Balancing Imbalance: Enriching Protein Language Model Predictions of Fitness

Zachary Aristei
zaristei3@gatech.edu

Rishi Banerjee
rbanerjee41@gatech.edu

Rahi Kotadia
rahikotadia@gmail.com

1 Abstract

Predicting protein fitness from imbalanced datasets is ongoing problem in the field of protein engineering. We adapted curated datasets from ProteinGym to determine the effect of imbalance reduction methods on protein fitness benchmarks. We attempt to improve existing models by applying Rank-N-Contrast methodology for addressing fitness imbalance using an ESM-2 language model. We find that while the training process has significant improvements in maintaining the top embeddings, the gains learned by the model are subject to the initial training conditions and provide insight to finetuning protein language models for protein fitness prediction tasks.

2 Introduction

Protein engineering is a well established field with many applications. It involves the creation and synthesis of novel proteins for application in fields like drug discovery. One of the key parts is being able to predict the function of a protein based on its structure, specifically the amino acid sequence. However, protein fitness datasets used for training are imbalanced due to there being significantly more unfit proteins stemming from mutations. Thus, addressing this imbalance may result in better fitness prediction performance.

In the context of protein engineering, fitness refers to an evaluative scalar metric corresponding to a real-world structural characteristic or functional result based on experimentation for a protein. For instance, protein fitness may relate to how effectively an enzyme is able to bind to a substrate, or the ratio of how much of a chemical reaction result is produced from an initial substrate.

In our work, we investigate multiple metrics of protein fitness based on various paper datasets along with evaluating embeddings generated from these datasets on regression tasks in order to determine the general adaptability of the embeddings we employ. This involves taking low rank projections of the embeddings using TSNE and analyzing the distribution of the unbalanced data. We then apply methods to address data imbalance such as Rank-N-Contrast. We evaluate the success of our model using metrics such as such as Spearman rank correlation and Normalized Discounted Cumulative Gain (NCDG).

3 Related Work

There are avenues of research being pursued to make a joint mapping between protein structure and function. One approach involves constructing a joint embedding space for untested proteins that utilizes multiple structural and functional aspects of known proteins utilizing a triplet loss function and and a Mixture-of-Experts (MoE) [2]. Hamasy et. al. have demonstrated that ensemble-style Mixture of Experts models for creating an embedding space have proven effective for information retrieval tasks.

Another method tries to predict the fitness of a protein to a specific function based on its DNA sequence using an LSTM approach [5]. Finally, several novel protein folding solutions such as AlphaFold utilize Reinforcement Learning to predict protein structure from amino-acid sequence [3],

which can later be mapped to specific functions using known protein structure-function mappings for training. However, none of these approaches focus on determining methods to compensate for the lopsided prior distribution in existing protein fitness datasets [10].

The development of robust LLMs has led to the emergence of some hybrid models that rely on existing protein featurization methods along with LLM next-token prediction to try better encoded latent patterns of fitness within protein sequences [6]. In order to avoid a general bias in favor of the training data for protein fitness prediction, exploratory AI methods such as reinforcement learning [12] and genetic algorithms [9] are also investigated for predicting fitness from sequences or tertiary structures.

Methods to address protein fitness imbalance are increasing, especially in recent years. Mardikoraem et al. compare protein fitness predictions from protein language models and traditional one hot encoding/phsyiochemical encoding by testing undersampling, random oversampling, and SMOTE to resolve the data imbalance [6].

Methods outside of the computational biology field can also potentially be adapted. Yang et al. outline a method to address deep regression imbalance, different from traditional methods that address categorical variable imbalance. They propose a feature distribution smoothing involving a symmetric kernel for feature smoothing that can be added as a layer on top of existing networks [15]. The Rank-N-Contrast [16] method is able to utilize contrastive learning in order to transform the representation such that it structurally maintains the order of the target variable within the representation space.

