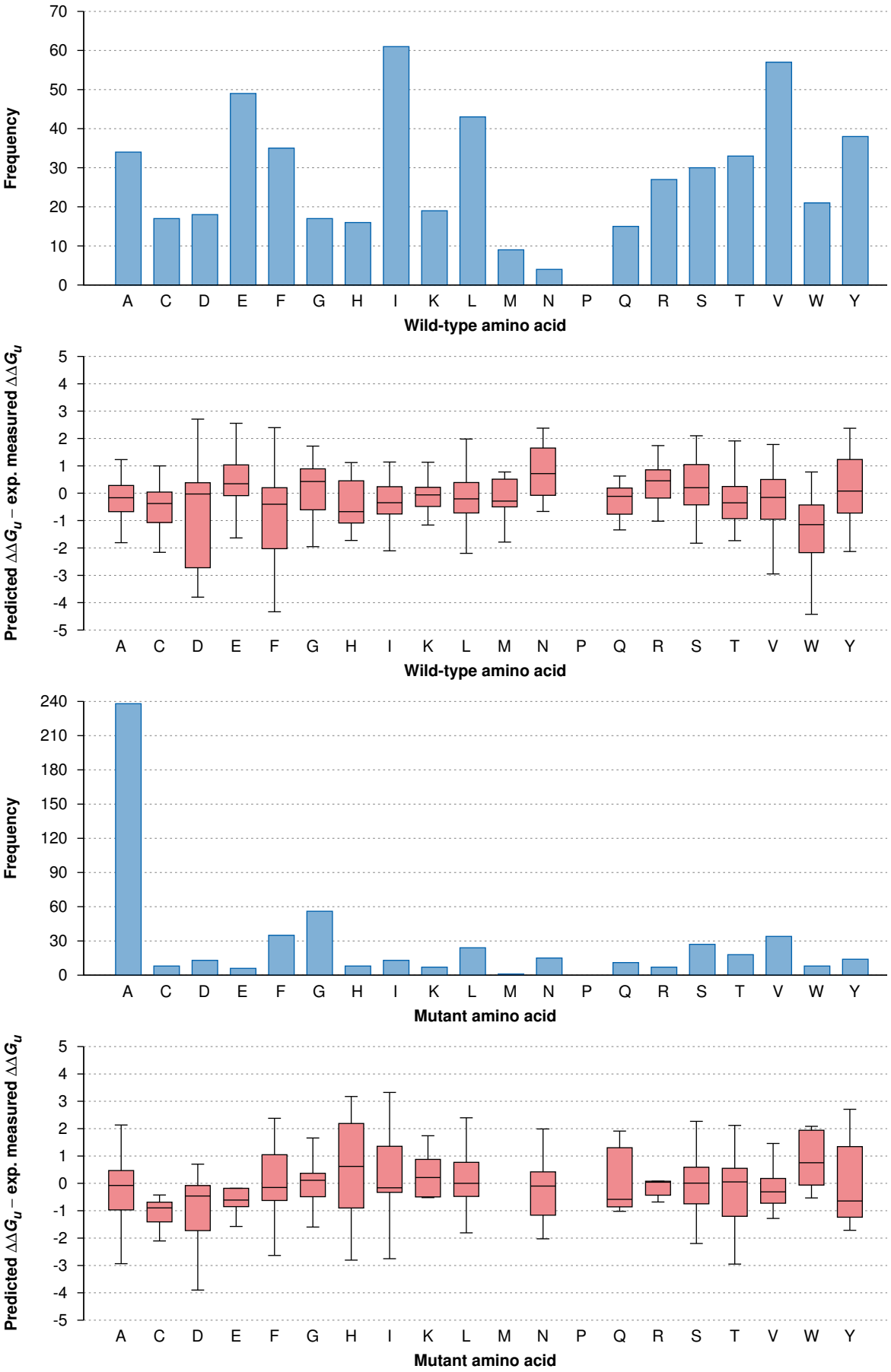

Figure S7: Frequencies and EASE-MM's prediction errors for different wild-type and mutant amino acid types from the S543 dataset.

