REPUBLIC OF TURKEY
ACIBADEM MEHMET ALİ AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

SYSTEMATIC UNDER-SAMPLING OF MUTATION DATASETS AND COMPARATIVE ASSESSMENT OF PROTEIN STABILITY PREDICTORS

NAROD KEBABCI

MASTER THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Assoc. Prof. Emel TİMUÇİN

ISTANBUL – 2021

REPUBLIC OF TURKEY
ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

SYSTEMATIC UNDER-SAMPLING OF MUTATION DATASETS AND COMPARATIVE ASSESSMENT OF PROTEIN STABILITY PREDICTORS

NAROD KEBABCI

MASTER THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Assoc. Prof. Emel TİMUÇİN

ISTANBUL – 2021

DECLARATION

I declare that this thesis study is my own study; I had no unethical behavior at any stage from planning to writing, I obtained all the information in this thesis following academic and ethical guidelines, I cited all the information and comments that were not acquired with this thesis work, and I provided resources in the list of references. I also declare no violation of any patents and copyrights during this thesis's study and writing.

14.06.2021
Narod Kebabcı

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Assoc. Prof. Emel Timucin, not only for being a great mentor but also for being a generous and sincere friend. She was always there with a solution when I doubted myself, struggling against errors. She always encouraged me to make my own decisions on this long path. Without her inspiration and guidance, I would not reveal my capabilities in our work. Besides learning how to conduct a study and question the findings, she also taught me how to be a reasonable and modest team leader. I will always admire her point of view. Her company was one of the greatest chances I had during my post-graduate studies.

I further extend my gratitude to the head of our department, Prof. Ugur Sezerman. He accepted me to the program firsthand. He believed in my wet-lab and computational experiment capabilities and offered me a chance to work in Eternans Ltd during my post-graduate studies. He has always provided moral support in the most challenging situations during this journey. He is one of the reasons I like machine learning algorithms.

I cannot forget my fellow labmates' help during this journey. Together, we teamed up to study lectures, brainstormed to solve challenging issues, and conducted experiments within our projects. Thank you for all the fun we have had in the last three years. It was a great pleasure to have you as friends.

And lastly, I would like to express my deepest thanks to my parents and sister; they always supported me in my decisions and faith in me through this journey.

TABLE OF CONTENTS

LIST OF ABBREVATIONS AND SYMBOLS

ΔΔG	Gibbs Free Energy
ALI	Aliphatic Amino Acids
ARO	Aromatic Amino Acids
ASA	Relative solvent accessible surface area
AUC	Area Under the Curve
CHA	Charged Amino Acids
DSC	Differential Scanning Calorimetry
MAE	Mean Absolute Error
MCC	Mathews Correlation Coefficient
MSE	Mean Signed Error
PCC	Pearson Correlation Coefficient
POL	Polar Amino Acids
PDB	Protein Data Bank
ROC	Receiver Operating Characteristic
SS	Secondary Structure
SVM	Support Vector Machine

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Predicting how an amino acid substitution affects protein stability is an important task, refining our understanding of protein folding. Although mutagenesis studies performed either in a wet- or a dry-lab contribute to this knowledge, computational methods stand out as more economical and less time-consuming. Thus, numerous computational methods have been developed in the last decade to predict the change in folding free energy upon mutation (ΔΔG). Alongside their advantageous features, computational ΔΔG predictors still suffer from certain limitations, such as overfitting. Overfitting of ΔΔG predictors, *i.e.*, the tendency to produce biased results toward their training sets, becomes a critical issue for developing accurate predictors. Indeed, the existing biases in mutation datasets, such as redundancy of destabilizing mutations and/or certain amino acids, are recognized as a primary source of overfitting. In this study, we developed a systematic under-sampling methodology to eliminate redundancies in the mutation datasets. PON-tstab, composed of 1564 mutations, was used as the toy dataset. We used two different reduced alphabets to group the mutant amino acids based on their sidechain biochemistry and secondary structure. This reduced set led to a more uniform sampling of each amino acid and of destabilizing-stabilizing mutations. We ultimately tested the performances of 11 different ΔΔG predictors on the entire PON-tstab and the constructed subset. We note slight differences in performances of almost every predictor, reflecting the influence of redundancy elimination. Overall, we underline the importance of producing high-quality mutation datasets for tackling overfitting issues and a reliable assessment of ΔΔG predictors. (The code is available on GitHub: https://github.com/narodkebabci/gRoR)

Keywords: Bias, Mutation, Protein Folding, Protein Stability, PON-Tstab

ÖZET

Mutasyon Veri Setlerinin Sistematik Olarak Örneklenmesi ve Protein Kararlılık Prediktörlerinin Karşılaştırmalı Değerlendirmesi

Protein katlanmalarını daha iyi anlamak için amino asit değişimlerinin protein stabilitesi üzerine etkisini doğru şekilde tahmin etmek çok önemlidir. Birçok mutagenez çalışması bu konuya katkıda bulunsa da deneysel yöntemler, hesaplamalı yöntemlere kıyasla daha uzun soluklu ve pahalıdırlar. Bu sebeple, geçtiğimiz 20 yılda, yanlış anlamlı mutasyonun sebep olduğu protein stabilitesindeki değişimi hesaplamak için birçok prediktör geliştirilmiştir. Bu yöntemlerin birçok avantajlı özelliği olmasına rağmen modellendikleri eğitim ve test verisindeki düzensizlikler aşırı öğrenme gibi çeşitli problemlerle karşılaşmalarına sebep olmaktadır. Veri setinde, özellikle proteinin stabilizasyonunu azaltan mutasyonların sayıca daha çok olması ve belirli amino asit mutantlarının fazlalığı aşırı öğrenmeye sebep olmaktadır. Bu çalışmada, öncelikle, geliştirdiğimiz örnekleme yöntemi ile 1564 mutasyondan oluşan PON-tstab veri setindeki fazlalıkları azaltmayı hedefledik. İki farklı alfabe kullanarak amino asitleri biyokimyasal ve ikincil yapı özelliklerine göre yeniden grupladık. Oluşturduğumuz veri setleri ile mutant amino asit tiplerinin daha eşit dağılmasını sağladık. Daha sonra, proteinin kararlılığını tahmin eden 11 farklı prediktörün performansını, PON-tstab ve oluşturduğumuz indirgenmiş veri setleri üzerine test ettik. İndirgenmiş veri setlerinin orijinal veri setine kıyasla daha iyi performans gösterdiğini kaydettik. Özetle, bu çalışma ile aşırı öğrenmeyi engellemek ve prediktörleri doğru bir şekilde test etmek için kaliteli veri setleri oluşturmanın önemini vurguladık. (Yöntem GitHub'da mevcuttur: https://github.com/narodkebabci/gRoR)

Anahtar Sözcükler: Mutasyon, PON-Tstab, Protein Katlanması, Protein Stabilizasyonu, Yanlılık

1. BACKGROUND AND AIM OF THE STUDY

Accurate prediction of protein stability changes (ΔΔG) caused by amino acid substitution is essential to clarify the relationship between the structure, function, and dynamic of that protein. Many mutagenesis studies performed either in a wet- or a dry-lab contributes to this knowledge. Traditional studies can reliably measure the thermodynamics effects of mutations via thermal or chemical denaturation experiments. However, these experimental studies often arduous, require much time and cost, especially for large-scale mutations. On the other hand, computational approaches stand out as more economical and less time-saving alternatives to experimental methods. As a result of their advantages, numerous computational methods have been developed over the last decade for predicting the change in free energy of folding (ΔΔG) upon amino acid substitution.

Recently, computational approaches utilizing machine learning applications have been developed to train models on mutation datasets and calculate the ΔΔG value. Besides their beneficial features, computational predictors still suffer from various limitations. Besides their beneficial features, machine learning-based predictors still suffer from various limitations. Overfitting issues due to an imbalance in training datasets and insufficiency in predictive features used to build the model significantly increase the risk of producing biased results. Consequently, without addressing a solution to overfitting problems, it's pointless to develop a new method. Furthermore, it's equally critical to comparatively assess the existing stability predictors to uncover the weaknesses of some and the strength of others. Yet, such evaluation studies can lead to overestimated or underestimated performance of the predictors, as blind datasets are not utilized for evaluation. Previously, some problems were associated with mutation datasets, such as the asymmetry between destabilizing and stabilizing mutations and/or the relatively higher abundance of particular amino acids than others (1–3). The dominance of destabilizing mutations within the training sets encourages the tendency to compute destabilizing mutations better than those that stabilize them

(4–7). Moreover, an exceptionally high abundance of alanine in the mutant positions of almost all datasets would generate a similar dependence of the predictors on the type of amino acids in the native or mutant position. Some predictors' performance may vary depending on the kind of protein and quality of the wild-type structure used (8–10). Given the fact that the existing biases in the mutation datasets are recognized as primary sources of overfitting, many efforts were dedicated to eliminating those redundancies in the variation datasets (9,11–13). One recent and successful example is the generation of the symmetrical dataset of S^{sym}, which overcomes the asymmetry in the frequencies of destabilizing and stabilizing mutations (9). Studies to identify bias in current predictors highlighted the importance of symmetry, stating that having both the direct and inverse version of each mutation may balance the dataset. However, they reveal that reverse mutations are not clearly the opposite outcomes of the forward mutations. Another study uses a reduced amino acid alphabet (14) to measure the bias against poorly represented mutation types (13). Still, randomly selecting data points from the reduced groups to create a subset was unsuitable for distributing the mutation types equally. Moreover, the non-uniform distribution of mutants' features can prevent a diligent curation of the dataset when generating a random subset. Aside the contribution of these relatively biased-free datasets, we stress that there still needs a systematic approach that can sample already obtained mutation datasets to include the minimum number of redundant mutations.

Here, we describe a systematic way of under-sampling to eliminate redundancy regarding the stability, biochemical and structural properties of mutations in the datasets. Using a reduced amino acid alphabet led to a more uniform sampling of ΔΔG, specific amino acids and destabilizing-stabilizing mutations, which are recognized as significant overfitting sources. We tested 11 protein stability predictors' performance on the PON-tstab and our balanced subsets constructed from the PON-tstab dataset (15). Eventually, our study highlights the importance of compiling high-quality mutation datasets to battle overfitting issues and a reliable assessment of ΔΔG predictors.

2. INTRODUCTION

2.1. Protein Stability and Folding

Protein stability is explained by the free energy difference between the folded and unfolded states. A newly synthesized or denatured polypeptide chain in unfolded state shifts randomly among various conformations before reaching its folded state. Of the many possible conformations with similar energy, the polypeptide chain forms the most stable conformation, namely native structure, which has the lowest Gibbs free energy (16,17). Although the polypeptide chain forms secondary structures to gain its native conformation, some parts of the folded chain may remain unfolded to act more flexibly. Such flexibility is restricted in the native state due to the stringent covalent bonds. Yet, plentiful noncovalent bonds allow a certain degree of flexibility, make protein marginally stable (18).

Numerous mechanisms play a role in protein folding, such as hydrogen bonds, van der Waals interactions, electrostatic interactions among ionized side chains, and salt bridges. Protein gains enthalpy from the atom-field effects within covalent and noncovalent interactions as atoms share electrons by hydrogen or van der Waals interactions (19). Besides intramolecular interactions, changes in entropy also induce the folding state of a protein. Since hydrophobic side chains are easily exposed to solvent when the protein is in the unfolded state, hydrogen bonds built a solvation layer of water around the protein. Such a situation lowers the entropy of the water due to the change in the freedom of movement (20). As the protein folds, these hydrogen bonds are broken and released water molecules. In short, the enthalpy in the native state becomes favorable when the polypeptide chain is folded and the gain in solvent entropy almost compensates for the loss of conformational entropy.

Previous studies show that hydrophobic effects are the major driver of protein folding by solidly packing protein cores. On the other hand, the interaction of oppositely charged groups may either stabilize or destabilize the folded state of proteins, depending on the environment's polarity level. Therefore, folding can be defined as the balance between forces and interactions within the protein to minimize solvent exposure of the nonpolar regions.

2.1.1. Folding of barnase

Barnase is one of the smallest globular proteins (110 amino acid residue) that has usually been used in thermal denaturation studies (21). It is an α + β protein with three α-helices and five strands of antiparallel β sheet. One of the helices out of three is the main secondary structure in the hydrophobic core, which is packed against the sheet. Especially, hydrogen bonds between CO groups and NH groups of residues from 6 to 18 drive the packing of major α-helix (22). Furthermore, two Asp residues in the solvent-exposed face of the major α-helix form salt-bridge with Arg110. This electrostatic interaction is critical since unpaired charges may significantly destabilize the structure (23). On the other hand, the formation of the β-sheet leads to the burial of solvent-accessible hydrophobic surface area (22).

Several thermodynamics and mutagenesis studies calculated free energy of folding by thermal and chemical denaturation methods. Overall, hydration of polar groups and increase of configurational entropy destabilize the structure of barnase while formation of van der Waals interactions and hydration of nonpolar groups stabilize it (21).

2.2. Relationship Between Point Mutations and Protein Stability

Many impulsive mutations are point mutations, defined as alterations in a single base pair in the DNA sequence. A point mutation can be nonsynonymous or synonymous. Mutations that introduce a stop codon are called a non-sense mutation, while a single base variation resulting in a single amino acid change is called a missense mutation. Additionally, frameshift mutations such as deletion or insertion of bases into a gene may alter the reading frame. Overall, these three types of mutations are called nonsynonymous mutations. On the other hand, synonymous mutations, which are also called silent mutations, do not change the resulting amino acid sequence but the codon (e.g., GUU to GUC; both encode Valine) (24).

Point mutations, especially those that cause single amino acid substitution, can affect the folding state and stability of the protein through various reasons such as changes in backbone strain, over-packing, reduction in hydrophobicity, and loss of electrostatic interaction or a disulfide bridge (25,26). For example, a point mutation that result in a substitution into a hydrophilic residue in the protein's hydrophobic core may change the interaction among amino acids within a protein or effects the system's entropy (26). Therefore, it's vital to examine the stability effects of mutations to understand the dreadful alterations in the protein's flexibility and conformational dynamics.

The change in free energy of folding ($\Delta\Delta G$) between a native protein and its mutant is quantified as a metric to explain whether the mutation stabilizes or destabilizes. Most point mutations are destabilizing and may lead to the development of several diseases such as genetic disorders, cancers, and neurodegenerative diseases (27). Indeed, substitution from a charged to a polar amino acid or from an aromatic to an aliphatic enhances the destabilization level. Besides, some point mutations can also lead to disease, although they elevate the stability (28).

In summary, nonsynonymous mutations may drastically change a protein's molecular function, such as ligand binding, catalysis, and allosteric regulation. Furthermore, point mutations can reduce the concentration of the protein as a result of unsuccessfully folded structures. Thus, it's essential to understand the stability effects of point mutations to examine the relationship between the structure and the function of a protein.

2.3. Experimental Methods to Predict the Folding Free Energy

Thermal and solvent denaturation are two main experimental methods to measure the change in protein's free energy of folding upon mutation. A folded protein can denature and lose its tertiary structure to a polypeptide chain when introduced with agents such as heat, urea, and guanidinium chloride (GdmCl) or with applied changes in pressure or pH in the environment (29). Next, when conditions switch to a more favorable environment, protein may regain its native structure; still, in some cases, even though the denaturant has been removed, this process can lead to aggregation or precipitation of the protein. Overall, change in stability is accurately calculated by the difference in Gibbs free energy between the native and denatured states of the protein.

The main methods to calculate the thermodynamic stability in wet-lab are Circular Dichroism, Differential Scanning Calorimetry (DSC), chemical denaturation, and several spectroscopic measurements. Before applying thermal or chemical unfolding to assess the protein stability, more than one step should be performed in many cases, such as protein expression, purification, and mutagenesis (30). Nevertheless, these methods require much time and cost for the entire process, making them arduous and less favorable (31). Therefore, indeed it’s critical to use computational algorithms to screen the $\Delta\Delta G$ effects upon mutation.

2.4. Computational Methods to Predict the Folding Free Energy

In the past 20 years, numerous computational methods that predict the change in the thermodynamic stability of proteins upon missense mutations have been developed. Among these tools, FoldX is one of the oldest empirical potential approaches that use an energy function acquired by a weighted combination of physical energy terms such as hydrogen bonds and electrostatic interactions, statistical energy terms, and structural descriptors (32,33). Although FoldX is still widely used to predict ΔΔG effects of mutations, various tools developed within previous years using other energy functions (physical potential approaches, statistical potential approaches) and machine learning approaches. The physical potential approaches derived from molecular mechanics use atomic force fields to calculate ΔΔG, making it inapplicable for large-scale mutation datasets as the simulating requires massive computational power (34,35). On the other hand, statistical potential approaches perform statistical analysis on environmental propensities, substitution, and occurrence frequencies of the 20 amino acids to make predictions from the database of known protein structures (36,37). A well-known method that uses statistical potential is SDM. It computes the change in protein stability with the environment-specific amino acid substitution tables.

As the availability level of large mutation databases that consist of experimentally measured ΔΔGs extended, the number of newly developed machine learning approaches grows. Numerous subsets of ProTherm (38) are used to train these methods to create a model for computing ΔΔG prediction. One of the earliest such predictors is I-Mutant2.0 (39). It uses a support vector machine (SVM) that considers amino acid substitutions and accessible surface area. Similarly, other predictors combine support vector machines with the random-forest algorithm or use graph-based structural signatures to encode distance patterns between atoms. One of the known random forest algorithm-based predictor is PROTS-RF, which used 41 different features, including secondary information and solvent accessibility, to develop its model (40). Recently,

a neural network-based predictor, DeepDDG, is developed, which considers the local environment of the mutation as a reflection of advancement in deep neural network techniques (41).

Moreover, meta-predictors that integrate tools for predicting protein stability changes are developed to enhance machine learning-based approaches accuracy. DUET (42), iStable (43), and iDeepDDG (41) are such examples of it. DUET combines the predictions of mCSM and SDM in an optimized predictor using a support vector machine. iStable also uses a support vector machine to combine predictions from eleven different protein stability predictors. iDeepDDG only integrates the outputs of mCSM, SDM, and DUET into the concatenation layer of its network.

In summary, computational methods are more economical and rapid alternatives to experimental studies. Despite their drawbacks, predictors are also user-friendly and accessible; besides, most of them measure the ΔΔG value rather than classifying the stability effect of the mutation. Nevertheless, we still require experimental training data sets as they substantially affect the quality of the prediction methods.

2.5. Circularity and Overfitting in the Mutation Datasets

The primary reasons for the overfitting are to use a limited number of mutations in the training dataset and select noninformative features to train prediction models (9,12). Almost all stability predictors are trained on the subsets of ProTherm (44), a public database containing thermodynamic data for native and mutant proteins. Generally, using these subsets containing a few thousands of mutations to train and test the predictors is insufficient, considering 380 different types of single amino acid substitutions. Thus, using too many features to train a model with an inadequate data amount may result in false estimations (45,46).

Utilizing high-quality and balanced data to train a machine learning approach is essential as the predictor's accuracy depends on the dataset. Yet, the experimentally calculated ΔΔG value of the same mutation may differ depending on the denaturation method used for the measurement. Due to the variability among experimental studies, it is unclear how to decide the correct value of ΔΔG upon mutation (4). On the other hand, when more than one crystal structure is present for a specific protein, it's again become uncertain to decide which PDB structure to select. Therefore, filtering steps are crucial to eliminate these redundancies and construct a reliable dataset.

Besides thermodynamic and structure related issues in the dataset, predictors also suffer from several unbalance problems. Such an imbalance example is the overrepresentation of alanine substitutions. While many amino acids are covered neither in the native nor in the mutant positions in the training sets (13), substitutions to several amino acids such as alanine, valine, and glycine are highly abundant. Therefore, the performance of each predictor on underrepresented mutation types varies due to the distinction in each predictor's subset and features that train its model. Furthermore, the asymmetry between the distribution of stabilizing and destabilizing mutations within the dataset encourages predicting a destabilizing mutant better than a stabilizing one (4–6).

Overlaps between training and testing datasets used to develop machine learning-based predictors lead to circularity. Therefore, it's essential to use an evaluation subset different from the dataset used to train the prediction models to prevent overfitting. Notably, during the assessment of meta-predictors, which combine prediction calculations of more than one method, one should be rigorous in constructing the training set. Consequently, different predictors are trained with various mutations from diverging datasets, which increases the risk of overlapping between those training sets and the meta-predictor's evaluation dataset. In addition, such circularity could lead to high prediction accuracy during the development of the predictor (15), which later may negatively affect the performance of the predictor's future estimates.

Most predictors don't consider informative features relevant to protein stability changes. In contrast, a single amino acid substitution introduces a new set of covalent bonds, volume, ionic strength, and hydrophobicity to protein. If such properties of mutant residue significantly differ from the native residue, conformation of the protein and the non-covalent bond between the mutant residue and neighbor residues may change to provide minimum energy difference (12). Overall, the thermodynamic stability of a globular protein is in the range of -5 to -15 kcal/mol, while the energy of a single hydrogen bond is 2-5 kcal/mol (47). Thus, a net gain or loss of a hydrogen bond of a mutant over its native residue can significantly stabilize or destabilize the mutant protein. Besides hundreds of hydrogen bonds occur in a folded protein, non-covalent bonds also impact protein stability. Consequently, for accurately predicting protein stability changes upon mutations, several features should be considered, such as amino acid types and counts of neighbor residues, mutation type, changes in the space surrounding the mutation site, and the experimental conditions, parallel to a manually curated and symmetric dataset.

3. MATERIALS AND METHODS

3.1. Mutation Datasets

Three mutation datasets, S2648, PON-tstab and S^{Sym} were retrieved from the VariBench database (48) (http://structure.bmc.lu.se/VariBench). The ΔΔG sign convention within these datasets is set to indicate stabilizing mutations with a negative sign and destabilizing mutations with a positive sign. We associated certain ΔΔG outcomes with mutation types such as destabilizing mutations referred to those with ΔΔG more than 0.5 kcal/mol, while stabilizing term is used to refer those with ΔΔG less than -0.5 kcal/mol. Neutral mutations are collectively defined as those with ΔΔG in the range of [-0.5-0.5]. All three datasets were investigated by means of their ΔΔG and amino acid frequency distributions.

3.2. Under-sampling Strategy

The manually curated PON-tstab was systematically under sampled in 3 steps: (i) duplet generation, (ii) constraints, (iii) selection. In the first step, each mutation in the curated PON-tstab dataset, was converted to a 2-letter representation by simply eliminating the mutation identifier. This representation was referred to as duplet as it is the combination of one-letter codes of the native and mutant amino acids. During this conversion, we applied two different amino acid alphabets, the common 20-letter alphabet that can give rise to 380 different groups and a 4-letter reduced alphabet that can give rise to 16 different groups. The reduced alphabet was generated based on the side chain biochemistry: *Ali* for amino acids with aliphatic (A, I, L, V, M, P, G), *Aro* for aromatic (F, Y, W), *Pol* for polar (S, T, C, N, Q) and *Cha* for charged sidechains (D, E, H, K, R). Next, the group formed by the same duplets were further divided into sub-groups that fall into the ΔΔG range of 2 kcal/mol, referred as stability constraint.

Beside stability, we also introduced 2 different structural constraints, namely secondary structure (SS) and relative accessible surface area (ASA) which were determined by DSSP (49,50) and Bio.PDB of Biopython (51), respectively. For secondary structure, 3 labels were generated by combining all helical structures such as α-helix, 3-10 helix, and Π-helix into the single label of helix; all structures associated with extended configurations such as isolated β-bridges, β-strands, β-sheets into sheet; and all of the disordered structures such as turns and coils into loop. Similarly, we have created 3 different ASA labels; none-to-low exposed regions with ASA values less than 0.10, low-to-medium exposed regions with ASA values ranging between [0.10-0.50] and medium-to-high exposed regions with the ASA values higher than 0.50. When structural constraints were applied, the duplets were divided into sub-groups based on their secondary structure and/or ASA labels. The last step is the under-sampling step in which 3 mutants from the each (sub-) group were selected while the rest of the group members were eliminated. Selections were made to allow maximal variation of the stability such that 3 mutations that have the maximum, minimum and median ΔΔG of the 2 kcal/mol interval were selected to represent the sub-groups, while the rest of the mutations within that group were eliminated.

3.3. ΔΔG Predictors

We have tested 11 different ΔΔG predictors on the curated PON-tstab dataset. These predictors are namely DeepDDG (41), mCSM (52), INPS-3D (5), I-Mutant 2.0 (39), I-Mutant 3.0 (1), SDM (7), MAESTRO (53), PoPMuSiC (54), DUET (42), iStable (43) and iDeepDDG (41). The predictions were performed by the web interface of each predictor. Among 11 predictors, 2 of them, I-Mutant 2.0 and I-Mutant 3.0, utilized protein sequences as input, while the rest of the predictors take 3D structures as input. Predictions by I-Mutant 2.0 and I-Mutant 3.0 were done by using default parameters of temperature (25°C) and pH (7.0). Predictions were conducted by providing a list of mutations for the predictors of DeepDDG, mSCM, INPS-3D, SDM, and MAESTRO while for PoPMuSiC, I-Mutant 2.0, I-Mutant 3.0, DUET and iStable,

only single-mutation was accepted as input. To perform predictions with iDeepDDG, we have used the ΔΔG predictions of mCSM, SDM, and DUET.

3.4. Performance Assessment

Performances of 11 ΔΔG predictors were assessed on the curated and under-sampled datasets by using multiple metrics including Pearson correlation coefficient (PCC), mean absolute error (MAE), and mean signed error (MSE).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\Delta\Delta G_{pred.} - \Delta\Delta G_{exp.}\right|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta\Delta G_{pred.} - \Delta\Delta G_{exp.}\right)$$

For calculating ROC and Mathew's correlation coefficient (MCC), a contingency matrix was generated in which a true positive (TP) and true negative (TN) is defined as the correct prediction of positive and negative ΔΔG outcomes respectively. While false positive (FP) and false negative (FN) terms refer to the incorrect predictions of a negative ΔΔG value as positive and a positive ΔΔG value as a negative respectively.

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

To quantify the agreement between the experimental and the computational methods, enhanced Bland-Altman plots were obtained. and from these plots, the bias (δ) of the predictors and their limit of agreement (LOA) are calculated as the mean difference between the experimental and predicted scores.

4. RESULTS

4.1. Protein Stability Datasets

Table 1. The most used datasets and subsets derived from ProTherm Database.