4 Data

Protein fitness is measured by different metrics from paper to paper. Mardikoraem et al. use two metrics: Binding Affinity and Stability. Stability is measured by melting temperature (Tm) and was evaluated on regression tasks and classification task, where < 35 °C is low stability and > 65 °C is high stability [6]. In this case, they adapt two datasets, Affibody and Novozymes Enzyme Stability Prediction (NESP) for the two metrics respectively.

However, a new dataset from Notin et al. called ProteinGym was developed specifically for protein fitness and lends itself well to the task. The data contains 250 deep mutational scanning assays (DMS) with 2.8 million mutations and curated clinical datasets with 66,000 mutants which covers 200 protein families. The data is separated by substitutions vs. indels, indicating an amino acid being swapped for another and insertions or deletions of an amino acid. The DMS data is given a score based on features like viral replication or thermostability based on the research they originally come from. The DMS scores are binarized using a method for cutoff, most often the median [7].

ProteinGym takes data from many papers and consolidates them, but the DMS Score value is not consistent over all of the data samples since it corresponds to a specific fitness measure from a specific paper.

Thus, we chose to use three specific datasets from ProteinGym.

- Chen et al. which measures thermostability in around 182,000 protein sequences [1].
- In particular, we used Tsuboyama et al. which measures thermodynamic folding stability for around 776,000 sequences [13].
- Olson et al. which measures fitness related to structural stability after mutations [8].

Figures 1 through 6 demonstrate the spread of the data and the issues with data distribution.

5 Methods

5.1 Methodology and Algorithm

When utilizing representations made from large-scale protein databases such as those made by protein language models, it is necessary to build models that are able to utilize these existing representations to analyze sequence-function relationships. There are currently universal protein

