STABILITY ORACLE: A STRUCTURE-BASED GRAPH-TRANSFORMER FOR IDENTIFYING STABILIZING MUTATIONS

Daniel J. Diaz[1, 2, 3], Chengyue Gong[1], Jeffrey Ouyang-Zhang[1], James M. Loy[2, 5], Jordan Wells[4], David Yang[5], Andrew D. Ellington[5], Alex Dimakis[6], Adam R. Klivans[1]

[1] Department of Computer Science, UT Austin [2] Intelligent Proteins, LLC [3] Department of Chemistry, UT Austin
[4] McKetta Department of Chemical Engineering, UT Austin [5] Department of Molecular Biosciences, UT Austin
[6] Chandra Family Department of Electrical and Computer Engineering, UT Austin

May 15, 2023

ABSTRACT

Stabilizing proteins is a fundamental challenge in protein engineering and is almost always a prerequisite for the development of industrial and pharmaceutical biotechnologies. Here we present Stability Oracle: a structure-based graph-transformer framework that achieves state-of-the-art performance on predicting the effect of a point mutation on a protein's thermodynamic stability ($\Delta\Delta G$). A strength of our model is its ability to identify *stabilizing* mutations, which often make up a small fraction of a protein's mutational landscape. Our framework introduces several data and machine learning innovations to overcome well-known challenges in data scarcity and bias, generalization, and computation time. Stability Oracle is first pretrained on over 2M masked microenvironments and then fine-tuned using a novel data augmentation technique, Thermodynamic Permutations (TP), applied to a $\sim$120K curated subset of the mega-scale cDNA display proteolysis dataset. This technique increases the original 120K mutations to over 2M thermodynamically valid $\Delta\Delta G$ measurements to generate the first structure training set that samples and balances all 380 mutation types. By using the masked microenvironment paradigm, Stability Oracle does not require a second mutant structure and instead uses amino acid structural embeddings to represent a mutation. This architectural design accelerates training and inference times: we can both train on 2M instances with just 119 structures and generate deep mutational scan (DMS) predictions from only the wildtype structure. We benchmark Stability Oracle with both experimental and AlphaFold structures of all proteins on T2837, a test set that aggregates the common test sets (SSym, S669, p53, and Myoglobin) with all additional experimental data from proteins with over a 30% sequence similarity overlap. We used TP augmented T2837 to evaluate performance for protein engineering: Stability Oracle correctly identifies 48% of stabilizing mutations ($\Delta\Delta G$ < -0.5 kcal/mol) and 74% of its stabilizing predictions are indeed stabilizing (18% and 8% of predictions were neutral and destabilizing, respectively). For a fair comparison between sequence and structure-based fine-tuned deep learning models, we build on the Prostata framework and fine-tune the sequence embeddings of ESM2 on our training set (Prostata-IFML). A head-to-head comparison demonstrates that Stability Oracle outperforms Prostata-IFML on regression and classification even though the model is 548 times smaller and is pretrained with 4000 times fewer proteins, highlighting the advantages of learning from structures.

1 Introduction

The ability to predict and understand the change in protein thermodynamic stability ($\Delta\Delta G$) for an amino acid substitution is a core task for the development of protein-based biotechnology such as industrial biocatalyst [1, 2, 3] and pharmaceutical biologics [4, 5, 6, 7]. Proteins with enhanced thermodynamic stability are less prone to unfolding and aggregation and more engineerable [8, 9]; stabilizing the scaffold of a protein enables downstream exploration of potentially destabilizing mutations that may improve a target function [8]. Thermodynamic stability is measured by

the Gibbs free energy change (ΔG) between the native and unfolded state and reflects the underlying integrity of the global structure. Engineering proteins is a very laborious process and bottlenecks the development of protein-based biotechnologies[10, 11], making the development of computational methods that can accurately predict $\Delta\Delta G$ of a point mutation, and in turn identify stabilizing mutations, a highly active research area[12, 13, 14, 15, 16].

Deep learning is currently revolutionizing many physical and biological disciplines[17, 18, 19, 20, 21], with AlphaFold2 leading the way as the scientific breakthrough of the year in 2021[22, 23] and sparking an entire wave of deep learning structure prediction tools[24, 25]. Although a variety of sequence-based [26, 27] and structure-based [28, 29, 30] deep learning frameworks for stability prediction have been reported, the lack of data and machine learning engineering issues have prevented deep learning algorithms from having a similarly revolutionary impact on protein stability prediction (See Appendix C for details on the current deep learning frameworks).

Systematic analyses [15, 13, 16, 31] of the state-of-the-art computational stability predictors published over the last 15 years highlight data scarcity, variation, bias, and leakage, and poor metrics for evaluating model performance as the critical issues hindering meaningful progress (See Appendix B for summary on the Primary data issues). As a result, the community still primarily relies on knowledge-based methods, such as Rosetta [32] and FoldX [33], or shallow machine learning approaches [34, 35, 36, 37, 38, 39, 40, 41, 42]. Furthermore, third-party analyses emphasize that only about 20% of predicted stabilizing mutations from the current state-of-the-art computational tools are actually stabilizing, even though 75-80% accuracy is often reported upon publication [14, 31, 13]. This poor generalization to new datasets highlights how pervasive the data leakage between training-test splits is in the computational stability prediction community [16, 31, 14, 13]. Finally, current computational tools are evaluated with Pearson correlation and error as the primary metrics. Due to the significant imbalance between stabilizing (<30%) and destabilizing mutations in training and test sets and the innate variations associated with measuring $\Delta\Delta G$, improvements in these metrics do not necessarily translate into improvements for identifying stabilizing mutations [31, 13]. Thus, metrics such as precision, recall, area under receiver operating characteristic (AUROC), and Matthew's correlation coefficient (MCC) should be prioritized to monitor improvements for protein engineering applications[31, 13, 14].(See Appendix B for summary on the metric issues)

To address these issues and develop a robust stability prediction model for protein engineering, we present Stability Oracle: a structure-based deep learning framework that makes use of several innovations in data and machine learning engineering specific to stability prediction. Stability Oracle has a graph-transformer architecture that treats atoms as tokens and utilizes their pairwise distances to inject a structural inductive bias into the attention mechanism. The input to Stability Oracle consists of the local chemistry surrounding a residue with the residue deleted (masked microenvironment) and two amino acid embeddings to represent a specific point mutation. This design decision enables Stability Oracle to generate all 380 possible point mutations from a single microenvironment, circumventing the need for computationally generated mutant structures and making deep mutational scanning inference computationally inexpensive.

To improve the generalization capability of Stability Oracle, we introduce a data augmentation technique—Thermodynamic Permutations (TP). Thermodynamic Permutations, similar to Thermodynamic Reversibility (TR)[12], are based on the state function property of Gibbs free energy [43]. For a specific position in a protein, TP expands n empirical $\Delta\Delta G$ measurements into $n(n-1)$ thermodynamically valid measurements, thus expanding the dataset by up to an order of magnitude based on the number of residues that have multiple amino acids experimentally characterized.

TP enables all 380 mutation types to be sampled in a training and test set for the first time, alleviating common data biases, such as those produced by alanine scanning experiments as well as other biases. Furthermore, unlike TR, it generates a balanced set of stabilizing and destabilizing mutations to non-wildtype amino acids. Thus, augmenting training and test set to better assess generalization for protein engineering (where you are mutating away from wildtype).

We present three new datasets (C2878, cDNA120K, T2837) to address data leakage issues. MMseqs2 [44] was used to generate the three datasets and ensure the maximum overlap between proteins in the training and test set is below 30% sequence similarity. Concat 2878 (C2878) and Test 2837 (T2837) datasets are new training and test splits from previously established training and test sets, respectively. The third (cDNA120K) is a curated subset of the cDNA display proteolysis dataset #1[45], a dataset of $\sim$850K thermodynamic folding stability measurements across 354 natural and 188 de novo mini-protein domains (40-72 amino acids).

Finally, we generate a sequence-based counterpart for Stability Oracle by fine-tuning ESM2 on our curated training and test sets using the Prostata framework [26] and train Prostata-IFML. With Prostata-IFML, we conduct head-to-head comparisons to demonstrate the advantage of structure over sequence-based methods. Overall, Stability Oracle and Prostata-IFML are state-of-the-art structure and sequence-based frameworks for computational stability prediction.

