REPORT

OPEN ACCESS

www.the-innovation.org

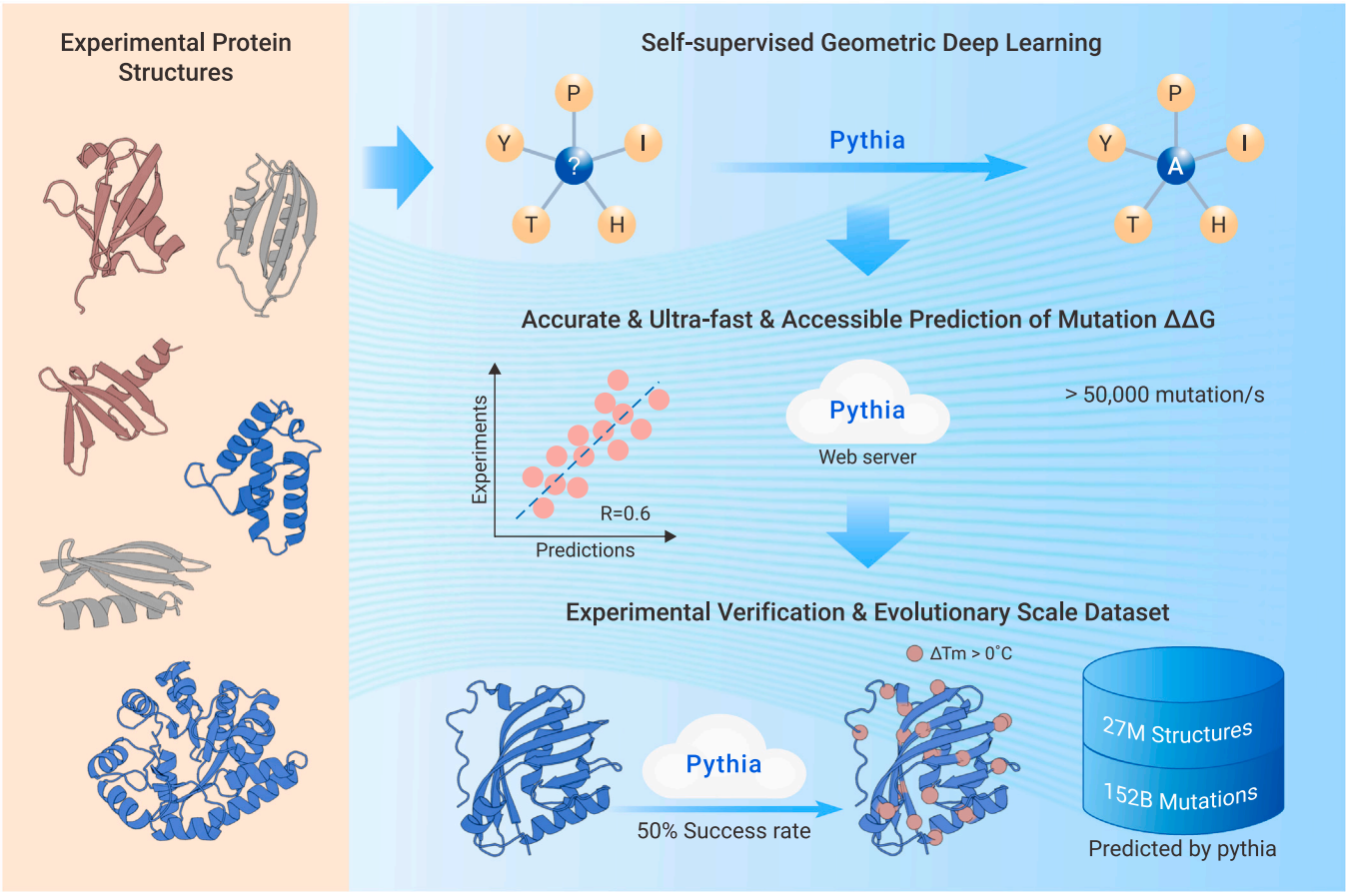
Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation

Jinyuan Sun,[1,2] Tong Zhu,[1,2] Yinglu Cui,[1,*] and Bian Wu[1,*]

*Correspondence: cuiyinglu@im.ac.cn (Y.C.); thebianwu@outlook.com (B.W.)