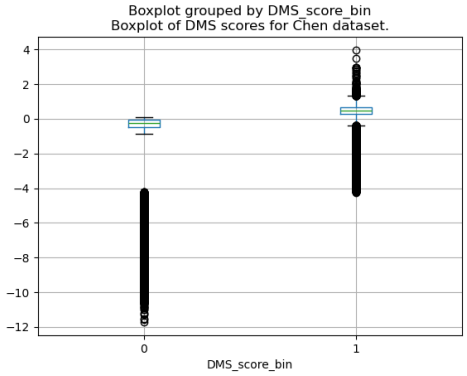

Figure 1: Chen Box Plot

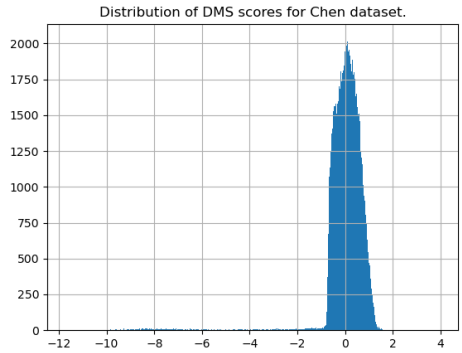

Figure 2: Chen Histogram

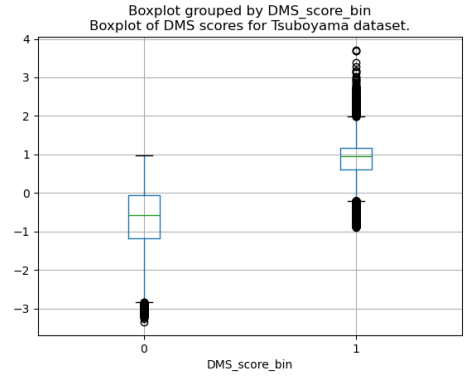

Figure 3: Tsuboyama Box Plot

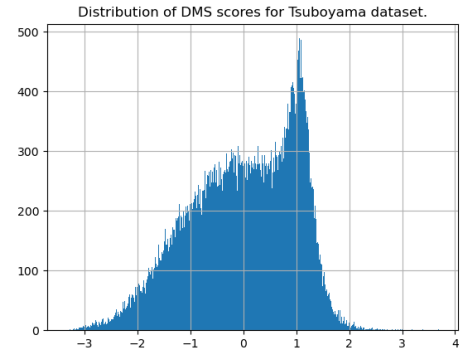

Figure 4: Tsuboyama Histogram

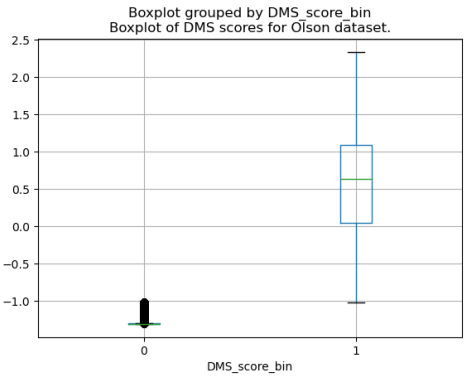

Figure 5: Olson Box Plot

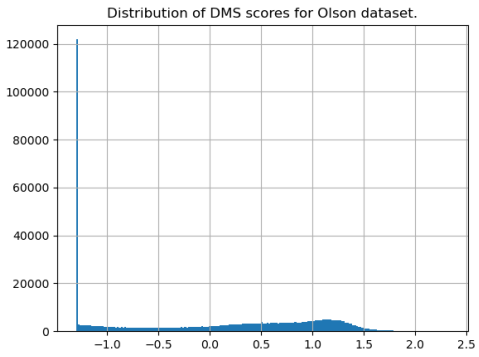

Figure 6: Olson Histogram

language models such as ESM [11] that have proven success on a variety of structural and functional tasks. However, these universal representations do not guarantee success in specific protein fitness tasks. One of the major reasons for this is that these representations fail to capture the continuous nature of the fitness data, such as the order of the data.

There are multiple methods that achieve the property of smoothing for a variety of imbalanced datasets [16, 15]. The Rank-N-Contrast [16] method is one that we apply that attempts to smooth the distribution space in order to obtain better prediction results on imbalanced data. Our algorithm

utilizes this Rank-N-Contrast loss function in order to finetune the representations of the data in order to generate a richer representation against the fitness prediction problem.

Taking the protein sequences from the datasets, we use the protein language model ESM-2 [14] to obtain feature embeddings for the sequences. The Protein Language Model that is used is ESM-2-8M, with the embedding being 320 dimensions. We finetune this model on the datasets and apply various metrics for dealing with the imbalanced data: Label Distribution Smoothing and Rank-N-Contrast. We then apply a linear regression model to generate the prediction based on the representation and the protein fitness statistic.

ProteinGym uses three metrics (among many) to evaluate their model: Spearman ρ coefficient, top-k recall, and NDCG. Top-k recall and NDCG are used to identify the proteins with the highest fitness values and are not only for modelling the distribution. In their paper, they select the k in top-k recall to be the top 10% of DMS scores and NDCG works by giving a higher weighting to models that correctly give the highest DMS scores [7].

6 Experiments

6.1 Baselines

- The main baseline that was used was to analyze the increase in performance between the functions was the linear regression between the PLM embeddings and the fitness scores for each of the problems. Due to the prohibitive time complexity of traditional upsampling methods for imbalanced regression such as SMOTER or SMOGN, these methods were not investigated due to the size of the datasets involved for Olson et al and Chen et al.
- Along with the base linear regression, we also investigated the baseline of weighted linear regression using label distribution smoothing, as advised in [15]. This is done in order to increase the correlation between the test-error distribution and the empirical label distribution to get a more accurate estimate of the label distribution, Doing this also allows you to build models, such as weighted linear regression, that take into account the sparsity of certain data fields in the forming of a line of best fit. The actual smoothing involves applying a symmetric kernel and reweighing the loss function by multiplying by the inverse of the new probability density calculated from the convolution.
- Lastly, we use the Rank-N-Contrast method as outlined in Zha et al [16]. This method involves applying a loss function which ranks and contrasts samples to those close it, forcing similar samples to be ranked highly. This loss function was applied to the datasets and compared to the two baselines.