Datasets	Total Mutations	Stabilizing Mutations	Neutral Mutations	Destabilizing Mutations
Dataset 1	1784	222	631	931
Dataset 2	2156	230	710	1225
S1615	1615	211	493	911
S2648	2648	295	755	1598
AUTO-MUTE Dataset (S1925)	1925	283	614	1028
PON-tstab Dataset	1564	232	467	865
I-Mutant2.0 Dataset (S1948)	1948	301	605	1042
S388	388	23	78	287
S^{sym}	684	225	234	225
p53	42	2	19	21

Experimentally calculated ΔΔG effects upon single amino acid substitutions are collected in several protein stability databases. The most known database, ProTherm, consisted of 17,113 mutations from 771 different proteins. All the contents of the mutations, such as experimental conditions, change in free energy of folding, PDB ID, protein name, are retrieved from more than 1500 scientific articles. Besides single amino acid substitutions, ProTherm also contains thermodynamic data from double and multiple mutants (44). Regrettably, this database is currently unavailable due to inconsistencies, including missing ΔΔG values, repetitions, etc. Nonetheless, a new and fresh version of the ProThermDB (55) is published with more than thirty thousand entries. In addition, the updated version of ProTherm includes thermodynamic data obtained for different organisms and cell lines.

ProTherm was last updated back in 2013 before it's updated in 2021. Therefore, to date, various subsets and/or databases are constructed by data validated in reference articles. ThermoMutDB (56) is one of the databases which manually curated the thermodynamic data from numerous referenced papers in ProTherm. In general, information on the protein, experimental methods and conditions, mutation details, and literature information of 16.018 mutations are included in this database. Although there are 380 fundamental mutation types, 10 of these mutations are absent in ThermoMutDB, including W to G, W to P, etc. Moreover, mutations to alanine are significantly overrepresented even though the database is precisely curated. Another database, namely FireProtDB (57), was published just after ThermoMutDB with similar properties. FireProtDB is manually curated the thermostability data from ProTherm and ProtaBank (58) to construct a database of single amino acid substitutions. Besides the ΔΔG value, UniProt ID, sequence features, and retrieved organism, changes in melting temperatures upon mutation are also added as a data entry. In total, 43% of the mutations are destabilizing, while there are only 14% stabilizing and 43% neutral mutations. Given that an abundance of destabilizing mutations may lead to the underestimated predictions, FireProtDB will be an insufficient database to battle the overfitting.

The VariBench is a benchmark database that contains training and testing datasets of various machine learning-based protein stability predictors such as I-Mutant2.0, PON-tstab, PoPMuSiC. Today, 22 different protein stability datasets are shared in VariBench, although only four were published in 2013. Most of these curated benchmark datasets (**Table 1**) are subsets of ProTherm. The first dataset collected in VariBench was consists of 2156 single amino acid substitutions from two different sources (59). Mutations within this dataset were filtered to avoid duplication, mismatches between the sequence and PDB structure, and missing residue in the PDB structure. Despite a variety of corrections noted, the inconsistency in the current subsets still ongoing. One of the main limitations is the number of stabilizing mutations within the datasets. Lower than 15% of all mutations in the most used datasets correspond to stabilizing, while more than 50% of all mutations are

destabilizing (**Table 1**). The second main limitation is the disagreement in the experimental studies. Applying different environmental conditions and denaturation methods during ΔΔG measurement of the same mutant may produce dissimilar stability measures. Therefore, predictors have to deal with the uncertainty caused by repeated experimental ΔΔG values for the same mutation. Here, we investigated the overall ΔΔG value distributions of the most used single variant datasets to observe the variability in a detailed manner.

4.1.1. ΔΔG distribution of three widely used mutation datasets

Three protein stability datasets, namely S2648, PON-tstab and S^{sym}, were extracted from the VariBench (48). The S2648 dataset contains 2648 mutations from 131 proteins (60), the PON-tstab contains 1564 mutations from 99 proteins (15) and the S^{sym} has 684 mutations from 15 proteins (9). Although all three datasets are either entirely or partially subsets of the ProTherm (38), they differ in size and curation status. The S2648 is the largest of three datasets without any reported curations while the PON-tstab is the medium-sized one and has been reported to be free of large redundancies (15). The smallest S^{sym} is purposefully generated to overcome generally observed asymmetry between destabilizing and stabilizing mutations. Exploring the characteristics of these 3 datasets which currently comprise a decent sample of the stability data can help us understand the possible sources of bias in single variant datasets.

Figure 1 shows ΔΔG distributions of three datasets. The first 2 sets, S2648 and PON-tstab, have asymmetric frequency distribution across mutation types such that more than half of the mutations within these datasets are destabilizing. ΔΔG distributions showed that S2648 and PON-tstab datasets have positive means of ~1.0 kcal/mol and are skewed towards destabilizing mutations while the S^{sym} has a mean of 0.0 kcal/mol and display no skewness (**Table 2**). Given the fact that S^{sym} has been

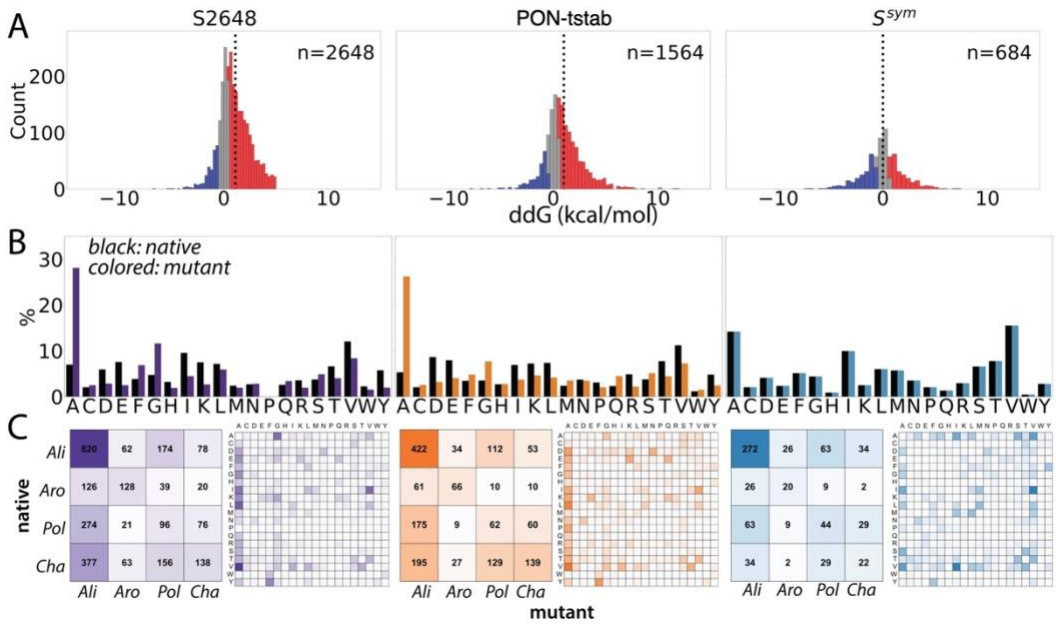

Figure 1. Distribution of ΔΔG values from S2648, PON-tstab and S^{sym} datasets.

(A) Blue bars indicate stabilizing; red bars indicate destabilizing mutations and grey bars indicate neutral mutations with ΔΔG values ranging between -0.5 and 0.5 kcal/mol. Dotted line marks the mean. (B) Amio acid frequency distributions in which black bars indicate the relative frequency of amino acids in the native position while the colored bars, S2648 (purple), PON-tstab (orange) and Ssym (blue), indicate the relative frequencies for the mutant position. (C) Left plots show the absolute frequencies of 4 amino acid groups based on side-chain biochemistry; the abbreviations refer to aliphatic, aromatic, polar and charged. Right plots show the frequencies of amino acid duplets for which vertical and horizontal axes indicate native and mutant position, respectively

generated to overcome the asymmetry between the frequencies of destabilizing and stabilizing mutations, its frequency distribution is, as anticipated, perfectly symmetric, i.e., half of the mutations is stabilizing, and the other half is destabilizing (**Figure 2**). The ΔΔG distribution of the PON-tstab has the highest skew among three selected datasets (**Table 2**).

Figure 1 shows the relative frequency of the amino acids in the native (black bars) and mutant (colored bars) positions. Similar to the ΔΔG distributions, the first two datasets showed asymmetric frequencies across native and mutant positions, while the S^{sym} dataset displayed symmetric distributions across native and mutant positions. However, none of the datasets including the S^{sym} had uniform amino acid distributions.

The amino acids, Ala and Val are highly abundant in all datasets, while others such as His, Pro and Trp are among the least frequent ones (**Figure 1B**). Despite being the largest dataset, the S2648 did not contain any Pro neither in the native nor in the mutant position. By using a 4-letter alphabet based on the biochemistry of amino acid side chains, a highly abundant class of aliphatic-to-aliphatic mutations were found in all datasets while charged-to-aromatic conversions were the least frequent (**Figure 1C**). Notwithstanding its small size and symmetrical distribution, the S^{sym} had the most unbalanced frequencies particularly across reduced alphabet groups (**Figure 1C**). Approximately, 40% of the S^{sym} dataset (272/684) were aliphatic-to-aliphatic conversions while this ratio was lower in the other datasets; 30% for S2648 and 20% for PON-tstab. Charged-to-aromatic conversions comprised only 2 out 684 (0.3%) mutations in the S^{sym} dataset while this ratio slightly rose to 0.8% for the S2648 and PON-tstab datasets. Ssym had the most unbalanced frequencies across reduced alphabet conversions due to the relatively larger frequencies of Ala, Val and Ile amino acids, particularly in the native position. Moreover, these three amino acids are abundant in the S2648 and PON-tstab datasets but only in the mutant positions, not in the native positions.

4.1.2. Classification of the neutral mutations

Often a binary classification is applied to mutations, i.e., destabilizing/stabilizing, while some predictors tend to label neutral mutations as well (1,15,33,61). Here we adopted an additional label for which we have labeled mutations as neutral if they had ΔΔG values ranging between -0.5–0.5 kcal/mol (33). We consider adoption of an additional label for neutral mutations is essential. Mainly because the datasets, even the curated and anti-symmetric ones, contain neutral mutations with their ΔΔG values equal to zero. 1-to-2% of all 3 datasets comprised to those zeros (**Figure 2**). Assignment of these zeros to any of the destabilizing or stabilizing classes is fundamentally incorrect. Therefore, instead of a binary classification, a ternary classification more accurately group mutations.

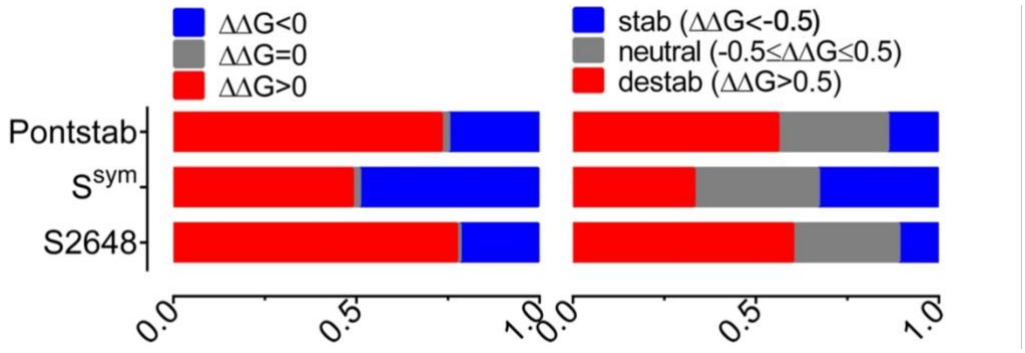

Figure 2. Relative frequencies of the stabilizing, neutral and destabilizing mutations within the datasets are displayed.

All of the three datasets showed unimodal ΔΔG distribution with nodes around the neutral ΔΔG range (**Figure 1**). Explicitly, we found 726, 468 and 235 mutations in the ΔΔG range of -0.5-0.5 kcal/mol in the S2648, PON-tstab and Ssym, respectively. Neutral mutations approximately correspond to one third of the datasets, (**Figure 2**) and they are concentrated in a narrow ΔΔG range while other mutation types, even highly populated destabilizing mutations are diluted over a much wider ΔΔG range. Explicitly, for the S2648, destabilizing mutations, corresponding to 60% span a range of 6.29 kcal/mol, while stabilizing mutations, comprising to 15%, cover a range of 4.49 kcal/mol. As a result, the concentration of neutral mutations (726 mutations per 1 kcal/mol) are higher than that of the destabilizing (~254 mutations per 1 kcal/mol) and much higher than that of the stabilizing mutations (~ 60 mutations per 1 kcal/mol) in the S2648. More extreme observations were made for other datasets. The symmetrical Ssym dataset had a concentration of 235 neutral mutations per 1 kcal/mol while it was reduced to only 32 mutations per 1 kcal/mol for both destabilizing and stabilizing mutations. For the PON-tstab, 468 neutral mutations were found for every 1 kcal/mol while only 31 stabilizing and 51 destabilizing mutations were found for the same range.

The concentrations of neutral mutations can be compared by the kurtosis parameter of the ΔΔG distributions. In parallel with the arguments that are put forward for the desired characteristics of the amino acid frequencies in mutation datasets (13),

an ideal (training) dataset should have a platykurtic ΔΔG distribution, i.e. negative kurtosis. A uniform frequency distribution of amino acids would ensure even sampling of amino acid types and similarly a platykurtic ΔΔG distribution would ensure a more uniform representation of mutation types, than mesokurtic or leptokurtic distributions with steep peaks around neutral mutations. Concentrated neutral mutations as probed by positive excess kurtosis values of the ΔΔG distributions was our particular observation for all datasets (**Table 2**). The PON-tstab dataset had the highest kurtosis, resulting in the most leptokurtic distribution among all 3 datasets. The S^{sym} similarly showed a leptokurtic distribution while the S2648 distribution relatively had a better shape with slightly lower kurtosis value than the other two datasets.

Table 2. Characteristics of ΔΔG distributions of the datasets used and generated in this study

Dataset	n	n_p	Median	Min.	Max.	Skewness	Kurtosis
S2648	2648	131	0.82	-5.0	6.8	0.148	0.749
PON-tstab	1564	99	0.70	-17.4	8.0	1.153	6.951
S^{sym}	684	15	0.00	-7.5	7.5	0	1.599

n_p is the number of proteins; median, min. and max. are in kcal/mol

4.1.3. Manual curation of the PON-tstab dataset

Every mutation in the PON-tstab was examined and were encountered some inconsistencies within this dataset (**Table 3**). Among 269 problematic mutations, 113 mutations were eliminated from the dataset. The removals were done due to 3 major issues; repetition, mismatch or PDB related. We found 125 repeated mutations of which 116 are duplicated and 9 are triplicated. We randomly removed 58 of the duplications and 6 of the triplications. Among 64 of the mutations that showed a mismatch between the sequence and PDB structure, 50 of them were corrected as they were derived from inconsistencies in the residue numbering of PDB structures. The remaining 14 mutations with unresolved mismatches were removed from the dataset.

Table 3. Curation of the PON-tstab Dataset

	Corrected	Removed	Total
Mismatch	48	4	52
Duplicated	55	10	65
Triplicated	3	2	5
PDB Related (NoID)	47	21	68
PDB Related (Quality)	0	14	14
Mismatch, Duplicated	2	0	2
PDB Related (NoID), Mismatch	0	10	10
PDB Related (NoID), Duplicated	1	48	49
PDB Related (NoID), Triplicated	0	4	4
Total	**156**	**113**	**269**

Given the accumulation of structure-based predictors, only mutations with valid PDB information were considered for our curated dataset. Therefore, the mutations with problematic PDB information were eliminated. Among 145 of the PDB-related issues, 131 mutations had their PDB IDs missing in the original dataset. 48 of them were corrected by cross-checking other entries of the same protein while the remaining 83 of them were eliminated. 33 of the eliminated mutations in this group were obtained from the cold shock protein (62). Although other mutation entries for this protein are present, their PDB information (1CSP or 1C9O) does not match with any of these 33 mutations. The reference study clarified that these mutations were obtained from a variant of cold shock protein B, namely CspB-TB which holds a Lys at the 12^{th} position. Although these variants were originated from a different protein variant, their entries in the downloaded PON-tstab dataset listed by the same protein name. In fact, these mutations were distinguished from other CspB entries only from their sequence identifiers. Because this uncertainty could not be resolved, they were linked to an issue during our curation and removed from the dataset. Apart from this set of mutations, the eliminated mutations due to a missing PDB ID were mainly obtained from Interleukin 6, Subtilisin, TEM beta-lactamase, Anthranilate Isomerase which had only a few numbers of mutations in this dataset. Furthermore, 14 mutations were eliminated due to the low PDB quality of the mutated positions. Specifically, 12 mutations from the lambda repressor whose provided PDB structure (PDB ID: 1RLP) only contained Cα atoms and 2 mutations from the tryptophan synthase alpha-subunit whose structure

has a missing region for the mutations (PDB ID: 1WQ5) were eliminated to low PDB quality.

In summary, the eliminated 113 mutations were obtained from 23 different proteins while only 5 of these proteins accounted for almost half of the eliminations. These 5 proteins were similarly listed by a non-standard sequence identifier which in turn hampered their curation. Therefore, we highlight the importance of compiling not just correct stability information but also correct sequence and structure information for stability datasets. Otherwise, correcting the missing or incorrect information in an already formed dataset can be difficult, particularly for plenty of cases. Consequently, 64 repeated and 49 problematic variations discarded from the original PON-tstab has led to a total of 1451 mutations from 89 different proteins in the curated version (**Appendix 1**).

4.2. Systematic Under-sampling Strategy for Elimination of Similar Mutations

The medium-sized PON-tstab had the most skewed and peaked ΔΔG distribution along with a non-uniform amino acid distribution (**Figure 1** and **Table 2**) which, among the 3 datasets, would least align with the desired characteristics of an ideal mutation dataset. Thus, we recruited the PON-tstab as the toy dataset to test how our under-sampling strategy would affect the distribution of the PON-tstab.

By following the steps in our workflow (**Figure 3**), we applied two different amino acid alphabets, the common 20-letter alphabet, giving rise to 380 different groups and a 4-letter reduced alphabet that can give rise to 16 different groups. However, when we inspect the duplet groups formed by the common 20-letter alphabet, we observed that not all duplets are formed. In contrast, instead of 380, we had only 294 duplets and 95 of these duplets are not represented more than 1 data point

(**Figure *6***). As a result, further dividing these duplet groups using additional structural constraints would become impractical and render under-sampling ineffective. Therefore, we did not apply structural constraints to the 20-letter alphabet duplets. Instead, as previously suggested for increasing the number of data points in each group (13,63), we have reduced the alphabet size to generate populated duplet groups.

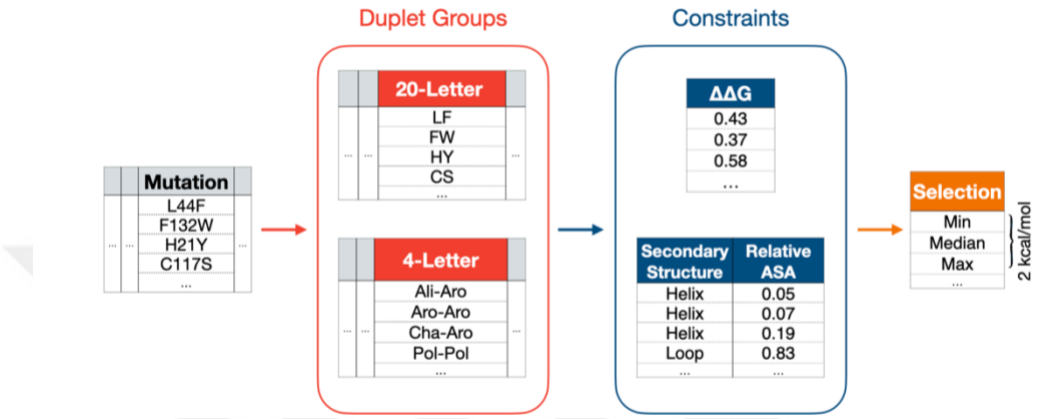

Figure 3. The under-sampling workflow.

First duplet groups are generated by either the common 20-letter alphabet or 4-letter reduced one. Second, constraints are applied to further re-group the duplet groups from the first step. Stability constraint re-grouped the duplets in every 2 kcal/mol interval. Optionally, two structural constraints based on secondary structure and relative ASA values are applied. Third, 3 mutations with the largest ΔΔG variation in the final group are selected.

After generation of duplet groups by using 4-letter reduced alphabet, we introduced stability constraints to re-group the duplets based on their ΔΔG similarity. As such, the mutation group formed by the same duplets were further divided into sub-groups that fall into the same ΔΔG range of 2 kcal/mol. Beside stability, we further introduced two different structural features, namely secondary structure (SS) and relative accessible surface area (ASA) as constraints. When structural constraints are considered, the duplets were further divided into sub-groups based on their SS and/or ASA labels.

The labels used for secondary structure (SS) and relative accessible surface area (ASA) were illustrated on three different proteins (**Figure 4**). The regions corresponding to the helix, sheet, and coil of the proteins are successfully displayed.

Buried residues were perfectly hidden inside the protein's core as we considered them between 0 to 0.10. Moreover, this classification leads to the production of a balanced set as buried residues are 33% of all mutations and exposed residues are 26% of all mutations. Selected mutations to construct 4L/SS/ASA subset were colored in dark purple, while the unselected mutations were colored in light purple on the right panel of **Figure 4.** Although we just picked three mutations from the sub-groups, our selection shows diversity along the structure.

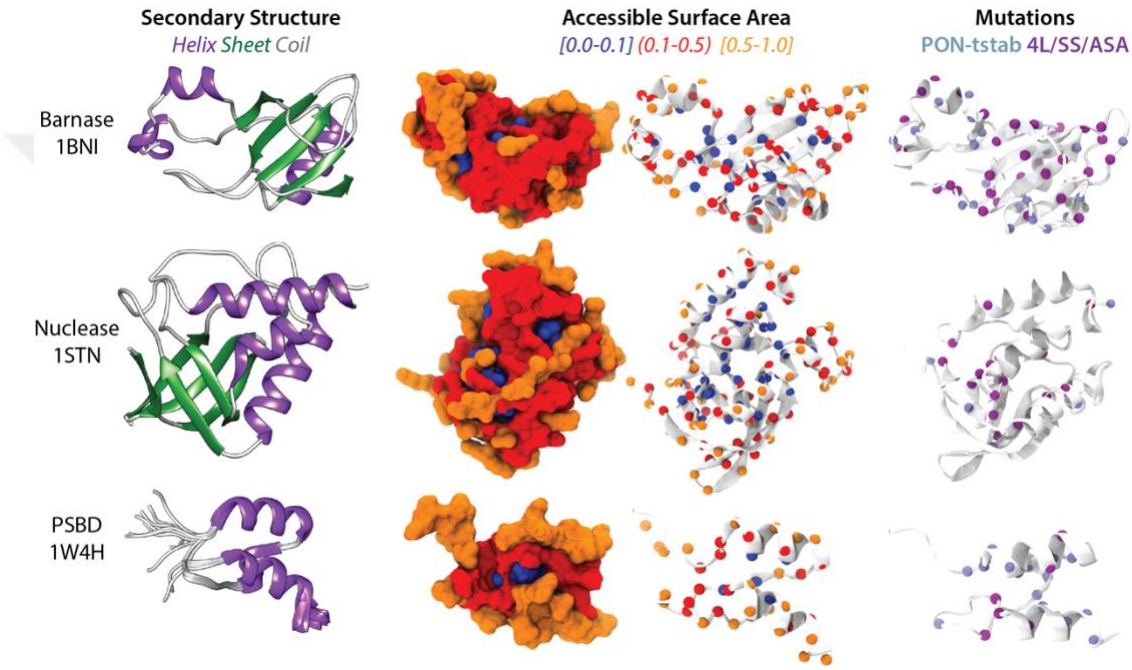

Figure 4. Visualization of Secondary structure and ASA labels on three different proteins.

If all the secondary structure and ASA labels are found in a given duplet group, then this duplet group formed the same duplets will be divided into 3 sub-groups based on the secondary structure of the mutations and further into 9 sub-groups based on the ASA of the mutations. For example, all of the Ali-Ali duplets that are found on the helical and a buried region are grouped together and next, 3 mutations were selected from each 2 kcal/mol interval. Visualization SS and ASA features is displayed in **Figure 5.**

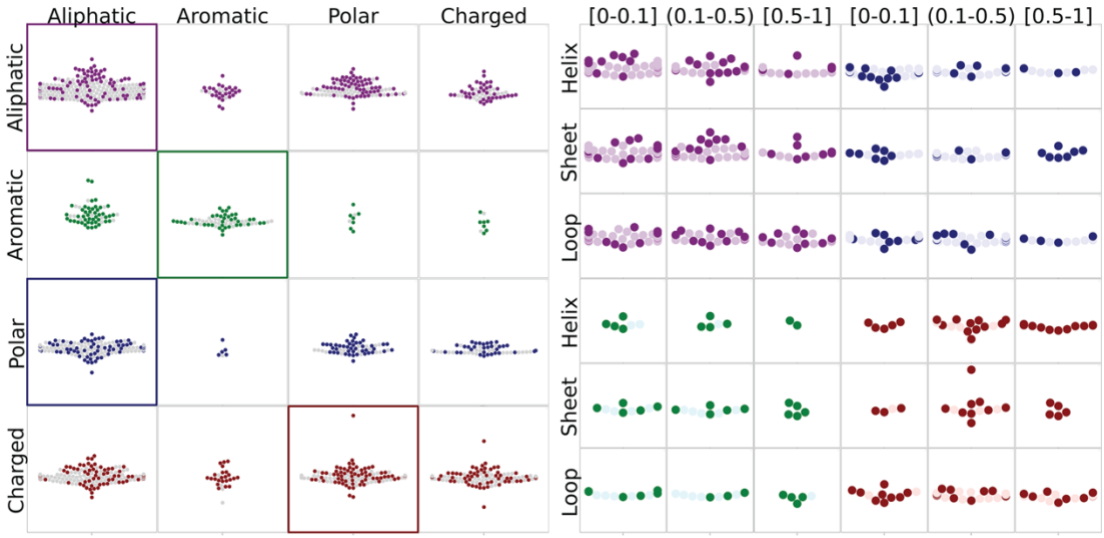

Figure 5. The ΔΔG distribution across reduced alphabet groups.

All possible 9 sub-groups from the most populated 4 reduced alphabet groups were shown on the left panel by using a color code for each group. Selected mutations in the 4L/SS/ASA subset were marked by a darker color in all plots.

4.2.1. Distribution of 5 distinct subsets from the manually curated PON-tstab

A collection of 1451 mutations from 89 proteins in the curated PON-tstab was utilized as the toy dataset to implement our under-sampling strategy. From this parent dataset, we have generated five different subsets by application of the workflow in **Figure 3** via combining two different alphabets and constraints. The under-sampled datasets are analyzed by means of their ΔΔG and amino acid frequency distributions (**Figure 6**). Each vertical panel in the **Figure 6** shows an under-sampled subset of the PON-tstab. The first 2 panels show the distributions of the subsets constructed by only applying the duplet groups formed by 20L and 4L alphabet, respectively. The third and fourth panels which are labeled as 4L/SS and 4L/ASA were built by applying either secondary structure or relative solvent accessible area to the 4L subset. The fifth subset, namely 4L/SS/ASA is formed by considering both structural constraints.