Received: February 4, 2024; Accepted: December 2, 2024; Published Online: January 6, 2025; https://doi.org/10.1016/j.xinn.2024.100750

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Predicting mutation-driven changes in protein stability using a self-supervised deep learning model.
- The model achieved state-of-the-art prediction accuracy across various benchmarks with exceptional speed.
- Experimental verification of Pythia-predicted mutations demonstrated a higher success rate than previous predictors.
- Large-scale mutation analysis across the protein universe revealed a correlation between protein stability and evolutionary information.

OPEN ACCESS

REPORT

The Innovation

Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation

Jinyuan Sun,[1,2] Tong Zhu,[1,2] Yinglu Cui,[1,*] and Bian Wu[1,*]

[1]AIM Center, College of Life Sciences and Technology, Beijing University of Chemical Technology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[2]University of Chinese Academy of Sciences, Beijing, China

*Correspondence: cuiyinglu@im.ac.cn (Y.C.); thebianwu@outlook.com (B.W.)

Received: February 4, 2024; Accepted: December 2, 2024; Published Online: January 6, 2025; https://doi.org/10.1016/j.xinn.2024.100750

Citation: Sun J., Zhu T., Cui Y., et al., (2025). Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation. The Innovation **6(1)**, 100750.

Predicting free energy changes (ΔΔG) is essential for enhancing our understanding of protein evolution and plays a pivotal role in protein engineering and pharmaceutical development. While traditional methods offer valuable insights, they are often constrained by computational speed and reliance on biased training datasets. These constraints become particularly evident when aiming for accurate ΔΔG predictions across a diverse array of protein sequences. Herein, we introduce Pythia, a self-supervised graph neural network specifically designed for zero-shot ΔΔG predictions. Our comparative benchmarks demonstrate that Pythia outperforms other self-supervised pretraining models and force field-based approaches while also exhibiting competitive performance with fully supervised models. Notably, Pythia shows strong correlations and achieves a remarkable increase in computational speed of up to 10^5-fold. We further validated Pythia's performance in predicting the thermostabilizing mutations of limonene epoxide hydrolase, leading to higher experimental success rates. This exceptional efficiency has enabled us to explore 26 million high-quality protein structures, marking a significant advancement in our ability to navigate the protein sequence space and enhance our understanding of the relationships between protein genotype and phenotype. In addition, we established a web server at https://pythia.wulab.xyz to allow users to easily perform such predictions.

INTRODUCTION

Proteins, often described as the molecular workhorses of life, carry out a diverse range of essential biochemical functions.[1,2] Despite their vital roles, most natural proteins exhibit only marginal stability, with Gibbs free energy differences between their native and unfolded states as low as 5 kcal/mol[1,3] or even less.[4,5] This narrow margin of stability renders them particularly susceptible to environmental changes and genetic mutations.[6] Even subtle alterations, such as single-point mutations, can disrupt this delicate balance, resulting in protein inactivation, misfolding, or aggregation. The destabilizing or stabilizing effects of these changes have broad implications for health, disease mechanisms, drug discovery, biotechnology, and our understanding of protein evolution.[7]

The modern era is characterized by endeavors to transcend the limitations of natural protein repertoires, and protein engineering has emerged as a promising avenue.[8] Protein sequences have been designed to enhance stability and solubility and to tailor activities to meet the demands of industrial applications.[2,9] Advancements in protein design computational tools have employed model-based and data-driven methodologies.[10] Among the model-based approaches, energy calculation is used to predict ΔΔG (the difference in ΔG between the wild type and mutant) resulting from amino acid substitutions, which helps to identify thermostabilizing mutations.[11] Several studies have successfully leveraged well-established energy functions in models such as Rosetta,[12] FoldX,[13] and ABACUS2[14] to design thermostable enzymes.[15] However, these methods are still limited by imbalanced parametrization of the energy functions and insufficient sampling of conformational space.[11]

Recent advances in machine learning (ML) present a promising avenue for solutions. One particularly compelling approach involves training ML models on experimental data that capture stability changes resulting from mutations, while leveraging features that are known *a priori* to influence stability. These models typically rely on carefully curated evolutionary features, such as BLOSUM62[16] and probabilities derived from multiple sequence alignments (MSAs),[17] as well as structural features that include accessible surface area,[18] predicted hydrogen bonds,[19] atomic charges,[20] and energy terms from Rosetta/FoldX-modeled mutation structures[21] and other calculations.[20,22]

In addition to feature engineering, various architectures have been explored, including 3D convolutional neural networks (CNNs),[21] graph neural networks (GNNs),[23] Bayesian neural networks,[20] and Transformers.[24] While these supervised methods are attractive because they can directly learn from experimental data and provide improved processing speed, they are often constrained by the limited availability of experimentally measured ΔΔG training data and the biases that can be present in these datasets.[25–28] Such challenges are common in biology due to the labor-intensive nature of wet lab experiments.[29]