6.2 Results

Through the process described in Method, the chosen metrics were calculated for the described datasets. The results are as follows in 1.

7 Conclusion and Discussion

7.1 Analysis

As seen in 1, the Rank-N-Contrast embeddings result in significant improvements in all of the metrics shown compared to the baselines. It seems that the Label Distribution Smoothed Linear model performed at par or worse than the dataset without any smoothing or weighting. One reason for this is that the sample weighting may make the function more sensitive to outliers rather than all of the datapoints, which may significantly affect the ranking of the datapoints at large. This is why there have been significant losses in the global Spearman ρ statistic but not for the Top-k Recall and NDCG statistics, which deal with only the ranking of the highest-valued points in the dataset.

Not only does the Rank-N-Contrast embedding show a significant improvement for the Chen and Tsuboyama test sets compared to the baselines for the global Spearman ρ statistic, there

Simple Linear Regression			
Dataset	Pearson Rho	Top K Recall	NDCG
Chen	.923	.668	.990
Olson	**.876**	**.511**	**.927**
Tsuboyama	.610	.281	.877
Label Distribution Smoothing			
Dataset	Pearson Rho	Top K Recall	NDCG
Chen	.894	.668	.990
Olson	.764	.511	.927
Tsuboyama	.601	.281	.878
Rank N Contrast			
Dataset	Pearson Rho	Top K Recall	NDCG
Chen	**.959**	**.754**	**.993**
Olson	.462	.231	.757
Tsuboyama	**.707**	**.346**	**.902**

Table 1: Fitness Measures across datasets for each method

were significant improvements in the Top-k Recall and NDCG values. This shows that the Rank-N-Contrast models is promising in finetuning embeddings that are better tuned towards choosing the most fit proteins in the dataset.

However, for the dataset from Olson et al. the linear model with the Rank-N-Contrast embeddings experienced a significant drop in the pearson rho-statistic. Figure 7 shows the difference in the residual distribution between the Rank-N-Contrast regressor and the original model.

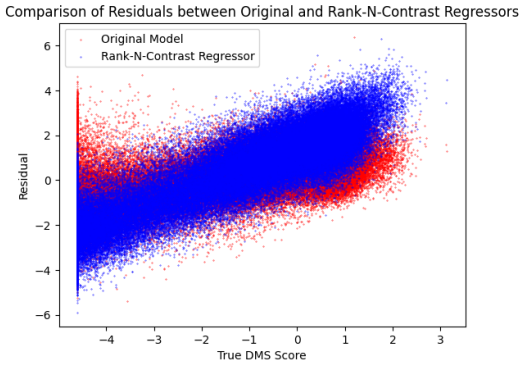

Figure 7: Residuals between Rank-N-Contrast and Original Regressor

From this, we have found that the underestimation between the higher value predictions and the true values is the most significant effect on the high-value rankings, which causes the most significant drops in the Top-k Recall and the NDCG. As a result, it warrants further investigation what distributional factors affect this performance degradation, as there are many hyperparameters during the Rank-N-Contrast training process that could be tailored to specific distributions.

7.2 Understanding Distributional Changes

In order to visualize the effect of the Rank-N-Contrast training process on the representations of sequences, we can use the t-SNE dimensionality reduction process to view the nature of the representation. Figures 8 to 13 show the changes in the significant dimensions of the representations. With regards to the DMS scores, the values of the representations become more correlated with the DMS-score after the Rank-N-Contrast fine-tuning process, which would lend itself to having a

regression model perform better. However, due to the inconsistencies with training, there are still sections of the plots that do not fit well in this paradigm.