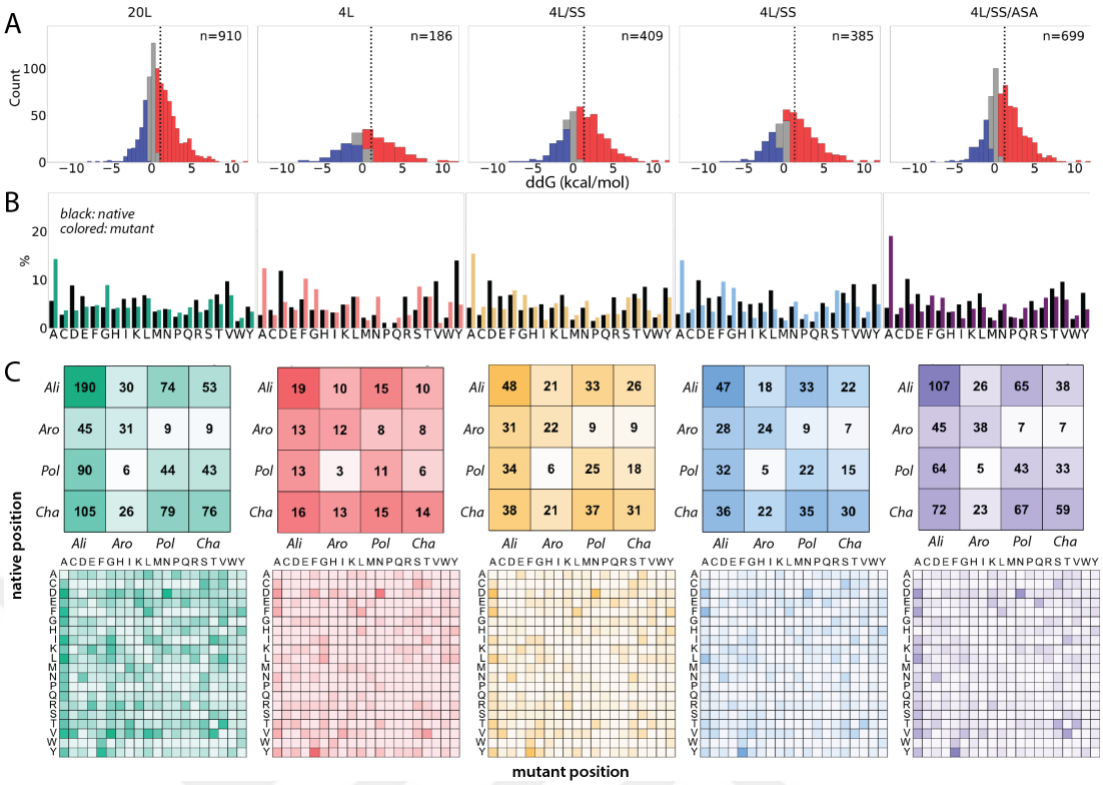

Figure 6. Five different under-sampled dataset examples.

Each vertical panel represents a different under-sampling trial. (A) ΔΔG distributions and (B) amino acid distributions along with checkerboard plots are shown, analogue to Figure 1.

In contrast to the original PON-tstab dataset (**Figure 1B-C**), a more uniform distribution across subsets is observed (**Figure 6).** Even applying one of the structural constraints reduced the skewness and kurtosis of the ΔΔG values in all of the under-sampled datasets (**Table 4**). Furthermore, frequency distribution of amino acids was also improved (**Figure 6**). Compared to the complete dataset, under-sampled datasets had more uniformly distributed frequencies across duplet groups based on both alphabets (**Figure 6B-C**). Our strategy that sampled 3 data points from every 2 kcal/mol interval did not take any measure to balance the mutation types. Given their highest concentration in the parent dataset, the neutral mutations were under sampled the most, resulting in an apparent reduction in the kurtosis of ΔΔG distributions while they were followed by destabilizing mutations. As a result, we observed a slightly improved ratio of destabilizing to stabilizing mutations. Among these five datasets, we note that the last subset (4L/SS/ASA) has the most symmetric distribution across

mutation types. In addition, compared to the other subsets, its less peaked and a more balanced subset of the PON-tstab (**Table 4**).

Table 4. Characteristics of ΔΔG distributions of the under-sampled datasets generated by a unique workflow

Dataset	n	n_p	Median	Skewness	Kurtosis	Skewness*	Kurtosis*
cur. PON-tstab	1451	89	0.70	1.187	6.931	0.825	3.825
20L	910	68	0.78	1.025	4.964	0.682	2.508
4L	186	49	0.75	0.822	2.917	0.335	0.786
4L/SS	409	61	1.19	0.756	3.380	0.379	1.415
4L/ASA	385	70	1.10	0.768	3.027	0.411	1.200
4L/SS/ASA	699	80	1.00	0.938	4.242	0.587	1.946

All datasets have the same range of ΔΔG distributions -17.40 to 8.00 kcal/mol
* Marks the skewness and kurtosis after removal of the outliers

4.3. Protein Stability Predictors

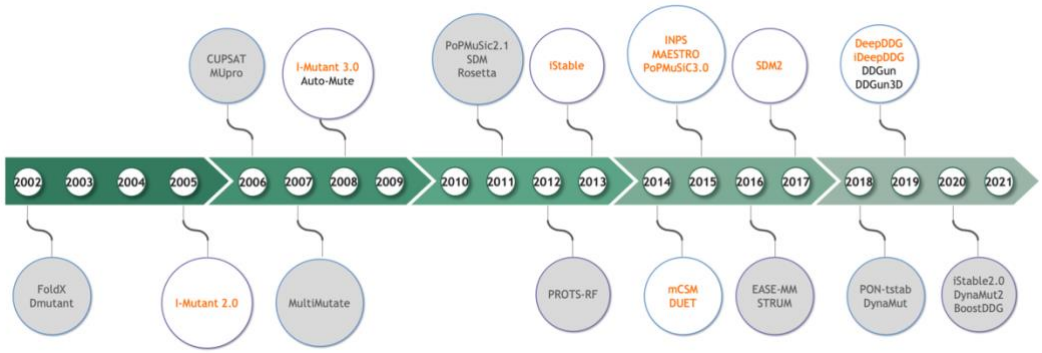

Figure 7. Timeline of protein stability predictors.

Numerous computational predictors have been developed in the last 20 years to calculate the change in free energy of folding upon mutation. The diversity in the protein stability predictors is due to the difference in algorithms predictors use, the dataset they are trained on, availability levels, and performance evaluation metrics. Orange-colored predictors in **Figure 7** are the ones that are comparatively assessed in this study. Almost all evaluated methods are easy to use and accessible web-based

predictors that apply machine learning approaches. Only SDM and PoPMuSiC differ due to their prediction strategy, which applies statistical potential approaches.

4.4. Comparative Assessment of 11 Protein Stability Predictors

Performance of the 11 different protein stability predictors is evaluated on manually curated PON-tstab dataset and under-sampled subsets by calculating several statistical metrics. Prediction performance was previously reported through correlation between the experimental and predicted scores (5,61,64). Although PCC is a widely used performance metric for ΔΔG predictors, it's not enough to accurately classify them (65). Thus, we further measure MCC and ROC-AUC to explain the predictor's performance on classification task. Moreover, MAE, MSE and Bland-Altman are measured to analyze bias.

4.4.1. Performance of 11 predictors on curated PON-tstab dataset

Based on PCC calculations (**Figure 8)**, highest PCC was observed for the I Mutant2.0 scores as 0.74, which was followed by iStable with a PCC of 0.70. On the other hand, SDM was found to be the worst predictor with a PCC of 0.42. SDM's performance was only slightly surpassed by INPS and Maestro that yielded PCC values of 0.48 and 0.50 respectively. DUET, a meta-predictor, that combines SDM and mCSM predictions was slightly less accurate than iDeepDDG which integrates the outputs of mCSM, SDM, and DUET as well. Apparently, iDeepDDG's performance exceeded the performance of DUET, one of its integrated tools on this dataset. PoPMuSiC which uses a statistical potential based scoring function rather than machine learning approaches produced a PCC value of 0.56.

Overall, we did not report a trend in the PCC-based scoring power of predictors associated with the underlying prediction methodology. However, we note that the best predictor I-Mutant2.0 is a sequence-based predictor. I-Mutant2.0 and 3.0 share similarities such that they are both developed by the same group (1,39) and both predictors accept protein sequence as input. I-Mutant2.0 can distinguish between destabilizing and stabilizing mutations while I-Mutant3.0 can also quantify the effect of neutral mutations. Moreover, to balance reverse mutations were included during training of I-Mutant3.0. Notwithstanding the similarities between two methods, their performance noticeably differed such that I-Mutant3.0 scores showed a much lower PCC of 0.58 than that of I-Mutant2.0.

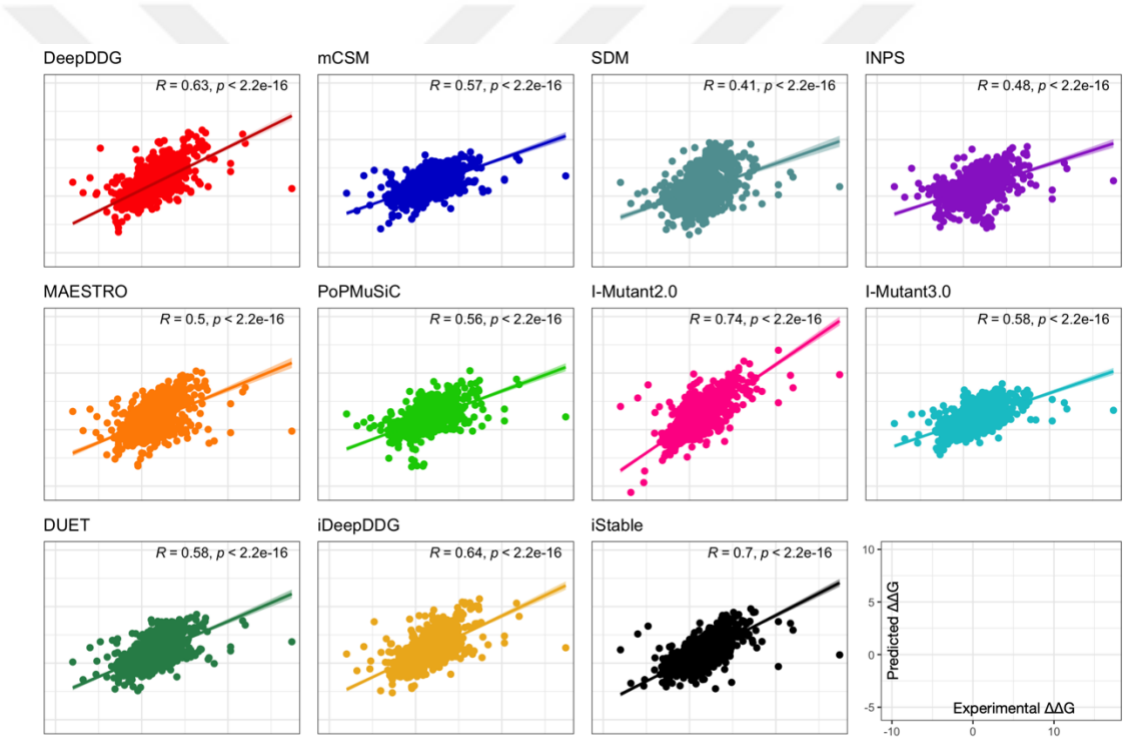

Figure 8. Performance analysis of 11 predictors on the curated PON-tstab.

Each tools performance is calculated by PCC and plotted by a scatter plot against the experimental ΔΔG values. For all plots, the experimental ΔΔG is plotted in the horizontal axis, while vertical axis represents the scores obtained from predictors.

We also calculated PCC for every protein in the curated PON-tstab, which had more than two mutations (**Figure 9**). Proteins with a high number of mutations displayed a relatively high PCC than those represented only by a few mutations. For

example, performance on barnase mutations corresponding to ~10% of the total dataset is mainly similar to the tools' performance on the complete dataset. Notably, the proteins with a high number of mutations reflected the entire dataset with a high number of destabilizing and neutral mutations. In contrast, under-represented proteins may have only stabilizing mutations and thus showed a weaker correlation.

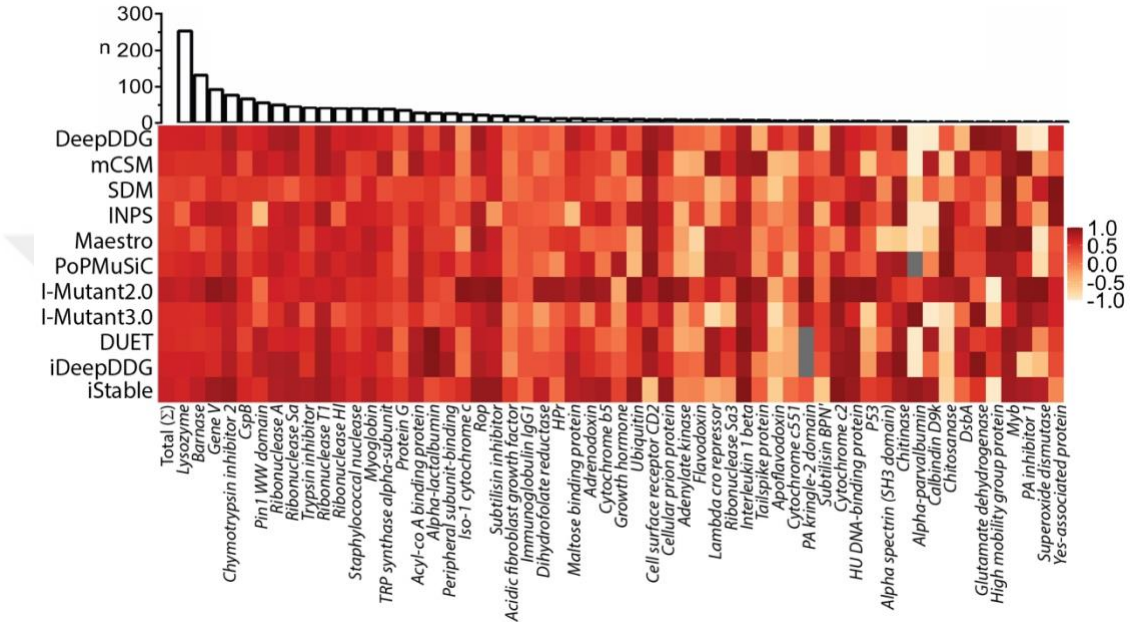

Figure 9. PCC calculations based on proteins in the curated PON-tstab.

57 different proteins with more than 2 mutations were ranked according to their number of mutations as illustrated by the bar plot. Grey cells indicate failed predictions.

Prediction errors were estimated by MAE and MSE (first horizontal panel of **Figure 11**). In line with PCC, I-Mutant2.0 had the lowest mean absolute error of 0.77 kcal/mol while SDM had the highest error (1.31 kcal/mol). In line with MAE, the MSE borders were defined by the performances of I-Mutant2.0 and SDM. The former had the lowest and the single positive MSE (0.05) while the latter predictor had the largest signed error of -0.53. Although all of the predictors except from I-Mutant2.0 showed negative MSE, none of the tools had showed a large error, producing an overall balanced MSE for the entire dataset.

Given the trend observed in error analyses, we have further analyzed the dependence of ΔΔG on the error. For every 1 kcal/mol interval, we assessed the performance of each tool by PCC and calculated error by MAE, MSE (**Figure 10**). This analysis, particularly the MSE heatmap, has revealed that predictors consistently overestimated the stabilizing mutations and underestimated the destabilizing mutations.

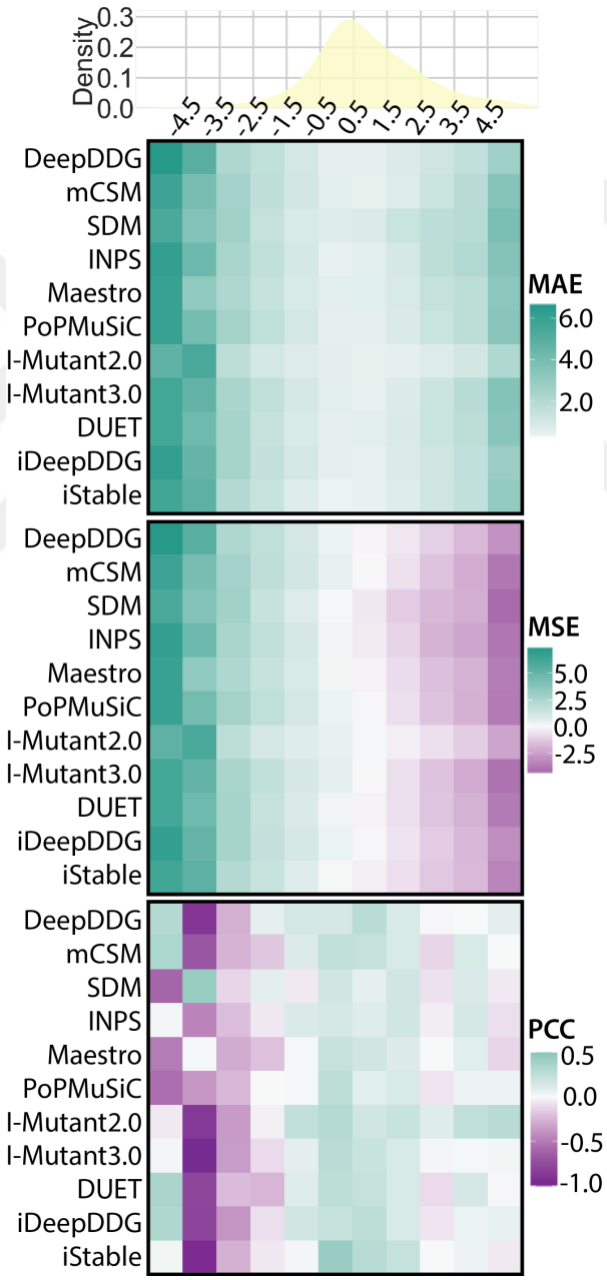

Figure 10. MAE, MSE and PCC calculations based on ΔΔG.

4.4.2. Performance of 11 predictors on under-sampled subsets

We summarized the performance of the eleven predictors for the under sampled datasets in **Figure 11** by calculating three different evaluation metrics. Since the correlation between the experimental and predicted ΔΔG values is the most widely used measure to compare the methods' performances, we visualized it as a PCC-heatmap in **Figure 11.** Almost every predictor shows similar metrics as the parent data set, with slightly higher correlations, reflecting the influence of redundancy elimination. Based on PCC calculations, I-Mutant2.0 displayed the highest average R of 0.76 for all under-sampled datasets, followed by iStable with an average R of 0.69. SDM and INPS, two of the worst performers for the complete dataset, exceeded PCC values of 0.42 and 0.48, respectively, plus measured as an average of 0.44 and 0.54 for the under-sampled datasets. On the other hand, PCC values of Maestro decreased for the under-sampled datasets compared to its performance on the full dataset. Overall, the scoring performance of the predictors slightly surpassed as we introduced the secondary structure constraints to the under-sampling methodology.

MAE and MSE are calculated to evaluate the error of the predictions. Similarly, to PCC, I-Mutant2.0 had the lowest error of 0.77 kcal/mol while SDM had the highest absolute error of 1.31 kcal/mol (curated PON-tstab panel in the **Figure 11**). However, this error increased to 0.95 for I-Mutant2.0 and 1.61 for SDM as we applied the under-sampling strategy. Notably, all of the tools showed negative signed error but close to zero. Only I-Mutant2.0 showed a positive error (0.05). SDM gave the largest signed error of -0.53, which represented the tendency to predict too much stabilization. Generally, for under-sampled datasets, MSE followed similar observations as parent dataset.

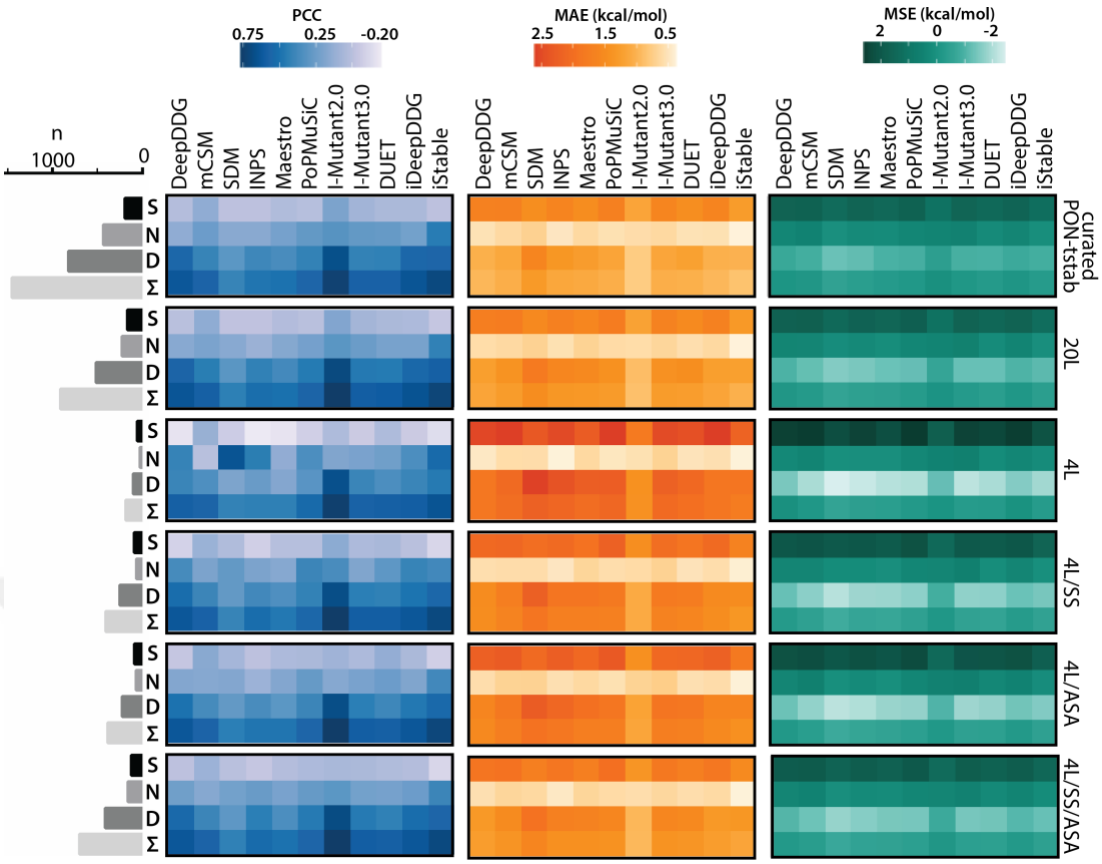

Figure 11. Performance analysis of 11 predictors on the curated PON-tstab and five under sampled subsets.

Furthermore, we have analyzed the agreement between the predictions and experimental score by Bland-Altman plots (66) which is a better way to compare the relationship between predicted and experimental values to understand how mutation bias affects the predictors' performance. The mean difference between experimental and predicted values is not zero, and this means that, on average, all predicted values measure fewer than the experimental values. This negative bias seems to be due to ΔΔG measurements over 6 kcal/mol (highly destabilizing mutations) and under -4 kcal/mol (highly stabilizing mutations), while for neutral mutations, data are closer to each other **Figure 12**. Overall, likewise to MSE analysis, Bland-Altman plots showed that almost all of the neutral mutations were predicted within the limit of agreement while the outliers were stabilizing and destabilizing mutations. Yet, most destabilizing mutations were within the limit of agreement and were less skewed compared to

stabilizing mutations. I-Mutant2.0 was the best predictor with the lowest bias (mean difference) and narrowest limit of agreement values.

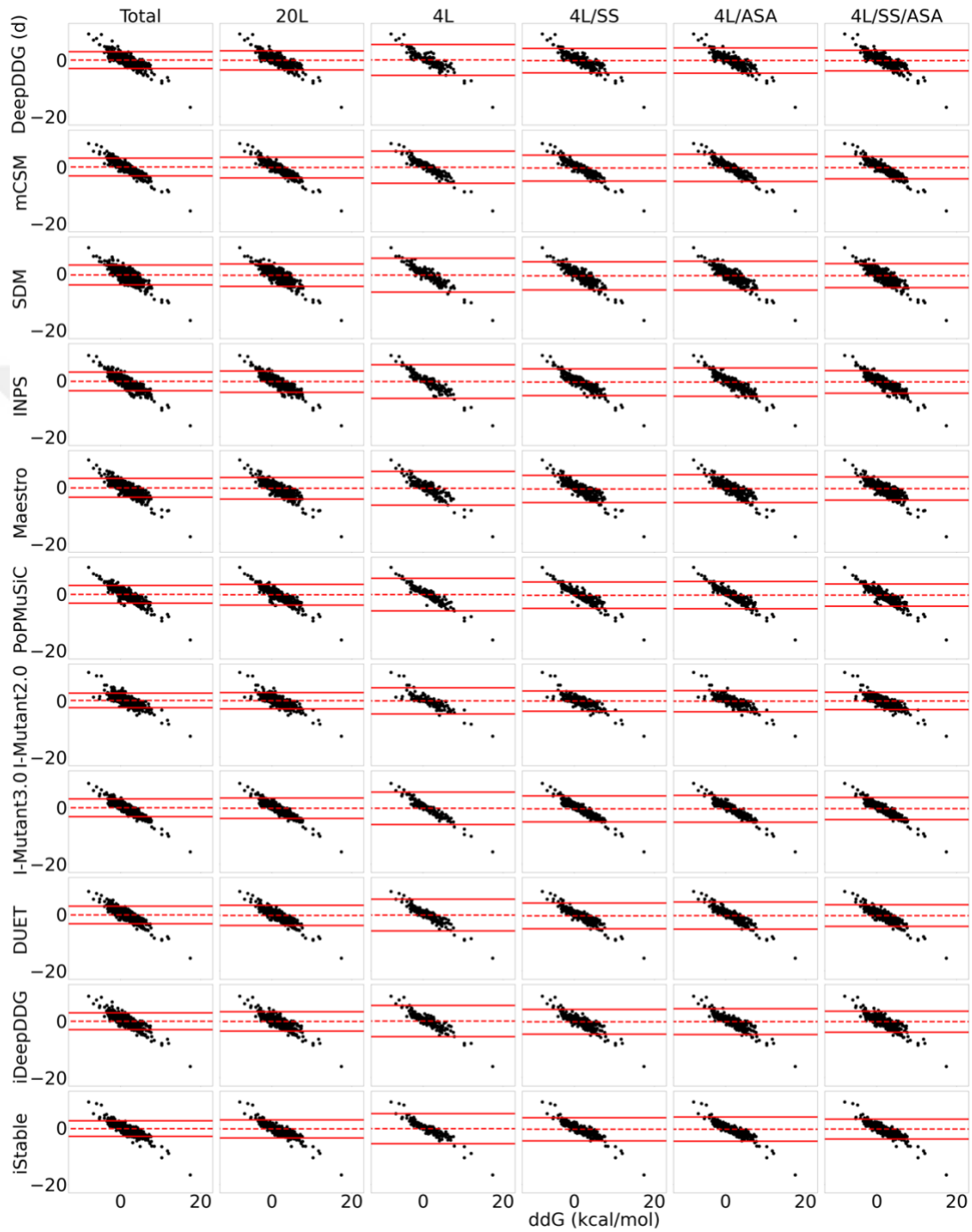

Figure 12. Bland-Altman analysis of the predictors for the curated and under-sampled datasets.