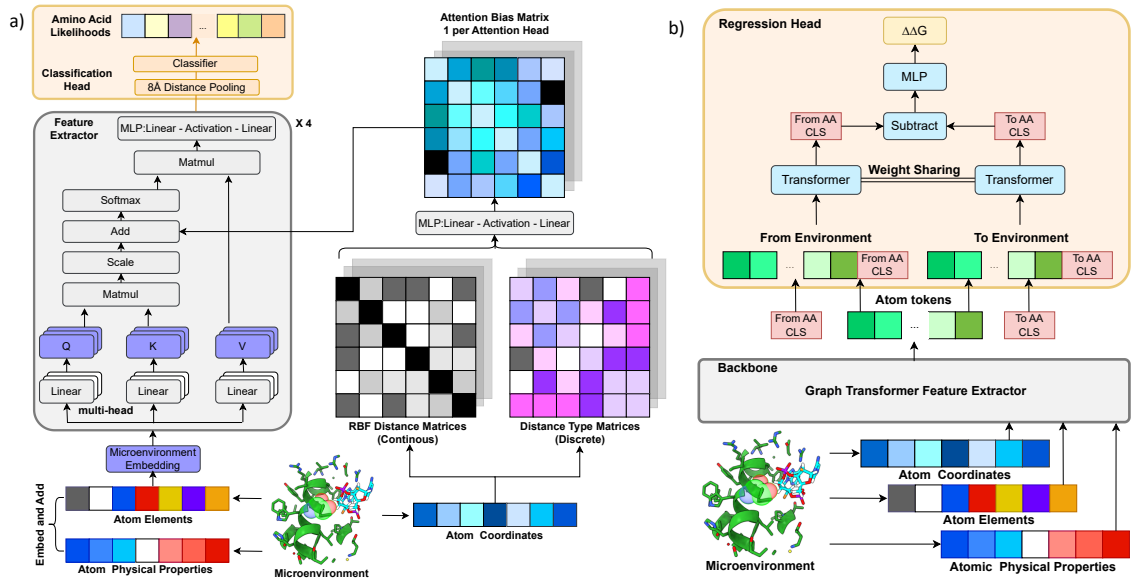

Figure 1: Overview of model architectures. a) Self-supervised pretraining graph-transformer architecture (MutComputeXGT). b) Fine-tuning pretrained graph transformer for stability regression (Stability Oracle).

2 Results

2.1 Designing a Graph-Transformer Framework for Structure-based Protein Engineering

In prior work we have experimentally shown that representations learned via self-supervised deep learning models on masked microenvironments (MutCompute and MutComputeX) can identify amino acid residues incongruent with their surrounding chemistry. These models can be used to "zero-shot" predict gain-of-function point mutations [46, 47, 48]. Self-supervised models, however, generate designs that are not biased towards a particular phenotype and do not update predictions based on experimental data [19, 49].

The MutCompute framework uses a voxelized molecular representation [46, 48]. For protein structures, voxelization is a suboptimal molecular representation, as most voxels consist of empty space, and rotational invariance is not encoded. Further, the MutCompute frameworks use convolution-based architectures, which lag behind modern attention-based architectures in terms of capacity and predictive power.

To develop a more powerful supervised framework, we first built MutComputeXGT, a graph-transformer version of MutComputeX as displyed in Figure 1 (a). Each atom is represented as a node with atomic elements, partial charges, and SASA values as features with pairwise atomic distances labeling edges. Our graph-transformer architecture converts the pairwise distances into continuous and categorical attention-biases to provide a structure-based inductive bias for the attention mechanism. To generate likelihoods for the masked amino acid, we average the embeddings of all atomic tokens within 8Å of the masked Cα. The design decision to narrow the pooling to atoms in the first contact shell of the masked amino acid is based on insights from systematic variation of the microenvironment volume when training self-supervised CNNs [49]. With a similar number of parameters and the same training-test splits, MutComputeXGT demonstrates superior representation learning capacity than MutComputeX, achieving WT accuracy of 92.98%$\pm$0.26% compared to $\sim$85% [48].

The Stability Oracle architecture makes use of both the feature-extractor and classifier of MutComputeXGT for supervised fine-tuning Figure 1(b). Previous structure-based stability predictors require two structures, one for the wildtype and one for the mutant structure. Instead, Stability Oracle uses only one structure encoded via a MutComputeXGT representation, and two amino acid level embeddings, one for the "from" and one for the "to" amino acid. This architectural decision alleviates the computational burden of generating a mutant structure for each $\Delta\Delta G$ prediction: for a typical 300 amino acid protein, prior work would generate 5700 computational structures (from Rosetta [32] or AlphaFold [22]) in order to predict the $\Delta\Delta G$ of every possible single point mutation. Stability Oracle, on the other hand, requires only one structure. Details on graph generation and training and inference times can be found in the Appendix A.

The "from" and "to" amino acid embeddings are the weights from the final layer of MutComputeXGT classifier. This design decision is based on the insight that the weights of these 20 neurons represent the similarity of a microenvironment's features to each of the 20 amino acids prior to being normalized into a likelihood distribution. Thus, they are structure-based contextualized embeddings of the 20 amino acids learned from the entire Protein Data Bank (PDB) (we will refer to them as structural amino acid embeddings). We show that structural amino acid embeddings improve performance compared to the naïve one-hot representation (See Table 1 in Appendix A).

We highlight several design decisions of the regression head used in Stability Oracle. Of note is the use a Siamese attention architecture that treats the mutation embeddings as two classification (CLS) tokens as shown in Figure 1(b). CLS tokens are commonly used in the NLP community to capture the global context for downstream tasks in contrast to word-level tokens. Since atoms and amino acids are objects at different scales, we designed the regression head to contextualize the "from" and "to" CLS tokens with the atomic tokens of a particular microenvironment. Once contextualized for a given microenvironment, the two CLS tokens are subtracted from each other and decoded into a $\Delta\Delta G$ prediction. We subtract the two CLS tokens to enforce the state function property of Gibbs free energy [50], providing the proper inductive bias for Thermodynamic Reversibility and "self-mutations" (where $\Delta\Delta G$ = 0 kcal/mol).

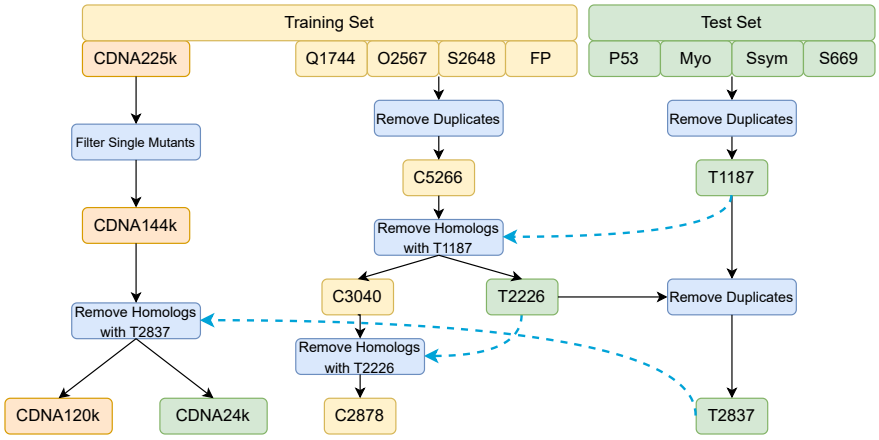

Figure 2: Training and test set generation pipeline.

2.2 Training Stability Oracle to Generalize across Protein Chemistry

The Stability Oracle framework was designed to generalize across all 380 mutation types at all positions within a protein structure. The development of such a model has historically been limited by data scarcity, bias, and leakage. To address these issues, we curated training and test datasets and created a new data augmentation technique called Thermodynamic Permutations (TP).

It is well known that a major issue with prior works is the inclusion of similar proteins in both the training and test set ("data leakage"). It has been demonstrated that training-test splits at the mutation, residue, or protein level result in data leakage, and strict sequence clustering is required to mitigate this effect [14, 51, 13, 31]. Here we created new training-test splits based on a 30% sequence similarity threshold computed by MMSeqs2 [44]. First, we designed T2837: a test set that combines the common test sets (SSym [50], S669[12], Myoglobin, and P53) with any additional experimental data in common training sets (q1744[28], o2567 [16], s2648[38], and FireProtDB [52]) with a sequence similarity greater than 30%. After establishing the T2837 test set, we generate our training set (C2878) from the remaining experimental data by filtering out proteins with over 30% sequence similarity to T2837. The same procedure was used to construct the cDNA120K training set from the single mutant subset that had experimental structures available of the recently published cDNA-display proteolysis Dataset #1 [45] (See Figure 2).

Even with the T2837 expanded test set, we are still unable to accurately estimate generalization error since 14% of all 380 mutation types are not represented in T2837. Additionally, T2837 is heavily biased with mutations to alanine (see Figure 3 (c), further hindering our ability to evaluate generalization. So far, the community has relied on Thermodynamic Reversibility (TR) augmentation to help alleviate the mutation type sampling issue. However, this additional data from TR biases the model towards primarily stabilizing mutations as the "to" amino acid becomes wildtype. For structure-based methods, especially those that use self-supervised pretraining, the microenvironment itself can leak the wildtype amino acid, as predicting wildtype from the masked microenvironment is the pretraining task. Thus, TR does not help gauge performance to non-wildtype mutations, nor does it improve sampling of the 380 mutation types throughout different structural motifs.

