Enhancing predictions of protein stability changes induced by single mutations using MSA-based Language Models

Francesca Cuturello *[1], Marco Celoria [1,2], Alessio Ansuini [1], and Alberto Cazzaniga [1]

[1]AREA Science Park, Località Padriciano 99, 34149, Trieste, Italy
[2]CINECA National Supercomputing Center, Casalecchio di Reno, 40033, Bologna, Italy

Abstract

Protein Language Models offer a new perspective for addressing challenges in structural biology, while relying solely on sequence information. Recent studies have investigated their effectiveness in forecasting shifts in thermodynamic stability caused by single amino acid mutations, a task known for its complexity due to the sparse availability of data, constrained by experimental limitations. To tackle the problem, we fine-tune various pre-trained models using a recently released mega-scale dataset. Our approach employs a stringent policy to reduce the widespread issue of overfitting, by removing sequences from the training set when they exhibit significant similarity with the test set. The MSA Transformer emerges as the most accurate among the models under investigation, given its capability to leverage co-evolution signals encoded in aligned homologous sequences. Moreover, the optimized MSA Transformer outperforms existing methods and exhibits enhanced generalization power, leading to a notable improvement in predicting changes in protein stability resulting from point mutations. The code and data associated with this study are available at https://github.com/marco-celoria/PLM4Muts.

1 Introduction

The emergence of Protein Language Models (PLMs) marks a substantial milestone in our ability to unlock the intricate language encrypted within amino acid sequences. Leveraging transformer-based natural language processing techniques [1, 2], PLMs are instrumental in large-scale pre-training, enabling the generation of representations that are tailored for transfer learning on specific tasks. This pre-training strategy enhances the model's ability to capture nuanced features and patterns inherent in protein sequences, thereby facilitating more effective transfer to downstream tasks such as protein structure prediction and function annotation [3–5]. Their versatility spans a wide range of applications, such as understanding protein fitness and evolutionary dynamics [6–9]. Inspired by the tight interconnection between evolutionary adaptability and structural stability, we explore the potential of these models in predicting thermodynamic stability changes induced by single amino acid mutations. The variation in stability from a wild-type protein to its mutated counterpart is determined by the difference in unfolding free energy between them. Assessing protein stability is crucial in molecular biology, given that proteins serve as fundamental components of living systems, executing diverse functions primarily dictated by their precise three-dimensional structure. Even subtle alterations in their sequence can perturb this structure, thereby impacting their function. Accurately predicting how mutations alter stability provides valuable insights into the interplay between the protein structure and function [10, 11], while enhancing our understanding of molecular evolution mechanisms [12, 13]. Moreover, in the context of enzyme engineering and drug development, the ability to regulate stability changes induced by mutations is essential for optimizing protein functionality to meet specific design requirements [14–16]. Considering that many pharmaceutical agents target proteins, understanding how non-synonymous variants influence stability facilitates the identification of promising drug targets and prediction of pharmacological treatment efficacy. Additionally, mutations in proteins can contribute to the onset of various diseases [17–19], including cancer [20, 21]. Investigating the effects of these mutations on protein structure stability unveils critical insights into disease mechanisms, potentially uncovering novel therapeutic targets. The exploration of mutations' impact on structural stability has traditionally been approached through classical methods, which rely on energy-based models or statistical potentials [22–25]. Machine learning models have also increasingly contributed, predominantly exploiting structure information [26–37], while only a few algorithms focus solely on protein sequences. [38, 39]. These techniques employ a variety of different models, ranging from Convolutional Neural Networks to Random Forest or Support Vector Machine regressions, often integrating graph-based representations of protein structures. However, the adoption of structure-based approaches can introduce a break in symmetry within the architectures, leading to a pronounced imbalance in prediction accuracy between direct and reverse mutations [40–42]. In this context, recent studies have highlighted the efficiency of transfer-learning from massively pre-trained protein language models [43–48]. Despite PLMs remarkable capabilities, the challenge of forecasting stability

*Corresponding author. francesca.cuturello@areascienceprk.it

effects from sequence-only information is compounded by the limited availability of experimental data [49, 50], given the resource-intensive nature of experiments. The scarcity of experimental data largely influences the robustness of machine learning models, which often encounter biases arising from overlaps between training and test sets [51–53]. Addressing these evaluation issues, a recent study [42] introduces a novel test dataset specifically designed to minimize similarity with sequences present in the commonly used training sets. The assessment of state-of-the-art models reported in that work sets a benchmark for future research, highlighting the substantial impact of shared homology on the generalization capability of the models. Notably, the study reveals that all compared approaches experience drastic drops in accuracy when measured on this newly introduced test set, providing evidence of the importance of factoring out sequence similarity in predictive modeling. In response to these challenges, we fine-tune three distinct models - MSA Transformer, ESM2, and ProstT5 [54–56] - by incorporating a regression head into the architecture. Moreover, we adopt a strategic combination of a rigorous pre-processing pipeline, drastically reducing the number of training sequences, and we leverage a recently published mega-scale dataset [57]. While the mega-scale training enhances performances compared to older datasets, we also observe meaningful improvements in training with a small dataset encompassing a larger number of proteins. The results of our experiments show the superior predictive capability of the fine-tuned MSA Transformer compared to other PLMs. Overall, the optimized MSA Transformer consistently outperforms existing methods, highlighting the efficacy of incorporating information about sequence evolution to understand the effect of mutations on protein stability.

2 Data

Assessing the impact of a single mutation on protein stability using machine learning techniques is hindered by the limited amount of available experimental data, encompassing mutations associated with only a few hundred proteins. The recently released mega-scale dataset [57], generated through an innovative high-throughput experimental assay, allows the screening of thousands of single mutations, encouraging further exploration in this domain. Since our method implies the use of deep learning models, for its evaluation it is crucial to address overfitting by avoiding overlaps between training and test set data. Within the context of protein sequences, the shared level of similarity between the two sets plays a critical role, affecting data leakage. In order to account for this effect, we employ BLASTp [58] to perform local pairwise alignments between the test and training sequences. A protein is removed from the training set if there exists an element in the test set with sequence identity exceeding 25% (relative to the alignment length), e-value below 0.01, and alignment overlap surpassing 50% (defined as the portion of the alignment length relative to the query length), following the typical cutoffs enforced in the context of remote homology detection. We ensure that the criterion for selecting training sequences is consistently met across all analyzed test sets. The initial datasets, prior to similarity filtering, are obtained from a recent study [43], and are curated by including PDB-id, wild type and mutated sequences, the position and type of amino acid substitution, and the corresponding experimental measure of stability variation. The elements of the datasets described in detail below are labeled with the number of point mutations they contain.