In contrast to supervised learning, which is restricted by the availability of labeled data, self-supervised learning (SSL) can glean insights from vast amounts of unlabeled data.[30] A particularly prominent SSL strategy is masked language modeling (MLM), which trains models to predict a masked or substituted token based on its contextual surroundings.[31] MLM has found widespread application across protein sequences,[32] MSAs,[33] and protein structures,[34] especially in predicting mutation fitness. For example, ESM-1v,[35] which was trained using MLM on 150 million sequences from the UniRef90 database, achieved exceptional zero-shot fitness prediction results on 41 deep mutation scanning datasets with an average Spearman's rho of 0.509. Furthermore, SSL approaches based on structure have been explored.

ProteinSolver was trained on both protein structure data and homologous sequences for protein design, and the probabilities assigned to individual residues have demonstrated a correlation with the stability of mutants.[36] In a similar vein, ABACUS-R[37] was developed using a high-quality subset of protein structure data for *de novo* protein design based on Transformer architecture and has shown superior predictive correlation for mutant stability compared with ProteinSolver. There have been concerted efforts to predict stability changes resulting from mutations by employing SSL to enhance structural feature extraction. A pretrained CNN that utilizes spherical convolutions was used to predict amino acid propensities, with the log-likelihoods of both wild-type and mutant sequences serving as features for supervised ΔΔG prediction through a neural network-based regressor.[38] Recent studies have reported improved prediction correlations compared with earlier efforts by incorporating predicted labels generated through Rosetta for data augmentation[25] or leveraging larger datasets derived from high-throughput experiments[39,40] in combination with more advanced deep learning models.[41–44] These recent advancements yield promising results, further highlighting SSL's potential in addressing molecular fitness challenges, including mutation stability.

Drawing on the foundational principles of SSL and insights from previous research, we have developed Pythia, a self-supervised model specifically designed for predicting ΔΔG of mutations based on protein structures. This model is constructed to decode intrinsic patterns among residues within given proteins, enabling precise predictions of mutational effects. Pythia operates independently of evolutionary information and manually crafted features derived from energy functions, learning stability directly from the protein structures themselves. Its evaluations against thousands of reliable experimental ΔΔG datasets and a recent mega-scale dataset, Pythia demonstrated superior prediction accuracy compared with other self-supervised models and energy functions. Its performance was comparable with, or even better than, that of supervised models across various benchmarks, while boasting significantly faster prediction throughput ranging from 700 to 100,000 mutations per second, depending on the hardware used. By focusing on limonene epoxide hydrolase (LEH), we empirically showcased Pythia's capacity to identify a greater share of effective thermostabilizing mutations. Moreover, we emphasized Pythia's potential for extensive exploration within the protein universe by calculating all single mutations in the high-quality predicted structures available in the AlphaFold database,[45]

amounting to over 26 million predicted protein structures. Pythia's source code is freely accessible at https://github.com/Wublab/pythia.

RESULTS

Model architecture and training of Pythia

Over the past few decades, numerous methods have been developed to investigate the relationship between the free energy landscape and the internal structure of proteins. However, the accuracy of these approaches appears to be constrained by the approximations and assumptions inherent in the models used. In this context, we propose that the energy of a protein in its unfolded state is largely unaffected by mutations,[16] given that there are virtually no stable specific interactions between the side chains of a protein when it is unfolded: $\Delta\Delta G \sim \Delta G^{folded}_{MUT} - \Delta G^{folded}_{WT}$. According to the Boltzmann hypothesis of protein folding and energy, the probability of a rotamer is determined by the energy, which is influenced by atomic interactions with neighboring residues. By summing the probabilities of all rotamers for a particular amino acid, we can derive the probability of that amino acid and, subsequently, its free energy. From this analysis, we can conclude that, for a specific position within a folded protein structure, the free energy difference attributable to amino acid substitutions dictates the relative probabilities of the various amino acids:

$$-\ln\frac{P_{AA_j}}{P_{AA_i}} = \frac{1}{k_B T}\Delta\Delta G_{AA_i \to AA_j}$$

$\Delta\Delta G_{AA_i \to AA_j}$ is the difference in the folding free energy change of AA_i to AA_j, P_{AA_i} is the probability of amino acid type i, k_B is the Boltzmann constant, and T is the temperature. The prediction of ΔΔG can be achieved by estimating the probabilities of each amino acid (represented as P_{AA}) at a specific position within a given structure. While stability has been found to correlate with the likelihood derived from MSAs,[39,40] this correlation is relatively weak. Moreover, the likelihood is influenced not only by folding stability but also by various other factors, including function, solubility, and aggregation.[44] Since energy is determined by the atomic interactions present in folded structures, we draw upon previous successes of statistical potentials in protein structure assessment[46] and *de novo* design[14] to suggest that the P_{AA} can be better estimated from structure data to gain a better prediction of ΔΔG.