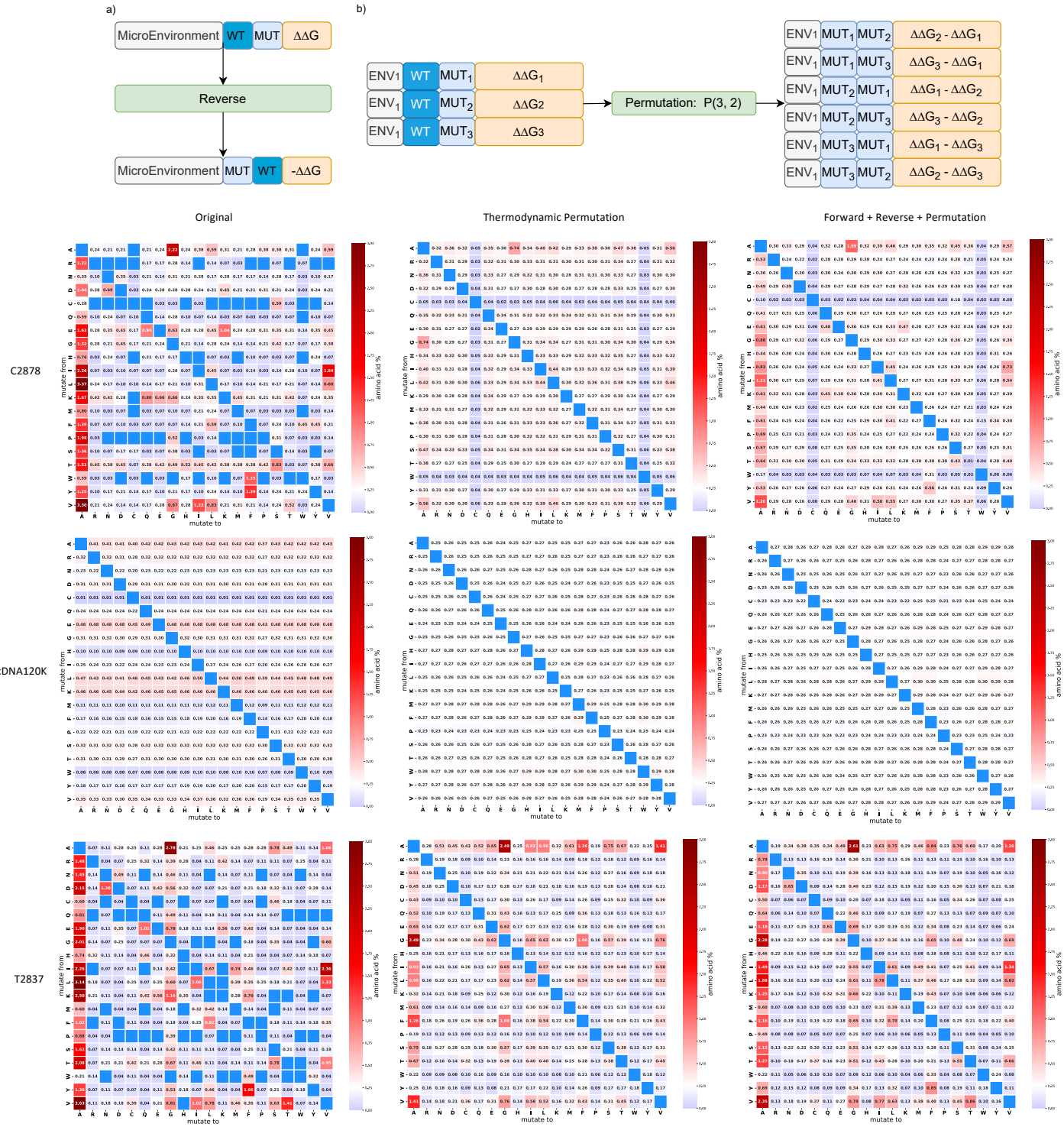

Figure 3: Thermodynamic data augmentation techniques applied to our datasets. A) Diagram describing how Thermodynamic Reversibility works (TR) B) Diagram describing how Thermodynamic Permutations works (TP) C) Heatmap representation of the mutation type distribution present in C2878, cDNA120K, T2837. For C2878, the original, TP, and original + TP + TR dataset consist of 2878, 18912, and 24668 mutations that sample 86.8%, 100%, and 100% of the mutation types, respectively (first row). For cDNA120K, the original, TP, and original + TP + TR datasets consist of 119690, 2071122, and 2190812 mutations, respectively, and each dataset samples 100% of the mutation types (second row). For T2837, the original, TP, and original + TP + TR datasets consist of 2837, 7720, and 13394 mutations that sample 85.5%, 100%, and 100% of the mutation types, respectively (third row).

To address these issues and improve Stability Oracle's ability to generalize, we introduce our main augmentation technique: Thermodynamic Permutations (TP). TP takes advantage of Gibbs free energy being a state function and enables the generation of thermodynamically accurate point mutations at residues with multiple amino acids experimentally characterized (see Figure 3). With TP, we can generate an additional 2.1M, 18.9K, and 7.7K point mutations that sample all 380 mutation types for cDNA120K, C2878, and T2837, respectively (these augmented datasets will be labeled with a TP suffix). Additionally, TP significantly helps alleviate several sampling biases in all 3 datasets (Figure 3 middle column). First, it provides mutational data for the 13.2% and 14.5% mutation types lacking any data in C2878 and T2837, thus providing the first training and test sets of experimentally characterized $\Delta\Delta G$ measurements (the cDNA display proteolysis dataset does not directly measure $\Delta\Delta G$ but instead derives $\Delta\Delta G$ values from Next-Gen sequencing data). Second, since order matters in TP, all augmented data generated is a balanced distribution of $\Delta\Delta G$ measurements (stabilizing vs destabilizing) that do not include mutations incorporating the wildtype amino acid of a particular microenvironment (Figure 3). This is in contrast to TR, which innately produces a $\Delta\Delta G$ distribution heavily biased to stabilizing mutations to the wildtype amino acid of that microenvironment. Thus, with TP we can augment an original training or test set with a balanced set of mutations that are not reversion to the wildtype amino acid. TP also generates $\Delta\Delta G$ measurements for mutation types in microenvironments that would rarely be experimentally characterized, expanding the microenvironment/mutation type combinations available for training and testing to improve generalizability across structural motifs found within a protein scaffold. Finally, TP alleviates the mutation type imbalance in all three datasets. This is visually reflected in Figure 3 as a softening of red (oversampled) and blue (undersampled) towards white (balanced sampling). This is most apparent for the "to" alanine bias plaguing the C2878 and T2837 datasets. As for cDNA120K, TP completely balances the distribution of the 380 mutation types with each one representing approximately 0.26% of the dataset (100% / 380), depicted in Figure 3(c) as uniformly white. Thus, we provide the first large-scale $\Delta\Delta G$ dataset (>1M) that evenly samples all 380 mutation types across over 100 protein domains.

We fine-tuned Stability Oracle on cDNA120K(TP) and C2878(TP) and evaluated its performance on T2837, T2837TP, and all common test sets (as demonstrated in Figure 4 and 5). As shown in Figure 5 (a), training on cDNA120K-TP provided the best performance overall on regression and classification metrics across the test sets. This was expected due to its sheer size and mutation-type balance compared to C2878TP. For a literature comparison, we report Pearson correlation coefficients on all test sets against several community predictors in Figure 4, with additional literature comparisons on the traditional regression metrics in Table 2, Appendix A. It is worth noting that Stability Oracle outperforms other predictors in the literature despite their documented issues with data leakage. To prevent inflation of Stability Oracle's binary classification performance, we focus on T2837 + TP and exclude all TR augmentation mutations for T2837, since these mutations are heavily biased with stabilizing mutations to the wild-type amino acid. Here, Stability Oracle demonstrated a recall of 0.69, a precision of 0.70, and an AUROC of 0.83 (Figure 5 (d)). Surprisingly, further fine-tuning with C2878TP did not improve performance on T2837 or T2837TP. Our analysis reveals a 100% homology overlap (>30% sequence similarity) between the proteins in cDNA120K and C2878, providing a rationale for the lack of performance improvement. However, C2878 fine-tuning improves performance on the interface subset of T2837: C2878 fine-tuning improves the Pearson correlation from 0.30 to 0.35, and the Spearman correlation from 0.29 to 0.35 for all interface microenvironements. This improvement is expected since the cDNA dataset consists of monomeric single-domain structures lacking interfaces with other proteins, ligands, or nucleotides. However, meaningful improvements is limited due to the limited amount of protein-protein (127 mutations), protein-ligand (94 mutations), and protein-nucleotide (9 mutations) data in C2878.

We explored the benefit of starting from a pretrained MutComputeXGT compared to randomly initializing feature extractor and structural amino acid embeddings. Unsurprisingly, fine-tuning Stability Oracle with MutComputeXGT's feature extractor and structural amino acid embeddings on C2878 significantly improved Stability Oracle's classification and regression performance on T2837 and T2837TP (see Table 2 in Appendix A).

Since experimental structures are often unavailable, we examined Stability Oracle's ability to generalize to structures generated by AlphaFold. We used ColabFold [53] to generate template-free predicted structures for each protein in T2837 and T2837TP. ColabFold failed to fold 1 protein and 2 structures were removed due to US-align having TM-scores < 0.5, which resulted in the removal of 50 mutations from T2837 [54, 55]. We observed no changes in classification metrics and slightly lower performance on regression metrics on T2837TP such as Pearson, Spearman, and RMSE (Figure 5 (c)). Overall, these results demonstrate the ability of Stability Oracle to generalize to AlphaFold scaffolds when an experimental structure is unavailable.