Y axis shows the difference between experimental values and predicted values while the X axis represents experimental values. Bias was represented by dashed lines and limit of agreement positions were shown by solid lines.

As MCC is a more balanced and comprehensive metric than specificity or sensitivity, we calculated it to examine the performance of the predictors additionally to PCC (**Figure 13**). Similar to PCC observations, MCC values of I-Mutant2.0 and iStable showed better performance than the rest of the predictors. Although adding structural constraints to the under-sampling workflow improves the I-Mutant's correlation, it didn't work in the same way for the MCC measure. Moreover, PCC of DeepDDG displayed better performance than iDeepDDG for under-sampled datasets, which is not valid for MCC. The worst predictors observed with an MCC lower than 0.30 were SDM and INPS-3D.

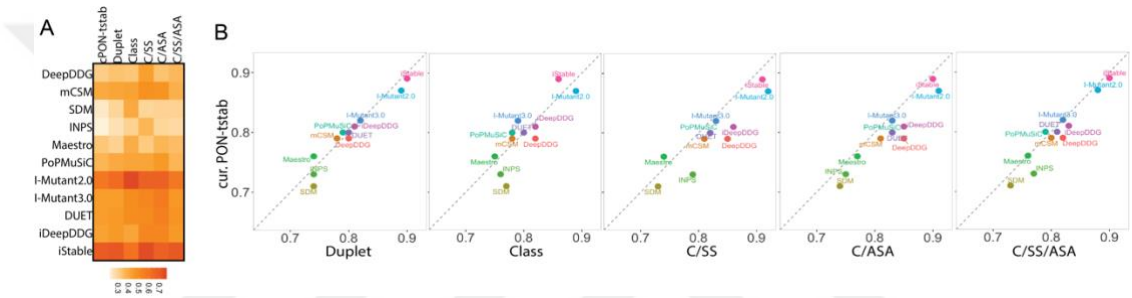

Figure 13. MCC and AUC-AUC classification of the predictors for the curated and under-sampled datasets.

Elimination of redundant values within the curated dataset improved the goodness for predictors. Based on the ROC-AUC calculations, we indicate I-Mutant2.0 as a perfect classifier against under-sampled datasets, which was again followed by iStable (**Figure 13**). In general, SDM, INPS, and Maestro showed a better ranking potential for under-sampled datasets than curated datasets; however, their performance wasn't adequate for a classification compared to other methods.

In summary, based on different evaluation metrics, I-Mutant2.0 and iStable were observed as more significant predictors compared to other methods. Moreover, predictors performed better in 4L/SS/ASA subset than the manually curated PON-tstab dataset, which exhibits the success of our workflow of reducing redundancy in the dataset.

5. DISCUSSION AND CONCLUSION

In this study, we took a close look at web-based protein stability predictors' training datasets and prediction performance to discover a way to eliminate overfitting issues.

We have inspected three mutation datasets in detail to observe whether an imbalance exists between destabilizing and stabilizing mutations and determining the overrepresentation of specific amino acids over others. Mutation datasets derived from a database such as ProTherm, ThermoMutDB display bias due to their nature since most experimental studies focused on destabilizing mutations. Consequently, datasets, namely S2648 and PON-tstab, are dominated by destabilizing and neutral mutants, while S^{sym} has symmetric distribution between destabilizing and stabilizing mutants. Such naturally occurred asymmetry in the datasets creates an additional challenge for machine learning-based stability predictors with respect to the prediction of stabilizing mutations. Furthermore, the frequency of mutants in the datasets also reveals several natural biases that need to be addressed (67,68). Such an example is the overrepresentation of alanine substitutions due to the general consideration of Ala as the best replacement amino acid to study the effects of mutations (69). Likewise, aliphatic amino acids apart from Met show naturally high frequencies, while aromatic ones, particularly Trp, are among the least frequently occurred amino acids (70). Nonetheless, an ideal mutation dataset is expected to be free of natural or unnatural biases and sample all types of instances evenly.

Uniformity is also essential for generating sensitive and specific models that can distinguish the natural phenomenon from error. Hence, we calculated skewness and kurtosis metrics of $\Delta\Delta G$ distributions to display whether a dataset has a uniform distribution or not. None of these three datasets showed a negative kurtosis and thus a

platykurtic ΔΔG distribution. Moreover, even the symmetric dataset, S^{sym}, did not show a uniform frequency distribution of amino acids.

PCC-based performance assessment led us to consider that predictors work better on destabilizing mutations than neutral and stabilizing ones, yet MAE and MSE calculations didn't show similar results. Mean signed error clarifies that all of the predictors tend to produce a score closer to zero, with a negative MSE in case of destabilizing mutations and a positive MSE for stabilizing mutations. In other words, each predictor overestimated the ΔΔG for stabilizing mutations while they all underestimated the score of destabilizing mutations. However, every predictor showed to most accurately predict the neutral mutations for which the error is close to zero.

Furthermore, as the ΔΔG values move from zero to either toward stabilizing or destabilizing values in intervals of 1 kcal/mol, the predictors fail to predict the ΔΔG values accurately. This error is a reflection of sparse data or result of the leptokurtic ΔΔG distribution. In summary, we consider that neutral mutation could be another reason for overfitting because they are extremely concentrated over a much narrow ΔΔG range than destabilizing or stabilizing mutations.

Today, many curated single variant databases are present (55–57); however, the imbalance and inconsistencies within the subsets derived from these databases lead to overestimated or underestimated performances. In this context, evaluating current protein stability predictors on a large, balanced, and novel blind dataset that is dissimilar with the training sets of the predictors is critical for developing new predictors and benchmarking studies.

6. REFERENCES

1. Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics. 2008;9(Suppl 2):S6.

2. Montanucci L, Savojardo C, Martelli PL, Casadio R, Fariselli P. On the biases in predictions of protein stability changes upon variations: the INPS test case. Valencia A, editor. Bioinformatics. 2019 Jul 15;35(14):2525–7.

3. Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Computational and Structural Biotechnology Journal. 2020;18:1968–79.

4. Thiltgen G, Goldstein RA. Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency. Deane CM, editor. PLoS ONE. 2012 Oct 29;7(10):e46084.

5. Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics. 2015 Sep 1;31(17):2816–21.

6. Pucci F, Bernaerts K, Teheux F, Gilis D, Rooman M. Symmetry Principles in Optimization Problems: an application to Protein Stability Prediction★. IFAC-PapersOnLine. 2015;48(1):458–63.

7. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Research. 2017 Jul 3;45(W1):W229–35.

8. Kepp KP. Towards a "Golden Standard" for computing globin stability: Stability and structure sensitivity of myoglobin mutants. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2015 Oct;1854(10):1239–48.

9. Pucci F, Bernaerts KV, Kwasigroch JM, Rooman M. Quantification of biases in predictions of protein stability changes upon mutations. Valencia A, editor. Bioinformatics. 2018 Nov 1;34(21):3659–65.

10. Caldararu O, Blundell TL, Kepp KP. A base measure of precision for protein stability predictors: structural sensitivity. BMC Bioinformatics. 2021 Dec;22(1):88.

11. Usmanova DR, Bogatyreva NS, Ariño Bernad J, Eremina AA, Gorshkova AA, Kanevskiy GM, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. Valencia A, editor. Bioinformatics. 2018 Nov 1;34(21):3653–8.

12. Fang J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Briefings in Bioinformatics. 2020 Jul 15;21(4):1285–92.

13. Caldararu O, Mehra R, Blundell TL, Kepp KP. Systematic Investigation of the Data Set Dependency of Protein Stability Predictors. J Chem Inf Model. 2020 Oct 26;60(10):4772–84.

14. Etchebest C, Benros C, Bornot A, Camproux A-C, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J. 2007 Nov;36(8):1059–69.

15. Yang Y, Urolagin S, Niroula A, Ding X, Shen B, Vihinen M. PON-tstab: Protein Variant Stability Predictor. Importance of Training Data Quality. IJMS. 2018 Mar 28;19(4):1009.

16. Baldwin RL. Energetics of Protein Folding. Journal of Molecular Biology. 2007 Aug;371(2):283–301.

17. Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proceedings of the National Academy of Sciences. 2006 Nov 7;103(45):16623–33.

18. Fersht AR, Matouschek A, Serrano L. The folding of an enzyme. Journal of Molecular Biology. 1992 Apr;224(3):771–82.

19. Lazar T. Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. By K. A. Dill, S. Bromberg. Macromol Chem Phys. 2003 Sep;204(14):1800–1800.

20. Privalov PL, Makhatadze GI. Contribution of Hydration to Protein Folding Thermodynamics. Journal of Molecular Biology. 1993 Jul;232(2):660–79.

21. Griko YV, Makhatadze GI, Privalov PL, Hartley RW. Thermodynamics of barnase unfolding: Thermodynamics of barnase. Protein Science. 1994 Apr;3(4):669–76.

22. Serrano L, Jr JTK, Cann P, Matouschek A, Fersht AR. II. Substructure of Barnase and the Contribution of Different Interactions to Protein Stability. :22.

23. Horovitz A, Serrano L, Avron B, Bycroft M, Fersht AR. Strength and co-operativity of contributions of surface salt bridges to protein stability. Journal of Molecular Biology. 1990 Dec;216(4):1031–44.

24. Lodish HF. Molecular cell biology. Eighth edition. New York: W.H. Freeman-Macmillan Learning; 2016. 1170 p.

25. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat. 2001 Apr;17(4):263–70.

26. Yue P, Li Z, Moult J. Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. Journal of Molecular Biology. 2005 Oct;353(2):459–73.

27. Thomas PJ, Qu B-H, Pedersen PL. Defective protein folding as a basis of human disease. Trends in Biochemical Sciences. 1995 Nov;20(11):456–9.

28. Taverna DM, Goldstein RA. Why are proteins so robust to site mutations? Journal of Molecular Biology. 2002 Jan;315(3):479–84.

29. Luque I, Leavitt SA, Freire E. The Linkage Between Protein Folding and Functional Cooperativity: Two Sides of the Same Coin? Annu Rev Biophys Biomol Struct. 2002 Jun;31(1):235–56.

30. Bartlett AI, Radford SE. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. Nat Struct Mol Biol. 2009 Jun;16(6):582–8.

31. Kucukkal TG, Petukh M, Li L, Alexov E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Current Opinion in Structural Biology. 2015 Jun;32:18–24.

32. Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. Journal of Molecular Biology. 2002 Jul;320(2):369–87.

33. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Research. 2005 Jul 1;33(Web Server):W382–8.

34. Pokala N, Handel TM. Energy Functions for Protein Design: Adjustment with Protein–Protein Complex Affinities, Models for the Unfolded State, and Negative Design of Solubility and Specificity. Journal of Molecular Biology. 2005 Mar;347(1):203–27.

35. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. Predicting free energy changes using structural ensembles. Nat Methods. 2009 Jan;6(1):3–4.

36. Topham CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Engineering Design and Selection. 1997 Jan 1;10(1):7–21.

37. Gilis D, Rooman M. Prediction of stability changes upon single-site mutations using database-derived potentials. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta). 1999 Feb 15;101(1–3):46–50.

38. Bava KA. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Research. 2004 Jan 1;32(90001):120D – 121.

39. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research. 2005 Jul 1;33(Web Server):W306–10.

40. Li Y, Fang J. PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. Srinivasan N, editor. PLoS ONE. 2012 Oct 15;7(10):e47247.

41. Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model. 2019 Apr 22;59(4):1508–14.

42. Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Research. 2014 Jul 1;42(W1):W314–9.

43. Chen C-W, Lin J, Chu Y-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinformatics. 2013 Jan;14(S2):S5.

44. Kumar MDS. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Research. 2006 Jan 1;34(90001):D204–6.

45. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012;13(Suppl 4):S2.

46. Geng C, Vangone A, Folkers GE, Xue LC, Bonvin AMJJ. iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. Proteins. 2019 Feb;87(2):110–9.

47. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The Stability Effects of Protein Mutations Appear to be Universally Distributed. Journal of Molecular Biology. 2007 Jun;369(5):1318–32.

48. Nair PS, Vihinen M. VariBench: A Benchmark Database for Variations. Human Mutation. 2013 Jan;34(1):42–9.

49. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22(12):2577–637.

50. Touw WG, Baakman C, Black J, te Beek TAH, Krieger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 2015 Jan;43(Database issue):D364-368.

51. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422–3.

52. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014 Feb 1;30(3):335–42.

53. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO - multi agent stability prediction upon point mutations. BMC Bioinformatics. 2015 Dec;16(1):116.

54. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011 Dec;12(1):151.

55. Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Research. 2021 Jan 8;49(D1):D420–4.

56. Xavier JS, Nguyen T-B, Karmarkar M, Portelli S, Rezende PM, Velloso JPL, et al. ThermoMutDB: a thermodynamic database for missense mutations. Nucleic Acids Research. 2021 Jan 8;49(D1):D475–9.

57. Stourac J, Dubrava J, Musil M, Horackova J, Damborsky J, Mazurenko S, et al. FireProtDB: database of manually curated protein stability data. Nucleic Acids Research. 2021 Jan 8;49(D1):D319–24.

58. Wang CY, Chang PM, Ary ML, Allen BD, Chica RA, Mayo SL, et al. ProtaBank: A repository for protein design and engineering data: ProtaBank: A Protein Engineering Database. Protein Science. 2018 Jun;27(6):1113–24.

59. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Engineering Design and Selection. 2009 Sep 1;22(9):553–60.

60. 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 Oct 1;25(19):2537–43.

61. Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat. 2010 Mar 15;31(6):675–84.

62. Gribenko AV, Makhatadze GI. Role of the Charge–Charge Interactions in Defining Stability and Halophilicity of the CspB Proteins. Journal of Molecular Biology. 2007 Feb;366(3):842–56.

63. Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins. 2006 Feb 27;63(4):986–95.

64. Iqbal S, Li F, Akutsu T, Ascher DB, Webb GI, Song J. Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. Briefings in Bioinformatics. 2021 May 31;bbab184.

65. Giavarina D. Understanding Bland Altman analysis. Biochem Med. 2015;25(2):141–51.

66. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999 Apr;8(2):135–60.

67. Gilis D, Massar S, Cerf NJ, Rooman M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol. 2001;2(11):research0049.1.

68. Itzkovitz S, Alon U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 2007 Apr;17(4):405–12.

69. Cunningham B, Wells J. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science. 1989 Jun 2;244(4908):1081–5.

70. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. UniProt archive. Bioinformatics. 2004 Nov 22;20(17):3236–7.

7. APPENDICES

Appendix 1. Manually curated version of PON-tstab dataset

Curated	Status	Protein Name	PDB	Mutation (3D)	Mutation (seq)	ΔΔG
1	No change	Acidic fibroblast growth factor	2AFG	C117S	C132S	0.26
1	ID Corrected	Acidic fibroblast growth factor	2AFG	C16S	C31S	2.73
1	ID Corrected	Acidic fibroblast growth factor	2AFG	C83S	C98S	1.84
1	Mismatch	Acidic fibroblast growth factor	2AFG	F108Y	F123Y	-0.33
1	Mismatch	Acidic fibroblast growth factor	2AFG	H102Y	H117Y	-0.10
1	Mismatch	Acidic fibroblast growth factor	2AFG	H21Y	H36Y	-0.39
1	ID Corrected	Acidic fibroblast growth factor	2AFG	H93G	H108G	1.60
1	Mismatch	Acidic fibroblast growth factor	2AFG	L44F	L59F	-0.47
1	Mismatch	Acidic fibroblast growth factor	2AFG	V109I	V124Y	0.11
1	No change	Acyl-coenzyme a binding protein	2ABD	A34G	A35G	1.57
1	No change	Acyl-coenzyme a binding protein	2ABD	A9G	A10G	1.81
1	No change	Acyl-coenzyme a binding protein	2ABD	D21A	D22A	-0.42
1	No change	Acyl-coenzyme a binding protein	2ABD	D21H	D22H	0.56
1	No change	Acyl-coenzyme a binding protein	2ABD	E67A	E68A	0.36
1	No change	Acyl-coenzyme a binding protein	2ABD	F5A	F6A	2.52
1	No change	Acyl-coenzyme a binding protein	2ABD	K32A	K33A	1.02
1	No change	Acyl-coenzyme a binding protein	2ABD	K32E	K33E	1.68
1	No change	Acyl-coenzyme a binding protein	2ABD	K32R	K33R	1.83
1	No change	Acyl-coenzyme a binding protein	2ABD	K52M	K53M	-0.18

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Acyl-coenzyme a binding protein	2ABD	K54A	K55A	0.86
1	No change	Acyl-coenzyme a binding protein	2ABD	K54M	K55M	-0.27
1	No change	Acyl-coenzyme a binding protein	2ABD	L15A	L16A	3.10
1	No change	Acyl-coenzyme a binding protein	2ABD	L25A	L26A	1.02
1	No change	Acyl-coenzyme a binding protein	2ABD	L80A	L81A	3.70
1	No change	Acyl-coenzyme a binding protein	2ABD	P19A	P20A	1.07
1	No change	Acyl-coenzyme a binding protein	2ABD	P44A	P45A	1.40
1	No change	Acyl-coenzyme a binding protein	2ABD	Q33A	Q34A	3.66
1	No change	Acyl-coenzyme a binding protein	2ABD	T35A	T36A	1.09
1	No change	Acyl-coenzyme a binding protein	2ABD	V12A	V13A	1.69
1	No change	Acyl-coenzyme a binding protein	2ABD	V77A	V78A	1.14
1	No change	Acyl-coenzyme a binding protein	2ABD	Y28A	Y29A	2.47
1	No change	Acyl-coenzyme a binding protein	2ABD	Y28F	Y29F	1.06
1	No change	Acyl-coenzyme a binding protein	2ABD	Y28N	Y29N	2.23
1	No change	Acyl-coenzyme a binding protein	2ABD	Y31N	Y32N	1.52
1	No change	Acyl-coenzyme a binding protein	2ABD	Y73A	Y74A	4.83
1	No change	Acyl-coenzyme a binding protein	2ABD	Y73F	Y74F	-0.27
1	No change	Adenylate kinase	1ANK	D84H	D84H	1.40
1	No change	Adenylate kinase	1ANK	F86L	F86L	0.80
1	No change	Adenylate kinase	1ANK	G85V	G85V	2.40
1	No change	Adenylate kinase	2AKY	I213F	I215F	1.91
1	No change	Adenylate kinase	2AKY	N169D	N171D	2.15

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Adenylate kinase	1ANK	R88G	R88G	0.20
1	No change	Adenylate kinase	2AKY	T77H	T79H	2.15
1	No change	Adenylate kinase	2AKY	V8I	V10I	1.20
1	No change	Adrenodoxin	1AYF	C95S	C153S	-0.96
1	No change	Adrenodoxin	1AYF	D76E	D134E	-0.45
1	No change	Adrenodoxin	1AYF	H56Q	H114Q	1.66
1	No change	Adrenodoxin	1AYF	H56R	H114R	1.16
1	No change	Adrenodoxin	1AYF	H56T	H114T	1.19
1	No change	Adrenodoxin	1AYF	T54A	T112A	1.31
1	No change	Adrenodoxin	1AYF	T54S	T112S	-0.05
1	No change	Adrenodoxin	1AYF	Y82F	Y140F	0.01
1	No change	Adrenodoxin	1AYF	Y82L	Y140L	0.36
1	No change	Adrenodoxin	1AYF	Y82S	Y140S	0.33
1	No change	Adrenodoxin	1AYF	Y82W	Y140W	0.21
1	No change	Alpha spectrin (SH3 domain)	1SHG	F52Y	F1015Y	0.44
1	No change	Alpha spectrin (SH3 domain)	1SHG	K18F	K981F	-2.33
1	No change	Alpha spectrin (SH3 domain)	1SHG	K59F	K1022F	0.92
1	No change	Alpha spectrin (SH3 domain)	1SHG	K59Y	K1022Y	-0.84
1	No change	Alpha-lactalbumin	1HFZ	A106S	A125S	1.05
1	No change	Alpha-lactalbumin	1HFZ	H107A	H126A	0.79
1	No change	Alpha-lactalbumin	1HFZ	H107W	H126W	1.72
1	No change	Alpha-lactalbumin	1HFZ	H107Y	H126Y	0.19

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Alpha-lactalbumin	1HFZ	H32A	H51A	2.13
1	No change	Alpha-lactalbumin	1HFZ	H32Y	H51Y	-0.07
1	No change	Alpha-lactalbumin	1HFZ	I59W	I78W	0.93
1	No change	Alpha-lactalbumin	1HFZ	K114E	K133E	0.65
1	No change	Alpha-lactalbumin	1HFZ	K114N	K133N	-2.65
1	No change	Alpha-lactalbumin	1HFZ	K114Q	K133Q	0.60
1	No change	Alpha-lactalbumin	1HFZ	L110E	L129E	0.19
1	No change	Alpha-lactalbumin	1HFZ	L110H	L129H	-1.39
1	No change	Alpha-lactalbumin	1HFZ	L110R	L129R	-0.43
1	No change	Alpha-lactalbumin	1HFZ	Q117A	Q136A	0.96
1	No change	Alpha-lactalbumin	1HFZ	Q54A	Q73A	0.41
1	No change	Alpha-lactalbumin	1HFY	T29I	T48I	-3.23
1	No change	Alpha-lactalbumin	1HFY	T29V	T48V	-1.86
1	No change	Alpha-lactalbumin	1HFY	T38A	T57A	1.60
1	No change	Alpha-lactalbumin	1HFZ	V42A	V61A	0.93
1	No change	Alpha-lactalbumin	1HFZ	V42G	V61G	1.15
1	No change	Alpha-lactalbumin	1HFZ	V42N	V61N	0.24
1	No change	Alpha-lactalbumin	1HFZ	W104Y	W123Y	2.44
1	No change	Alpha-lactalbumin	1HFZ	W118H	W137H	0.60
1	No change	Alpha-lactalbumin	1HFZ	W118Y	W137Y	1.17
1	No change	Alpha-lactalbumin	1HFZ	Y103A	Y122A	2.39
1	No change	Alpha-lactalbumin	1HFZ	Y103P	Y122P	0.22

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Alpha-parvalbumin	1RTP	A21P	A22P	0.50
1	No change	Alpha-parvalbumin	1RTP	H26P	H27P	-1.25
1	No change	Alpha-parvalbumin	1RTP	K80S	K81S	0.29
0	PDB Related	Anthranilate Isomerase	no ID	S212C	S212C	0.96
0	Duplicated	Apoflavodoxin	1FTG	D75K	D127K	-1.00
1	No change	Apoflavodoxin	1FTG	D150K	D151K	0.28
1	No change	Apoflavodoxin	1FTG	D65K	D66K	0.26
1	No change	Apoflavodoxin	1FTG	D75K	D76K	-1.00
0	Duplicated	Apoflavodoxin	1FTG	E20K	E41K	-2.21
1	No change	Apoflavodoxin	1FTG	E20K	E21K	-1.60
1	No change	Apoflavodoxin	1FTG	E61K	E62K	-0.88
1	No change	Apoflavodoxin	1FTG	E72K	E73K	-1.28
1	No change	Arc repressor	1ARR	P8L	P8L	-2.48
1	No change	Barnase	1BNI	A32C	A79C	1.00
1	No change	Barnase	1BNI	A32D	A79D	0.70
1	No change	Barnase	1BNI	A32E	A79E	0.60
1	No change	Barnase	1BNI	A32F	A79F	0.70
1	No change	Barnase	1BNI	A32G	A79G	0.90
1	No change	Barnase	1BNI	A32H	A79H	0.80
1	No change	Barnase	1BNI	A32I	A79I	0.80
1	No change	Barnase	1BNI	A32K	A79K	0.20
1	No change	Barnase	1BNI	A32L	A79L	0.30

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Barnase	1BNI	A32M	A79M	0.30
1	No change	Barnase	1BNI	A32N	A79N	0.70
1	No change	Barnase	1BNI	A32P	A79P	4.10
1	No change	Barnase	1BNI	A32Q	A79Q	0.50
1	No change	Barnase	1BNI	A32R	A79R	0.10
1	No change	Barnase	1BNI	A32S	A79S	0.40
1	No change	Barnase	1BNI	A32T	A79T	0.80
1	No change	Barnase	1BNI	A32V	A79V	0.90
1	No change	Barnase	1BNI	A32W	A79W	1.00
1	No change	Barnase	1BNI	A32Y	A79Y	0.80
1	No change	Barnase	1BNI	D12A	D59A	-1.10
0	Duplicated	Barnase	1BNI	D12A	D59A	0.31
1	No change	Barnase	1BNI	D12G	D59G	1.20
1	No change	Barnase	1BNI	D44E	D91E	-0.10
1	No change	Barnase	1BNI	D54A	D101A	3.15
1	No change	Barnase	1BNI	D54N	D101N	2.53
1	No change	Barnase	1BNI	D75N	D122N	-4.80
1	No change	Barnase	1BNI	D8A	D55A	0.80
1	No change	Barnase	1BNI	D93N	D140N	4.11
1	No change	Barnase	1BNI	E29G	E76G	1.80
1	No change	Barnase	1BNI	E29Q	E76Q	0.00
1	No change	Barnase	1BNI	E73A	E120A	2.50

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Barnase	1BNI	E73G	E120G	5.40
1	No change	Barnase	1BNI	F7L	F54L	4.60
1	No change	Barnase	1BNI	G52A	G99A	5.30
1	No change	Barnase	1BNI	G52V	G99V	8.40
1	No change	Barnase	1BNI	G53A	G100A	3.40
1	No change	Barnase	1BNI	G53V	G100V	7.80
1	No change	Barnase	1BNI	G65S	G112S	-0.50
1	No change	Barnase	1BNI	H18K	H65K	1.20
1	No change	Barnase	1BNI	H18Q	H65Q	1.60
1	No change	Barnase	1BNI	I109A	I156A	2.10
1	No change	Barnase	1BNI	I109V	I156V	0.80
1	No change	Barnase	1BNI	I25A	I72A	3.50
1	No change	Barnase	1BNI	I25V	I72V	1.10
1	No change	Barnase	1BNI	I4A	I51A	1.40
1	No change	Barnase	1BNI	I4V	I51V	0.60
1	No change	Barnase	1BNI	I51A	I98A	4.70
1	No change	Barnase	1BNI	I51V	I98V	1.83
1	No change	Barnase	1BNI	I55A	I102A	1.10
1	No change	Barnase	1BNI	I55G	I102G	3.10
1	No change	Barnase	1BNI	I55T	I102T	0.60
1	No change	Barnase	1BNI	I55V	I102V	0.30
1	No change	Barnase	1BNI	I76A	I123A	1.90