The energy of a protein is determined by the interactions among neighboring residues, which led us to adopt a widely recognized GNN known for its effectiveness in protein structure prediction[47] and sequence design.[48] A protein local structure was transformed into a graph representation using a k-nearest neighbor graph (Figure 1A). In this graph, each amino acid acts as a node and is connected to its 32 nearest amino acids, determined by the Euclidean distance of the C-alpha atom. The features of each node include one-hot encoding for the amino acid type, along with the backbone dihedral angles (ϕ, ψ, and ω) represented using sine and cosine functions. To maintain SE(3) invariance, we incorporated the distances between five backbone atoms—C-alpha, C, N, O, and C-beta (when available)—in our edge encoding, with distance measured in Ångström (Å). In addition, we introduced supplementary features such as the relative positional encoding of amino acids in the sequence and chain identity encoding. The chain identity encoding assigned a binary value of 1 if two amino acids belong to the same chain, or 0 if they do not (Figure 1B). The training objective was to predict the natural amino acid type of the central node using information from the nodes and edges (Figure 1C).

Pythia employs the message-passing neural network (MPNN) architecture,[49] specifically designed with an attention-based message-passing and readout function (Figure 1D). By integrating attention mechanisms into the MPNN, this approach, referred to as the attention message-passing layer (AMPL), allows the model to focus more effectively on substructures critical to the desired interaction properties during the learning process. In each layer of the AMPL, the vertex representation is updated using an attention block, which is then concatenated with the edge representation to derive the message representation (Figure 1E). This message representation subsequently serves as a query to further refine the node representation through an additional attention block (Figure 1F). The final model consists of three AMPLs, each operating with a hidden dimension of 128.

During the model training phase, we evaluated several hyperparameters, including the masking ratio of the central nodes and the noise level of the backbone coordinates, as outlined in Table S1. The physical unit of noise is Å aligning with the input distance features. To enhance robustness and generalizability, we developed two distinct models. One model was trained on specifically defined protein domains obtained from the CATH database,[50] while the other was developed using a nonredundant protein structure dataset constructed in this study by clustering high-resolution bioassemblies from the RCSB PDB database.[51] The final prediction of Pythia is computed using the averaged outputs from these two models. We have launched a web server at https://pythia.wulab.xyz to facilitate predictions (Figure 1G).

Benchmark evaluation of Pythia in ΔΔG prediction

Pythia was evaluated alongside a diverse array of pretrained protein models and three widely used energy function methods on the S2648 dataset,[52] which is recognized as a standard training set for supervised ML models for predicting ΔΔG of mutations due to its high quality. In this assessment, Pythia achieved a Spearman's rho of 0.616 and Pearson's r of 0.598 (Figure 2A), outperforming all models tested across six critical performance metrics: Spearman's rho, Pearson's r, accuracy, F1-score, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) (Figure 2B). Notably, all structure-based pretrained models demonstrated higher correlation compared with sequence-based and MSA-based pretrained models, with the state-of-the-art model (ESM2-t33[53]) failing to exceed a correlation of 0.4. Furthermore, larger protein language models did not consistently outperform their smaller counterparts, consistent with previous findings that suggest larger models trained on more extensive datasets may estimate the density of sequence data more effectively without necessarily improving fitness estimations.[54] This highlights the importance of incorporating structural information, as it provides valuable insights into inter-residue interactions, making it a more effective strategy for predicting the thermodynamic properties and mutation effects. Our findings reinforced our idea that probabilities derived from energy assessments are more accurately determined from structural data rather than sequence data. However, while structure-based models remain suboptimal when compared with energy function-based methods, Pythia stands out as the only model to achieve a higher correlation. Pythia demonstrated an improved ability to transfer its learning to the single mutation prediction task and, for the first time, outperformed force field-based methods among pretrained models in predicting ΔΔG. Remarkably, Pythia accomplishes this with just 1.3 million parameters, which is one-third of the parameter count of the second smallest model (Figure 2C).