- **Test sets**: For evaluating the models, we employ the well-established *S669* dataset [42, 50] and the *ssym* dataset [41] as test sets. The *S669* dataset, consisting of 94 proteins and 669 single mutations, is deliberately designed to minimize similarity with widely used training sets, such as *S2648* [27] and VariBench [59]. On the other hand, the *ssym* dataset, encompassing 15 proteins and 342 mutations, is specifically constructed to measure prediction biases of the methods towards destabilizing mutations.

- **Small training set**: The small training set that we consider, also denoted by *S1465*, is derived from the asymmetrical *Q3421* dataset in STRUM [60] by applying the similarity filtering based on sequence alignment against the test sets (Figure 1(a)). It includes 1465 point mutations and comprises 99 proteins sampled with an average of 15 mutations. Moreover, we create a reverse and symmetrical counterpart of this dataset to estimate the impact of the training labels distribution asymmetry on the model learning efficiency.

- **Large training set**: The large training set that we consider, denoted by *S157479*, is generated by 1) identifying direct amino acid substitutions within the mega-scale database, initially including more than 300 thousand mutations, 2) incorporating similarity filtering (Figure 1(b))) excluding proteins lacking any identified homologous sequence in UniClust30. The dataset comprises a total of 151 proteins and 157479 mutations. Each protein features about one thousand mutations, providing a valuable source for the uniform screening of point mutations across proteins.

- **Designed training set** The small and large datasets display a notable disparity in the sampled mutations, despite representing the same number of proteins. This prompts an investigation into the influence

of protein heterogeneity on the efficiency of the learning process. We include all of *S1465* and augment it with a randomly chosen set of 15 mutations from each protein in *S157479*, to construct a subset of the mega-scale database ensuring a mutation count comparable to that of *S1465*. With this procedure, we obtain a set of 3730 mutations, encompassing 250 proteins (*S3730*).

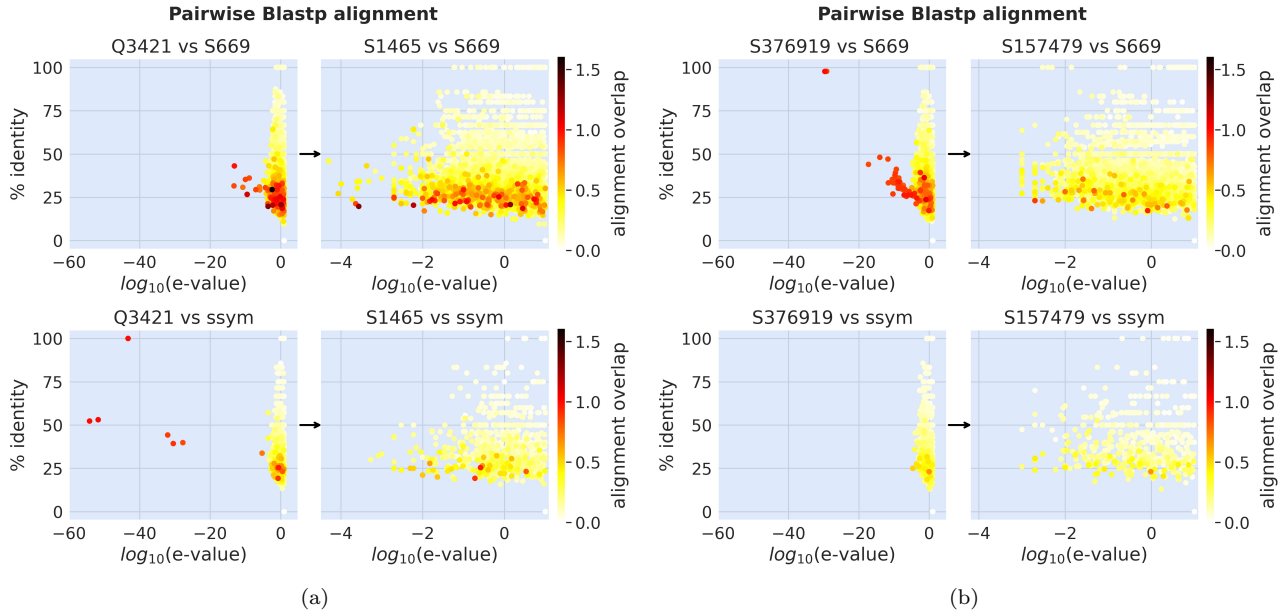


Figure 1: Pairwise BLASTp alignments (% identity, e-value and alignments overlap) of the training set against the *S669* (**top**) and *ssym* (**bottom**) test sets, before (**left**) and after (**right**) filtering by similarity. Each arrow indicates the transition from the original dataset to the filtered one. 1(a): from *Q3421* to *S1465*. 1(b): from *S376919* (the mega-scale dataset) to *S157479*.

3 Methods

We present a framework for refining pre-trained Protein Language Models through a supervised approach, harnessing their ability to encode biologically meaningful sequence representations. Our model is specifically designed to predict thermodynamic stability variations resulting from single amino acid substitutions in proteins. This is defined as the difference in the unfolding free energy between the mutated (M) and wild type (W) conformations, measured in kcal/mol:

$$\Delta\Delta G = \Delta G_M - \Delta G_W \tag{1}$$

We conduct end-to-end fine-tuning of PLMs, with the PLM module independently processing mutated and wild-type sequences. Subsequently, a Multi-Layer Perceptron (MLP) takes in input the differences between mutated and wild-type representations extracted from the last hidden layer of the PLM. The model is trained to minimize the Mean Absolute Error between predicted and experimental $\Delta\Delta G$ values. When training with the small dataset, we fine-tune three different protein language models: ESM-2 [55], ProstT5 [56], and MSA Transformer [54] (Figure 2), renaming them consequently. The initial pre-training weights are obtained from the HuggingFace Transformer Library [61]. The evaluation metrics are the Pearson's correlation coefficient (r), the Mean Absolute Error (MAE), and the Root Mean Square Error (RMSE).