2.3 Evaluating Stability Oracle's Ability to Identify Stabilizing Mutations

For computational stability predictors to accelerate protein engineering, it is critical that their predictions correctly identify stabilizing mutations. That is, precision is more important than recall. However, it is well documented that

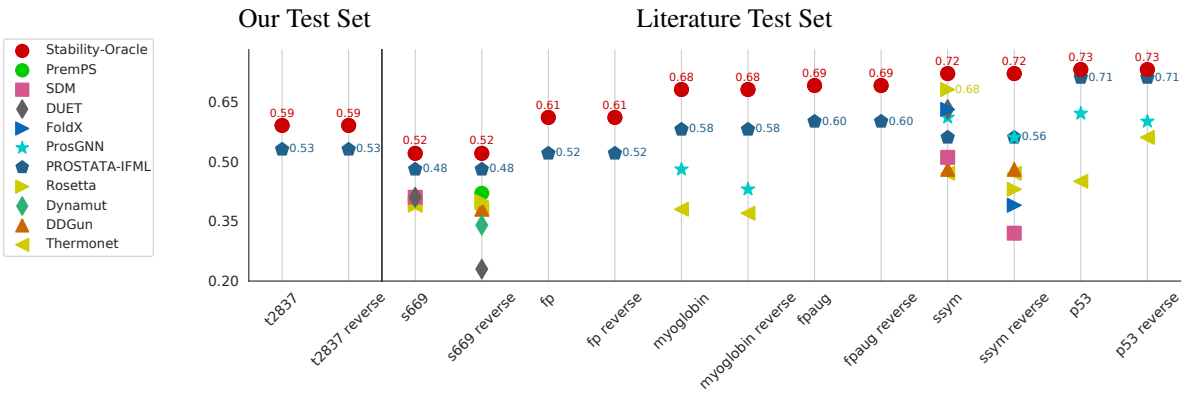

Figure 4: Pearson Correlation Coefficient across several test sets. We show the performance of Stability Oracle and Prostata-IFML on T2837 (forward and reverse). We compare Stability Oracle and Prostata-IFML to the common test sets used by the computational stability prediction community.

state-of-the-art stability predictors can correctly predict stabilizing mutations with only a 20% success rate and most stabilizing predictions are actually neutral or destabilizing [31, 13]. While molecular dynamic-based methods, such as Free Energy Perturbation (FEP), have demonstrated a 50% success rate at identifying stabilizing mutations, their computational demand prevents them from scaling to entire protein applications like computational deep mutational scans (DMS) [31, 56]. Thus, there is a strong need for a method that can match the performance of FEP while being computationally inexpensive.

To evaluate Stability Oracle's ability to identify stabilizing mutations, we filtered its predictions on T2837 and T2837TP at different $\Delta\Delta G$ thresholds and assessed the distribution of experimental stabilizing ($\Delta\Delta G$ < -0.5 kcal/mol), neutral ($|\Delta\Delta G| \leq$ 0.5 kcal/mol), and destabilizing ($\Delta\Delta G$ > 0.5 kcal/mol) mutations. The 0.5 kcal/mol cutoff was chosen based on the average experimental error [57]. With the $\Delta\Delta G$ < -0.5 kcal/mol prediction threshold, 1770 mutations were filtered with an experimental distribution of 74.0% stabilizing, 17.8% neutral, and 8.2% destabilizing and 48.1% of all stabilizing mutations were correctly identified (recall). With a more stringent threshold, the precision increased approaching 0.9 with a -2 kcal/mol $\Delta\Delta G$ threshold. Additional benchmarks on different augmentations and test sets are provided in Appendix A. The success rate of stabilizing mutations surpasses what is typically observed with FEP methods ($\sim$50%) [31, 56] with orders of magnitude less computational cost. In fact, Stability Oracle can generate a stability DMS of the COVID-19 homo-trimer spike protein (PDB: 6vsb, 3357 residues, 67,140 inferences) with 156 seconds (46 ms per residue) for microenvironment pre-processing and 150 seconds (46 ms per residue) for model inference on a single CPU core and A100 GPU, respectively. We further examine Stability Oracle's ability to identify stabilizing mutations by amino acid (in Figure 6). Here, we observe that Stability Oracle is able to correctly predict stabilizing mutations across most amino acids, whether mutating "from" or "to". However, several amino acids lack sufficient "from" or "to" stabilizing predictions to draw meaningful conclusions. This is even more apparent when looking at "from"-"to" pairs (see Figure 8, Appendix A), highlighting how data scarcity still hinders proper model evaluation.

It has been pointed out by the community that the subset of experimentally validated stabilizing surface mutations are heavily skewed towards mutations to hydrophobic amino acids. An analysis by Broom et al. of the ProTherm database[58] indicates that surface stabilizing mutation bias typically increases side chain hydrophobicity ($\Delta\Delta G_{solvation}$) with a median change of 0.8 kcal/mol. This hydrophobicity bias is equivalent to an alanine-to-valine mutation on the protein surface [31]. We examined if this underlying hydrophobic bias persisted within our training pipeline by computing the precision and recall of polar and hydrophobic mutations as a function of relative solvent accessibility (RSA) of the wildtype residue. Our precision and recall results across different RSA bins of T2837(TP) indicate that there is no meaningful between stabilizing predictions to hydrophobic and polar amino acids across a protein structure (see Figure 7). This demonstrates that the TP-augmented cDNA120K training set mitigates a well-known known bias in computational stability prediction.

2.4 Comparing Against Sequence and Structure Fine-Tuned Stability Predictors

Over the last three years, self-supervised Large Language Models for proteins (pLLMs or "sequence models") have had a tremendous impact on the protein community [59, 60, 61, 62, 63, 64, 65, 66, 25]. Understanding the advantages and disadvantages of sequence versus structure-based prediction models continues to be an active area of research in protein design and engineering [67, 19]. As such, we evaluated how Stability Oracle compares to a state-of-the-art sequence-based deep learning model. For protein stability prediction, we considered Prostata [26], a model that

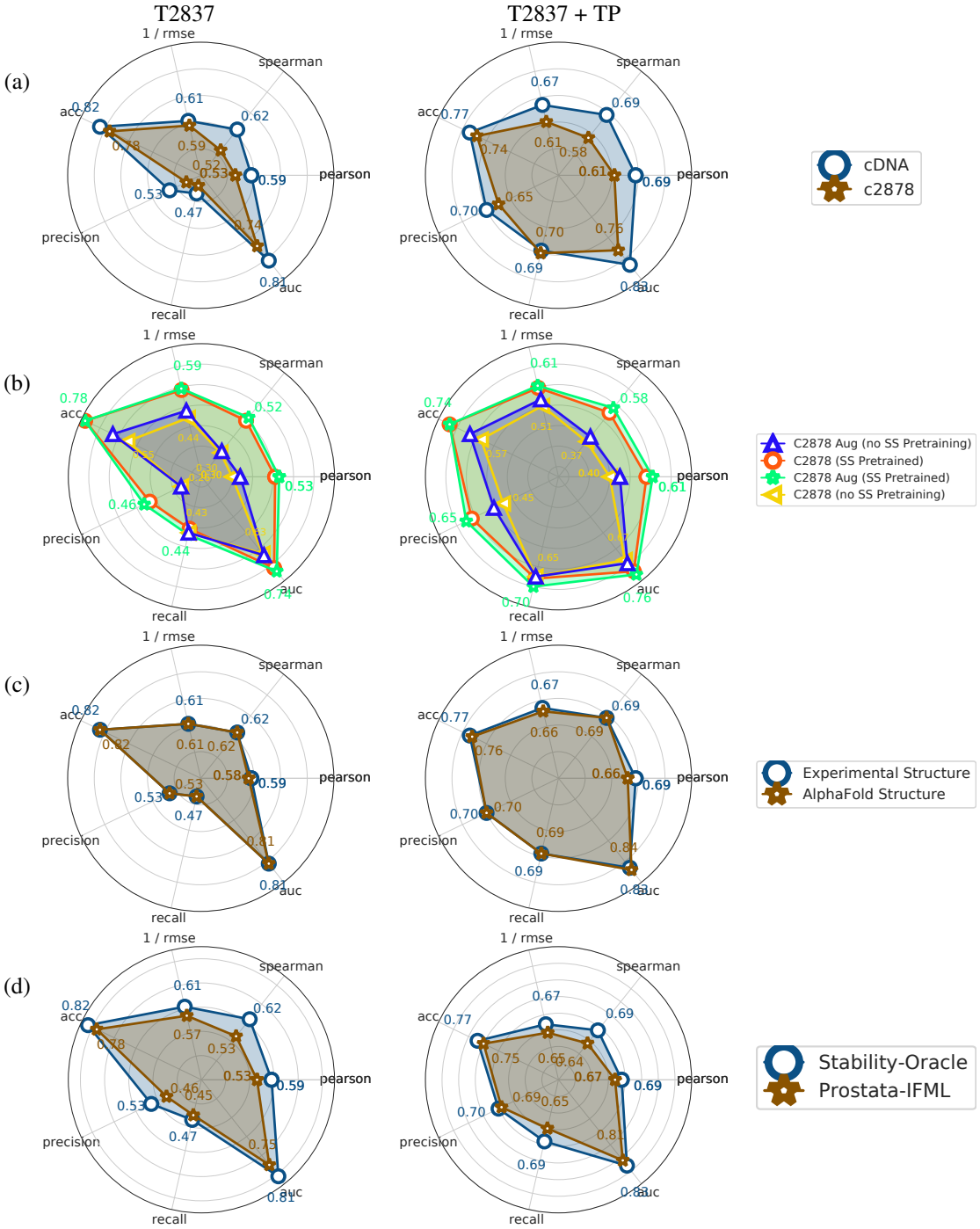

Figure 5: Regression and classification performance of different models on T2837 and T2837TP. (a) Comparison between the C2878 and cDNA120K training sets. (b) Comparison of Stability Oracle performance when trained on C2878 with and without a pre-trained backbone and structural amino acid embeddings and with and without TP augmentation. (c) Comparison of Stability Oracle performance when tested on experimental or AlphaFold structures of T2837. (d) Comparison of Stability Oracle and Prostata-IFML performances.