We further explored the performance of Pythia in comparison with supervised ML models. A direct comparison with ML-based predictors presents challenges due to varying training datasets, which may lead to data leakage and biases.[55] To mitigate this issue, we utilized a dataset known as S669, which has not been used in training any supervised ML models and shares sequence identities of less than 25% with S2648 and the VariBench dataset.[26] As shown in Figures 2D and 2E, the prediction performances yield a Spearman's rho ranging from 0.28 to 0.63 for supervised ML models[25,56–61] and 0.28 to 0.59 for statistical methods.[62,63] Pythia outperformed all evaluated methods on the S669 dataset across all metrics, achieving a Spearman's rho of 0.66. One significant challenge for supervised ΔΔG predictors is their inability to maintain the symmetry between direct and inversed mutations.[26] In contrast, Pythia does not depend on any labeled ΔΔG data during training. It addresses the symmetry issue, at least partially, by utilizing a fixed protein backbone configuration, while still achieving the highest Spearman's rho in inverse predictions from remodeled structures of mutants (Table S3).

In addition to its impressive prediction accuracy, Pythia offers a significant advantage in computational speed. Force field-based methods often require sampling of local side chain or even backbone structural conformation to achieve more accurate predictions, but face constraints in computational speed, particularly when handling proteins of large sequence length. Even with a fixed backbone, these methods can manage only about 10 mutations per minute. Among them, FoldX, the highest-performing option, is particularly slow, averaging just 1 mutation per minute on a CPU core (E3-2678v3) due to its elaborate sampling methodology, which necessitates multiple independent runs and subsequent averaging. In comparison, when tasked with computing 380,741 mutations for 131 proteins in the S2648 dataset, Pythia completes the initialization and computations in merely 20 s on 24 CPU cores, achieving an approximate rate of 50,000 mutations per minute on a single core. This remarkable efficiency surpasses that