- **ESM2_ddG**: ESM-2 is one of the largest architectures among single sequence models and stands out for its role in structure prediction. We adopt the architecture with 640 million parameters and 36 layers. After the last hidden layer, we concatenate the representations of the full sequence average and the mutated amino acid positions. The difference between these vectors for wild-type and mutated sequences is then utilized as input for the subsequent MLP. A comparable transfer learning approach involving the ESM-2 model is explored in a recent study [43], although it employs different training datasets.

- **ProstT5_ddG**: ProstT5, a bilingual protein language model featuring structure-aware embeddings, undergoes training on Foldseek's 3Di structures [62] and amino acid sequences. Designed with an encoder-decoder transformer architecture, ProstT5 is tailored for translating between sequences and structures. In our methodology, we pre-process wild-type and mutated sequences by translating them into their corresponding 3Di structures. The ProstT5 encoder module is fine-tuned by processing both the sequences

and the 3Di structures. We concatenate the resulting representations, feeding the difference between wild type and mutated ones into the regression head.

- **MSAesm_ddG**: The MSA Transformer introduces a structural modification compared to single sequence models, by processing a set of homologous aligned sequences. It achieves superior accuracy in contact prediction compared to individual sequence methods, while utilizing significantly fewer parameters. Notably, this model plays a pivotal role in the AlphaFold2 structure prediction algorithm [63]. We construct Multiple Sequence Alignments using HHblits [64], searching for wild-type homologous sequences in the Uniclust30 database [65]. Subsequently, we substitute the mutated amino acid into the wild type to generate the mutated alignments. Only the query sequence representations are used to input the regression, similarly to the ESM-2_ddG architecture.

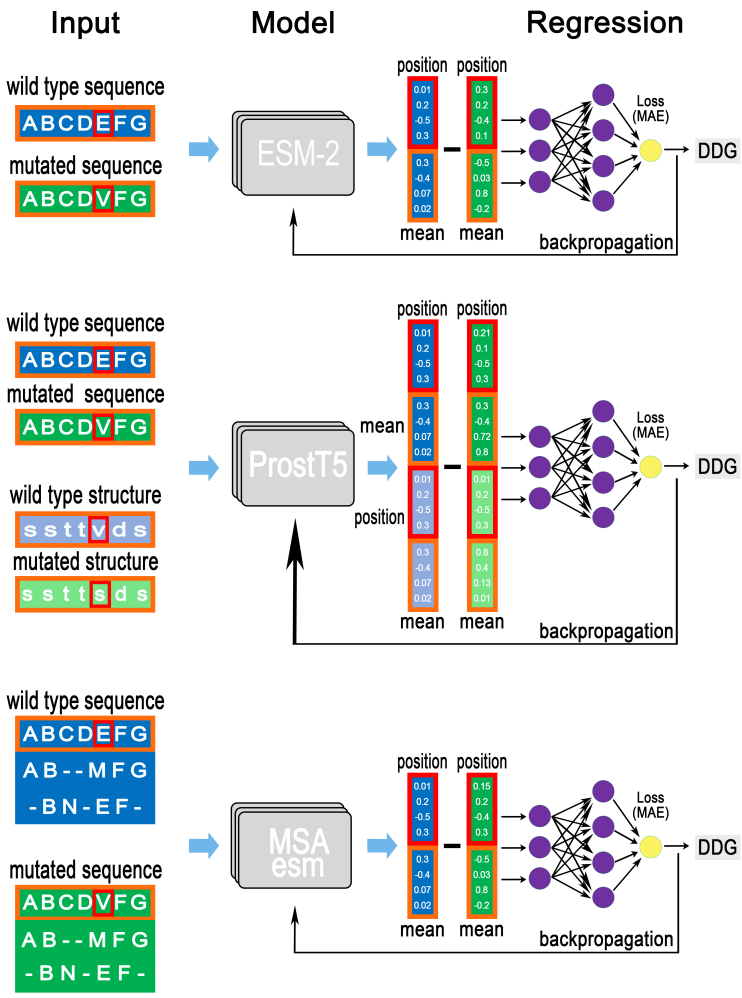


Figure 2: Sketch of models architectures: ESM2_ddG, ProstT5_ddG and MSAesm_ddG.

4 Results

4.1 Models, Symmetry, and Datasets

We fine-tune ESM-2, ProstT5, and MSA Transformer models, trained with the small dataset (*S1465*), and evaluate performances on the *S669* test set (Table 1). This experimental setup serves as a reference for identifying the most effective protein language model among the assessed options. In the comparison, the MSA Transformer emerges as the most promising model, while both the ESM2_ddG and the ProstT5_ddG methods lag behind it in terms of Pearson's correlation coefficient. It is worth to mention that the optimized ESM_ddG demonstrates results on par with previous findings [43], if considering the more rigorous sequence similarity filtering applied to our training set. Ultimately, given the superior correlation outcome and its considerably lighter architecture compared to the other models under consideration, we decide to employ the fine-tuned MSA Transformer for subsequent analyses.

Table 1: Performance of MSAesm_ddG, ESM2_ddG, and ProstT5_ddG, trained with the small *S1465* dataset.