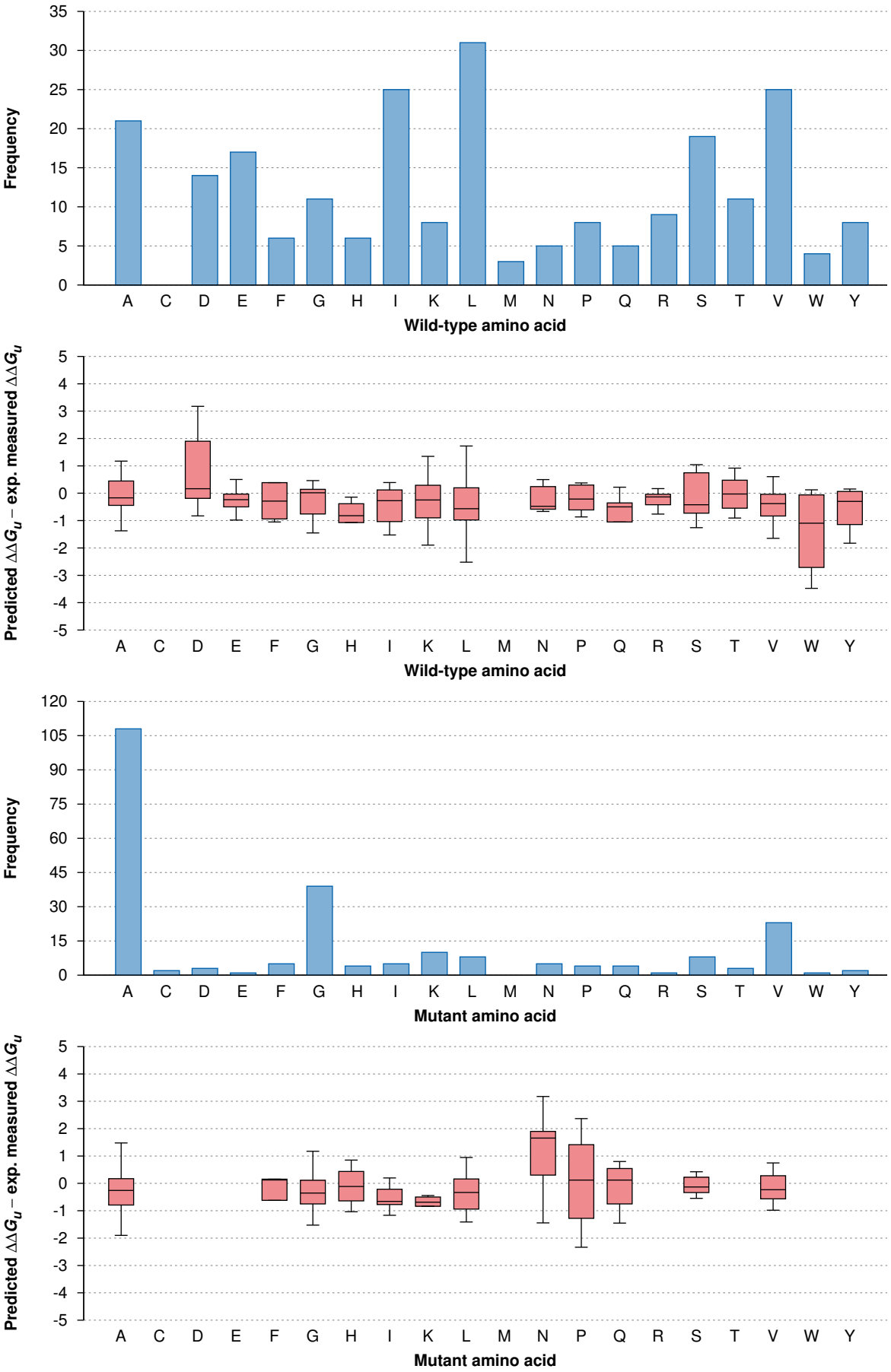

Figure S8: Frequencies and EASE-MM's prediction errors for different wild-type and mutant amino acid types from the S236 dataset.