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Barnase	1BNI	I76V	I123V	0.80
1	No change	Barnase	1BNI	I88A	I135A	4.21
1	No change	Barnase	1BNI	I88G	I135G	7.40
1	No change	Barnase	1BNI	I88L	I135L	0.30
1	No change	Barnase	1BNI	I88V	I135V	1.37
1	No change	Barnase	1BNI	I96A	I143A	3.46
1	No change	Barnase	1BNI	I96G	I143G	5.70
1	No change	Barnase	1BNI	I96V	I143V	1.02
1	No change	Barnase	1BNI	K108R	K155R	-0.90
1	No change	Barnase	1BNI	K19R	K66R	-0.20
1	No change	Barnase	1BNI	K27G	K74G	0.40
1	No change	Barnase	1BNI	K62R	K109R	0.45
1	No change	Barnase	1BNI	K66A	K113A	-0.20
1	No change	Barnase	1BNI	L14A	L61A	4.53
1	No change	Barnase	1BNI	L33Q	L80Q	1.30
1	No change	Barnase	1BNI	L89G	L136G	7.00
1	No change	Barnase	1BNI	L89T	L136T	2.50
1	No change	Barnase	1BNI	L89V	L136V	0.30
1	No change	Barnase	1BNI	L95G	L142G	4.70
1	No change	Barnase	1BNI	N23A	N70A	2.20
1	No change	Barnase	1BNI	N41D	N88D	2.50
1	No change	Barnase	1BNI	N58A	N105A	2.70

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Barnase	1BNI	N58D	N105D	-0.50
1	No change	Barnase	1BNI	N5A	N52A	1.90
1	No change	Barnase	1BNI	N77A	N124A	1.60
1	No change	Barnase	1BNI	N84A	N131A	2.00
1	No change	Barnase	1BNI	Q104A	Q151A	0.20
1	No change	Barnase	1BNI	Q15A	Q62A	0.20
1	No change	Barnase	1BNI	Q15G	Q62G	1.60
1	No change	Barnase	1BNI	Q15I	Q62I	-1.00
1	No change	Barnase	1BNI	Q31A	Q78A	-0.10
1	No change	Barnase	1BNI	Q31S	Q78S	0.20
1	No change	Barnase	1BNI	R110A	R157A	0.40
1	No change	Barnase	1BNI	R69K	R116K	3.13
1	No change	Barnase	1BNI	R69M	R116M	2.12
1	No change	Barnase	1BNI	R69S	R116S	2.72
1	No change	Barnase	1BNI	R72G	R119G	2.50
1	No change	Barnase	1BNI	R83K	R130K	4.13
1	No change	Barnase	1BNI	R83Q	R130Q	-2.05
1	No change	Barnase	1BNI	S28E	S75E	-0.40
1	No change	Barnase	1BNI	S85A	S132A	0.12
1	No change	Barnase	1BNI	S91A	S138A	1.48
1	No change	Barnase	1BNI	S92A	S139A	2.82
1	ID Corrected	Barnase	1BNI	S92G	S139G	2.95

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Barnase	1BNI	T100G	T147G	2.80
1	No change	Barnase	1BNI	T105V	T152V	2.20
1	No change	Barnase	1BNI	T16R	T63R	-0.42
1	No change	Barnase	1BNI	T16S	T63S	1.79
1	No change	Barnase	1BNI	T26A	T73A	2.05
1	No change	Barnase	1BNI	T26E	T73E	0.05
1	No change	Barnase	1BNI	T26G	T73G	1.72
1	No change	Barnase	1BNI	T26N	T73N	1.29
1	No change	Barnase	1BNI	T26Q	T73Q	1.72
1	No change	Barnase	1BNI	T26S	T73S	0.56
1	No change	Barnase	1BNI	T26V	T73V	2.31
1	No change	Barnase	1BNI	T6A	T53A	2.29
1	No change	Barnase	1BNI	T6D	T53D	-0.11
1	No change	Barnase	1BNI	T6E	T53E	0.27
1	No change	Barnase	1BNI	T6G	T53G	1.27
1	No change	Barnase	1BNI	T6N	T53N	1.27
1	No change	Barnase	1BNI	T6Q	T53Q	1.87
1	No change	Barnase	1BNI	T6S	T53S	0.22
1	No change	Barnase	1BNI	T79V	T126V	-0.30
1	No change	Barnase	1BNI	T99V	T146V	2.70
1	No change	Barnase	1BNI	V10A	V57A	3.40
1	No change	Barnase	1BNI	V10T	V57T	2.50

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Barnase	1BNI	V36A	V83A	1.30
1	No change	Barnase	1BNI	V36T	V83T	1.10
1	No change	Barnase	1BNI	V45A	V92A	1.80
1	No change	Barnase	1BNI	V45T	V92T	2.40
1	No change	Barnase	1BNI	Y103F	Y150F	0.00
1	No change	Barnase	1BNI	Y13A	Y60A	3.30
1	No change	Barnase	1BNI	Y13G	Y60G	6.40
1	No change	Barnase	1BNI	Y17A	Y64A	2.00
1	No change	Barnase	1BNI	Y17G	Y64G	4.90
1	No change	Barnase	1BNI	Y24F	Y71F	0.00
1	No change	Barnase	1BNI	Y78F	Y125F	1.45
1	No change	Barnase	1BNI	Y97G	Y144G	6.60
0	PDB Related, Duplicated	Barnase	no ID		D59G	0.72
0	PDB Related, Mismatch	Barnase	no ID		D55G	1.00
1	No change	Barstar	1BTA	C82A	C83A	0.48
0	Mismatch	Beta lactamase	3BLS	Y150F	Y166F	0.70
1	No change	Calbindin D9k	1IGV	D19N	D23N	-0.72
1	No change	Calbindin D9k	1IGV	E17Q	E21Q	-0.38
1	No change	Calbindin D9k	1IGV	E26Q	E30Q	-0.10
1	No change	Catabolite activator protein	1G6N	S128A	S129A	0.22
1	No change	Catabolite activator protein	1G6N	S128P	S129P	-0.14
1	No change	Cell surface receptor protein CD2	1CDC	A40G	A62G	1.71

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Cell surface receptor protein CD2	1CDC	I18V	I40V	1.17
1	No change	Cell surface receptor protein CD2	1CDC	I57V	I79V	0.43
1	No change	Cell surface receptor protein CD2	1CDC	L16V	L38V	2.20
1	No change	Cell surface receptor protein CD2	1CDC	L50V	L72V	-0.37
1	No change	Cell surface receptor protein CD2	1CDC	L89V	L111V	0.14
1	No change	Cell surface receptor protein CD2	1CDC	L95V	L117V	0.75
1	No change	Cell surface receptor protein CD2	1CDC	V30A	V52A	4.88
1	No change	Cell surface receptor protein CD2	1CDC	V78A	V100A	3.01
1	No change	Cellular prion protein	1AG2	E200K	E199K	0.14
1	No change	Cellular prion protein	1AG2	F198S	F197S	2.46
1	No change	Cellular prion protein	1AG2	M129V	M128V	0.33
1	No change	Cellular prion protein	1AG2	Q217R	Q216R	2.13
1	No change	Cellular prion protein	1AG2	R208H	R207H	1.43
1	No change	Cellular prion protein	1AG2	T183A	T182A	4.62
1	No change	Cellular prion protein	1AG2	T190V	T189V	-0.17
1	No change	Cellular prion protein	1AG2	V180I	V179I	0.50
1	No change	Cellular prion protein	1AG2	V210I	V209I	0.26
1	No change	Chitinase	1KFW	G253P	G291P	0.62
1	No change	Chitinase	1KFW	G405Q	G443Q	0.62
1	No change	Chitinase	1KFW	G92P	G130P	-0.53
1	No change	Chitinase	1KFW	N197K	N235K	-0.84
1	No change	Chitosanase	1CHK	W101F	W141F	4.27

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Chitosanase	1CHK	W227F	W267F	2.51
1	No change	Chitosanase	1CHK	W28F	W68F	4.43
1	Mismatch	Chymotrypsin inhibitor	2CI2	A35G	A36G	1.09
1	No change	Chymotrypsin inhibitor	2CI2	A77G	A78G	1.88
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	A77G	A78G	1.88
1	Mismatch	Chymotrypsin inhibitor	2CI2	D42A	D43A	0.96
1	Mismatch	Chymotrypsin inhibitor	2CI2	D64A	D65A	0.80
1	Mismatch	Chymotrypsin inhibitor	2CI2	D71A	D72A	3.41
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	E26A	E27A	0.47
1	Mismatch	Chymotrypsin inhibitor	2CI2	E26A	E27A	0.47
1	Mismatch	Chymotrypsin inhibitor	2CI2	E26Q	E27Q	0.62
1	Mismatch	Chymotrypsin inhibitor	2CI2	E33D	E34D	0.52
1	Mismatch	Chymotrypsin inhibitor	2CI2	E33N	E34N	0.70
1	Mismatch	Chymotrypsin inhibitor	2CI2	E33Q	E34Q	0.29
1	Mismatch	Chymotrypsin inhibitor	2CI2	E34D	E35D	0.74
1	Mismatch	Chymotrypsin inhibitor	2CI2	E34N	E35N	1.07
1	Mismatch	Chymotrypsin inhibitor	2CI2	E34Q	E35Q	0.47
0	Duplicated	Chymotrypsin inhibitor	2CI2	E45A	E46A	0.32
1	No change	Chymotrypsin inhibitor 2	2CI2	E45A	E46A	0.57
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	E60A	E61A	0.70
1	No change	Chymotrypsin inhibitor	2CI2	E60A	E61A	0.68
1	Mismatch	Chymotrypsin inhibitor	2CI2	F69A	F70A	3.84

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	Mismatch	Chymotrypsin inhibitor	2CI2	F69L	F70L	2.11
1	Mismatch	Chymotrypsin inhibitor	2CI2	F69V	F70V	2.39
1	No change	Chymotrypsin inhibitor	2CI2	I39V	I40V	1.29
1	No change	Chymotrypsin inhibitor	2CI2	I48A	I49A	3.84
1	No change	Chymotrypsin inhibitor	2CI2	I48V	I49V	1.10
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	I49A	I50A	2.12
1	No change	Chymotrypsin inhibitor	2CI2	I49A	I50A	2.12
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	I49G	I50G	3.52
1	No change	Chymotrypsin inhibitor	2CI2	I49G	I50G	3.52
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	I49T	I50T	1.34
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	I49V	I50V	0.08
1	No change	Chymotrypsin inhibitor	2CI2	I49T	L50T	1.34
1	No change	Chymotrypsin inhibitor	2CI2	I49V	I50V	-0.08
1	Mismatch	Chymotrypsin inhibitor	2CI2	I56A	I57A	0.03
1	No change	Chymotrypsin inhibitor	2CI2	I76A	I77A	4.25
1	No change	Chymotrypsin inhibitor	2CI2	I76V	I77V	-0.20
1	Mismatch	Chymotrypsin inhibitor	2CI2	K21A	K22A	0.55
1	Mismatch	Chymotrypsin inhibitor	2CI2	K21M	K22M	0.67
1	Mismatch	Chymotrypsin inhibitor	2CI2	K30A	K31A	-0.42
1	Mismatch	Chymotrypsin inhibitor	2CI2	K36A	K37A	0.49
1	Mismatch	Chymotrypsin inhibitor	2CI2	K36G	K37G	2.32
1	Mismatch	Chymotrypsin inhibitor	2CI2	K37A	K38A	-0.22

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	Mismatch	Chymotrypsin inhibitor	2CI2	K37G	K38G	0.98
0	Duplicated	Chymotrypsin inhibitor	2CI2	K43A	K44A	0.65
1	No change	Chymotrypsin inhibitor 2	2CI2	K43A	K44A	0.64
1	Mismatch	Chymotrypsin inhibitor	2CI2	K43G	K44G	3.19
1	Mismatch	Chymotrypsin inhibitor	2CI2	K72N	K73N	0.00
1	No change	Chymotrypsin inhibitor	2CI2	L27A	L28A	2.66
1	Mismatch	Chymotrypsin inhibitor	2CI2	L40A	L41A	1.33
1	Mismatch	Chymotrypsin inhibitor	2CI2	L40G	L41G	1.38
1	Mismatch	Chymotrypsin inhibitor	2CI2	L51A	L52A	2.37
1	Mismatch	Chymotrypsin inhibitor	2CI2	L51I	L52I	0.26
1	Mismatch	Chymotrypsin inhibitor	2CI2	L51V	L52V	0.50
1	No change	Chymotrypsin inhibitor	2CI2	L68A	L69A	3.82
1	Mismatch	Chymotrypsin inhibitor	2CI2	N75A	N76A	0.83
1	Mismatch	Chymotrypsin inhibitor	2CI2	N75D	N76D	1.21
1	Mismatch	Chymotrypsin inhibitor	2CI2	P25A	P26A	1.57
0	Duplicated	Chymotrypsin inhibitor	2CI2	P44A	P45A	1.76
1	No change	Chymotrypsin inhibitor 2	2CI2	P44A	P45A	1.93
1	Mismatch	Chymotrypsin inhibitor	2CI2	P52A	P53A	0.17
1	Mismatch	Chymotrypsin inhibitor	2CI2	P80A	P81A	3.34
1	Mismatch	Chymotrypsin inhibitor	2CI2	Q41A	Q42A	0.02
1	Mismatch	Chymotrypsin inhibitor	2CI2	Q41G	Q42G	0.60
1	Mismatch	Chymotrypsin inhibitor	2CI2	R62A	R63A	0.58

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	T22A	T23A	0.85
1	Mismatch	Chymotrypsin inhibitor	2CI2	S31A	S32A	0.89
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	T22G	T23G	1.16
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	T22V	T23V	0.32
1	Mismatch	Chymotrypsin inhibitor	2CI2	S31G	S32G	0.80
1	No change	Chymotrypsin inhibitor	2CI2	T22A	T23A	0.85
1	No change	Chymotrypsin inhibitor	2CI2	T22G	T23G	1.16
1	No change	Chymotrypsin inhibitor	2CI2	T22V	T23V	0.32
1	Mismatch	Chymotrypsin inhibitor	2CI2	T55A	T56A	-0.23
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	T58A	T59A	0.72
0	Duplicated	Chymotrypsin inhibitor	no ID	T58D	T59D	-0.02
1	Mismatch	Chymotrypsin inhibitor	2CI2	T55S	T56S	0.02
1	Mismatch	Chymotrypsin inhibitor	2CI2	T55V	T56V	0.76
1	No change	Chymotrypsin inhibitor	2CI2	T58A	T59A	0.69
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V53A	V54A	0.64
1	No change	Chymotrypsin inhibitor	2CI2	T58D	T59D	-0.04
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V53G	V54G	2.43
1	No change	Chymotrypsin inhibitor	2CI2	V38A	V39A	0.48
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V53T	V54T	1.03
1	No change	Chymotrypsin inhibitor	2CI2	V53A	V54A	0.64
1	No change	Chymotrypsin inhibitor	2CI2	V53G	V54G	2.43
1	No change	Chymotrypsin inhibitor	2CI2	V53T	V54T	1.03

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Chymotrypsin inhibitor	2CI2	V57A	V58A	1.47
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V79A	V80A	1.51
1	No change	Chymotrypsin inhibitor	2CI2	V66A	V67A	4.88
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V79G	V80G	3.24
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V79T	V80T	0.38
1	No change	Chymotrypsin inhibitor	2CI2	V70A	V71A	1.95
1	No change	Chymotrypsin inhibitor	2CI2	V79A	V80A	1.51
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V82A	V83A	1.45
1	No change	Chymotrypsin inhibitor	2CI2	V79G	V80G	3.24
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V82G	V83G	3.50
1	No change	Chymotrypsin inhibitor	2CI2	V79T	V80T	0.38
0	PDB related, Duplicated	Chymotrypsin inhibitor	no ID	V82T	V83T	1.15
1	No change	Chymotrypsin inhibitor	2CI2	V82A	V83A	1.45
1	No change	Chymotrypsin inhibitor	2CI2	V82G	V83G	3.50
1	No change	Chymotrypsin inhibitor	2CI2	V82T	V83T	1.15
1	No change	Cold shock protein	1CSP	A46E	A46E	0.60
1	ID Corrected	Cold shock protein	1CSP	A46L	A46L	-0.81
1	No change	Cold shock protein B	1CSP	D24K	D24K	0.48
1	No change	Cold shock protein B	1CSP	D24N	D24N	0.72
1	No change	Cold shock protein B	1CSP	D25K	D25K	1.43
1	No change	Cold shock protein B	1CSP	D25Q	D25Q	0.81
0	PDB Related, Mismatch	Cold shock protein	no ID	K20Q	K20Q	0.00

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Cold shock protein B	1CSP	E12K	E12K	0.62
1	No change	Cold shock protein B	1CSP	E19K	E19K	0.57
1	No change	Cold shock protein B	1CSP	E19Q	E19Q	0.19
1	No change	Cold shock protein	1C9O	E21A	E21A	0.29
1	No change	Cold shock protein	1C9O	E21K	E21K	0.26
0	PDB Related, Duplicated	Cold shock protein B	no ID	D24K	D24K	0.33
1	No change	Cold shock protein B	1CSP	E21Q	E21Q	0.31
1	No change	Cold shock protein	1CSP	E3L	E3L	-1.60
1	No change	Cold shock protein B	1CSP	E3Q	E3Q	-0.98
1	No change	Cold shock protein	1CSP	E3R	E3R	-2.65
1	ID Corrected	Cold shock protein	1CSP	E3V	E3V	-1.77
0	PDB Related, Duplicated	Cold shock protein B	no ID	D24N	D24N	0.26
1	No change	Cold shock protein B	1CSP	E42K	E42K	-0.02
1	No change	Cold shock protein B	1CSP	E42Q	E42Q	0.12
1	No change	Cold shock protein B	1CSP	E43K	E43K	-0.17
1	No change	Cold shock protein B	1CSP	E43Q	E43Q	-0.05
1	No change	Cold shock protein	1C9O	E46A	E46A	0.22
1	No change	Cold shock protein	1C9O	E46K	E46K	0.65
1	No change	Cold shock protein B	1CSP	E50K	E50K	0.57
1	No change	Cold shock protein B	1CSP	E50Q	E50Q	1.22
1	No change	Cold shock protein B	1CSP	E53K	E53K	0.19
1	No change	Cold shock protein B	1CSP	E53Q	E53Q	-0.19

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	ID Corrected	Cold shock protein	1CSP	E66K	E66K	-2.18
1	No change	Cold shock protein	1CSP	E66L	E66L	-2.10
1	ID Corrected	Cold shock protein	1CSP	E66V	E66V	-1.72
1	ID Corrected	Cold shock protein	1CSP	F15A	F15A	1.67
1	ID Corrected	Cold shock protein	1CSP	F17A	F17A	1.34
1	ID Corrected	Cold shock protein	1CSP	F27A	F27A	0.69
0	PDB Related	Cold shock protein B	no ID	D25N	D25N	0.72
0	PDB Related, Triplicated	Cold shock protein B	no ID	E21K	E21K	0.96
0	PDB Related, Duplicated	Cold shock protein B	no ID	E21Q	E21Q	0.38
0	PDB Related, Duplicated	Cold shock protein B	no ID	K13E	K13E	0.31
0	PDB Related, Duplicated	Cold shock protein B	no ID	K5E	K5E	1.91
0	PDB Related, Duplicated	Cold shock protein B	no ID	K5Q	K5Q	0.65
0	PDB Related, Duplicated	Cold shock protein B	no ID	R3E	R3E	3.83
0	PDB Related	Cold shock protein B	no ID	R3Q	R3Q	1.00
0	Mismatch	Cold shock protein B	1CSP	E12K	K12E	-0.43
0	Triplicated	Cold shock protein B	1CSP	E21K	E21K	0.12
0	PDB Related	Cold shock protein B	no ID	D10K	D10K	2.73
1	No change	Cold shock protein	1CSP	F38A	F38A	-0.31
1	No change	Cold shock protein	1C9O	G23Q	G23Q	0.29
1	No change	Cold shock protein	1C9O	H29E	H29E	0.77
0	PDB Related, Duplicated	Cold shock protein B	no ID	D25K	D25K	0.91
1	No change	Cold shock protein B	1CSP	K13E	K13E	0.29

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Cold shock protein B	1CSP	K13Q	K13Q	0.14
1	No change	Cold shock protein B	1CSP	K39E	K39E	0.26
1	No change	Cold shock protein B	1CSP	K39Q	K39Q	0.02
1	No change	Cold shock protein B	1CSP	K5E	K5E	3.11
1	No change	Cold shock protein B	1CSP	K5Q	K5Q	1.96
1	No change	Cold shock protein B	1CSP	K65E	K65E	1.70
1	No change	Cold shock protein B	1CSP	K65Q	K65Q	0.77
1	ID Corrected	Cold shock protein	1CSP	L2R	L2R	0.36
1	No change	Cold shock protein	1C9O	L66E	L66E	1.24
1	No change	Cold shock protein B	1CSP	N10D	N10D	-0.31
0	PDB Related, Duplicated	Cold shock protein B	no ID	E43K	E43K	0.22
1	No change	Cold shock protein B	1CSP	N10K	N10K	1.51
0	PDB Related, Duplicated	Cold shock protein B	no ID	E43Q	E43Q	0.14
0	PDB Related	Cold shock protein B	no ID	E48K	E48K	1.34
0	PDB Related	Cold shock protein B	no ID	E48Q	E48Q	0.45
1	No change	Cold shock protein	1C9O	N11S	N11S	-0.33
0	PDB Related, Duplicated	Cold shock protein B	no ID	E50K	E50K	-0.40
1	No change	Cold shock protein B	1CSP	N55D	N55D	-0.50
0	PDB Related, Duplicated	Cold shock protein B	no ID	E50Q	E50Q	-0.24
1	No change	Cold shock protein B	1CSP	N55K	N55K	0.10
0	PDB Related, Duplicated	Cold shock protein B	no ID	E53K	E53K	-0.07
1	No change	Cold shock protein	1C9O	Q2L	Q2L	-0.55

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

0	PDB Related, Duplicated	Cold shock protein B	no ID	E53Q	E53Q	-0.31
0	PDB Related	Cold shock protein B	no ID	K12Q	K12Q	-0.53
1	No change	Cold shock protein	1C9O	Q53E	Q53E	0.10
1	No change	Cold shock protein	1C9O	R3A	R3A	1.89
0	PDB Related, Mismatch	Cold shock protein B	no ID	K20E	K20E	1.32
0	PDB Related	Cold shock protein B	no ID	K20V	K20V	-2.18
1	No change	Cold shock protein	1C9O	R3E	R3E	2.75
0	PDB Related, Duplicated	Cold shock protein B	no ID	K39E	K39E	-0.19
1	No change	Cold shock protein	1C9O	R3K	R3K	0.19
0	PDB Related, Duplicated	Cold shock protein B	no ID	K39Q	K39Q	-0.43
0	PDB Related	Cold shock protein B	no ID	K42E	K42E	0.22
0	PDB Related	Cold shock protein B	no ID	K42Q	K42Q	-0.07
0	PDB Related	Cold shock protein B	no ID	K55E	K55E	-0.81
0	PDB Related	Cold shock protein B	no ID	K55Q	K55Q	-0.79
1	No change	Cold shock protein	1C9O	R3L	R3L	0.93
1	No change	Cold shock protein	1C9O	R56E	R56E	-0.77
1	No change	Cold shock protein B	1CSP	R56Q	R56Q	-0.38
1	No change	Cold shock protein	1C9O	S24D	S24D	-0.22
0	PDB Related	Cold shock protein B	no ID	K7E	K7E	2.42
0	PDB Related	Cold shock protein B	no ID	K7Q	K7Q	1.22
1	No change	Cold shock protein B	1CSP	S48E	S48E	-0.10
1	ID Corrected	Cold shock protein	1CSP	S48K	S48K	-1.41

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

0	Duplicated	Cold shock protein B	1CSP	S48K	S48K	-0.74
1	No change	Cold shock protein	1C9O	T31S	T31S	-0.17
1	ID Corrected	Cold shock protein	1CSP	T64R	T64R	-1.08
1	No change	Cold shock protein B	1CSP	V20K	V20K	1.94
1	No change	Cold shock protein B	1CSP	V20Q	V20Q	1.82
1	No change	Cold shock protein	1C9O	V64T	V64T	0.26
1	No change	Cold shock protein	1C9O	Y15F	Y15F	0.05
1	No change	Cytochrome b5	1CYO	F35H	F40H	2.82
1	No change	Cytochrome b5	1CYO	F35L	F40L	1.87
1	No change	Cytochrome b5	1CYO	F35Y	F40Y	-0.79
1	No change	Cytochrome b5	1B5M	P81A	P79A	1.15
1	No change	Cytochrome b5	1CYO	V45E	V50E	2.68
1	No change	Cytochrome b5	1CYO	V45H	V50H	1.34
1	No change	Cytochrome b5	1CYO	V45Y	V50Y	1.53
1	No change	Cytochrome b5	1CYO	V61E	V66E	1.10
1	No change	Cytochrome b5	1CYO	V61H	V66H	1.63
1	No change	Cytochrome b5	1CYO	V61K	V66K	2.34
1	No change	Cytochrome b5	1CYO	V61Y	V66Y	1.29
1	No change	Cytochrome c	1YCC	N52I	N58I	-3.14
1	No change	Cytochrome c2	1C2R	G34S	G55S	2.20
1	No change	Cytochrome c2	1C2R	K12D	K33D	1.30
1	No change	Cytochrome c2	1C2R	K14E	K35E	1.30