	***S669* test set**		
Models trained on *S1465*	**Pearson r**	**MAE**	**RMSE**
MSAesm_ddG	0.49	1.02	1.45
ESM2_ddG	0.48	1.02	1.44
ProstT5_ddG	0.42	1.09	1.52

Furthermore, we leverage the smallest dataset to explore the impact on predictions of the $\Delta\Delta G$ distribution symmetry within the training data. We invert the order of wild type and mutated sequences and change the sign of the $\Delta\Delta G$ values to construct the reverse training set. For the symmetrical set, we include both the direct and reverse data. Training with these datasets, we note no significant impact on the measured accuracy (Table 6 in Supporting Information). This aligns with our expectations, as the inherent imbalance of the $\Delta\Delta G$ distribution towards destabilizing mutations typically affects structure-based procedures that violate symmetry in their architectures. Relying on these observations, we choose to focus exclusively on direct mutations in the subsequent analyses.
We trained the MSAesm_ddG model with both the large dataset (*S157479*) and the much smaller, designed dataset (*S3730*). Our findings demonstrate the appreciable advantage of training with the mega-scale data compared to the previously adopted *S1465* (Table 2). Interestingly, despite its considerably smaller size, training with the designed dataset yields comparable accuracy. This can be due to its extended range and variety of proteins, with respect to the small and large training sets. For subsequent comparisons with established methods we prioritize the MSAesm ddG model, trained on the large dataset. The motivation for this choice is twofold: 1) the large training dataset offers the opportunity to evaluate the performance on the *ssym* data, not used during validation and 2) MSAesm_ddG shows superior efficacy with respect to the other models under investigation. This statement is supported by the outcome of the optimized ESM-2 (Table 7 in Supporting Information), showing a lower gain when training with the mega-scale data.

Table 2: Performance of MSAesm_ddG trained with the large *S157479*, designed *S3730* and small *S1465* datasets.

	***S669* test set**		
Training Set	**Pearson r**	**MAE**	**RMSE**
S157479	0.54	0.96	1.39
S3730	0.52	0.96	1.41
S1465	0.49	1.02	1.45

It should be emphasized that, while the end-to-end procedure performs very well, training solely the MLP block can still produce remarkably good results, as shown in Table 3. This underscores the effectiveness of pre-training in imparting the evolutionary knowledge required to adeptly address the task. This observation necessitates additional experiments to both identify and potentially explain the transformation in representations resulting from the fine-tuning process.

Table 3: Baseline performance of the model, without updating the MSA Transformer weights. Training with *S157479*, *S3730* and *S1465*.

	***S669* test set**		
Training Set	**Pearson r**	**MAE**	**RMSE**
S157479	0.53	1.01	1.43
S3730	0.51	0.99	1.41
S1465	0.48	1.02	1.44

4.2 Comparison of prediction methods

We evaluate the efficacy of the developed MSAesm_ddG model, trained extensively on the mega-scale dataset, by conducting a comprehensive comparison with various models as outlined in [42]. This analysis is performed using the *S669* test set, specifically designed to prevent overfitting for all methods (Table 4). Performance metrics for recently released algorithms are sourced from their respective manuscripts [37,43,48]. The techniques under comparison encompass a variety of approaches, including classical and machine-learning-based methods. Only a few methods focus solely on sequence information, while the majority incorporate structural data.

- **Machine Learning methods**. *Sequence-based*: ACDC-NN-Seq [38], INPS-Seq [39]. *Structure-based*: I-Mutant3.0 [26], PoPMuSiC [27], MAESTRO DUET [28], [29], INPS3D [30], Dynamut [31], PremPS [32], ThermoNet [33], ACDC-NN [34], Dynamut2 [35], RaSP [37]. **PLM-based methods**. *Sequence-based*: PROSTATA [43]. *Structure-based*: ThermoMPNN [48].
- **Classical methods**. *Sequence-based*: DDGun [25]. *Structure-based*: FoldX [22], Rosetta [23], SDM [24], DDGun3D [25].

Table 4: Methods performance on the *S669* test set.

	***S669* test set**		
Method	**Pearson r**	**MAE**	**RMSE**
MSAesm_ddG	**0.54**	**0.96**	**1.39**
PROSTATA	0.49	1.00	1.45
ACDC-NN	0.46	1.05	1.49
ACDC-NN-Seq	0.42	1.08	1.53
ThermoMPNN	0.43	-	1.52
DDGun3D	0.43	1.11	1.60
DDGun	0.41	1.25	1.72
PremPS	0.41	1.08	1.50
RaSP	0.39	1.14	1.63
ThermoNet	0.39	1.17	1.62
Rosetta	0.39	2.08	2.70
Dynamut	0.41	1.19	1.60
INPS3D	0.43	1.07	1.50
INPS-Seq	0.43	1.09	1.52
SDM	0.41	1.26	1.67
PoPMuSiC	0.41	1.09	1.51
MAESTRO	0.50	1.06	1.44
FoldX	0.22	1.56	2.30
DUET	0.41	1.10	1.52
I-Mutant3.0	0.36	1.12	1.52
Dynamut2	0.34	1.15	1.58

Additionally, we provide results on the *ssym* test set (Table 5), where we compare the performance of MSAesm_ddG with that of another PLM-based model [43], given the authors' assertion of no data leakage for the *ssym* test set.

Table 5: Performance of two PLM-based methods on the *ssym* test set.