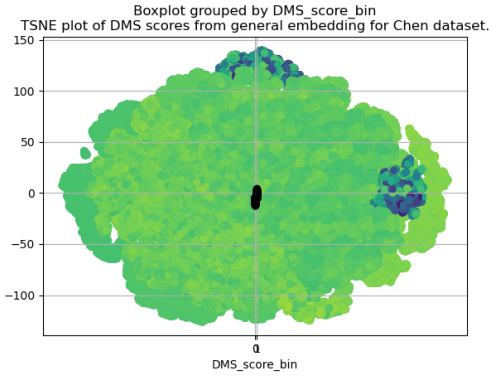

Figure 8: Chen ESM2 t-SNE (2 dim)

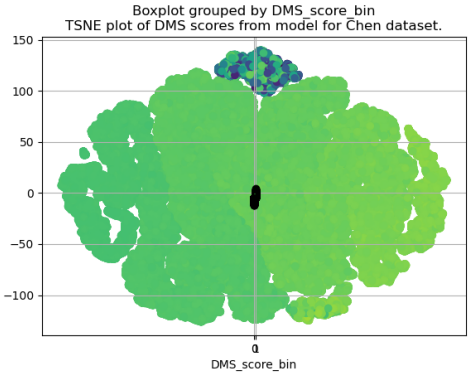

Figure 9: Chen Rank-N-Contrast t-SNE (2 dim)

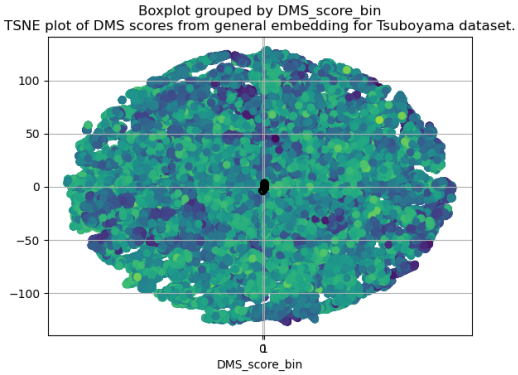

Figure 10: Tsuboyama ESM2 t-SNE (2 dim)

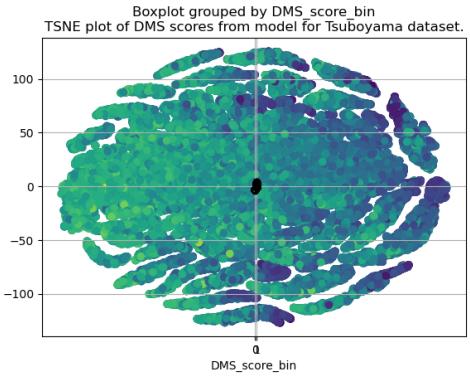

Figure 11: Tsuboyama Rank-N-Contrast t-SNE (2 dim)

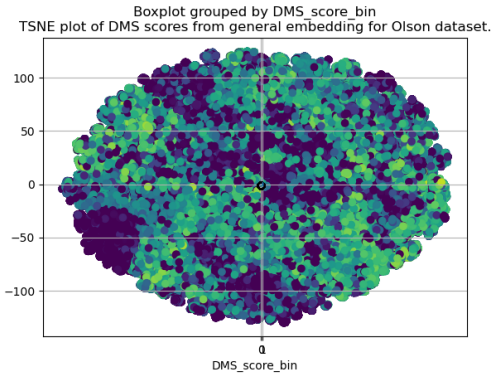

Figure 12: Olson ESM2 t-SNE (2 dim)

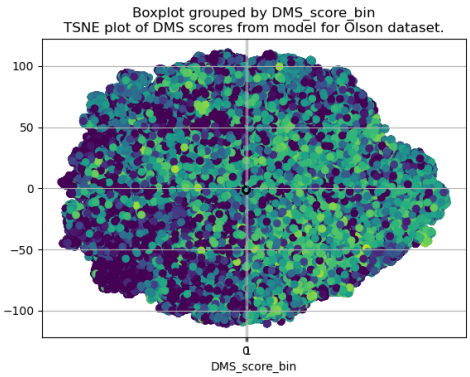

Figure 13: Tsuboyama Rank-N-Contrast t-SNE (2 dim)