Supplementary Tables

Table S1: Individual predictive features ranked by their correlation with experimentally measured stability changes ($\Delta\Delta G_u$) on the S1676 dataset

Feature name	r^a	p^a	**Definition**[b]
Δ bulkiness	0.348	6.7×10^{-49}	*amino acid parameter*, Gromiha *et al.* [3], Δ bulkiness = bulkiness$_{mt}$ − bulkiness$_{wt}$, Table S5
Δ hydrophobicity	0.339	3.1×10^{-46}	*amino acid parameter*, Meiler *et al.* [4], Δ hydrophobicity = hydrophobicity$_{mt}$ − hydrophobicity$_{wt}$, Table S5
Δ steric parameter	0.328	1.9×10^{-43}	*amino acid parameter*, Meiler *et al.* [4], Δ steric parameter = steric parameter$_{mt}$ − steric parameter$_{wt}$, Table S5
Δ sheet tendency	0.309	1.7×10^{-38}	*amino acid parameter*, Meiler *et al.* [4], Δ sheet tendency = sheet tendency$_{mt}$ − sheet tendency$_{wt}$, Table S5
Δ polarisability	0.279	2.6×10^{-31}	*amino acid parameter*, Meiler *et al.* [4], Δ polarisability = polarisability$_{mt}$ − polarisability$_{wt}$, Table S5
Δ PSSM	0.271	1.2×10^{-29}	*evolutionary feature*, PSSM was generated with PSI-BLAST [5]; Δ PSSM = PSSM$_{mt}$ − PSSM$_{wt}$, PSSM$_{wt}$ and PSSM$_{mt}$ are the probabilities of the wild-type and mutant amino acids at the mutation site, respectively
rASA	0.268	6.0×10^{-29}	*predicted structural property*, relative accessible surface area of the mutated residue was predicted with SPIDER [2]
Δ volume	0.265	2.0×10^{-28}	*amino acid parameter*, Meiler *et al.* [4], Δ volume = volume$_{mt}$ − volume$_{wt}$, Table S5
Δ flexibility	−0.202	6.5×10^{-17}	*amino acid parameter*, Vihinen *et al.* [6], Δ flexibility = flexibility$_{mt}$ − flexibility$_{wt}$, Table S5
PSSM$_{wt}$	−0.179	1.8×10^{-13}	*evolutionary feature*, PSSM was generated with PSI-BLAST [5]; PSSM$_{wt}$ is the probability of the wild-type amino acid at the mutation site
sheet probability	−0.132	5.3×10^{-8}	*predicted structural property*, probability that the mutation site is located in a sheet was predicted with SPIDER [2]
Δ ionisation[c]	−0.131	6.7×10^{-8}	*amino acid parameter*, Gromiha *et al.* [3], Δ ionisation = ionisation$_{mt}$ − ionisation$_{wt}$, Table S5
property entropy	−0.122	4.9×10^{-7}	*evolutionary feature*, overall conservation of the mutation site expressed as property entropy with respect to six amino acid 'property' groups [7]; the property entropy was calculated from a multiple sequence alignment of the 30 most similar sequences ranked by *e*-value with PSI-BLAST [5] (see Materials and Methods)
Δ compressibility	−0.091	1.8×10^{-4}	*amino acid parameter*, Gromiha *et al.* [3], Δ compressibility = compressibility$_{mt}$ − compressibility$_{wt}$, Table S5
coil probability	0.084	5.9×10^{-4}	*predicted structural property*, probability that the mutation site is located in a coil was predicted with SPIDER [2]
Δ isoelectric point	0.067	6.0×10^{-3}	*amino acid parameter*, Meiler *et al.* [4], Δ isoelectric point = isoelectric point$_{mt}$ − isoelectric point$_{wt}$, Table S5
Δ helix tendency	0.054	0.026	*amino acid parameter*, Meiler *et al.* [4], Δ helix tendency = helix tendency$_{mt}$ − helix tendency$_{wt}$, Table S5
helix probability	0.040	0.100	*predicted structural property*, probability that the mutation site is located in a helix was predicted with SPIDER [2]
disorder probability	0.022	0.361	*predicted structural property*, probability that the mutation site is in a disordered region of the protein was predicted with SPINE-D [8]