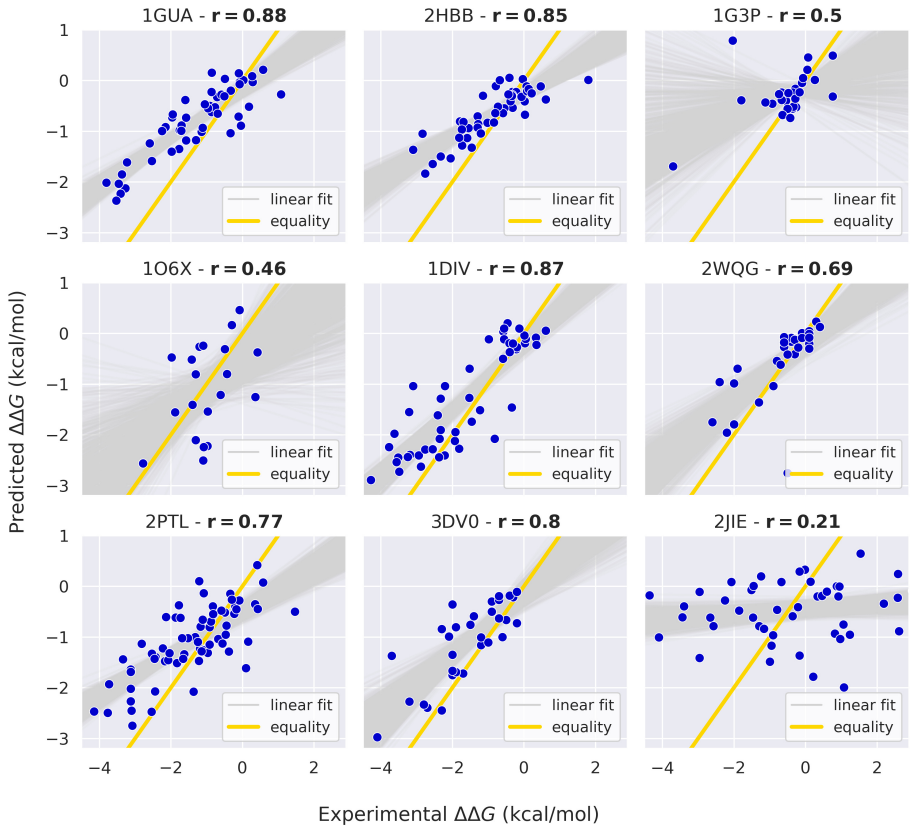
	***ssym* test set**		
Method	**Pearson r**	**MAE**	**RMSE**
MSAesm_ddG	**0.60**	**0.87**	**1.27**
PROSTATA	0.51	-	1.40

The results underscore the robustness and effectiveness of the developed method, establishing it as a leading approach for predicting the impact of mutations on protein stability. Furthermore, MSAesm_ddG outperforms

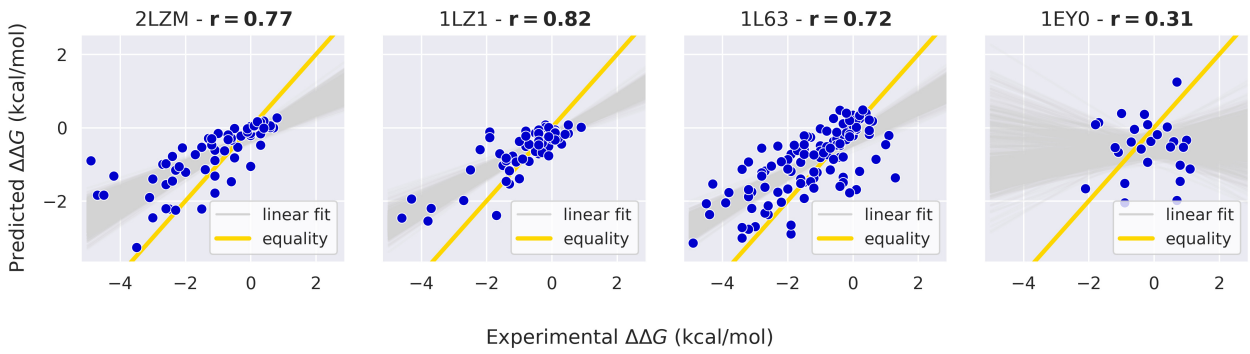
a similar PLM-based approach that involves fine-tuning the single-sequence ESM-2 model, emphasizing the crucial role of integrating evolutionary information into the model architecture.

4.3 Impact of protein fold class on predictive performance

We present illustrative results for proteins with more than 20 mutations in both the *S669* (Figure 3(a)) and *ssym* (Figure 3(b)) datasets. The majority of predictions for these proteins exhibit a Pearson correlation coefficient above 0.6, while only 30% show a lower correlation with the experimental $\Delta\Delta G$ values. Remarkably, the protein with the lowest correlation (2JIE) comprises a large number of residues, forcing to restrict the inference to only a limited number of homologues due to memory usage constraints. This limitation could potentially contribute to the poorer performance of MSAesm_ddG for this specific protein.



(a)



(b)

Figure 3: Predicted and experimental $\Delta\Delta G$ of proteins presenting more than 20 mutations. 3(a): *S669* test set. 3(b): *ssym* test set. The linear fit over 1000 bootstrapped samples in *gray* shows the trend of the shift between predictions and experiments.

Examining the effective slope of the linear fit between predictions and experiments, we measure a consistent shift between the two across most depicted proteins. This effect arises from the diverse nature of the training

and test $\Delta\Delta G$ distributions, stemming from different experimental conditions. In particular, the training distribution exhibits less prominent tails compared to the test data (Figure 4 in Supporting Information), which reflects into a narrower range of predicted values with respect to the experimental $\Delta\Delta G$ in the test set. Overall, these findings highlight how the model's accuracy is distinctly influenced by the specific protein under consideration, prompting further investigation in this regard. To delve deeper into the factors influencing predictive performance, we focus on proteins with a minimum of 5 mutations in the test sets and categorize them through the CATH fold class annotation [66]. Predictions are classified as positive cases if they demonstrate a correlation above the Pearson correlation coefficient measured on the full test set (21 cases), and negative if they fall below it (15 cases). Results reveal that among the analyzed cases, positive examples consist of 9.5% mainly beta, 38% mainly alpha, and 52.5% alpha/beta secondary structure motifs. Conversely, negative proteins comprise 47% mainly beta, 27% mainly alpha, and 27% alpha/beta secondary structure motifs. Of particular note is the observation that the number of superfamilies annotated as mainly alpha structures in CATH is nearly twice as large as those annotated as mainly beta. From this observation, we hypothesize that the prevalence of mainly alpha secondary structures among protein families could contribute to higher uncertainty in the model's predictions for mainly beta annotated proteins, which are less represented in the pre-training of the MSA Transformer.