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Cytochrome c2	1C2R	K32E	K53E	-0.20
1	No change	Cytochrome c2	1C2R	P35A	P56A	2.00
1	No change	Cytochrome c551	451C	E43Y	E65Y	-0.60
1	No change	Cytochrome c551	451C	F34Y	F56Y	-1.90
1	No change	Cytochrome c551	451C	F7A	F29A	-1.10
1	No change	Cytochrome c551	451C	Q37R	Q59R	-0.50
1	No change	Cytochrome c551	451C	V13M	V35M	-0.40
1	No change	Cytochrome c551	451C	V78I	V100I	-1.00
1	No change	Dihydrofolate reductase	1RX4	E139K	E139K	2.36
1	No change	Dihydrofolate reductase	1RX4	E139Q	E139Q	1.36
1	No change	Dihydrofolate reductase	1RX4	L28R	L28R	-0.50
1	No change	Dihydrofolate reductase	1RX4	V75A	V75A	0.20
1	No change	Dihydrofolate reductase	1RX4	V75C	V75C	0.20
1	No change	Dihydrofolate reductase	1RX4	V75H	V75H	1.90
1	No change	Dihydrofolate reductase	1RX4	V75I	V75I	2.00
1	No change	Dihydrofolate reductase	1RX4	V75R	V75R	2.80
1	No change	Dihydrofolate reductase	1RX4	V75S	V75S	1.90
1	No change	Dihydrofolate reductase	1RX4	V75Y	V75Y	2.30
1	No change	Dihydrofolate reductase	1RX4	V88A	V88A	-0.20
1	No change	Dihydrofolate reductase	1RX4	V88I	V88I	1.73
1	No change	DsbA	1A23	H32L	H51L	-5.30
1	No change	DsbA	1A23	H32S	H51S	-5.20

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	DsbA	1A23	H32Y	H51Y	-6.80
1	No change	Fibroblast growth factor 1	2AFG	C117I	C132I	1.20
1	ID Corrected	Fibroblast growth factor 1	2AFG	C117V	C132V	-1.50
1	Mismatch	Fibroblast growth factor 1	2AFG	F132W	F147W	-0.29
1	Mismatch	Fibroblast growth factor 1	2AFG	F85W	F100W	0.07
0	Duplicated	Fibroblast growth factor 1	2AFG	L44F	L59F	-0.72
1	No change	Fibroblast growth factor 1	2AFG	L44P	L59P	-0.69
1	No change	Fibroblast growth factor 1	2AFG	L44W	L59W	-3.40
1	No change	Fibroblast growth factor 1	2AFG	L73V	L88V	1.46
1	No change	Fibroblast growth factor 1	2AFG	V109L	V124L	0.57
1	ID Corrected	Fibroblast growth factor 1	2AFG	V31I	V46I	-4.00
1	ID Corrected	Flavodoxin	1FLV	A101L	A102L	-0.12
1	ID Corrected	Flavodoxin	1FLV	A101V	A102V	0.29
1	ID Corrected	Flavodoxin	1FLV	G87A	G88A	-0.04
1	No change	Flavodoxin	1FLV	G87L	G88L	0.64
1	No change	Flavodoxin	1FLV	G87V	G88V	0.13
1	ID Corrected	Flavodoxin	1FLV	V18I	V19I	-0.69
1	ID Corrected	Flavodoxin	1FLV	V18L	V19L	-0.20
1	ID Corrected	Flavodoxin	1FLV	V83I	V84I	-0.02
1	No change	Gene V	1VQB	A86T	A86T	0.66
1	No change	Gene V	1VQB	A86V	A86V	-0.47
1	No change	Gene V	1VQB	C33A	C33A	0.50

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Gene V	1VQB	C33I	C33I	0.90
1	No change	Gene V	1VQB	C33L	C33L	2.60
1	No change	Gene V	1VQB	C33M	C33M	3.49
1	No change	Gene V	1VQB	C33S	C33S	4.30
1	No change	Gene V	1VQB	C33T	C33T	4.60
1	No change	Gene V	1VQB	C33V	C33V	0.18
1	No change	Gene V	1VQB	D36C	D36C	2.10
1	No change	Gene V	1VQB	D36N	D36N	1.00
1	No change	Gene V	1VQB	D50H	D50H	1.60
1	No change	Gene V	1VQB	E30F	E30F	-2.00
1	No change	Gene V	1VQB	E30M	E30M	-0.60
1	No change	Gene V	1VQB	E30N	E30N	1.10
1	No change	Gene V	1VQB	E40C	E40C	1.60
1	No change	Gene V	1VQB	E40T	E40T	0.40
1	No change	Gene V	1VQB	F13T	F13T	0.67
1	No change	Gene V	1VQB	F68L	F68L	4.30
1	No change	Gene V	1VQB	F68V	F68V	5.00
1	No change	Gene V	1VQB	F73W	F73W	-0.80
1	No change	Gene V	1VQB	H64C	H64C	-0.50
1	No change	Gene V	1VQB	I47A	I47A	7.10
1	No change	Gene V	1VQB	I47C	I47C	5.30
1	No change	Gene V	1VQB	I47F	I47F	2.00

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Gene V	1VQB	I47L	I47L	0.70
1	No change	Gene V	1VQB	I47M	I47M	2.20
1	No change	Gene V	1VQB	I47T	I47T	7.60
1	No change	Gene V	1VQB	I47V	I47V	2.60
1	No change	Gene V	1VQB	I6V	I6V	0.68
1	No change	Gene V	1VQB	I78C	I78C	4.40
1	No change	Gene V	1VQB	I78T	I78T	6.60
1	No change	Gene V	1VQB	I78V	I78V	1.30
1	No change	Gene V	1VQB	K24V	K24V	-0.80
1	No change	Gene V	1VQB	K69H	K69H	1.30
1	No change	Gene V	1VQB	K69M	K69M	-0.10
1	No change	Gene V	1VQB	L28V	L28V	-1.10
1	No change	Gene V	1VQB	L32H	L32H	0.90
1	No change	Gene V	1VQB	L32R	L32R	1.60
1	No change	Gene V	1VQB	L32W	L32W	-2.80
1	No change	Gene V	1VQB	L32Y	L32Y	-1.04
1	No change	Gene V	1VQB	L37A	L37A	7.70
1	No change	Gene V	1VQB	L37C	L37C	4.60
1	No change	Gene V	1VQB	L37I	L37I	1.40
1	No change	Gene V	1VQB	L37T	L37T	5.20
1	No change	Gene V	1VQB	L37V	L37V	3.50
1	No change	Gene V	1VQB	L49A	L49A	6.10

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Gene V	1VQB	L49C	L49C	4.10
1	No change	Gene V	1VQB	L49I	L49I	1.90
1	No change	Gene V	1VQB	L49T	L49T	5.70
1	No change	Gene V	1VQB	L49V	L49V	2.90
1	No change	Gene V	1VQB	L65P	L65P	1.47
1	No change	Gene V	1VQB	L81C	L81C	3.70
1	No change	Gene V	1VQB	L81T	L81T	5.10
1	No change	Gene V	1VQB	L81V	L81V	0.20
1	No change	Gene V	1VQB	M77A	M77A	2.10
1	No change	Gene V	1VQB	M77C	M77C	0.00
1	No change	Gene V	1VQB	M77F	M77F	0.20
1	No change	Gene V	1VQB	M77I	M77I	-1.60
1	No change	Gene V	1VQB	M77L	M77L	1.20
1	No change	Gene V	1VQB	M77T	M77T	0.80
1	No change	Gene V	1VQB	M77V	M77V	-1.20
1	No change	Gene V	1VQB	R82C	R82C	1.50
1	No change	Gene V	1VQB	S67C	S67C	3.70
1	No change	Gene V	1VQB	S67T	S67T	1.60
1	No change	Gene V	1VQB	T48C	T48C	0.80
1	No change	Gene V	1VQB	T48V	T48V	0.00
1	No change	Gene V	1VQB	T62C	T62C	0.70
1	No change	Gene V	1VQB	T62V	T62V	-1.30

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Gene V	1VQB	V19C	V19C	0.30
1	No change	Gene V	1VQB	V19T	V19T	0.60
1	No change	Gene V	1VQB	V35A	V35A	2.30
1	No change	Gene V	1VQB	V35C	V35C	1.50
1	No change	Gene V	1VQB	V35F	V35F	3.20
1	No change	Gene V	1VQB	V35I	V35I	0.70
1	No change	Gene V	1VQB	V35L	V35L	2.70
1	No change	Gene V	1VQB	V35M	V35M	1.10
1	No change	Gene V	1VQB	V35T	V35T	5.30
1	No change	Gene V	1VQB	V43C	V43C	2.10
1	No change	Gene V	1VQB	V43T	V43T	1.60
1	No change	Gene V	1VQB	V45A	V45A	2.10
1	No change	Gene V	1VQB	V45C	V45C	0.00
1	No change	Gene V	1VQB	V45L	V45L	3.00
1	No change	Gene V	1VQB	V45T	V45T	3.50
1	No change	Gene V	1VQB	V63C	V63C	4.10
1	No change	Gene V	1VQB	V63T	V63T	5.00
1	No change	Gene V	1VQB	V70C	V70C	3.30
1	No change	Gene V	1VQB	V70P	V70P	5.10
1	No change	Gene V	1VQB	V70T	V70T	3.50
1	No change	Gene V	1VQB	Y26R	Y26R	0.40
1	No change	Gene V	1VQB	Y41A	Y41A	0.40

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Gene V	1VQB	Y41F	Y41F	0.60
1	No change	Glutamate dehydrogenase	1B26	E231A	E232A	0.17
1	No change	Glutamate dehydrogenase	1B26	K193A	K194A	-0.23
1	No change	Glutamate dehydrogenase	1B26	R190A	R191A	0.34
1	No change	Growth hormone	3HHR	E74A	E74A	0.49
1	No change	Growth hormone	3HHR	E74D	E74D	0.55
1	No change	Growth hormone	3HHR	E74L	E74L	0.79
1	No change	Growth hormone	3HHR	E74Q	E74Q	0.18
1	No change	Growth hormone	3HHR	E74S	E74S	0.91
1	No change	Growth hormone	3HHR	E74T	E74T	0.55
1	No change	Growth hormone	3HHR	S71A	S71A	0.97
1	No change	Growth hormone	3HHR	S71Q	S71Q	1.28
1	No change	Growth hormone	3HHR	S71T	S71T	-0.33
1	No change	Growth hormone	3HHR	S71V	S71V	1.19
1	No change	High mobility group protein	1HME	G35H	G123H	-0.47
1	No change	High mobility group protein	1HME	I34H	I122H	0.00
1	No change	High mobility group protein	1HME	S33H	S121H	1.50
1	ID Corrected	HPr	2HPR	G49A	G49A	-0.80
1	ID Corrected	HPr	2HPR	G49E	G49E	-1.40
1	ID Corrected	HPr	1POH	K49A	K49A	-1.60
1	ID Corrected	HPr	1POH	K49D	K49D	-1.80
1	ID Corrected	HPr	1POH	K49E	K49E	-2.10

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	ID Corrected	HPr	1POH	K49G	K49G	-1.20
1	ID Corrected	HPr	1POH	K49M	K49M	-1.00
1	ID Corrected	HPr	1POH	K49N	K49N	-0.40
1	ID Corrected	HPr	1POH	K49Q	K49Q	-1.50
1	ID Corrected	HPr	1POH	K49R	K49R	0.20
1	ID Corrected	HPr	1POH	K49S	K49S	-1.20
0	Duplicated	HPr protein	2HPR	S46D	S46D	-0.70
1	No change	HPr protein	1POH	S46D	S46D	-1.05
1	No change	HU DNA-binding protein	1HUE	A27S	A27S	1.19
1	No change	HU DNA-binding protein	1HUE	A56S	A56S	0.13
0	PDB Related, Mismatch	HU DNA-binding protein	no ID	E15G	E15G	-1.55
0	PDB Related, Mismatch	HU DNA-binding protein	no ID	I42V	I42V	-0.52
1	No change	HU DNA-binding protein	1HUE	M69I	M69I	0.00
0	PDB Related, Mismatch	HU DNA-binding protein	no ID	S27A	S27A	-0.73
1	No change	HU DNA-binding protein	1HUE	S31T	S31T	-0.41
1	No change	HU DNA-binding protein	1HUE	V42I	V42I	0.82
1	No change	Immunoglobulin IgG1	1FC1	D399A	D282A	0.70
1	No change	Immunoglobulin IgG1	1FC1	F405A	F288A	2.50
1	No change	Immunoglobulin IgG1	1FC1	K370A	K253A	1.10
1	No change	Immunoglobulin IgG1	1FC1	K392A	K275A	0.40
1	No change	Immunoglobulin IgG1	1FC1	K409A	K292A	2.40
1	No change	Immunoglobulin IgG1	1FC1	L351A	L234A	1.30

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Immunoglobulin IgG1	1FC1	L368A	L251A	2.20
1	No change	Immunoglobulin IgG1	1FC1	L398A	L281A	0.10
1	No change	Immunoglobulin IgG1	1FC1	P395A	P278A	3.20
1	No change	Immunoglobulin IgG1	1FC1	Q347A	Q230A	1.10
1	No change	Immunoglobulin IgG1	1FC1	T350A	T233A	0.10
1	No change	Immunoglobulin IgG1	1FC1	T366A	T249A	2.30
1	No change	Immunoglobulin IgG1	1FC1	T394A	T277A	0.60
1	No change	Immunoglobulin IgG1	1FC1	V397A	V280A	0.60
1	No change	Immunoglobulin IgG1	1FC1	Y349A	Y232A	0.70
1	No change	Immunoglobulin IgG1	1FC1	Y407A	Y290A	2.10
1	No change	Interleukin 1 beta	1IOB	K97G	K213G	1.20
1	No change	Interleukin 1 beta	1IOB	K97R	K213R	0.50
1	No change	Interleukin 1 beta	1IOB	K97V	K213V	-0.80
1	No change	Interleukin 1 beta	1IOB	T9A	T125A	0.80
1	No change	Interleukin 1 beta	1IOB	T9G	T125G	2.60
1	No change	Interleukin 1 beta	1IOB	T9L	T125L	0.70
1	No change	Interleukin 1 beta	1IOB	T9Q	T125Q	1.90
0	PDB Related	Interleukin 6	no ID	H31A	H57A	0.30
0	PDB Related	Interleukin 6	no ID	W34A	W60A	-3.10
1	No change	Iso-1 cytochrome c	1YCC	C102A	C108A	-2.90
1	No change	Iso-1 cytochrome c	1YCC	C102S	C108S	-2.80
1	No change	Iso-1 cytochrome c	1YCC	C102T	C108T	-0.80

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Iso-1 cytochrome c	1YCC	F82Y	F88Y	0.70
1	No change	Iso-1 cytochrome c	1YCC	K73I	K79I	0.40
1	No change	Iso-1 cytochrome c	1YCC	K73V	K79V	-0.10
1	No change	Iso-1 cytochrome c	1YCC	K73W	K79W	1.60
1	No change	Iso-1 cytochrome c	1YCC	L85A	L91A	2.80
1	No change	Iso-1 cytochrome c	1YCC	N52A	N58A	-0.63
1	No change	Iso-1 cytochrome c	1YCC	N52H	N58H	1.27
1	No change	Iso-1 cytochrome c	1YCC	N52L	N58L	-2.56
1	No change	Iso-1 cytochrome c	1YCC	N52M	N58M	-2.25
1	No change	Iso-1 cytochrome c	1YCC	N52Q	N58Q	0.08
1	No change	Iso-1 cytochrome c	1YCC	N52S	N58S	1.29
1	No change	Iso-1 cytochrome c	1YCC	N52T	N58T	0.53
1	No change	Iso-1 cytochrome c	1YCC	N52V	N58V	-1.67
0	Mismatch	Iso-1 cytochrome c	1YCC	N57I	N58I	-4.20
1	No change	Iso-1 cytochrome c	1YCC	P76G	P82G	0.78
1	No change	Iso-1 cytochrome c	1YCC	P76L	P82L	3.44
1	No change	Iso-1 cytochrome c	1YCC	P76R	P82R	0.97
1	No change	Iso-1 cytochrome c	1YCC	P76S	P82S	1.89
1	No change	Iso-1 cytochrome c	1YCC	P76V	P82V	1.07
1	No change	Iso-1 cytochrome c	1YCC	P76W	P82W	2.38
1	No change	Iso-1 cytochrome c	1YCC	P76Y	P82Y	2.07
1	No change	Lambda cro repressor	5CRO	Y26C	Y26C	-2.20

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lambda cro repressor	5CRO	Y26D	Y26D	-2.70
1	No change	Lambda cro repressor	5CRO	Y26F	Y26F	-0.40
1	No change	Lambda cro repressor	5CRO	Y26H	Y26H	-1.90
1	No change	Lambda cro repressor	5CRO	Y26L	Y26L	-1.10
1	No change	Lambda cro repressor	5CRO	Y26Q	Y26Q	-1.40
1	No change	Lambda cro repressor	5CRO	Y26V	Y26V	-0.90
1	No change	Lambda cro repressor	5CRO	Y26W	Y26W	0.10
0	PDB Related	Lambda repressor	1LRP	G46A	G47A	-0.66
0	PDB Related	Lambda repressor	1LRP	A49V	A50V	1.22
0	PDB Related	Lambda repressor	1LRP	A66T	A67T	2.99
0	PDB Related	Lambda repressor	1LRP	G48A	G49A	-0.87
0	PDB Related	Lambda repressor	1LRP	G48N	G49N	-0.79
0	PDB Related	Lambda repressor	1LRP	G48S	G49S	-0.68
0	PDB Related	Lambda repressor	1LRP	I84S	I85S	2.25
0	PDB Related	Lambda repressor	1LRP	K4Q	K5Q	-0.43
0	PDB Related	Lambda repressor	1LRP	Q33Y	Q34Y	-1.32
0	PDB Related	Lambda repressor	1LRP	Q44Y	Q45Y	0.02
0	PDB Related	Lambda repressor	1LRP	Y22H	Y23H	2.25
0	PDB Related	Lambda repressor	1LRP	Y88C	Y89C	-2.40
1	No change	Lysozyme	2LZM	A129L	A129L	1.30
1	No change	Lysozyme	2LZM	A129M	A129M	1.90
0	Triplicated	Lysozyme	1L63	A129M	A129M	1.90

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

0	Triplicated	Lysozyme	1L63	A129M	A129M	2.80
1	No change	Lysozyme	2LZM	A146T	A146T	1.50
1	No change	Lysozyme	2LZM	A160T	A160T	1.65
1	No change	Lysozyme	4LYZ	A31I	A49I	-1.40
1	No change	Lysozyme	4LYZ	A31L	A49L	-1.80
1	No change	Lysozyme	4LYZ	A31V	A49V	-1.20
1	No change	Lysozyme	2LZM	A41D	A41D	-0.29
1	No change	Lysozyme	2LZM	A41V	A41V	-0.30
1	No change	Lysozyme	2LZM	A42K	A42K	3.70
1	No change	Lysozyme	2LZM	A74P	A74P	4.70
1	No change	Lysozyme	2LZM	A82P	A82P	-0.80
1	No change	Lysozyme	2LZM	A93P	A93P	-0.10
1	No change	Lysozyme	1EL1	A93S	A93S	0.26
1	No change	Lysozyme	2LZM	A93T	A93T	-0.06
1	No change	Lysozyme	2LZM	C54T	C54T	-0.30
1	No change	Lysozyme	2LZM	C54V	C54V	0.70
1	No change	Lysozyme	4LYZ	C94A	C112A	-6.00
1	No change	Lysozyme	4LYZ	D101A	D119A	-0.76
1	No change	Lysozyme	4LYZ	D101E	D119E	0.00
1	No change	Lysozyme	4LYZ	D101F	D119F	-0.72
1	No change	Lysozyme	4LYZ	D101G	D119G	-0.45
1	No change	Lysozyme	4LYZ	D101K	D119K	-0.19

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	4LYZ	D101N	D119N	-0.04
1	No change	Lysozyme	4LYZ	D101Q	D119Q	0.08
1	No change	Lysozyme	4LYZ	D101R	D119R	-0.27
1	No change	Lysozyme	4LYZ	D101S	D119S	-0.87
1	No change	Lysozyme	1LZ1	D102N	D120N	0.67
1	No change	Lysozyme	1LZ1	D120N	D138N	0.67
1	No change	Lysozyme	1LZ1	D18N	D36N	1.70
1	No change	Lysozyme	2LZM	D20A	D20A	0.30
1	No change	Lysozyme	2LZM	D20N	D20N	-1.30
1	No change	Lysozyme	2LZM	D20S	D20S	-0.70
1	No change	Lysozyme	2LZM	D20T	D20T	-0.90
1	No change	Lysozyme	2LZM	D47A	D47A	0.95
1	No change	Lysozyme	1LZ1	D49N	D67N	1.00
1	No change	Lysozyme	1LZ1	D67N	D85N	2.25
1	No change	Lysozyme	2LZM	D72P	D72P	2.70
1	No change	Lysozyme	2LZM	D89A	D89A	0.50
1	No change	Lysozyme	2LZM	D92N	D92N	1.40
1	No change	Lysozyme	2LZM	E108V	E108V	-0.70
1	No change	Lysozyme	2LZM	E11A	E11A	-1.10
1	No change	Lysozyme	2LZM	E11F	E11F	-1.70
1	No change	Lysozyme	2LZM	E11H	E11H	-0.10
1	No change	Lysozyme	2LZM	E11M	E11M	-1.60

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	E11N	E11N	0.10
1	No change	Lysozyme	2LZM	E22K	E22K	-0.57
1	No change	Lysozyme	2LZM	E45A	E45A	-0.01
1	No change	Lysozyme	1LZ1	E7Q	E25Q	1.43
1	No change	Lysozyme	1L63	F104A	F104A	2.70
1	No change	Lysozyme	1L63	F104M	F104M	0.40
1	No change	Lysozyme	2LZM	F153A	F153A	3.80
0	Triplicated	Lysozyme	1L63	F153A	F153A	3.80
0	Triplicated	Lysozyme	1L63	F153A	F153A	3.40
1	No change	Lysozyme	2LZM	F153I	F153I	0.20
1	No change	Lysozyme	2LZM	F153L	F153L	-0.35
0	Duplicated	Lysozyme	1L63	F153M	F153M	0.60
1	No change	Lysozyme	2LZM	F153M	F153M	0.60
1	No change	Lysozyme	2LZM	F153V	F153V	1.80
1	No change	Lysozyme	4LYZ	F34Y	F52Y	-0.19
1	No change	Lysozyme	4LYZ	F3Y	F21Y	0.45
1	No change	Lysozyme	4LYZ	G102R	G120R	-0.38
1	No change	Lysozyme	4LYZ	G102V	G120V	0.04
1	No change	Lysozyme	2LZM	G113A	G113A	-0.30
1	No change	Lysozyme	2LZM	G113E	G113E	-0.30
1	No change	Lysozyme	2LZM	G156D	G156D	2.30
1	No change	Lysozyme	2LZM	G30A	G30A	-0.10

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	G30F	G30F	1.50
1	No change	Lysozyme	4LYZ	G49N	G67N	-2.00
1	No change	Lysozyme	2LZM	G77A	G77A	-0.40
1	No change	Lysozyme	4LYZ	H15L	H33L	-0.76
1	No change	Lysozyme	1EL1	H21G	H21G	0.48
1	No change	Lysozyme	1AM7	H31D	H31D	1.60
1	No change	Lysozyme	2LZM	H31N	H31N	4.00
1	No change	Lysozyme	1AM7	H48N	H48N	5.00
1	No change	Lysozyme	1L63	I100A	I100A	2.50
1	No change	Lysozyme	2LZM	I100M	I100M	1.60
1	No change	Lysozyme	2LZM	I100V	I100V	0.40
1	No change	Lysozyme	1L63	I17A	I17A	2.30
1	No change	Lysozyme	1L63	I17M	I17M	2.20
1	No change	Lysozyme	1L63	I27M	I27M	3.10
1	No change	Lysozyme	2LZM	I3A	I3A	0.70
1	No change	Lysozyme	2LZM	I3C(S-S)	I3C	1.20
1	No change	Lysozyme	2LZM	I3D	I3D	3.20
1	No change	Lysozyme	2LZM	I3E	I3E	2.00
1	No change	Lysozyme	2LZM	I3F	I3F	1.10
1	No change	Lysozyme	2LZM	I3G	I3G	2.10
1	No change	Lysozyme	2LZM	I3L	I3L	-0.40
1	No change	Lysozyme	2LZM	I3M	I3M	0.90

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	I3S	I3S	1.70
1	No change	Lysozyme	2LZM	I3T	I3T	2.30
1	No change	Lysozyme	2LZM	I3V	I3V	0.40
1	No change	Lysozyme	2LZM	I3W	I3W	2.80
1	No change	Lysozyme	2LZM	I3Y	I3Y	2.30
1	No change	Lysozyme	1L63	I50A	I50A	1.60
1	No change	Lysozyme	1L63	I50M	I50M	0.40
1	No change	Lysozyme	4LYZ	I55A	I73A	4.40
1	No change	Lysozyme	4LYZ	I55F	I73F	2.46
1	No change	Lysozyme	4LYZ	I55L	I73L	0.45
1	No change	Lysozyme	4LYZ	I55M	I73M	2.27
1	No change	Lysozyme	4LYZ	I55T	I73T	4.96
1	No change	Lysozyme	4LYZ	I55V	I73V	0.91
1	No change	Lysozyme	1EL1	I56L	I56L	0.24
1	No change	Lysozyme	2LZM	I58T	I58T	3.40
1	No change	Lysozyme	1L63	I78A	I78A	1.20
1	No change	Lysozyme	2LZM	I78M	I78M	1.50
1	No change	Lysozyme	2LZM	I78V	I78V	0.80
1	No change	Lysozyme	2LZM	K124G	K124G	0.10
1	No change	Lysozyme	2LZM	K135E	K135E	1.00
1	No change	Lysozyme	4LYZ	K13D	K31D	-8.00
1	No change	Lysozyme	2LZM	K147E	K147E	0.70

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	K16E	K16E	-0.50
1	No change	Lysozyme	2LZM	K43A	K43A	1.07
1	No change	Lysozyme	2LZM	K48A	K48A	0.56
1	No change	Lysozyme	2LZM	K60H	K60H	0.10
1	No change	Lysozyme	2LZM	K60P	K60P	0.00
1	No change	Lysozyme	2LZM	K83H	K83H	0.40
1	No change	Lysozyme	2LZM	K85A	K85A	0.60
1	No change	Lysozyme	1L63	L118A	L118A	3.20
1	No change	Lysozyme	2LZM	L118I	L118I	1.20
0	Duplicated	Lysozyme	1L63	L118M	L118M	0.70
1	No change	Lysozyme	2LZM	L118M	L118M	0.70
0	Duplicated	Lysozyme	1L63	L121A	L121A	2.20
1	No change	Lysozyme	2LZM	L121A	L121A	2.25
0	Duplicated	Lysozyme	1L63	L121M	L121M	0.80
1	No change	Lysozyme	2LZM	L121M	L121M	0.80
1	No change	Lysozyme	2LZM	L133D	L133D	5.70
1	No change	Lysozyme	2LZM	L133F	L133F	0.30
1	No change	Lysozyme	2LZM	L133M	L133M	0.40
1	No change	Lysozyme	1L63	L33A	L33A	2.90
1	No change	Lysozyme	1L63	L33M	L33M	2.00
1	No change	Lysozyme	2LZM	L39A	L39A	0.90
1	No change	Lysozyme	2LZM	L46A	L46A	1.86