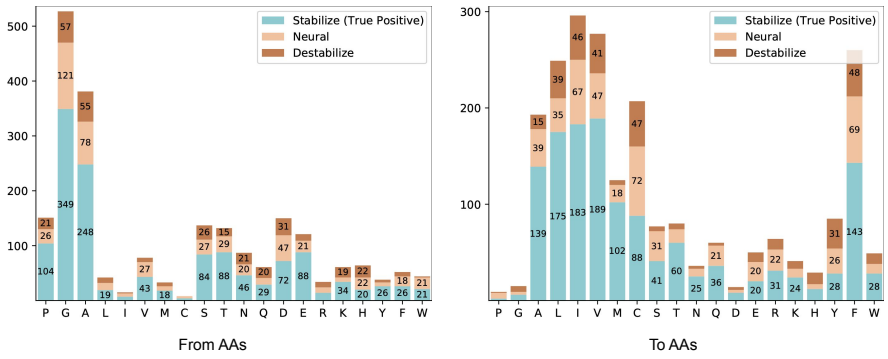

Figure 6: Experimental distribution of Stability Oracle's stabilizing predictions ($\Delta\Delta G$ < -0.5 kcal/mol) on T2837 + TP test set for different "from" and "to" amino acid types. Here, stabilizing, neutral, and destabilizing mutations are defined by $\Delta\Delta G$ < -0.5 kcal/mol, $|\Delta\Delta G| \leq$ 0.5 kcal/mol, and $\Delta\Delta G$ > 0.5 kcal/mol, respectively.

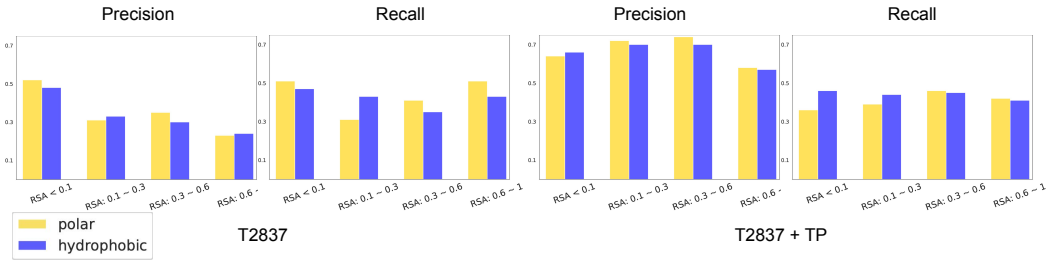

Figure 7: Evaluating Stability Oracle's precision and recall on different relative solvent accessibility (RSA) ranges of T2837 and T2837TP. The results demonstrates that there are no biases between polar and hydrophobic residues throughout a protein structure.

fine-tunes ESM2 embeddings (a SOTA pLLM) [25] on common training and test sets. For a fair evaluation and correction for data leakage, we fine-tuned Prostata models on the same training and test sets as Stability Oracle. We call our version of this sequence model Prostata-IFML.

Stability Oracle outperforms Prostata and Prostata-IFML on nearly every metric (see Figure 5 (d)), even though Stability Oracle has 548 times fewer parameters ($\sim$658.M vs $\sim$1.2M) and was pretrained with 2,000 times fewer proteins ($\sim$46M vs $\sim$23K) at the same sequence similarity (50%). Interestingly, Prostata-IFML achieves similar results for identifying stabilizing mutations: with a stability prediction thresholds of $\Delta\Delta G$ < -0.5 kcal/mol on T2837+TP, PROSTATA filtered 1798 mutations, with an experimental distribution of 73.7% stabilizing, 17.9% neutral and 8.4% destabilizing and correctly identified 48.4% of the stabilizing mutations. Overall, for finding stabilizing mutations, Stability Oracle slightly outperformed Prostata-IFML with a more efficient pre-trained backbone and fine-tuned regression head and Prostata-IFML is the state-of-the-art at $\Delta\Delta G$ prediction when an accurate structure is unavailable.

We also compare against the RaSP framework [68]: a structure-based CNN model that follows a similar training pipeline of self-supervised pretraining with voxelized masked microenvironments followed by fine-tuning on DMS $\Delta\Delta G$ datasets. Briefly, RaSP is pretrained with 2315 structures clustered at a 30% sequence similarity and then fine-tuned on 10 protein DMS datasets computationally generated by the Rosetta Cartesian $\Delta\Delta G$ program. Similar to Stability Oracle, RaSP uses a masked microenvironment paradigm to avoid the need to generate mutant structures. In contrast to Stability Oracle, RaSP represents mutations via one-hot encodings of the "from" and "to" amino acids. Additionally, the RaSP regression head does not make use of a Siamese architecture to explicitly model the state function property of $\Delta\Delta G$. Instead, it concatenates all inputs and decodes them directly into $\Delta\Delta G$ values using an MLP block. To compare with Stability Oracle, we used the RaSP Colab notebook to generate DMS predictions on every protein in T2837.

Stability Oracle significantly outperforms RaSP on both regression and classification tasks. With a stability prediction thresholds of $\Delta\Delta G$ < -0.5 kcal/mol on T2837, RaSP filtered 96 mutations with an experimental distribution of 30.1%

stabilizing, 33.3% neutral, and 36.6% destabilizing compared to Stability Oracle, which filtered 72 mutations with an experimental distribution of 51.4% stabilizing, 25% neutral, and 23.6% destabilizing. Out of the 274 stabilizing mutations in T2837, RaSP and Stability Oracle correctly identified 28 and 37 of them, respectively.

3 Discussion

Reliable prediction of stabilizing mutations is critical for the acceleration of protein-based biotechnologies. However, to date, all computational stability predictors prioritize improvements in the Pearson and RMSE metrics. Neither of these metrics, however, is appropriate for evaluating improvements in identifying stabilizing mutations; several studies explain this in great detail [31, 14, 13]. We report all of the above metrics but focus on improvements in AUROC, MCC, recall, precision, and Spearman to tune our parameters. By using a 30% sequence similarity threshold for training-test splits, we expect the performance of Stability Oracle and Prostata-IFML to have superior generalization performance compared to models trained using traditional training-test splits. Stability Oracle's ability to correctly identify stabilizing mutations (precision 0.51 and 0.72 on T2837 and T2837 + TP, respectively) matches and outperforms FEP-based methods while being several orders of magnitude faster.

The literature has recently adopted Thermodynamic Reversibility (TR) to address the imbalance between stabilizing and destabilizing mutations in the training and test sets. However, this biases stabilizing mutations to wildtype amino acids. For computational models that leverage evolutionary features and protein structures as input, mutations to wildtype amino acids leak information and are of limited use in a protein engineering context where mutating away from wildtype is the primary goal. Our data augmentation technique, Thermodynamic Permutations (TP), generates mutations with a balanced $\Delta\Delta G$ distribution and does not generate mutations to the wildtype residue (see Figure 3 (b)). This has several benefits. First, TP augmentation mitigates the above imbalance and expands the number of stabilizing mutations in the test set to non-wildtype amino acids. Secondly, TP augmented data mitigates the bias to wildtype inherent in self-supervised pre-training methods. During end-to-end fine-tuning, TP forces the feature extractor to find additional chemical signatures in the microenvironment, not just the ones most relevant for wildtype identification.

Stability Oracle is able to predict stabilizing mutations on both the surface of the protein and also in the core. This is something structure-based self-supervised deep learning models such as MutComputeX typically struggle with, as masked residue prediction is much easier in the core compared to the surface. This added capability has direct applications to biological developments since mutations on the surface can disrupt known epitopes on viral antigens.

Acknowledgments

This work was sponsored by grants from the Defense Threat Reduction Agency (HDTRA1201001) and the Welch Foundation (F-1654) and the NSF AI Institute for Foundations of Machine Learning (IFML). We would like to thank AMD for the donation of critical hardware and support resources from its HPC Fund.

References

[1] Shuke Wu, Radka Snajdrova, Jeffrey C. Moore, Kai Baldenius, and Uwe T. Bornscheuer. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angewandte Chemie - International Edition*, 60(1):88–119, 2021.

[2] Elizabeth L. Bell, William Finnigan, Scott P. France, Anthony P. Green, Martin A. Hayes, Lorna J. Hepworth, Sarah L. Lovelock, Haruka Niikura, Sílvia Osuna, Elvira Romero, Katherine S. Ryan, Nicholas J. Turner, and Sabine L. Flitsch. Biocatalysis. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

[3] Joseph P. Adams, Murray J.B. Brown, Alba Diaz-Rodriguez, Richard C. Lloyd, and Gheorghe Doru Roiban. Biocatalysis: A Pharma Perspective. *Advanced Synthesis and Catalysis*, 361(11):2421–2432, 2019.

[4] Steven M. Jay and Richard T. Lee. Protein engineering for cardiovascular therapeutics: Untapped potential for cardiac repair. *Circulation Research*, 113(7):933–943, 2013.

[5] Michaela Gebauer and Arne Skerra. Engineered protein scaffolds as next-generation therapeutics. *Annual Review of Pharmacology and Toxicology*, 60:391–415, 2020.