5 Discussion

The advent of Protein Language Models as robust instruments for deciphering the language embedded within amino acid sequences has profoundly influenced the landscape of structural biology. In this study, we delved into the efficacy of PLMs to forecast thermodynamic stability changes induced by single amino acid mutations. In order to actively address the pervasive issue of overfitting, we filter training sequences adopting a stringent approach based on their resemblance to test sequences. While training with a mega-scale dataset yielded optimal results, we observed comparable accuracy when employing a smaller dataset designed to encompass a broader range of proteins. This suggests that the injection of diversity into the training dataset, even with a smaller mutation count, significantly amplifies the model's generalization capabilities. Furthermore, our protein-wise analysis unveiled intriguing trends in prediction accuracy across diverse protein structures. Notably, we observed a discernible dependency of prediction accuracy on the specific protein under consideration, suggesting that structural peculiarities inherent to individual proteins play a crucial role in shaping predictive outcomes. Positive instances, characterized by a higher correlation between predicted and experimental stability changes, were primarily associated with alpha/beta secondary structure motifs, while negative instances, exhibiting lower correlation, were predominantly linked with primarily beta structures. This underscores the importance of considering protein-specific features in stability prediction models and warrants further investigation into the nuanced relationship between protein structure and prediction accuracy. Another pivotal aspect of our study lies in the comparison with established prediction methods. First of all, our findings reveal the superior performance of the optimized MSA Transformer model compared to other PLMs such as ESM-2 and ProstT5. Moreover, this model consistently outperformed classical and machine learning-based methods that span both sequence-based and structure-based approaches. A key advantage of the MSA Transformer lies in its proficiency at leveraging co-evolution signals inherent in homologous sequences within multiple sequence alignments. These signals, reflecting conserved residue interactions throughout evolutionary history, provide valuable insights into the functional and structural constraints governing proteins. By integrating this rich source of evolutionary information, the MSA Transformer effectively discriminates between stabilizing and destabilizing mutations, even in cases where conventional methods struggle. Moreover, the model benefits from its robust architecture, which is specifically designed to handle large-scale sequence data and capture long-range dependencies. The self-attention mechanism employed in the transformer architecture allows the model to focus on relevant regions of the sequence and determine the importance of different amino acids based on their contextual information within the alignment. This attention mechanism enables to effectively discern changes in sequence patterns that are indicative of stability-altering mutations. Unlike traditional machine learning approaches, reliant on domain-specific features or handcrafted representations, our model learns directly from raw sequence data, rendering it highly adaptable to diverse protein structures and functional contexts. This adaptability empowers the model to outperform existing methods, albeit contingent upon the availability of homologous sequences. In summary, our study advances the field by demonstrating the effectiveness of fine-tuning protein language models using evolutionary information for predicting stability changes induced by mutations. Looking ahead, further advancements in PLM-based architectures, along with the increasing availability of comprehensive sequence databases, are set to drive continued progress in this domain, facilitating the design of more stable and functional proteins for diverse biomedical and biotechnological applications.

6 Competing interests

No competing interest is declared.

7 Acknowledgments and Funding

The authors acknowledge the AREA Science Park supercomputing platform ORFEO made available for conducting the research reported in this paper and the technical support of the Laboratory of Data Engineering staff. F.C., M.C., A.A., and A.C. were supported by the European Union – NextGenerationEU within the project PNRR "PRP@CERIC" IR0000028 - Mission 4 Component 2 Investment 3.1 Action 3.1.1.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

[3] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17, 2019.

[4] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pages 2020–12, 2020.

[5] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. 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.

[6] Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, 2021.

[7] 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.

[8] Brian L Hie, Kevin K Yang, and Peter S Kim. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285, 2022.

[9] Nicole N Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J Rollins, Daniel Ritter, Chris Sander, Yarin Gal, and Debora S Marks. Learning from prepandemic data to forecast viral escape. *Nature*, pages 1–8, 2023.

[10] Nobuhiko Tokuriki, Francois Stricher, Luis Serrano, and Dan S Tawfik. How protein stability and new functions trade off. *PLoS computational biology*, 4(2):e1000002, 2008.

[11] Nobuhiko Tokuriki and Dan S Tawfik. Protein dynamism and evolvability. *Science*, 324(5924):203–207, 2009.

[12] Jesse D Bloom, Sy T Labthavikul, Christopher R Otey, and Frances H Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, 2006.

[13] Konstantin B Zeldovich, Peiqiu Chen, and Eugene I Shakhnovich. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences*, 104(41):16152–16157, 2007.

[14] James W Bryson, Stephen F Betz, Helen S Lu, Daniel J Suich, Hongxing X Zhou, Karyn T O'Neil, and William F DeGrado. Protein design: a hierarchic approach. *Science*, 270(5238):935–941, 1995.

[15] Michele Vendruscolo, Amos Maritan, and Jayanth R Banavar. Stability threshold as a selection principle for protein design. *Physical review letters*, 78(20):3967, 1997.

[16] Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.

[17] Zhen Wang and John Moult. Snps, protein structure, and disease. *Human mutation*, 17(4):263–270, 2001.

[18] Pauline C Ng and Steven Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.

[19] Peng Yue, Zhaolong Li, and John Moult. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology*, 353(2):459–473, 2005.

[20] AndreasC Joerger and Alan R Fersht. Structure–function–rescue: the diverse nature of common p53 cancer mutants. *Oncogene*, 26(15):2226–2242, 2007.

[21] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011.

[22] 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.

[23] Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

[24] Arun Prasad Pandurangan, Bernardo Ochoa-Montano, David B Ascher, and Tom L Blundell. Sdm: a server for predicting effects of mutations on protein stability. *Nucleic acids research*, 45(W1):W229–W235, 2017.

[25] Ludovica Montanucci, Emidio Capriotti, Giovanni Birolo, Silvia Benevenuta, Corrado Pancotti, Dennis Lal, and Piero Fariselli. Ddgun: an untrained predictor of protein stability changes upon amino acid variants. *Nucleic Acids Research*, 50(W1):W222–W227, 2022.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] Castrense Savojardo, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio. Inps-md: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics*, 32(16):2542–2544, 2016.

[31] Carlos HM Rodrigues, Douglas EV Pires, and David B Ascher. Dynamut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic acids research*, 46(W1):W350–W355, 2018.