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	1L63	L66A	L66A	3.30
1	No change	Lysozyme	1L63	L66M	L66M	1.00
1	No change	Lysozyme	1L63	L7A	L7A	2.30
1	No change	Lysozyme	1L63	L84A	L84A	3.70
0	Duplicated	Lysozyme	1L63	L84M	L84M	1.90
1	No change	Lysozyme	2LZM	L84M	L84M	1.90
1	No change	Lysozyme	1L63	L91A	L91A	2.60
0	Duplicated	Lysozyme	1L63	L91M	L91M	0.80
1	No change	Lysozyme	2LZM	L91M	L91M	0.80
1	No change	Lysozyme	2LZM	L99A	L99A	4.50
1	No change	Lysozyme	2LZM	L99F	L99F	0.30
0	Duplicated	Lysozyme	1L63	L99F	L99F	0.60
1	No change	Lysozyme	2LZM	L99G	L99G	6.30
1	No change	Lysozyme	2LZM	L99I	L99I	1.50
0	Duplicated	Lysozyme	1L63	L99M	L99M	0.60
1	No change	Lysozyme	2LZM	L99M	L99M	0.60
1	No change	Lysozyme	2LZM	L99V	L99V	2.00
1	No change	Lysozyme	1L63	M102A	M102A	2.90
1	No change	Lysozyme	2LZM	M102K	M102K	6.90
1	No change	Lysozyme	2LZM	M102L	M102L	0.70
1	No change	Lysozyme	4LYZ	M105T	M123T	1.00
1	No change	Lysozyme	1L63	M106A	M106A	1.90

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	M106I	M106I	-0.20
1	No change	Lysozyme	2LZM	M120Y	M120Y	0.10
1	No change	Lysozyme	4LYZ	M12L	M30L	-0.24
1	No change	Lysozyme	1L63	M6A	M6A	1.60
1	No change	Lysozyme	2LZM	M6I	M6I	1.38
1	No change	Lysozyme	2LZM	N101A	N101A	1.50
1	No change	Lysozyme	4LYZ	N103D	N121D	1.00
1	No change	Lysozyme	2LZM	N116D	N116D	-0.60
1	No change	Lysozyme	2LZM	N132F	N132F	-1.30
1	No change	Lysozyme	2LZM	N132I	N132I	-1.20
1	No change	Lysozyme	2LZM	N132M	N132M	-1.50
1	No change	Lysozyme	2LZM	N144D	N144D	-0.50
1	No change	Lysozyme	2LZM	N144E	N144E	-0.40
1	No change	Lysozyme	2LZM	N144H	N144H	-0.30
1	No change	Lysozyme	2LZM	N163D	N163D	0.21
1	No change	Lysozyme	4LYZ	N19K	N37K	1.06
1	No change	Lysozyme	2LZM	N40A	N40A	-0.32
1	No change	Lysozyme	2LZM	N40D	N40D	-0.44
1	No change	Lysozyme	2LZM	N55G	N55G	0.60
1	No change	Lysozyme	4LYZ	N77H	N95H	0.38
1	No change	Lysozyme	4LYZ	P70N	P88N	-3.50
1	No change	Lysozyme	2LZM	Q105A	Q105A	0.60

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	Q105E	Q105E	1.10
1	No change	Lysozyme	2LZM	Q105G	Q105G	1.50
1	No change	Lysozyme	2LZM	Q105M	Q105M	1.20
1	No change	Lysozyme	4LYZ	Q121H	Q139H	0.45
1	No change	Lysozyme	2LZM	Q123E	Q123E	-0.40
1	No change	Lysozyme	2LZM	Q69P	Q69P	2.90
1	No change	Lysozyme	4LYZ	R114H	R132H	-0.68
1	No change	Lysozyme	2LZM	R119E	R119E	0.04
1	No change	Lysozyme	2LZM	R119H	R119H	0.29
1	No change	Lysozyme	2LZM	R119M	R119M	-0.10
1	No change	Lysozyme	2LZM	R14K	R14K	0.03
1	No change	Lysozyme	2LZM	R154E	R154E	1.10
1	No change	Lysozyme	4LYZ	R21Q	R39Q	0.15
1	No change	Lysozyme	4LYZ	R68K	R86K	0.04
1	No change	Lysozyme	4LYZ	R73K	R91K	-0.23
1	No change	Lysozyme	2LZM	R80K	R80K	0.17
1	No change	Lysozyme	2LZM	R96A	R96A	2.00
1	No change	Lysozyme	2LZM	R96C	R96C	2.90
1	No change	Lysozyme	2LZM	R96D	R96D	3.50
1	No change	Lysozyme	2LZM	R96E	R96E	2.50
1	No change	Lysozyme	2LZM	R96F	R96F	4.20
1	No change	Lysozyme	2LZM	R96G	R96G	2.60

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	R96H	R96H	2.80
1	No change	Lysozyme	2LZM	R96I	R96I	2.90
1	No change	Lysozyme	2LZM	R96K	R96K	0.00
1	No change	Lysozyme	2LZM	R96L	R96L	3.20
1	No change	Lysozyme	2LZM	R96M	R96M	2.70
1	No change	Lysozyme	2LZM	R96N	R96N	3.00
1	No change	Lysozyme	2LZM	R96P	R96P	5.50
1	No change	Lysozyme	2LZM	R96Q	R96Q	0.30
1	No change	Lysozyme	2LZM	R96S	R96S	2.60
1	No change	Lysozyme	2LZM	R96T	R96T	2.80
1	No change	Lysozyme	2LZM	R96V	R96V	2.40
1	No change	Lysozyme	2LZM	R96W	R96W	4.50
1	No change	Lysozyme	2LZM	R96Y	R96Y	4.70
0	Duplicated	Lysozyme	2LZM	S117F	S117F	-1.10
1	No change	Lysozyme	2LZM	S117F	S117F	-1.10
1	No change	Lysozyme	2LZM	S117I	S117I	-1.70
1	No change	Lysozyme	2LZM	S117V	S117V	-2.00
1	No change	Lysozyme	2LZM	S38D	S38D	-0.60
1	No change	Lysozyme	2LZM	S38N	S38N	0.00
1	No change	Lysozyme	2LZM	S44A	S44A	-0.34
1	No change	Lysozyme	2LZM	S90H	S90H	1.10
1	No change	Lysozyme	4LYZ	S91A	S109A	0.15

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	4LYZ	S91D	S109D	2.31
1	No change	Lysozyme	4LYZ	S91T	S109T	-0.99
1	No change	Lysozyme	4LYZ	S91V	S109V	0.08
1	No change	Lysozyme	4LYZ	S91Y	S109Y	3.07
1	No change	Lysozyme	2LZM	T109D	T109D	-0.60
1	No change	Lysozyme	2LZM	T109N	T109N	-0.10
1	No change	Lysozyme	2LZM	T115E	T115E	-0.04
1	No change	Lysozyme	2LZM	T151S	T151S	-0.39
1	No change	Lysozyme	2LZM	T152A	T152A	1.50
1	No change	Lysozyme	2LZM	T152C	T152C	0.50
1	No change	Lysozyme	2LZM	T152I	T152I	0.40
1	No change	Lysozyme	2LZM	T152S	T152S	2.00
0	Duplicated	Lysozyme	1L63	T152V	T152V	-0.20
1	No change	Lysozyme	2LZM	T152V	T152V	-0.20
1	No change	Lysozyme	2LZM	T157I	T157I	1.20
1	No change	Lysozyme	2LZM	T26S	T26S	-0.57
1	No change	Lysozyme	4LYZ	T40I	T58I	2.20
1	No change	Lysozyme	4LYZ	T40S	T58S	0.27
1	No change	Lysozyme	2LZM	T59A	T59A	1.50
1	No change	Lysozyme	2LZM	T59D	T59D	1.20
1	No change	Lysozyme	2LZM	T59G	T59G	1.60
1	No change	Lysozyme	2LZM	T59N	T59N	1.10

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Lysozyme	2LZM	T59S	T59S	0.20
1	No change	Lysozyme	2LZM	T59V	T59V	1.50
1	No change	Lysozyme	1L63	V103A	V103A	1.60
1	No change	Lysozyme	2LZM	V103I	V103I	0.50
0	Duplicated	Lysozyme	1L63	V103M	V103M	1.20
1	No change	Lysozyme	2LZM	V103M	V103M	1.20
1	No change	Lysozyme	1EL1	V109K	V109K	0.43
1	No change	Lysozyme	2LZM	V111A	V111A	1.00
1	No change	Lysozyme	2LZM	V111F	V111F	1.40
1	No change	Lysozyme	2LZM	V111I	V111I	0.70
1	No change	Lysozyme	1L63	V111M	V111M	0.70
1	No change	Lysozyme	2LZM	V131A	V131A	-0.39
1	No change	Lysozyme	2LZM	V149A	V149A	3.15
1	No change	Lysozyme	2LZM	V149C	V149C	2.00
1	No change	Lysozyme	2LZM	V149G	V149G	4.90
1	No change	Lysozyme	2LZM	V149I	V149I	0.00
1	No change	Lysozyme	1L63	V149M	V149M	2.80
1	No change	Lysozyme	2LZM	V149S	V149S	4.40
1	No change	Lysozyme	2LZM	V149T	V149T	3.00
1	No change	Lysozyme	1L63	V87A	V87A	1.50
1	No change	Lysozyme	2LZM	V87I	V87I	0.30
1	No change	Lysozyme	1L63	V87M	V87M	2.30

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Maltose binding protein	3MBP	A276G	A302G	1.50
1	No change	Maltose binding protein	3MBP	D55N	D81N	0.00
1	No change	Maltose binding protein	3MBP	G19C	G45C	2.30
1	No change	Maltose binding protein	3MBP	P133A	P159A	0.30
1	No change	Maltose binding protein	3MBP	P133S	P159S	2.10
1	No change	Maltose binding protein	3MBP	P159A	P185A	1.80
1	No change	Maltose binding protein	3MBP	P159S	P185S	2.10
1	No change	Maltose binding protein	3MBP	P48A	P74A	-0.50
1	No change	Maltose binding protein	3MBP	P48S	P74S	-0.30
1	No change	Maltose binding protein	3MBP	T345I	T371I	-0.70
1	No change	Maltose binding protein	3MBP	V8G	V34G	1.10
1	No change	Maltose binding protein	3MBP	Y283D	Y309D	3.20
1	No change	Myb	1MBG	V103A	V103A	1.87
1	No change	Myb	1MBG	V103I	V103I	-0.72
1	No change	Myb	1MBG	V103L	V103L	-2.49
1	No change	Myoglobin	1BVC	A130K	A131K	2.10
1	No change	Myoglobin	1BVC	A130L	A131L	0.90
1	ID Corrected	Myoglobin	1BVC	D122A	D123A	0.10
1	ID Corrected	Myoglobin	1BVC	D20A	D21A	0.50
1	No change	Myoglobin	1BVC	D44A	D45A	-0.25
1	No change	Myoglobin	1BVC	D60A	D61A	0.15
1	No change	Myoglobin	1BVC	E109A	E110A	-0.17

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Myoglobin	1BVC	E109G	E110G	0.89
1	ID Corrected	Myoglobin	1BVC	E18A	E19A	0.95
1	No change	Myoglobin	1BVC	E4A	E5A	0.35
1	No change	Myoglobin	1BVC	F123K	F124K	2.10
1	No change	Myoglobin	1BVC	G129A	G130A	-1.10
1	No change	Myoglobin	1BVC	G23A	G24A	1.12
1	No change	Myoglobin	1BVC	H116A	H117A	-0.16
1	No change	Myoglobin	1BVC	H36Q	H37Q	0.80
1	No change	Myoglobin	1BVC	I142A	I143A	1.90
1	ID Corrected	Myoglobin	1BVC	K133A	K134A	0.05
1	No change	Myoglobin	1BVC	K140A	K141A	-0.35
1	No change	Myoglobin	1BVC	K56A	K57A	0.35
1	No change	Myoglobin	1BVC	K77A	K78A	-0.20
1	No change	Myoglobin	1BVC	L11A	L12A	0.44
1	No change	Myoglobin	1BVC	L137A	L138A	1.78
1	No change	Myoglobin	1BVC	L149A	L150A	1.60
1	No change	Myoglobin	1BVC	L9A	L10A	0.41
1	No change	Myoglobin	1BVC	M131A	M132A	2.20
1	No change	Myoglobin	1BVC	P88A	P89A	-0.59
1	No change	Myoglobin	1BVC	Q8A	Q9A	-0.88
1	No change	Myoglobin	1BVC	Q8G	Q9G	0.50
1	ID Corrected	Myoglobin	1BVC	R118A	R119A	0.65

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	ID Corrected	Myoglobin	1BVC	R139A	R140A	0.45
1	No change	Myoglobin	1BVC	S117A	S118A	0.26
1	No change	Myoglobin	1BVC	T51A	T52A	1.41
1	No change	Myoglobin	1BVC	T67A	T68A	0.26
1	No change	Myoglobin	1BVC	V114A	V115A	1.45
1	No change	Myoglobin	1BVC	V13A	V14A	0.67
1	No change	Myoglobin	1BVC	V66A	V67A	-0.75
1	No change	Myoglobin	1BVC	V68T	V69T	0.40
1	No change	Myoglobin	1BVC	W14F	W15F	1.10
1	No change	Myoglobin	1BVC	W7F	W8F	0.90
1	No change	Onconase	1ONC	M23L	M23L	-2.32
0	PDB Related, Mismatch	Parvalbumin	no ID	C72S	C73S	-0.72
1	No change	Parvalbumin	1RRO	P21A	P21A	0.74
1	No change	Parvalbumin	1RRO	P26A	P26A	0.74
1	No change	Peripheral subunit-binding domain	1W4H	A130G	A113G	0.04
1	No change	Peripheral subunit-binding domain	1W4H	A134G	A117G	0.37
1	No change	Peripheral subunit-binding domain	1W4H	A140G	A123G	1.02
1	No change	Peripheral subunit-binding domain	1W4H	A146G	A129G	0.58
1	No change	Peripheral subunit-binding domain	1W4H	A148G	A131G	0.43
1	No change	Peripheral subunit-binding domain	1W4H	A168G	A151G	0.07
1	No change	Peripheral subunit-binding domain	1W4H	D162N	D145N	3.10
1	No change	Peripheral subunit-binding domain	1W4H	H166A	H149A	-0.06

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Peripheral subunit-binding domain	1W4H	H166G	H149G	1.58
1	No change	Peripheral subunit-binding domain	1W4H	I135A	I118A	1.63
1	No change	Peripheral subunit-binding domain	1W4H	I135V	I118V	0.03
1	No change	Peripheral subunit-binding domain	1W4H	I149V	I132V	0.75
1	No change	Peripheral subunit-binding domain	1W4H	L131A	L114A	0.94
1	No change	Peripheral subunit-binding domain	1W4H	L138A	L121A	1.67
1	No change	Peripheral subunit-binding domain	1W4H	L139A	L122A	3.50
1	No change	Peripheral subunit-binding domain	1W4H	L144A	L127A	1.90
1	No change	Peripheral subunit-binding domain	1W4H	L158A	L141A	2.12
1	ID Corrected	Peripheral subunit-binding domain	1W4H	L167A	L150A	0.57
1	No change	Peripheral subunit-binding domain	1W4H	P133A	P116A	-0.38
1	No change	Peripheral subunit-binding domain	1W4H	S132G	S115G	0.47
1	No change	Peripheral subunit-binding domain	1W4H	T152A	T135A	1.36
1	No change	Peripheral subunit-binding domain	1W4H	T152S	T135S	0.29
1	No change	Peripheral subunit-binding domain	1W4H	T159S	T142S	1.63
1	No change	Peripheral subunit-binding domain	1W4H	V154G	V137G	0.84
1	No change	Peripheral subunit-binding domain	1W4H	V163A	V146A	2.02
1	No change	Pin1 WW domain	1PIN	A31G	A31G	1.49
1	No change	Pin1 WW domain	1PIN	E12A	E12A	0.48
1	No change	Pin1 WW domain	1PIN	E35A	E35A	0.73
1	No change	Pin1 WW domain	1PIN	F25A	F25A	2.14
1	No change	Pin1 WW domain	1PIN	F25L	F25L	1.37

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Pin1 WW domain	1PIN	F25Y	F25Y	-0.22
1	No change	Pin1 WW domain	1PIN	G10A	G10A	0.85
1	No change	Pin1 WW domain	1PIN	G20A	G20A	0.86
1	No change	Pin1 WW domain	1PIN	H27A	H27A	0.12
1	No change	Pin1 WW domain	1PIN	H27G	H27G	0.68
1	No change	Pin1 WW domain	1PIN	I28A	I28A	0.36
1	No change	Pin1 WW domain	1PIN	I28G	I28G	0.48
1	No change	Pin1 WW domain	1PIN	K13A	K13A	0.02
1	No change	Pin1 WW domain	1PIN	K6A	K6A	-0.15
1	No change	Pin1 WW domain	1PIN	L7A	L7A	1.72
1	No change	Pin1 WW domain	1PIN	L7I	L7I	0.85
1	No change	Pin1 WW domain	1PIN	L7N	L7N	0.90
1	No change	Pin1 WW domain	1PIN	L7V	L7V	1.26
1	No change	Pin1 WW domain	1PIN	M15A	M15A	0.59
1	No change	Pin1 WW domain	1PIN	N26D	N26D	1.88
1	No change	Pin1 WW domain	1PIN	N30A	N30A	0.47
1	No change	Pin1 WW domain	1PIN	N30G	N30G	-0.53
1	No change	Pin1 WW domain	1PIN	P8A	P8A	0.97
1	No change	Pin1 WW domain	1PIN	P8G	P8G	0.92
1	No change	Pin1 WW domain	1PIN	P9A	P9A	0.19
1	No change	Pin1 WW domain	1PIN	P9G	P9G	0.49
1	No change	Pin1 WW domain	1PIN	Q33A	Q33A	0.57

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Pin1 WW domain	1PIN	R14A	R14A	1.63
1	No change	Pin1 WW domain	1PIN	R17A	R17A	0.00
1	No change	Pin1 WW domain	1PIN	R17G	R17G	0.22
1	No change	Pin1 WW domain	1PIN	R21A	R21A	0.65
1	No change	Pin1 WW domain	1PIN	R21G	R21G	0.67
1	No change	Pin1 WW domain	1PIN	R36A	R36A	0.29
1	No change	Pin1 WW domain	1PIN	S16A	S16A	0.44
1	No change	Pin1 WW domain	1PIN	S16G	S16G	0.92
1	No change	Pin1 WW domain	1PIN	S18A	S18A	-0.03
1	No change	Pin1 WW domain	1PIN	S18G	S18G	-0.02
1	No change	Pin1 WW domain	1PIN	S19A	S19A	0.14
1	No change	Pin1 WW domain	1PIN	S19G	S19G	-0.01
1	No change	Pin1 WW domain	1PIN	S32A	S32A	0.26
1	No change	Pin1 WW domain	1PIN	S32G	S32G	0.81
1	No change	Pin1 WW domain	1PIN	S38A	S38A	-0.05
1	No change	Pin1 WW domain	1PIN	S38G	S38G	0.08
1	No change	Pin1 WW domain	1PIN	T29A	T29A	1.25
1	No change	Pin1 WW domain	1PIN	T29D	T29D	1.33
1	No change	Pin1 WW domain	1PIN	T29G	T29G	1.99
1	No change	Pin1 WW domain	1PIN	T29S	T29S	0.65
1	No change	Pin1 WW domain	1PIN	V22A	V22A	0.26
1	No change	Pin1 WW domain	1PIN	W11F	W11F	1.93

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Pin1 WW domain	1PIN	W34A	W34A	0.40
1	No change	Pin1 WW domain	1PIN	W34F	W34F	0.10
1	No change	Pin1 WW domain	1PIN	Y23A	Y23A	1.08
1	No change	Pin1 WW domain	1PIN	Y23F	Y23F	0.52
1	No change	Pin1 WW domain	1PIN	Y23L	Y23L	1.18
1	No change	Pin1 WW domain	1PIN	Y24F	Y24F	0.66
1	No change	Pin1 WW domain	1PIN	Y24W	Y24W	0.56
1	No change	Plasminogen activator inhibitor 1	1C5G	E350P	E350P	-0.30
1	No change	Plasminogen activator inhibitor 1	1C5G	E350R	E350R	0.30
1	No change	Plasminogen activator inhibitor 1	1C5G	R30E	R53E	1.20
1	No change	Plasminogen activator kringle-2 domain	1TPK	V65A	V280A	2.00
1	No change	Plasminogen activator kringle-2 domain	1TPK	V65I	V280I	-0.30
1	No change	Plasminogen activator kringle-2 domain	1TPK	V65L	V280L	1.90
1	No change	Plasminogen activator kringle-2 domain	1TPK	V65M	V280M	2.20
1	No change	Plasminogen activator kringle-2 domain	1TPK	V65S	V280S	2.40
1	No change	Plasminogen activator kringle-2 domain	1TPK	V65T	V280T	1.00
1	No change	Protein G	1PGA	A23P	A23P	0.30
1	No change	Protein G	1PGA	A24P	A24P	0.50
1	No change	Protein G	1PGA	A48P	A48P	0.70
1	No change	Protein G	1PGA	D36P	D36P	3.10

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Protein G	1PGA	D46A	D272A	1.36
1	No change	Protein G	1PGA	D47A	D273A	-0.36
1	No change	Protein G	1PGA	G9P	G9P	2.40
1	ID Corrected	Protein G	no ID	I6N	I232N	1.93
1	ID Corrected	Protein G	no ID	I6T	I232T	1.99
1	No change	Protein G	1PGA	K10P	K10P	0.20
1	No change	Protein G	1PGA	K50A	K276A	0.45
1	No change	Protein G	1PGA	T25P	T25P	2.80
1	No change	Protein G	1PGA	T2P	T2P	2.70
1	No change	Protein G	1PGA	T49A	T275A	0.86
1	No change	Protein G	1PGA	T53C	T279C	-0.78
1	No change	Protein G	1PGA	T53D	T279D	0.85
1	No change	Protein G	1PGA	T53E	T279E	-0.23
1	No change	Protein G	1PGA	T53F	T279F	0.28
1	No change	Protein G	1PGA	T53G	T279G	1.21
1	No change	Protein G	1PGA	T53H	T279H	-0.37
1	No change	Protein G	1PGA	T53I	T279I	0.09
1	No change	Protein G	1PGA	T53K	T279K	-0.35
1	No change	Protein G	1PGA	T53L	T279L	-0.45
1	No change	Protein G	1PGA	T53M	T279M	-0.90
1	No change	Protein G	1PGA	T53N	T279N	-0.52
1	No change	Protein G	1PGA	T53Q	T279Q	-0.38

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Protein G	1PGA	T53R	T279R	-0.40
1	No change	Protein G	1PGA	T53S	T279S	-0.87
0	Mismatch	Protein G	1PGA	T53T	T279A	-1.36
1	No change	Protein G	1PGA	T53V	T279V	0.42
1	No change	Protein G	1PGA	T53W	T279W	-1.04
1	No change	Protein G	1PGA	T53Y	T279Y	-0.27
1	No change	Protein G	1PGA	V21P	V21P	-0.50
1	No change	Protein G	1PGA	V29P	V29P	3.50
1	No change	Protein G	1PGA	V54A	V280A	0.10
1	No change	Ribonuclease A	1RTB	A109G	A135G	0.43
1	No change	Ribonuclease A	1RTB	A4S	A30S	0.51
1	No change	Ribonuclease A	1RTB	A5S	A31S	0.37
1	No change	Ribonuclease A	1RTB	A64G	A90G	0.43
1	No change	Ribonuclease A	1RTB	D121A	D147A	1.50
1	No change	Ribonuclease A	1RTB	D121N	D147N	1.10
1	No change	Ribonuclease A	1RTB	F46A	F72A	6.36
1	No change	Ribonuclease A	1RTB	F46L	F72L	3.20
1	No change	Ribonuclease A	1RTB	F46V	F72V	4.54
1	No change	Ribonuclease A	1RTB	H119A	H145A	0.90
1	No change	Ribonuclease A	1RTB	I106A	I132A	4.37
1	No change	Ribonuclease A	1RTB	I106L	I132L	1.79
1	No change	Ribonuclease A	1RTB	I106V	I132V	0.80

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease A	1RTB	I107A	I133A	2.84
1	No change	Ribonuclease A	1RTB	I107L	I133L	2.19
1	No change	Ribonuclease A	1RTB	I107V	I133V	0.08
1	No change	Ribonuclease A	1RTB	I81A	I107A	2.99
1	No change	Ribonuclease A	1RTB	I81G	I107G	4.80
1	No change	Ribonuclease A	1RTB	I81V	I107V	0.43
1	No change	Ribonuclease A	1RTB	P114A	P140A	3.20
1	No change	Ribonuclease A	1RTB	P114G	P140G	3.30
1	No change	Ribonuclease A	1RTB	P93A	P119A	2.70
1	No change	Ribonuclease A	1RTB	P93G	P119G	2.20
1	No change	Ribonuclease A	1RTB	P93S	P119S	2.10
1	No change	Ribonuclease A	1RTB	S123A	S149A	-0.46
1	No change	Ribonuclease A	1RTB	S75A	S101A	2.50
1	No change	Ribonuclease A	1RTB	S75T	S101T	2.80
1	No change	Ribonuclease A	1RTB	V108A	V134A	4.20
1	No change	Ribonuclease A	1RTB	V108G	V134G	7.28
1	No change	Ribonuclease A	1RTB	V108I	V134I	0.43
1	No change	Ribonuclease A	1RTB	V108L	V134L	0.70
1	No change	Ribonuclease A	1RTB	V116A	V142A	0.66
1	No change	Ribonuclease A	1RTB	V116G	V142G	1.18
1	No change	Ribonuclease A	1RTB	V118A	V144A	1.92
1	No change	Ribonuclease A	1RTB	V118G	V144G	2.77

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease A	1RTB	V47A	V73A	3.80
1	No change	Ribonuclease A	1RTB	V47G	V73G	7.34
1	No change	Ribonuclease A	1RTB	V54A	V80A	2.58
1	No change	Ribonuclease A	1RTB	V54G	V80G	4.87
1	No change	Ribonuclease A	1RTB	V54I	V80I	1.95
1	No change	Ribonuclease A	1RTB	V54L	V80L	1.09
1	No change	Ribonuclease A	1RTB	V57A	V83A	2.85
1	No change	Ribonuclease A	1RTB	V57G	V83G	5.52
1	No change	Ribonuclease A	1RTB	V57I	V83I	1.30
1	No change	Ribonuclease A	1RTB	V57L	V83L	2.37
1	No change	Ribonuclease A	1RTB	V63A	V89A	2.02
1	No change	Ribonuclease A	1RTB	V63G	V89G	3.50
1	No change	Ribonuclease A	1RTB	Y97A	Y123A	12.00
1	No change	Ribonuclease A	1RTB	Y97F	Y123F	3.54
1	No change	Ribonuclease A	1RTB	Y97G	Y123G	11.70
1	No change	Ribonuclease HI	2RN2	A125T	A125T	0.00
1	No change	Ribonuclease HI	2RN2	A24V	A24V	-0.88
1	No change	Ribonuclease HI	2RN2	D10A	D10A	-4.05
1	No change	Ribonuclease HI	2RN2	D10E	D10E	-1.00
1	No change	Ribonuclease HI	2RN2	D10H	D10H	-2.40
1	No change	Ribonuclease HI	2RN2	D10N	D10N	-2.00
1	No change	Ribonuclease HI	2RN2	D10S	D10S	-2.70