[a] r, Pearson correlation coefficient; p, probability that r is different from 0 due to random chance.
[b] *wt* and *mt* refer to the wild-type and mutant amino acids, respectively.
[c] equilibrium constant with reference to the ionisation property of COOH group

Table S2: Predictive features selected with the sequential forward floating selection algorithm for the five models of EASE-MM, ranked by their contributions to the respective models

		r decrease upon removing[b]		
Model	**Feature**[a]	**Relative**	**Absolute**	r **(single feature)**[c]
helix	rASA[d]	23.7%	0.117	0.295
	Δ helix tendency	9.2%	0.046	0.095
	Δ volume	4.4%	0.022	0.278
	Δ bulkiness	3.8%	0.019	0.321
	Δ compressibility	3.4%	0.017	0.160
	Δ isoelectric point	2.2%	0.011	0.193
	helix probability	0.3%	0.002	0.186
	coil probability	0.0%	0.000	0.182
	features combined			0.495
sheet	Δ PSSM[e]	5.3%	0.033	0.314
	Δ volume	4.7%	0.029	0.443
	Δ hydrophobicity	4.6%	0.029	0.449
	Δ compressibility	3.7%	0.023	0.109
	Δ helix tendency	1.7%	0.011	0.075
	sheet probability	0.9%	0.005	0.119
	coil probability	0.5%	0.003	0.002
	Δ steric parameter	0.2%	0.001	0.503
	disorder probability	0.2%	0.001	0.091
	Δ bulkiness	0.1%	0.000	0.533
	features combined			0.618
coil	Δ hydrophobicity	19.5%	0.087	0.233
	Δ flexibility	5.7%	0.026	0.227
	rASA[d]	3.4%	0.015	0.212
	Δ polarisability	2.2%	0.010	0.045
	Δ PSSM[e]	1.5%	0.007	0.143
	sheet probability	1.1%	0.005	0.129
	$PSSM_{wt}$ [e]	0.9%	0.004	0.063
	coil probability	0.5%	0.002	0.219
	Δ volume	0.3%	0.001	0.092
	features combined			0.449
buried	Δ isoelectric point	6.0%	0.037	0.089
	Δ bulkiness	5.6%	0.034	0.514
	Δ PSSM[e]	4.4%	0.027	0.274
	rASA[d]	2.7%	0.016	0.135
	Δ polarisability	1.7%	0.010	0.434
	Δ volume	1.5%	0.009	0.428
	Δ flexibility	1.0%	0.006	0.262
	Δ sheet tendency	0.8%	0.005	0.410
	features combined			0.612
exposed	Δ volume	19.2%	0.071	0.076
	helix probability	15.9%	0.059	0.008
	rASA[d]	6.8%	0.025	0.107
	Δ hydrophobicity	6.5%	0.024	0.141
	sheet probability	6.3%	0.023	0.075
	Δ helix tendency	4.4%	0.016	0.004
	Δ flexibility	2.2%	0.008	0.015
	$PSSM_{wt}$ [e]	1.3%	0.005	0.097
	features combined			0.370