[6] Gautam Kumar Meghwanshi, Navpreet Kaur, Swati Verma, Narendra Kumar Dabi, Abhishek Vashishtha, P. D. Charan, Praveen Purohit, H. S. Bhandari, N. Bhojak, and Rajender Kumar. Enzymes for pharmaceutical and therapeutic applications. *Biotechnology and Applied Biochemistry*, 67(4):586–601, 2020.

[7] Scott Banta, Kevin Dooley, and Oren Shur. Replacing antibodies: Engineering new binding proteins. *Annual Review of Biomedical Engineering*, 15:93–113, 2013.

[8] Keqin Chen and Frances H Arnold. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin e for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences*, 90(12):5618–5622, 1993.

[9] Jessica S. Ebo, Nicolas Guthertz, Sheena E. Radford, and David J. Brockwell. Using protein engineering to understand and modulate aggregation. *Current Opinion in Structural Biology*, 60:157–166, 2020.

[10] Frances H Arnold. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

[11] Lori Giver, Anne Gershenson, Per-Ola Freskgard, and Frances H Arnold. Directed evolution of a thermostable esterase. *Proceedings of the National Academy of Sciences*, 95(22):12809–12813, 1998.

[12] Corrado Pancotti, Silvia Benevenuta, Giovanni Birolo, Virginia Alberini, Valeria Repetto, Tiziana Sanavia, Emidio Capriotti, and Piero Fariselli. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Briefings in Bioinformatics*, 23(2):bbab555, 2022.

[13] Silvia Benevenuta, Giovanni Birolo, Tiziana Sanavia, Emidio Capriotti, and Piero Fariselli. Challenges in predicting stabilizing variations: An exploration. *Frontiers in Molecular Biosciences*, 9(January):1–10, 2023.

[14] Fabrizio Pucci, Martin Schwersensky, and Marianne Rooman. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Current opinion in structural biology*, 72:161–168, 2022.

[15] Tiziana Sanavia, Giovanni Birolo, Ludovica Montanucci, Paola Turina, Emidio Capriotti, and Piero Fariselli. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Computational and Structural Biotechnology Journal*, 18:1968–1979, 2020.

[16] Octav Caldararu, Rukmankesh Mehra, Tom L. Blundell, and Kasper P. Kepp. Systematic investigation of the data set dependency of protein stability predictors. *Journal of Chemical Information and Modeling*, 60(10):4772–4784, 2020.

[17] Nikita Fedik, Roman Zubatyuk, Maksim Kulichenko, Nicholas Lubbers, Justin S. Smith, Benjamin Nebgen, Richard Messerly, Ying Wai Li, Alexander I. Boldyrev, Kipton Barros, Olexandr Isayev, and Sergei Tretiak. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nature Reviews Chemistry*, 6(9):653–672, 2022.

[18] Andreas Bender, Nadine Schneider, Marwin Segler, W. Patrick Walters, Ola Engkvist, and Tiago Rodrigues. Evaluation guidelines for machine learning tools in the chemical sciences. *Nature Reviews Chemistry*, 6(6):428–442, 2022.

[19] Daniel J. Diaz, Anastasiya V. Kulikova, Andrew D. Ellington, and Claus O. Wilke. Using machine learning to predict the effects and consequences of mutations in proteins. *Current Opinion in Structural Biology*, 78:102518, 2023.

[20] Yuting Xu, Deeptak Verma, Robert P. Sheridan, Andy Liaw, Junshui Ma, Nicholas M. Marshall, John McIntosh, Edward C. Sherer, Vladimir Svetnik, and Jennifer M. Johnston. Deep Dive into Machine Learning Models for Protein Engineering. *Journal of Chemical Information and Modeling*, 60(6):2773–2790, 2020.

[21] Nicola Bordin, Christian Dallago, Michael Heinzinger, Stephanie Kim, Maria Littmann, Clemens Rauer, Martin Steinegger, Burkhard Rost, and Christine Orengo. Novel machine learning approaches revolutionize protein knowledge. *Trends in Biochemical Sciences*, 48(4):345–359, 2023.

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[23] Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*, 2022.

[24] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[25] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[26] Dmitriy Umerenkov, Tatiana I Shashkova, Pavel V Strashnov, Fedor Nikolaev, Maria Sindeeva, Nikita V Ivanisenko, and Olga L Kardymon. Prostata: Protein stability assessment using transformers. *bioRxiv*, pages 2022–12, 2022.

[27] S Benevenuta, C Pancotti, P Fariselli, G Birolo, and T Sanavia. An antisymmetric neural network to predict free energy changes in protein variants. *Journal of Physics D: Applied Physics*, 54(24):245403, 2021.

[28] Bian Li, Yucheng T Yang, John A Capra, and Mark B Gerstein. Predicting changes in protein thermodynamic stability upon point mutation with deep 3d convolutional neural networks. *PLoS computational biology*, 16(11):e1008291, 2020.

[29] H. Cao, J. Wang, L. He, Y. Qi, and J. Z. Zhang. Deepddg: Predicting the stability change of protein point mutations using neural networks. *J. Chem. Inf. Model.*, 59:1508–1514, 2019.

[30] Shuyu Wang, Hongzhou Tang, Peng Shan, and Lei Zuo. Pros-gnn: Predicting effects of mutations on protein stability using graph neural networks. *bioRxiv*, pages 2021–10, 2021.

[31] Aron Broom, Kyle Trainor, Zachary Jacobi, and Elizabeth M. Meiering. Computational Modeling of Protein Stability: Quantitative Analysis Reveals Solutions to Pervasive Problems. *Structure*, 28(6):717–726.e3, 2020.

[32] Elizabeth H Kellogg, Andrew Leaver-Fay, and David Baker. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*, 79(3):830–838, 2011.

[33] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

[34] Emidio Capriotti, Piero Fariselli, and Rita Casadio. I-mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research*, 33(suppl_2):W306–W310, 2005.

[35] Douglas EV Pires, David B Ascher, and Tom L Blundell. mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342, 2014.

[36] Douglas EV Pires, David B Ascher, and Tom L Blundell. Duet: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research*, 42(W1):W314–W319, 2014.

[37] Josef Laimer, Heidi Hofer, Marko Fritz, Stefan Wegenkittl, and Peter Lackner. Maestro-multi agent stability prediction upon point mutations. *BMC bioinformatics*, 16(1):1–13, 2015.

[38] Yves Dehouck, Jean Marc Kwasigroch, Dimitri Gilis, and Marianne Rooman. Popmusic 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC bioinformatics*, 12(1):1–12, 2011.

[39] Catherine L Worth, Robert Preissner, and Tom L Blundell. Sdm—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research*, 39(suppl_2):W215–W222, 2011.

[40] Gilad Wainreb, Lior Wolf, Haim Ashkenazy, Yves Dehouck, and Nir Ben-Tal. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics*, 27(23):3286–3292, 2011.

[41] Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proc. Natl. Acad. Sci. USA*, 110:E193–E201, 2013.

[42] Jianlin Cheng, Arlo Randall, and Pierre Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1125–1132, 2006.

[43] Peter Atkins, Peter William Atkins, and Julio de Paula. *Atkins' physical chemistry*. Oxford university press, 2014.

[44] Martin Steinegger and Johannes Soding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[45] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J Weinstein, Niall M Mangan, Sergey Ovchinnikov, and Gabriel J Rocklin. Mega-scale experimental analysis of protein folding stability in biology and protein design. *bioRxiv*, pages 2022–12, 2022.

[46] Raghav Shroff, Austin W Cole, Daniel J Diaz, Barrett R Morrow, Isaac Donnell, Ankur Annapareddy, Jimmy Gollihar, Andrew D Ellington, and Ross Thyer. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS synthetic biology*, 9(11):2927–2935, 2020.

[47] Inyup Paik, Phuoc HT Ngo, Raghav Shroff, Daniel J Diaz, Andre C Maranhao, David JF Walker, Sanchita Bhadra, and Andrew D Ellington. Improved bst dna polymerase variants derived via a machine learning approach. *Biochemistry*, 2021.

[48] Simon d'Oelsnitz, Daniel J Diaz, Daniel J Acosta, Mason W Schechter, Matthew B Minus, James R Howard, Hannah Do, James Loy, Hal Alper, and Andrew D Ellington. Synthetic microbial sensing and biosynthesis of amaryllidaceae alkaloids. *bioRxiv*, pages 2023–04, 2023.

[49] Anastasiya V. Kulikova, Daniel J. Diaz, James M. Loy, , Andrew D. Ellington, and Claus O. Wilke. Learning the local landscape of protein structures with convolutional neural networks. *J. Biol. Phys.*, 47:435–454, 2021.

[50] Fabrizio Pucci, Katrien V. Bernaerts, Jean Marc Kwasigroch, and Marianne Rooman. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, 34(21):3659–3665, 2018.

[51] Gilad Wainreb, Lior Wolf, Haim Ashkenazy, Yves Dehouck, and Nir Ben-Tal. Protein stability: A single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics*, 27(23):3286–3292, 2011.

[52] Jan Stourac, Juraj Dubrava, Milos Musil, Jana Horackova, Jiri Damborsky, Stanislav Mazurenko, and David Bednar. Fireprotdb: database of manually curated protein stability data. *Nucleic acids research*, 49(D1):D319–D324, 2021.