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease HI	2RN2	D134A	D134A	-1.51
1	No change	Ribonuclease HI	2RN2	D134E	D134E	-0.86
1	No change	Ribonuclease HI	2RN2	D134H	D134H	-1.92
1	No change	Ribonuclease HI	2RN2	D134I	D134I	-1.27
1	No change	Ribonuclease HI	2RN2	D134L	D134L	-1.51
1	No change	Ribonuclease HI	2RN2	D134N	D134N	-0.90
1	No change	Ribonuclease HI	2RN2	D134Q	D134Q	-1.32
1	No change	Ribonuclease HI	2RN2	D134S	D134S	-1.08
1	No change	Ribonuclease HI	2RN2	D134T	D134T	-1.08
1	No change	Ribonuclease HI	2RN2	D134V	D134V	-1.12
1	No change	Ribonuclease HI	2RN2	D70A	D70A	-1.10
1	No change	Ribonuclease HI	2RN2	D70E	D70E	-0.10
1	No change	Ribonuclease HI	2RN2	D70N	D70N	-1.60
1	No change	Ribonuclease HI	2RN2	D94E	D94E	0.40
1	No change	Ribonuclease HI	2RN2	E119V	E119V	-0.74
1	No change	Ribonuclease HI	2RN2	E135K	E135K	0.22
1	No change	Ribonuclease HI	2RN2	E48A	E48A	0.30
1	No change	Ribonuclease HI	2RN2	E48D	E48D	0.20
1	No change	Ribonuclease HI	2RN2	E48Q	E48Q	-0.30
1	No change	Ribonuclease HI	2RN2	G23A	G23A	-0.50
1	No change	Ribonuclease HI	2RN2	H62D	H62D	0.17
1	No change	Ribonuclease HI	2RN2	H62P	H62P	-1.10

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease HI	2RN2	H62R	H62R	-0.36
1	No change	Ribonuclease HI	2RN2	K117R	K117R	-0.03
1	No change	Ribonuclease HI	2RN2	K91R	K91R	0.00
1	No change	Ribonuclease HI	2RN2	K95A	K95A	-0.10
1	No change	Ribonuclease HI	2RN2	K95G	K95G	-1.90
1	No change	Ribonuclease HI	2RN2	K95N	K95N	-0.89
1	No change	Ribonuclease HI	2RN2	Q113P	Q113P	0.60
1	No change	Ribonuclease HI	2RN2	R41C	R41C	-0.44
1	No change	Ribonuclease HI	2RN2	V74A	V74A	3.40
1	No change	Ribonuclease HI	2RN2	V74I	V74I	-0.60
1	No change	Ribonuclease HI	2RN2	V74L	V74L	-0.90
1	No change	Ribonuclease Sa	1RGG	D17K	D17K	1.10
1	No change	Ribonuclease Sa	1RGG	D25H	D25H	-0.90
1	No change	Ribonuclease Sa	1RGG	D25K	D25K	-0.90
1	No change	Ribonuclease Sa	1RGG	D33A	D33A	6.20
1	No change	Ribonuclease Sa	1RGG	D79A	D79A	-2.90
1	No change	Ribonuclease Sa	1RGG	D79E	D79E	0.30
1	No change	Ribonuclease Sa	1RGG	D79F	D79F	-3.00
1	No change	Ribonuclease Sa	1RGG	D79H	D79H	-1.80
1	No change	Ribonuclease Sa	1RGG	D79I	D79I	-2.70
1	No change	Ribonuclease Sa	1RGG	D79K	D79K	-2.30
1	No change	Ribonuclease Sa	1RGG	D79L	D79L	-2.70

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease Sa	1RGG	D79N	D79N	-1.80
1	No change	Ribonuclease Sa	1RGG	D79R	D79R	-2.70
1	No change	Ribonuclease Sa	1RGG	D79W	D79W	-2.30
1	No change	Ribonuclease Sa	1RGG	D79Y	D79Y	-2.80
1	No change	Ribonuclease Sa	1RGG	E41K	E41K	0.70
1	No change	Ribonuclease Sa	1RGG	E54Q	E54Q	1.70
1	No change	Ribonuclease Sa	1RGG	E74K	E74K	-0.90
1	No change	Ribonuclease Sa	1RGG	H85Q	H85Q	0.00
1	No change	Ribonuclease Sa	1RGG	N39A	N39A	2.20
1	No change	Ribonuclease Sa	1RGG	N39D	N39D	1.50
1	No change	Ribonuclease Sa	1RGG	N39S	N39S	2.30
1	No change	Ribonuclease Sa	1RGG	Q38A	Q38A	-1.00
1	No change	Ribonuclease Sa	1RGG	Q94K	Q94K	-0.30
1	No change	Ribonuclease Sa	1RGG	R65A	R65A	1.00
1	No change	Ribonuclease Sa	1RGG	T16V	T16V	-0.30
1	No change	Ribonuclease Sa	1RGG	T18V	T18V	1.40
1	No change	Ribonuclease Sa	1RGG	T56V	T56V	1.90
1	No change	Ribonuclease Sa	1RGG	T59V	T59V	1.70
1	No change	Ribonuclease Sa	1RGG	T5V	T5V	0.00
1	No change	Ribonuclease Sa	1RGG	T67V	T67V	0.00
1	No change	Ribonuclease Sa	1RGG	T72V	T72V	0.20
1	No change	Ribonuclease Sa	1RGG	T82V	T82V	1.70

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease Sa	1RGG	V2T	V2T	0.90
1	No change	Ribonuclease Sa	1RGG	V36T	V36T	1.30
1	No change	Ribonuclease Sa	1RGG	V43T	V43T	0.50
1	No change	Ribonuclease Sa	1RGG	V57T	V57T	4.40
1	No change	Ribonuclease Sa	1RGG	Y30F	Y30F	-0.40
1	No change	Ribonuclease Sa	1RGG	Y49F	Y49F	0.20
1	No change	Ribonuclease Sa	1RGG	Y51F	Y51F	2.30
1	No change	Ribonuclease Sa	1RGG	Y52F	Y52F	3.60
1	No change	Ribonuclease Sa	1RGG	Y55F	Y55F	0.60
1	No change	Ribonuclease Sa	1RGG	Y80F	Y80F	1.50
1	No change	Ribonuclease Sa	1RGG	Y81F	Y81F	1.20
1	No change	Ribonuclease Sa	1RGG	Y86F	Y86F	0.30
1	No change	Ribonuclease Sa3	1MGR	Y11F	Y53F	0.60
1	No change	Ribonuclease Sa3	1MGR	Y33F	Y75F	-0.50
1	No change	Ribonuclease Sa3	1MGR	Y54F	Y96F	2.60
1	No change	Ribonuclease Sa3	1MGR	Y55F	Y97F	2.10
1	No change	Ribonuclease Sa3	1MGR	Y58F	Y100F	0.70
1	No change	Ribonuclease Sa3	1MGR	Y83F	Y125F	1.50
1	No change	Ribonuclease Sa3	1MGR	Y84F	Y126F	1.00
1	No change	Ribonuclease Sa3	1MGR	Y89F	Y131F	0.00
1	No change	Ribonuclease T1	1RN1	A21D	A47D	0.71
1	No change	Ribonuclease T1	1RN1	A21E	A47E	0.69

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease T1	1RN1	A21H	A47H	0.17
1	No change	Ribonuclease T1	1RN1	A21Q	A47Q	0.40
1	No change	Ribonuclease T1	1RN1	A21S	A47S	0.40
1	No change	Ribonuclease T1	1RN1	D49H	D75H	-1.10
1	No change	Ribonuclease T1	1RN1	D76A	D102A	3.75
1	No change	Ribonuclease T1	1RN1	D76N	D102N	3.10
1	No change	Ribonuclease T1	1RN1	D76S	D102S	3.20
1	No change	Ribonuclease T1	1RN1	E58A	E84A	0.78
1	No change	Ribonuclease T1	1RN1	H40T	H66T	0.26
1	No change	Ribonuclease T1	1RN1	H92A	H118A	0.62
1	No change	Ribonuclease T1	1RN1	N36A	N62A	0.00
1	No change	Ribonuclease T1	1RN1	N44A	N70A	1.60
1	No change	Ribonuclease T1	1RN1	N44D	N70D	1.60
1	No change	Ribonuclease T1	1RN1	N44S	N70S	1.50
1	No change	Ribonuclease T1	1RN1	N81A	N107A	2.91
1	No change	Ribonuclease T1	1RN1	N9A	N35A	0.71
1	No change	Ribonuclease T1	1RN1	Q25K	Q51K	-0.94
1	No change	Ribonuclease T1	1RN1	S12A	S38A	1.08
1	No change	Ribonuclease T1	1RN1	S17A	S43A	-0.57
1	No change	Ribonuclease T1	1RN1	S64A	S90A	1.56
1	No change	Ribonuclease T1	1RN1	V16A	V42A	2.19
1	No change	Ribonuclease T1	1RN1	V16C	V42C	4.68

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ribonuclease T1	1RN1	V16S	V42S	4.71
1	No change	Ribonuclease T1	1RN1	V16T	V42T	4.10
1	No change	Ribonuclease T1	1RN1	V78A	V104A	4.08
1	No change	Ribonuclease T1	1RN1	V78C	V104C	3.67
1	No change	Ribonuclease T1	1RN1	V78S	V104S	4.73
1	No change	Ribonuclease T1	1RN1	V78T	V104T	3.10
1	No change	Ribonuclease T1	1RN1	V89C	V115C	3.54
1	No change	Ribonuclease T1	1RN1	V89S	V115S	4.87
1	No change	Ribonuclease T1	1RN1	V89T	V115T	3.07
1	No change	Ribonuclease T1	1RN1	W59Y	W85Y	0.93
1	No change	Ribonuclease T1	1RN1	Y11F	Y37F	2.03
1	No change	Ribonuclease T1	1RN1	Y24W	Y50W	-1.24
1	No change	Ribonuclease T1	1RN1	Y42F	Y68F	-1.15
1	No change	Ribonuclease T1	1RN1	Y42W	Y68W	0.14
0	PDB Related	Ribonuclease T1	1RN1	Y45W	Y71W	-0.74
1	No change	Ribonuclease T1	1RN1	Y56F	Y82F	0.71
1	No change	Ribonuclease T1	1RN1	Y57F	Y83F	0.44
1	No change	Ribonuclease T1	1RN1	Y68F	Y94F	1.35
1	No change	Ribose-binding protein	2DRI	A188C	A213C	1.80
1	No change	Ribose-binding protein	2DRI	L62C	L87C	2.70
1	No change	Rop	1ROP	D30A	D30A	-0.30
1	No change	Rop	1ROP	D30C	D30C	-0.80

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Rop	1ROP	D30E	D30E	-1.00
1	No change	Rop	1ROP	D30F	D30F	0.10
1	No change	Rop	1ROP	D30G	D30G	-2.00
1	No change	Rop	1ROP	D30H	D30H	-0.90
1	No change	Rop	1ROP	D30I	D30I	0.80
1	No change	Rop	1ROP	D30K	D30K	-0.90
1	No change	Rop	1ROP	D30L	D30L	0.10
1	No change	Rop	1ROP	D30M	D30M	-0.60
1	No change	Rop	1ROP	D30N	D30N	-0.80
1	No change	Rop	1ROP	D30P	D30P	1.60
1	No change	Rop	1ROP	D30Q	D30Q	-1.80
1	No change	Rop	1ROP	D30R	D30R	-0.80
1	No change	Rop	1ROP	D30S	D30S	-1.00
1	No change	Rop	1ROP	D30T	D30T	0.40
1	No change	Rop	1ROP	D30V	D30V	0.40
1	No change	Rop	1ROP	D30W	D30W	0.40
1	No change	Rop	1ROP	D30Y	D30Y	-0.20
1	No change	Rop	1ROP	L41A	L41A	6.12
1	No change	Rop	1ROP	L41V	L41V	2.53
1	No change	Sac7d	1AZP	V30I	V30I	-0.70
1	No change	Spectrin	1AJ3	W22F	W1784F	1.11
1	No change	Sso7d	1SSO	I29V	I30V	0.40

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Staphylococcal nuclease	1STN	A60G	A142G	1.50
1	No change	Staphylococcal nuclease	1STN	A60V	A142V	2.90
1	No change	Staphylococcal nuclease	1STN	A90S	A172S	2.37
1	No change	Staphylococcal nuclease	1STN	D77A	D159A	3.20
1	No change	Staphylococcal nuclease	1STN	D77G	D159G	2.20
1	No change	Staphylococcal nuclease	1STN	D83A	D165A	3.90
1	No change	Staphylococcal nuclease	1STN	D83G	D165G	2.80
1	No change	Staphylococcal nuclease	1STN	D95A	D177A	3.60
1	No change	Staphylococcal nuclease	1STN	D95G	D177G	3.10
1	No change	Staphylococcal nuclease	1STN	E73A	E155A	1.60
1	ID Corrected	Staphylococcal nuclease	1STN	F34C	F34C	2.70
1	No change	Staphylococcal nuclease	1STN	F61A	F143A	2.40
1	No change	Staphylococcal nuclease	1STN	F61G	F143G	4.70
1	No change	Staphylococcal nuclease	1STN	F76A	F158A	4.10
1	No change	Staphylococcal nuclease	1STN	F76G	F158G	4.70
1	No change	Staphylococcal nuclease	1STN	G79S	G161S	1.30
1	No change	Staphylococcal nuclease	1STN	G88V	G170V	-0.50
1	No change	Staphylococcal nuclease	1STN	H124L	H206L	4.07
1	No change	Staphylococcal nuclease	1STN	H46Y	H128Y	3.35
1	No change	Staphylococcal nuclease	1STN	I18G	I100G	2.60
1	No change	Staphylococcal nuclease	1STN	I18M	I100M	3.11
1	No change	Staphylococcal nuclease	1STN	L103A	L185A	4.60

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Staphylococcal nuclease	1STN	L25A	L107A	2.60
1	No change	Staphylococcal nuclease	1STN	L36A	L118A	3.60
1	No change	Staphylococcal nuclease	1STN	L36G	L118G	5.40
1	No change	Staphylococcal nuclease	1STN	L37A	L119A	1.80
1	No change	Staphylococcal nuclease	1STN	L37G	L119G	3.90
1	No change	Staphylococcal nuclease	1STN	L89F	L171F	1.67
1	No change	Staphylococcal nuclease	1STN	N118D	N200D	2.50
1	No change	Staphylococcal nuclease	1STN	S141A	S223A	0.12
1	No change	Staphylococcal nuclease	1STN	T33S	T115S	1.40
1	No change	Staphylococcal nuclease	1STN	T62A	T144A	2.40
1	No change	Staphylococcal nuclease	1STN	T62G	T144G	3.40
1	No change	Staphylococcal nuclease	1STN	V23F	V105F	1.43
1	No change	Staphylococcal nuclease	1STN	V66L	V148L	-0.80
1	No change	Staphylococcal nuclease	1STN	W140F	W140F	0.60
1	No change	Staphylococcal nuclease	1STN	W140H	W140H	0.40
1	No change	Staphylococcal nuclease	1STN	W140L	W140L	4.50
1	No change	Staphylococcal nuclease	1STN	W140Y	W140Y	1.10
1	No change	Staphylococcal nuclease	1STN	Y91S	Y173S	5.30
1	No change	Subtilisin BPN'	1SUP	G169A	G276A	-0.30
0	PDB Related, Mismatch	Subtilisin BPN'	no ID	K57E	K87E	1.00
1	No change	Subtilisin BPN'	1SUP	M50F	M157F	-0.48
1	No change	Subtilisin BPN'	1SUP	N218S	N325S	-1.07

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Subtilisin BPN'	1SUP	N76D	N183D	-0.45
1	No change	Subtilisin BPN'	1SUP	Q206C	Q313C	-1.25
0	PDB Related, Mismatch	Subtilisin BPN'	no ID	Q40L	Q70L	1.10
1	No change	Subtilisin BPN'	1SUP	Y217K	Y324K	-0.72
1	No change	Subtilisin inhibitor	3SSI	M103A	M134A	2.56
1	No change	Subtilisin inhibitor	3SSI	M103G	M134G	6.96
1	No change	Subtilisin inhibitor	3SSI	M103I	M134I	2.09
1	No change	Subtilisin inhibitor	3SSI	M103L	M134L	0.07
1	No change	Subtilisin inhibitor	3SSI	M103V	M134V	2.03
1	No change	Subtilisin inhibitor	3SSI	M73A	M104A	-0.55
1	No change	Subtilisin inhibitor	3SSI	M73D	M104D	-1.60
1	No change	Subtilisin inhibitor	3SSI	M73E	M104E	-1.05
1	No change	Subtilisin inhibitor	3SSI	M73G	M104G	0.06
1	No change	Subtilisin inhibitor	3SSI	M73I	M104I	0.94
1	No change	Subtilisin inhibitor	3SSI	M73K	M104K	-0.35
1	No change	Subtilisin inhibitor	3SSI	M73L	M104L	0.25
1	No change	Subtilisin inhibitor	3SSI	M73V	M104V	0.66
1	No change	Subtilisin inhibitor	3SSI	V13A	V44A	6.78
1	No change	Subtilisin inhibitor	3SSI	V13F	V44F	5.69
1	No change	Subtilisin inhibitor	3SSI	V13G	V44G	10.30
1	No change	Subtilisin inhibitor	3SSI	V13I	V44I	0.89
1	No change	Subtilisin inhibitor	3SSI	V13L	V44L	2.37

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Subtilisin inhibitor	3SSI	V13M	V44M	5.68
1	No change	Superoxide dismutase	1N0J	Q143N	Q167N	-0.06
1	No change	Superoxide dismutase	1N0J	Y34F	Y58F	-3.10
1	No change	Superoxide dismutase	1N0J	I58T	I82T	2.70
1	No change	Tailspike protein	1CLW	E309V	E310V	2.30
1	No change	Tailspike protein	1CLW	G177R	G178R	2.10
1	No change	Tailspike protein	1CLW	G244R	G245R	4.00
1	No change	Tailspike protein	1CLW	G323D	G324D	0.40
1	No change	Tailspike protein	1CLW	R285K	R286K	10.30
1	No change	Tailspike protein	1CLW	R382S	R383S	17.40
1	No change	Tailspike protein	1CLW	T235I	T236I	1.60
0	PDB Related	TEM beta-lactamase	no ID	E104K	E102K	-0.64
0	PDB Related	TEM beta-lactamase	no ID	G238S	G236S	1.25
0	PDB Related	TEM beta-lactamase	no ID	R164H	R162H	0.17
0	PDB Related	TEM beta-lactamase	no ID	R164S	R162S	0.61
0	PDB Related	TEM beta-lactamase	no ID	S235A	S233A	-0.58
1	No change	Thioredoxin	2TRX	L78K	L79K	3.90
1	No change	Thioredoxin	2TRX	L78R	L79R	4.00
1	No change	Trypsin inhibitor	1BPI	D3A	D38A	-0.20
1	No change	Trypsin inhibitor	1BPI	D50A	D85A	0.40
1	No change	Trypsin inhibitor	1BPI	E49A	E84A	0.20
1	No change	Trypsin inhibitor	1BPI	E7A	E42A	1.50

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Trypsin inhibitor	1BPI	F22A	F57A	2.00
1	No change	Trypsin inhibitor	1BPI	F4A	F39A	3.00
1	No change	Trypsin inhibitor	1BPI	G12A	G47A	1.80
1	No change	Trypsin inhibitor	1BPI	G28A	G63A	1.00
1	No change	Trypsin inhibitor	1BPI	G36A	G71A	2.10
1	No change	Trypsin inhibitor	1BPI	G36S	G71S	0.70
1	No change	Trypsin inhibitor	1BPI	G37A	G72A	2.30
1	No change	Trypsin inhibitor	1BPI	G56A	G91A	0.20
1	No change	Trypsin inhibitor	1BPI	G57A	G92A	0.20
1	No change	Trypsin inhibitor	1BPI	I18A	I53A	1.50
1	No change	Trypsin inhibitor	1BPI	I19A	I54A	2.10
1	No change	Trypsin inhibitor	1BPI	K15A	K50A	0.40
1	No change	Trypsin inhibitor	1BPI	K26A	K61A	0.00
1	No change	Trypsin inhibitor	1BPI	K41A	K76A	0.40
1	No change	Trypsin inhibitor	1BPI	K46A	K81A	-0.10
1	No change	Trypsin inhibitor	1BPI	L29A	L64A	0.00
1	No change	Trypsin inhibitor	1BPI	L6A	L41A	0.60
1	No change	Trypsin inhibitor	1BPI	M52A	M87A	1.70
1	No change	Trypsin inhibitor	1BPI	N24A	N59A	2.20
1	No change	Trypsin inhibitor	1BPI	N44A	N79A	3.30
1	No change	Trypsin inhibitor	1BPI	P13A	P48A	1.20
1	No change	Trypsin inhibitor	1BPI	P2A	P37A	1.30

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Trypsin inhibitor	1BPI	P8A	P43A	0.30
1	No change	Trypsin inhibitor	1BPI	P9A	P44A	0.80
1	No change	Trypsin inhibitor	1BPI	Q31A	Q66A	1.00
1	No change	Trypsin inhibitor	1BPI	R17A	R52A	0.30
1	No change	Trypsin inhibitor	1BPI	R1A	R36A	0.50
1	No change	Trypsin inhibitor	1BPI	R20A	R55A	1.80
1	No change	Trypsin inhibitor	1BPI	R39A	R74A	0.00
1	No change	Trypsin inhibitor	1BPI	R42A	R77A	0.50
1	No change	Trypsin inhibitor	1BPI	R53A	R88A	0.10
1	No change	Trypsin inhibitor	1BPI	S47A	S82A	1.60
1	No change	Trypsin inhibitor	1BPI	T11A	T46A	0.00
1	No change	Trypsin inhibitor	1BPI	T32A	T67A	0.10
1	No change	Trypsin inhibitor	1BPI	T54A	T89A	0.10
1	No change	Trypsin inhibitor	1BPI	V34A	V69A	1.20
1	No change	Trypsin inhibitor	1BPI	Y10A	Y45A	1.20
1	No change	Trypsin inhibitor	1BPI	Y35A	Y70A	1.10
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C118A	C118A	1.36
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C118S	C118S	2.27
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C118V	C118V	1.34
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C154A	C154A	1.03
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C154S	C154S	1.72
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C154V	C154V	1.12

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Tryptophan synthase alpha-subunit	1WQ5	C81A	C81A	0.69
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C81G	C81G	1.58
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C81S	C81S	1.41
1	No change	Tryptophan synthase alpha-subunit	1WQ5	C81V	C81V	1.32
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49A	E49A	0.08
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49C	E49C	-0.01
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49D	E49D	0.80
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49F	E49F	0.86
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49G	E49G	-0.08
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49H	E49H	-0.33
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49I	E49I	-0.46
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49K	E49K	-0.20
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49L	E49L	-0.44
1	ID Corrected	Tryptophan synthase alpha-subunit	1WQ5	E49M	E49M	-3.10
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49N	E49N	0.69
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49P	E49P	0.00
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49Q	E49Q	0.23
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49S	E49S	0.59
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49T	E49T	0.62
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49V	E49V	-0.14
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49W	E49W	0.97
1	No change	Tryptophan synthase alpha-subunit	1WQ5	E49Y	E49Y	0.17

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	ID Corrected	Tryptophan synthase alpha-subunit	1WQ5	F22L	F22L	0.90
1	ID Corrected	Tryptophan synthase alpha-subunit	1WQ5	G211E	G211E	0.40
1	No change	Tryptophan synthase alpha-subunit	1WQ5	G211R	G211R	-0.10
1	ID Corrected	Tryptophan synthase alpha-subunit	1WQ5	G234D	G234D	-0.10
1	ID Corrected	Tryptophan synthase alpha-subunit	1WQ5	G234K	G234K	0.30
1	No change	Tryptophan synthase alpha-subunit	1WQ5	P132A	P132A	0.92
1	No change	Tryptophan synthase alpha-subunit	1WQ5	P132G	P132G	0.78
1	No change	Tryptophan synthase alpha-subunit	1WQ5	P207A	P207A	1.64
0	PDB Related	Tryptophan synthase alpha-subunit	1WQ5	P57A	P57A	0.04
0	PDB Related	Tryptophan synthase alpha-subunit	1WQ5	P62A	P62A	0.52
1	No change	Tryptophan synthase alpha-subunit	1WQ5	P96A	P96A	2.00
1	No change	Tumor suppressor P53 complexed with DNA	1TUP	C242S	C242S	2.94
1	No change	Tumor suppressor P53 complexed with DNA	1TUP	R175H	R175H	3.01
1	No change	Tumor suppressor P53 complexed with DNA	1TUP	R248Q	R248Q	1.94
1	No change	Tumor suppressor P53 complexed with DNA	1TUP	R249S	R249S	1.95
1	No change	Tumor suppressor P53 complexed with DNA	1TUP	R273H	R273H	0.35
1	No change	Ubiquitin	1AAR	F45W	F121W	-0.32
1	No change	Ubiquitin	1AAR	H68E	H144E	-0.76
1	No change	Ubiquitin	1AAR	H68Q	H144Q	-0.55
1	No change	Ubiquitin	1AAR	K27Q	K103Q	1.91

Appendix 1. Manually curated version of PON-tstab dataset (cont.)

1	No change	Ubiquitin	1AAR	K29N	K105N	1.48
1	No change	Ubiquitin	1AAR	K29Q	K105Q	1.67
1	No change	Ubiquitin	1AAR	K6E	K82E	-0.53
1	No change	Ubiquitin	1AAR	K6Q	K82Q	-0.26
1	No change	Ubiquitin	1AAR	R42E	R118E	-1.63
1	No change	Ubiquitin	1AAR	R72Q	R148Q	0.33
1	No change	Yes-associated protein	1K9Q	A20R	A180R	-0.89
1	No change	Yes-associated protein	1K9Q	D34T	D194T	-0.35
1	No change	Yes-associated protein	1K9Q	L30Y	L190Y	-0.17

8. CURRICULUM VITAE