[a] Δ, the change between the mutant and wild-type amino acids
[b] Decrease in Pearson correlation coefficient (r) for the given data partition (*e.g.*, helix) upon removing the given feature from the given model (*e.g.*, *helix*)
[c] Pearson correlation coefficient (r) of a single feature for the given data partition (*e.g.*, helix)
[d] rASA, relative accessible surface area
[e] Δ PSSM = $PSSM_{mt} - PSSM_{wt}$; $PSSM_{wt}$, PSSM probability of the wild-type amino acids; $PSSM_{mt}$, PSSM probability of the mutant amino acids; PSSM, position-specific scoring matrix

Table S3: Comparison of the prediction performance when swapping the five different models of EASE-MM and their corresponding data partitions on the S1676 dataset

	S1676 data partition									
	helix		**sheet**		**coil**		**buried**		**exposed**	
Model	r^a	p^b	r^a	p^b	r^a	p^b	r^a	p^b	r^a	p^b
helix	**0.50**	—	0.55	2.5×10^{-3}	0.37	9.3×10^{-3}	—	—	—	—
sheet	0.38	1.0×10^{-4}	**0.62**	—	0.38	7.7×10^{-3}	—	—	—	—
coil	0.40	4.2×10^{-4}	0.53	1.8×10^{-4}	**0.45**	—	—	—	—	—
buried	—	—	—	—	—	—	**0.61**	—	0.21	1.1×10^{-6}
exposed	—	—	—	—	—	—	0.51	5.7×10^{-7}	**0.37**	

[a] r, Pearson correlation coefficient; correlation coefficients of the 'matching' models (*i.e.*, the *helix* model for the helix data partition) are highlighted in bold.
[b] p, probability that the correlation coefficients (r) of the given model and the 'matching' model (*i.e.*, the *helix* model for the helix data partition) are different due to random chance (Williams' test for comparing correlation coefficients).

Table S4: Comparison of the prediction performance of EASE-MM when the structural properties are predicted from the sequence with SPIDER, calculated from the structure with DSSP, or drawn randomly.

Method	**Dataset**	**SS[a] and ASA[a]**	r^a	p^a	**RMSE[a]**
EASE-MM	S543	SPIDER[b]	0.53	—	1.22
		DSSP[c]	0.53	0.973	1.24
		random[d]	0.36	1.1×10^{-7}	1.36
	S236	SPIDER[b]	0.59	—	1.03
		DSSP[c]	0.57	0.446	1.06
		random[d]	0.31	2.2×10^{-3}	1.27

[a] SS, secondary structure; ASA, accessible surface area; r, Pearson correlation coefficient; p, probability that the correlation coefficients (r) of the given method and that of EASE-MM based on SPIDER are different due to random chance (Williams' test for comparing correlation coefficients); RMSE, root mean square error.
[b] SS and ASA were predicted from the protein *sequence* using SPIDER [2].
[c] SS and ASA were calculated from the protein *structure* using DSSP [1].
[d] The tests were repeated ten times, each time with *randomly drawn* SS and ASA; results were averaged.

Table S5: Scaled values of the 11 amino acid parameters which were implemented as candidate predictive features