[53] Milot Mirdita, Konstantin Schutze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

[54] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[55] Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature methods*, 19(9):1109–1115, 2022.

[56] Guido Scarabelli, Eliud O. Oloo, Johannes K.X. Maier, and Agustina Rodriguez-Granillo. Accurate prediction of protein thermodynamic stability changes upon residue mutation using free energy perturbation. *Journal of Molecular Biology*, 434(2):167375, 2022.

[57] Emidio Capriotti, Piero Fariselli, Ivan Rossi, and Rita Casadio. A three-state prediction of single point mutations on protein stability changes. *BMC bioinformatics*, 9(2):1–9, 2008.

[58] K Abdulla Bava, M Michael Gromiha, Hatsuho Uedaira, Koji Kitajima, and Akinori Sarai. Protherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic acids research*, 32(suppl_1):D120–D121, 2004.

[59] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019.

[60] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[61] Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

[62] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

[63] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[64] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

[65] Pascal M Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora Marks. Trancepteve: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *bioRxiv*, pages 2022–12, 2022.

[66] Erik Nijkamp, Jeffrey Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022.

[67] Anastasiya V. Kulikova, Daniel J. Diaz, Tianlong Chen, T. Jeffrey Cole, Andrew D. Ellington, and Claus O. Wilke. Sequence and structure based deep learning models represent different aspects of protein biochemistry. *bioRxiv*, 2023.

[68] Lasse M. Blaabjerg, Maher M. Kassem, Lydia L. Good, Nicolas Jonsson, Matteo Cagiada, Kristoffer E. Johansson, Wouter Boomsma, Amelie Stein, and Kresten Lindorff-Larsen. Rapid protein stability prediction using deep learning representations. *bioRxiv*, page 2022.07.14.500157, 2022.

[69] Tianlong Chen, Chengyue Gong, Daniel Jesus Diaz, Xuxi Chen, Jordan Tyler Wells, Zhangyang Wang, Andrew Ellington, Alex Dimakis, Adam Klivans, et al. Hotprotein: A novel framework for protein thermostability prediction and editing. In *NeurIPS 2022 AI for Science: Progress and Promises*.

[70] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 2023.

[71] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

A Experiment Details

A.1 Experiment Settings

Pretraining Dataset We pretrain our graph-transformer backbone on proteins with structure files. We generate several different datasets for training, and all datasets are generated from protein crystal structures with up to a 3A resolution and protein sequences with up to 50% sequence similarity, providing us with 23,539 protein sequences to sample. In the following part, we give details descriptions of two different types of datasets, *Rand100* and *Interface*.

Rand100 microenvironment dataset followed the sampling methodology previous established by [46]. Briefly, given one target protein, we randomly sampled 100 residues or half of the number of overall residues, whichever number was smaller. We then did a 90/10 split on the data and extracted interface residues (within 5Åsimilar to below) to generate interface test sets. We will describe the test data details below.

To generate the *Interface* micro-environment dataset, for each protein sequence, we sampled all residues at the interface (within 5Å) of ligands with at least 3 carbon atoms, nucleotides, halogens, and cations and then randomly sampled to backfill up to 200 residues or half the length of the protein sequence. The intent was to ensure that we sampled the limited amount of protein:non-interface data in the PDB while not oversampling large proteins. Finally, we did a 90/10 split for each interface sampling and random sampling and combined them to create our training and test dataset.

Training Set We clean up a new dataset cDNA120K [45], which contains 144K labeled free energy data for small proteins. We first train on this dataset and fine-tune it on C2878.

Test Set To fully test different models, we collect and clean related datasets in the literature and make our new version test set. To get more comparisons, we report the numbers on all the literature datasets here. In summary, we test our model on S-sym, P53, Myoglobin, FireProtDB and S669. S-sym, P53, and Myoglobin datasets are proposed in ThermoNet [28] and are widely used by the follow-up works as the testing benchmarks. S669 includes curated data dissimilar at 25% sequence identity to S2648 training data. FireProtDB [52], which is a superset of published experiments records and is the most up-to-date dataset to our knowledge. We download the latest version of FireProtDB, clean the duplicates, and extract the $\Delta\Delta G$ values. Additionally, we create several more domains from the original test data. ① Swap the mutation and wild-type amino acids, and this new test domain is marked as *inverse*. ② Once we have multiple single-point mutations in the same position, we randomly pick two amino acids as the wild-type and the mutation. We mark this test domain as TP.

Baselines In literature, researchers have developed stability change predictors with sequence or structure models. ❶ PROSTATA [26] uses the pretrained ESM-2 model to transform two protein sequences, and predict the free energy change. ❷ Given the local environment for the wild-type and mutation, ThermoNet [28] and ProsGNN [30] project the local environment into latent features with 3D neural networks (3D CNNs, GraphNNs), subtract the two latent features, and predict the free energy change. ❸ When there is not enough labeled data, a pretrained language model or mask predictor (e.g. ESM models [59]) can serve as a zero-shot free energy change predictor. These works release the current best results on the free energy prediction benchmark, and thus we set up them as our baselines. ❹ Compared to the baseline models, we summarize our advantages. *Quality*: We achieve better classification and regression performance. *Efficiency*: The number of parameters is much fewer than in sequence models. Compare to ThermoNet and ProsGNN, our model only requires one local environment, and gets rid of folding the mutation local environment.

Evaluation Metrics We evaluate regression and classification metrics. First, for a fair comparison, we report Pearson correlation and RMSE for our regressor. Furthermore, in practice, instead of RMSE, we have more interest in the performance on stabilizing mutations. Therefore, we report the precision and recall for the stabilizing mutations and the AUROC values for the binary classification problems.

Architecture Mark the wild-type and the mutation as a_w and a_m, we get their embedding $e(a_w) \in \mathbb{R}^{[1\times H]}$ and $e(a_m) \in \mathbb{R}^{[1\times H]}$ in a lookup table. Mask the masked local environment as G, we pass the masked local environment graph G through our pretrained graph-transformer, and then have $h \in \mathbb{R}^{[N\times H]}$, where N and H represents the number of tokens and the hidden dimension. Concatenate h and one amino acid, we obtain the representation for a local environment with amino acid, *from* $=$ Concat(a_w, h), *to* $=$ Concat(a_m, h). Pass *to* and *from* through attention blocks, subtract the difference, and process the difference with a dense layer, we come to the output free energy change prediction.

Training Configuration During training, we shuffle the training data and use Huber loss with $\delta = 1$ as the training loss. When training on large-scale datasets, *e.g.*, cDNA120K, we end-to-end tune all the parameters, with learning rate 5×10^{-5} for regression head parameters and learning rate 2×10^{-5} for backbone parameters. We apply optimizer AdamW with batch size 960, weight decay 0.1, optimizer EMA with $\eta = 0.99$, number of iterations 750. To finetune the CDNA-trained model on the small datasets (*e.g.*, C2878), we freeze the backbone parameters and use AdamW with learning rate 5×10^{-7}, batch size $1,024$, weight decay 0.1, number of iterations 500. Train from scratch on the small datasets (*e.g.*, Q1744), we freeze the backbone parameters, and set optimizer AdamW with 5×10^{-5}, batch size $1,024$, weight decay 0.1, and number of iterations 500.

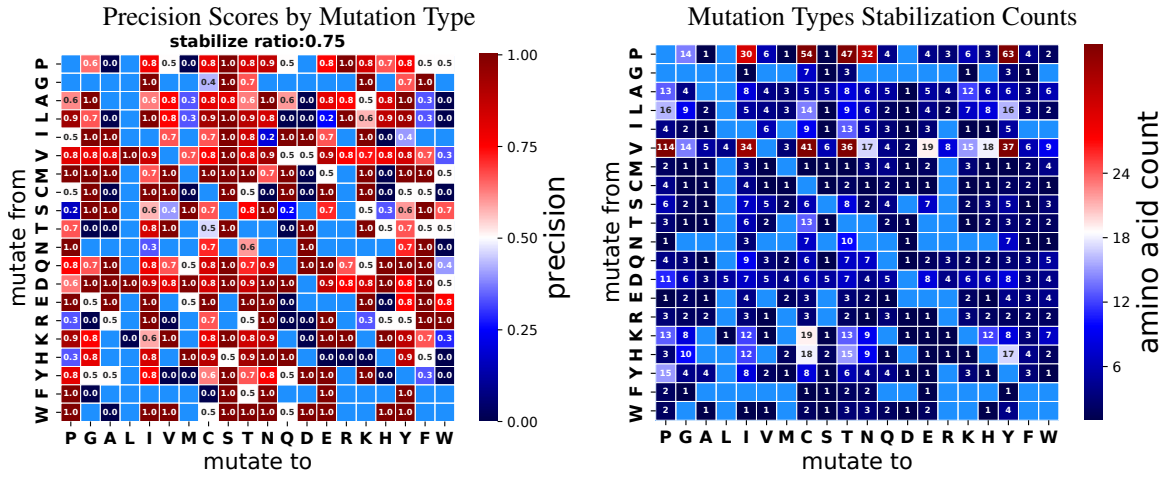

Figure 8: Stability Oracle precision on the T2837 + TP test set for the 380 mutation types pairs.

Evaluating our structural amino acid embeddings We compare the one-hot encodings vs our structural amino acid embeddings to represent a mutation.