[32] Yuting Chen, Haoyu Lu, Ning Zhang, Zefeng Zhu, Shuqin Wang, and Minghui Li. Premps: Predicting the impact of missense mutations on protein stability. *PLoS computational biology*, 16(12):e1008543, 2020.

[33] 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.

[34] 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.

[35] Carlos HM Rodrigues, Douglas EV Pires, and David B Ascher. Dynamut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science*, 30(1):60–69, 2021.

[36] Yunzhuo Zhou, Qisheng Pan, Douglas EV Pires, Carlos HM Rodrigues, and David B Ascher. Ddmut: predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Research*, page gkad472, 2023.

[37] 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. *Elife*, 12:e82593, 2023.

[38] Corrado Pancotti, Silvia Benevenuta, Valeria Repetto, Giovanni Birolo, Emidio Capriotti, Tiziana Sanavia, and Piero Fariselli. A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Genes*, 12(6):911, 2021.

[39] Castrense Savojardo, Pier Luigi Martelli, Rita Casadio, and Piero Fariselli. On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Briefings in bioinformatics*, 22(1):601–603, 2021.

[40] Fabrizio Pucci, Katrien Bernaerts, Fabian Teheux, Dimitri Gilis, and Marianne Rooman. Symmetry principles in optimization problems: an application to protein stability prediction. *IFAC-PapersOnLine*, 48(1):458–463, 2015.

[41] 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.

[42] 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.

[43] Dmitriy Umerenkov, Fedor Nikolaev, Tatiana I Shashkova, Pavel V Strashnov, Maria Sindeeva, Andrey Shevtsov, Nikita V Ivanisenko, and Olga L Kardymon. Prostata: a framework for protein stability assessment using transformers. *Bioinformatics*, 39(11):btad671, 2023.

[44] Felix Jung, Kevin Frey, David Zimmer, and Timo Mühlhaus. Deepstabp: A deep learning approach for the prediction of thermal protein stability. *International Journal of Molecular Sciences*, 24(8):7444, 2023.

[45] Yijie Zhang, Zhangyang Gao, Cheng Tan, and Stan Z Li. Efficiently predicting protein stability changes upon single-point mutation with large language models. *arXiv preprint arXiv:2312.04019*, 2023.

[46] Daniel J Diaz, Chengyue Gong, Jeffrey Ouyang-Zhang, James M Loy, Jordan Wells, David Yang, Andrew D Ellington, Alex Dimakis, and Adam R Klivans. Stability oracle: A structure-based graph-transformer for identifying stabilizing mutations. *bioRxiv*, pages 2023–05, 2023.

[47] Kit Sang Chu and Justin B Siegel. Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset. *bioRxiv*, pages 2023–11, 2023.

[48] Henry Dieckhaus, Michael Brocidiacono, Nicholas Z Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121(6):e2314853121, 2024.

[49] Rahul Nikam, A Kulandaisamy, K Harini, Divya Sharma, and M Michael Gromiha. Prothermdb: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic acids research*, 49(D1):D420–D424, 2021.

[50] Joicymara S Xavier, Thanh-Binh Nguyen, Malancha Karmarkar, Stephanie Portelli, Pâmela M Rezende, João PL Velloso, David B Ascher, and Douglas EV Pires. Thermomutdb: a thermodynamic database for missense mutations. *Nucleic acids research*, 49(D1):D475–D479, 2021.

[51] 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.

[52] Jianwen Fang. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Briefings in bioinformatics*, 21(4):1285–1292, 2020.

[53] 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.

[54] 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.

[55] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[56] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Martin Steinegger, and Burkhard Rost. Prostt5: Bilingual language model for protein sequence and structure. *bioRxiv*, pages 2023–07, 2023.

[57] 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, 2023.

[58] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[59] Preethy Sasidharan Nair and Mauno Vihinen. V ari b ench: A benchmark database for variations. *Human mutation*, 34(1):42–49, 2013.

[60] Lijun Quan, Qiang Lv, and Yang Zhang. Strum: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, 32(19):2936–2946, 2016.

[61] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[62] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pages 1–4, 2023.

[63] 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.

[64] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

[65] Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J Martin, Johannes Söding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research*, 45(D1):D170–D176, 2017.

[66] Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

Supplementary Information

Table 6: MSAesm_ddG model trained with the symmetrical, reverse and direct $\Delta\Delta G$ distributed *S1465*.

	***S669* test set**		
***S1465* symmetry**	**Pearson r**	**MAE**	**RMSE**
symmetrical	0.49	1.02	1.45
reverse	0.48	1.05	1.48
direct	0.49	1.02	1.45

Table 7: Performance of ESM2_ddG trained with the large *S157479*, designed *S3730* and small *S1465* datasets.

	***S669* test set**		
Training Set	**Pearson r**	**MAE**	**RMSE**
S157479	0.50	1.00	1.46
S3730	0.48	1.01	1.44
S1465	0.48	1.02	1.44

Table 8: Baseline performance of the model, without updating the ESM-2 weights. Training with the *S157479*, *S3730* and *S1465* datasets.

	***S669* test set**		
Training Set	**Pearson r**	**MAE**	**RMSE**
S157479	0.51	1.02	1.46
S3730	0.48	1.01	1.44
S1465	0.48	1.03	1.45