AA[a]	**H[b]**	**V[b]**	**P[b]**	**IP[b]**	**HT[b]**	**ST[b]**	**GSI[b]**	**F_0[b]**	**F_1[b]**	**F_2[b]**	**C[b]**	**B[b]**	**EC[b]**
Ala	−0.171	−0.677	−0.680	−0.170	0.900	−0.476	−0.350	−0.044	−0.234	−0.269	0.587	−0.099	0.829
Asp	−0.767	−0.281	−0.417	−0.900	−0.155	−0.635	−0.213	−0.103	0.900	0.014	−0.475	−0.082	0.247
Cys	0.508	−0.359	−0.329	−0.114	−0.652	0.476	−0.140	−0.642	−0.773	−0.035	−0.433	0.094	−0.388
Glu	−0.696	−0.058	−0.241	−0.868	0.900	−0.582	−0.230	0.347	0.480	0.021	−0.900	0.105	0.565
Phe	0.646	0.412	0.373	−0.272	0.155	0.318	0.363	−0.863	−0.504	−0.113	−0.673	0.721	0.035
Gly	−0.342	−0.900	−0.900	−0.179	−0.900	−0.900	−0.900	0.701	0.527	−0.050	0.378	−0.900	0.829
His	−0.271	0.138	0.110	0.195	−0.031	−0.106	0.384	−0.480	−0.186	−0.255	−0.297	0.115	−0.088
Ile	0.652	−0.009	−0.066	−0.186	0.155	0.688	0.900	−0.332	−0.662	−0.411	−0.288	0.879	−0.900
Lys	−0.889	0.163	0.066	0.727	0.279	−0.265	−0.088	0.339	0.844	0.900	−0.375	0.317	0.547
Leu	0.596	−0.009	−0.066	−0.186	0.714	−0.053	0.213	−0.590	−0.115	−0.064	−0.288	0.879	0.865
Met	0.337	0.087	0.066	−0.262	0.652	−0.001	0.110	−0.738	−0.900	−0.893	−0.205	0.370	0.724
Asn	−0.674	−0.243	−0.329	−0.075	−0.403	−0.529	−0.213	0.516	0.242	0.000	−0.166	0.031	0.265
Pro	0.055	−0.294	−0.900	−0.010	−0.900	0.106	0.247	0.059	0.868	0.014	0.900	0.487	0.212
Gln	−0.464	−0.020	−0.110	−0.276	0.528	−0.371	−0.230	0.870	0.416	−0.319	−0.403	0.192	0.529
Arg	−0.900	0.466	0.373	0.900	0.528	−0.371	0.105	−0.066	0.416	−0.206	0.430	0.175	−0.106
Ser	−0.364	−0.544	−0.637	−0.265	−0.466	−0.212	−0.337	0.900	0.575	−0.050	−0.024	−0.300	0.600
Thr	−0.199	−0.321	−0.417	−0.288	−0.403	0.212	0.402	0.192	0.599	0.028	−0.212	0.323	0.406
Val	0.331	−0.232	−0.285	−0.191	−0.031	0.900	0.677	−0.480	−0.385	−0.120	−0.127	0.896	0.794
Trp	0.900	0.900	0.900	−0.209	0.279	0.529	0.479	−0.900	−0.464	−0.900	−0.074	0.900	0.900
Tyr	0.188	0.541	0.417	−0.274	−0.155	0.476	0.363	−0.634	−0.361	−0.659	−0.738	0.546	0.582

[a] AA denotes an amino acid in the standard three-letter code.
[b] H, hydrophobicity; V, volume; P, polarisability; IP, isoelectric point; HT, helix tendency; ST, sheet tendency; GSI, graph shape index (steric parameter); F_0, flexibility with no rigid neighbours; F_1, flexibility with one rigid neighbour; F_2, flexibility with two rigid neighbours; C, compressibility; B, bulkiness; and EC, equilibrium constant with reference to the ionisation property of COOH group.

References

[1] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.

[2] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific Reports 5 (2015) 11476.

[3] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, Journal of Protein Chemistry 18 (5) (1999) 565–578.

[4] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, Molecular modeling annual 7 (9) (2001) 360–369.

[5] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research 25 (17) (1997) 3389.

[6] M. Vihinen, E. Torkkila, P. Riikonen, Accuracy of protein flexibility predictions, Proteins: Structure, Function, and Bioinformatics 19 (2) (1994) 141–149.

[7] L. A. Mirny, E. I. Shakhnovich, Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, Journal of Molecular Biology 291 (1) (1999) 177–196.

[8] T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, Y. Zhou, SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method, Journal of Biomolecular Structure and Dynamics 29 (4) (2012) 799–813.