Representation	Test Set	*Pearson*	*Spearman*	*RMSE*	*AUROC*	Precision	Recall	Accuracy
Continuous Embedding	T2837	0.59	0.62	1.64	0.81	0.47	0.53	0.82
One-Hot	T2837	0.48	0.44	1.90	0.68	0.33	0.45	0.66

Table 1: We present the comparison of two different protein sequence representations, Continuous Embedding and One-Hot, on the T2837 test set. The table shows the average values of Pearson correlation coefficient, Spearman correlation coefficient, Root Mean Squared Error (RMSE), Area Under the Receiver Operating Characteristic Curve (AUROC), Precision, Recall, and Accuracy for three trials. Models were trained on cDNA120K dataset.

Results on Different Test Sets Stability Oracle performance on different test sets are reported in Table 2, 3, & 4 .

Test Set	*Pearson*	*Spearman*	*RMSE*	*AUROC*	Precision	Recall	Accuracy
S-sym	0.68	0.68	1.22	0.87	0.69	0.52	0.83
S-sym Reverse	0.68	0.68	1.19	0.87	0.82	0.95	0.80
Myoglobin	0.62	0.62	0.89	0.81	0.56	0.36	0.71
Myoglobin Reverse	0.64	0.64	0.89	0.81	0.77	0.94	0.72
FireProtDB	0.58	0.62	1.48	0.79	0.61	0.41	0.77
FireProtDB Reverse	0.58	0.62	1.46	0.79	0.78	0.94	0.76
S669	0.52	0.53	1.43	0.75	0.58	0.37	0.75
S669 Reverse	0.52	0.53	1.43	0.75	0.80	0.92	0.75
P53	0.70	0.62	1.50	0.77	0.45	0.91	0.67
P53 Reverse	0.70	0.60	1.48	0.77	0.75	0.95	0.74

Table 2: We presents the performance of our ensemble of three models on various test sets. We report the Pearson and Spearman correlation coefficients, root mean square error (RMSE), area under the receiver operating characteristic (AUROC) curve, Precision, Recall, and Accuracy. The test sets are categorized based on the literature.

Description	Test Set	# Mutations	*Pearson*	*Spearman*	*RMSE*	*MCC*	*AUROC*	Precision	Recall	Accuracy
All Data	S669 Orig	667	0.5156	0.5314	1.4388	0.3354	0.7469	0.5849	0.3713	0.7763
All Data	S669 TR	667	0.5200	0.5309	1.4305	0.3245	0.7509	0.8039	0.9235	0.7748
All Data	S669 Orig + TR	1334	0.6662	0.6859	0.7755	0.5512	0.8415	0.7696	0.7846	0.7755
All Data	S669 TP	658	0.4675	0.4437	1.0069	0.2956	0.7060	0.6366	0.6718	0.6474
All Data	S669 Orig + TP	1325	0.5770	0.5806	1.3105	0.4023	0.7725	0.6517	0.5179	0.7440
All Data	S669 Orig + TP + TR	1992	0.6403	0.6405	1.3104	0.4852	0.8107	0.7339	0.7568	0.7425

Table 3: Stability Oracle's regression and classification metrics on S669 and its augmented datasets. TR: Thermodynamic Reversibility augmentation, TP: Thermodynamic Permutation augmentation. Note: 2 mutations failed during data processing.

Description	Test Set	#Mutations	*Pearson*	*Spearman*	*RMSE*	*MCC*	*AUROC*	Precision	Recall	Accuracy
All Data	T2837 Orig	2830	0.5874	0.6216	1.6549	0.3897	0.8088	0.5381	0.4677	0.8159
All Data	T2837 TR	2830	0.5904	0.6226	1.6464	0.3905	0.8138	0.8619	0.9186	0.8159
All Data	T2837 Orig + TR	5660	0.7426	0.7808	1.6527	0.6345	0.8951	0.8074	0.8290	0.8171
All Data	T2837 TP	3849	0.6556	0.6500	1.4359	0.4742	0.8107	0.7287	0.7410	0.7371
All Data	T2837 Orig + TP	10528	0.6829	0.6839	1.5300	0.5042	0.8356	0.6903	0.6787	0.7704
All Data	T2837 Orig + TP + TR	13358	0.7034	0.7156	1.5313	0.5399	0.8486	0.7597	0.7830	0.7698
Experimental $\Delta\Delta G < -0.5$	T2837 Orig	277	0.2184	0.3428	2.1292	-	-	1.0000	0.5109	0.5109
Experimental $\Delta\Delta G < -0.5$	T2837 TR	1786	0.6032	0.6027	1.8053	-	-	1.0000	0.9345	0.9345
Experimental $\Delta\Delta G < -0.5$	T2837 TP	1236	0.5757	0.6013	1.6146	-	-	1.0000	0.8398	0.8398
Experimental $\Delta\Delta G < -0.5$	T2837 Orig + TP	2720	0.5200	0.5530	1.8079	-	-	1.0000	0.4814	0.4814

Table 4: Stability Oracle's regression and classification metrics on T2837 and its augmented datasets. TR: Thermodynamic Reversibility augmentation, TP: Thermodynamic Permutation augmentation. Note: we cannot calculate MCC and AUROC for the experimental stabile dataset since there is only a single label.

B Data and Metric Issues

Systematic analysis of computational stability predictors published over the last 15 years has identified several important shortcomings hindering the field from taking advantage of deep learning algorithms. These shortcomings are present in every facet of the data pipeline: data scarcity, data bias, data leakage, poor metrics for evaluating performance, and computational cost. Here is a high-level, non-exhaustive overview of the most pressing issues. The limited amount of experimental data is heavily biased in several ways: 1) dominated by destabilizing mutations, resulting in only $\sim$20% of point mutations predicted to stabilize indeed increase stability [14, 31, 13]; 2) stabilizing mutations are narrowly concentrated between -2 to 0 kcal/mol, causing models to overfit to this distribution and underestimate and not generalize to strongly stabilizing mutations [14, 31, 13]; 3) stabilizing mutations are enriched with surface hydrophobic mutations, causing stabilizing predictions to often unintentionally lower protein solubility [14, 31, 13] 4) of the 380 types of amino acid substitutions $\sim$20-70% have no data in the common training and validation sets and the ones sampled are significantly biased toward mutations "from" or "to" alanine and valine [31]. Due to the hierarchical nature of proteins, splitting experimental data into training and validation sets can be done at the mutation, position, protein, and sequence cluster levels. All but sequence cluster result in data leakage, thus, most published computational tools report inflated evaluation metrics and display poor generalization.

The common metrics used to assess the performance of computational stability predictors (Pearson correlation, classification accuracy, and error) are not appropriate for the goal of accurately identifying stabilizing mutations and tend to be misleading [14, 31, 13]. The metrics naively formulate the problem as a regression or classification task and are not tailored to the actual goal: accurately predicting stabilizing mutations. This misalignment has resulted in metrics that do not prioritize improvement in accurately predicting stabilizing mutations and their game-ability has been demonstrated in a study that consisted of 21 established computational tools [31]. Here, the authors demonstrate how Spearman and Matthew correlation can correctly identify deficiencies in an overly simplified model while the original metrics couldn't. We argue that these metrics are still insufficient, and metrics specifically tailored toward accurately identifying strong stabilizing mutations are still required.

C Overview of Deep Learning for Stability Prediction

The lack of available stability data has prevented deep learning frameworks from significantly accelerating the field and the community still primarily relies on knowledge-based methods, such as Rosetta [32] and FoldX [33], or shallow machine learning approaches [34, 35, 36, 37, 38, 39, 40, 41, 42]. Nonetheless, several deep learning frameworks for stability predictions have been published and they can be broadly categorized by their input type: sequence or

structure. The current structure frameworks include DeepDDG [29], ThermoNet [28], ProSGNN [30] , and AC-DNN [27] and sequence frameworks include Prostata [26] and HotProteins [69]. While both structure and sequence frameworks are competitive with current state-of-the-art methods, the two types have their own benefits and limitations. The structure frameworks capture atomic interactions within a protein and with other biomolecules, such as ligands, nucleotides, and other proteins. Furthermore, the richness of atomic representations enables generalization while training with only 10^3-10^4 protein structures. However, all existing frameworks require an additional mutant structure to make a stability prediction. This adds a significant computational hurdle that scales linearly with the number of inferences. For example, to conduct a computational deep mutational scan (DMS) of a 300 amino acid protein, one must computationally generate 5700 mutant structures via AlphaFold2 or Rosetta(cite). For sequence-based frameworks their primary benefit is that they required little data engineering and have billions of primary sequences readily available for training. This gives them a low barrier to entry and makes them popular amongst the machine learning community, resulting in an explosion of transformer-based large language models (LLMs) across the protein community[59, 60, 25, 66, 26, 64, 61, 65, 70, 71]. There are two primary limitations to sequence models. First, LLMs require extensive training on large sequence databases (UniRef50 $\sim$46M) to learn meaningful representations for downstream applications. Training these models is very computational expensive and only affordable to well-funded academics and technology companies [59, 60, 25, 64, 65, 71]. Second, the self-supervised pre-training of transformer models constrains the training space to single protein sequences. As protein engineers, we intuitively understand that proteins are more than their primary sequence; proteins often form complexes, undergo post-translational modifications, and interact with cofactors, DNA, RNA, and oligosaccharides. This artificial domain shift limits their generalization in downstream task and has already been empirically demonstrated when predicting stability changes at protein-protein and protein-ligand interfaces [26].