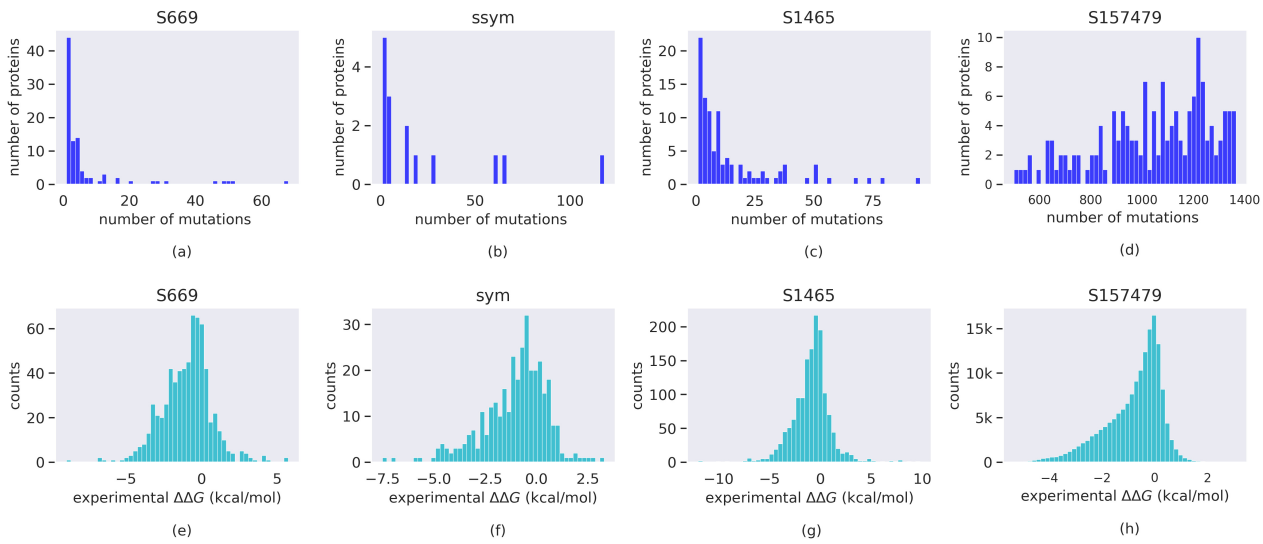


Figure 4: Number of proteins vs number of mutations (a-d) and $\Delta\Delta G$ distribution (e-h) for all datasets.

Tuning Hyperparameters

The Multi-Layer Perceptron block consists of one hidden layer with size 768, followed by a ReLU activation and a dropout layer (with probability p of an element to be zeroed set to $p = 0.2$) that feeds into the final linear layer. We adopt a PyTorch implementation with AdamW optimizer (`torch.optim.AdamW` with default parameters, apart from the learning rate) and OneCycleR learning-rate scheduler (`torch.optim.lr_scheduler.OneCycleLR`). The learning rate is selected for each model and training dataset by means of a grid search with `lr` $\in \{0.5\times10^{-5}, 10^{-5}, 0.5\times10^{-4}, 10^{-4}, 0.5\times10^{-3}, 10^{-3}, 0.5\times10^{-2}, 10^{-2}\}$. In order to reduce the memory footprint, the PyTorch Automatic Mixed Precision is enabled by means of `torch.autocast` (specifically using `dtype=torch.float16`) in combination with the gradient scaling function `torch.cuda.amp.GradScaler()`. The gradient norm is clipped using `torch.nn.utils.clip_grad_norm_` with `max_norm` = 10, to prevent possible exploding gradient problems. We set a maximum number of 20 epochs for the training in combination with an early stopping criterion. The latter consists in saving the model parameters associated to the epoch with lowest MAE on the validation set, and stopping the training if the MAE on the validation is not decreasing for the next 5 epochs. For training with both the small and designed datasets, we validate on the *ssym* set, which does not overlap with both the training and the test sequences (Figure 5(a)). When training with the large dataset, a customized validation set is constructed (*S1082*) by excluding from the small dataset sequences that are similar to those in the large dataset (Figure 5(b)), following the criterion stated in the main text. Using the PyTorch DistributedDataParallel class (`torch.nn.parallel.DistributedDataParallel`), the fine-tuning is performed on 16 nodes of the LEONARDO Booster partition at CINECA National Supercomputing Center, each consisting of a single socket 32 cores Intel Xeon Platinum 8358 (2.60GHz), together with 4 × NVIDIA A100 GPUs (64GB HBM2e and NVLink 3.0) and 512 GB DDR4 (3200 MHz) RAM. The internal network is provided by NVIDIA Mellanox Infiniband HDR DragonFly+ 200 Gbps. We consider a batch size $\mathcal{B} = 1$ for a single GPU, so the overall effective batch size $\mathcal{B}_{\text{eff}} = 64$. Finally, we load the model parameters for the inference on the test sets on a single GPU.

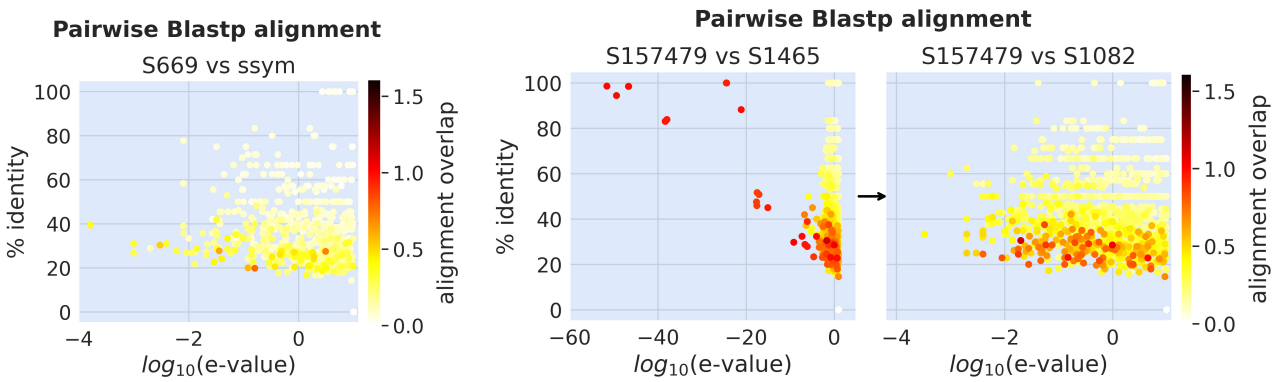


Figure 5: 5(a): pairwise BLASTp alignments (% identity, e-value and alignments overlap) of the *S669* test set against the ssym set. 5(b): pairwise BLASTp alignments of the *S157479* training set against the *S1465* dataset, before (**left**) and after (**right**) filtering by similarity to obtain the *S1082* dataset (used for validation in the *S157479* training).