7.3 Conclusion

As expected, the Rank-N Contrast methodology yielded better protein stability prediction results than simple linear regression and label distribution smoothing. However, our results indicate to us that the

feasiblity of these models is heavily contingent on the curation of the dataset. The Olson dataset in particular has a distribution that vastly skews the results to be poorer than the other two results, and indicates that not all forms and distributions for protein fitness will be correlated across experiments.

7.4 Limitations and Future Work

In addition to the methods that we performed above, there were a handful of methods that we wanted to try but did not have the time or compute resources yield results in time for. One such method involved resampling the dataset using a methodology known as the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) [4]. We also did not try the Feature Distribution Smoothing method from Yang et al., which was shown to be better than just Label Distribution Smoothing. We wanted to experiment with this technique because we noticed a significant imbalance in the stability distribution for our protein dataset. SMOGN combats this by combining random sampling with synthetic oversampling if which keeps samples if the K Nearest Neighbors of the sample falls within a threshold. When we ran this, we realized that the runtime complexity on the orders of our datasets was too big for it to be feasible within the context of this project. However, if we had more time, we would want to explore how much such as resampling paradigm can improve model performance.

It is also worth noting that differentiable sorting and ranking algorithms may be fairly sensitive to minor changes in model hyper-parameters, because a change in the ordering of the components may result in sudden changes and discontinuities in the calculated loss function. As a result, it may be worth considering different techniques from regression and gradient descent, such as reinforcement learning and evolutionary algorithms to be paired with Rank-N-Contrast learning in order to determine the next set of weights or parameters to evaluate. These methods tend to assume a nondifferentiable or nonconvex loss function, and so they may be better able to optimize this type of problem by better traversing the inherent tradeoff between exploration and exploitation.

8 Contributions

- Dataset Collection and Pre-processing - Zachary, Rahi
- Base PLM Implementation - Rishi
- Enriched Model design and implementation - Rahi, Zachary
- Model training and finetuning - Rishi, Zachary
- Results and Analysis - Rahi, Rishi

9 Source Code

The source code can be found at this Github page.

References

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

[2] Tymor Hamamsy, Meet Barot, James T. Morton, Martin Steinegger, Richard Bonneau, and Kyunghyun Cho. Learning sequence, structure, and function representations of proteins with language models. *bioRxiv*, 2023.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[4] Nicholas Kunz. SMOGN: Synthetic minority over-sampling technique for regression with gaussian noise, 2020.

[5] Yunan Luo, Guangde Jiang, Tianhao Yu, Yang Liu, Lam Vo, Hantian Ding, Yufeng Su, Wesley Wei Qian, Huimin Zhao, and Jian Peng. ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun.*, 12(1):5743, September 2021.

[6] Mehrsa Mardikoraem and Daniel Woldring. Protein fitness prediction is impacted by the interplay of language models, ensemble learning, and sampling methods. *Pharmaceutics*, 15(5), April 2023.

[7] Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023.

[8] C. Anders Olson, Nicholas C. Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, November 2014.

[9] Jan T Pedersen and John Moult. Genetic algorithms for protein structure prediction. *Current Opinion in Structural Biology*, 6(2):227–231, 1996.

[10] 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.

[11] 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. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[12] Haoran Sun, Liang He, Pan Deng, Guoqing Liu, Haiguang Liu, Chuan Cao, Fusong Ju, Lijun Wu, Tao Qin, and Tie-Yan Liu. Accelerating protein engineering with fitness landscape modeling and reinforcement learning. *bioRxiv*, pages 2023–11, 2023.

[13] 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 design. *Nature*, 620(7973):434–444, August 2023.

[14] Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, 2022.

[15] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression, 2021.

[16] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression, 2023.