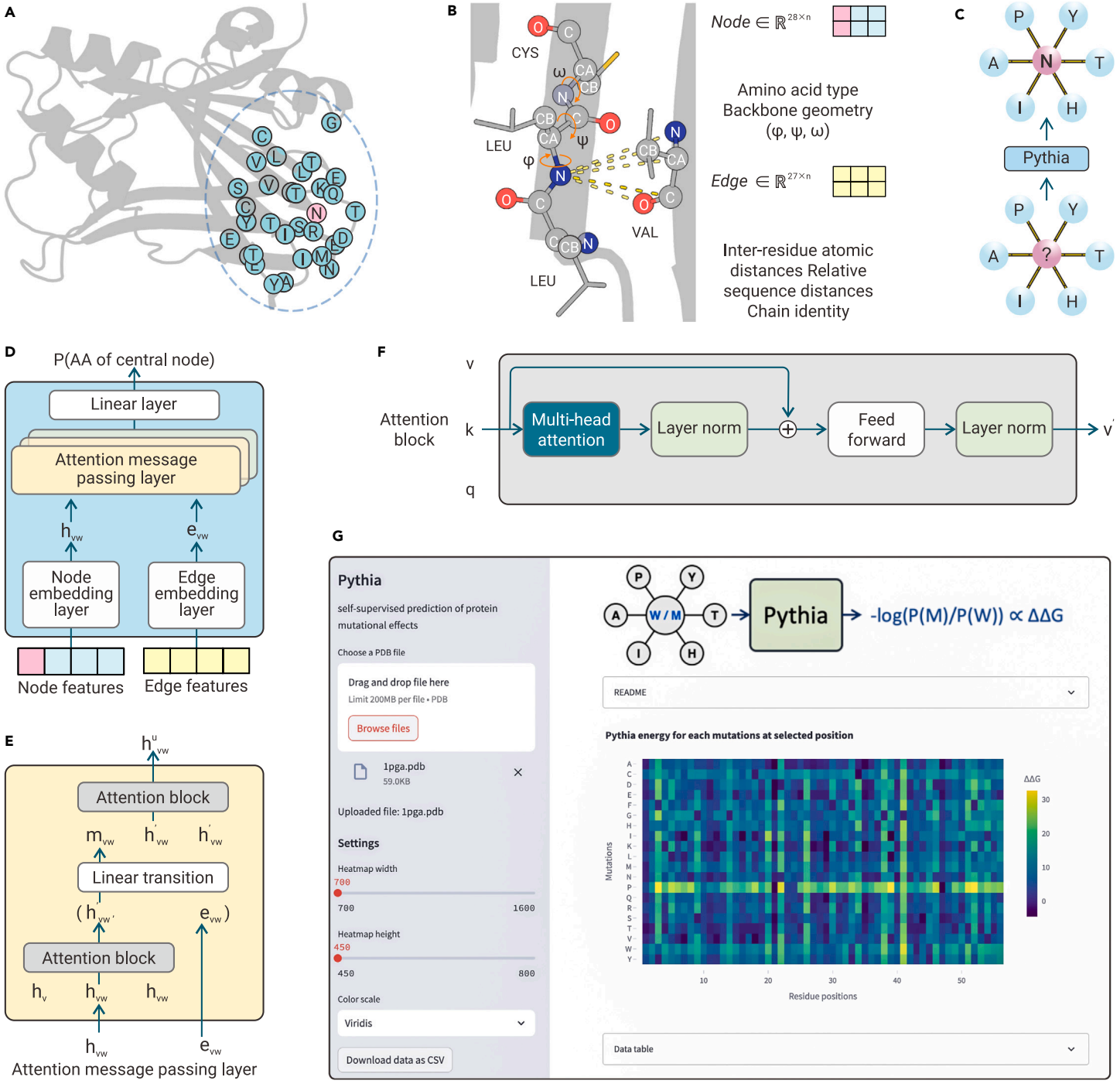


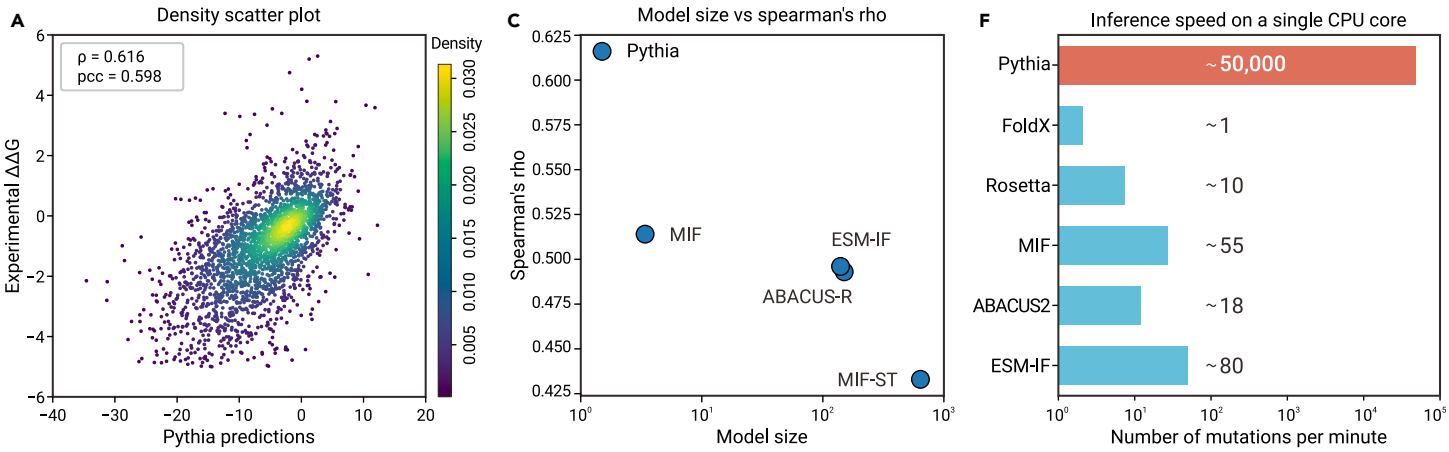
Figure 1. Overview of the Pythia model (A) Pythia processes a protein's local structure as a k-NN graph of C-alpha atoms, abstracting it into an amino acid graph. (B) Node features include amino acid type and three dihedral angles (ϕ, ψ, ω), while edge features consist of distances between main chain atoms, sequence positions, and chain information. (C) Pythia's training task is to predict the amino acid type of the central node. (D) The architecture of the Pythia model. The node and edge features independently traverse the embedding layer and enter the attention-based message-passing neural network. The output is the probabilities of the 20 amino acids. (E) Breakdown of the architecture of attention message-passing layer. Within this layer, the information of nodes is first updated using the attention block. The embeddings of edge (e_{vw}) are concatenated with the representation of nodes (h'_{vw}) to get m_{vw}. Subsequently, m_{vw} and h'_{vw} go through the attention module, resulting in the updated h^{u}_{vw}. (F) The structure of the attention block. (G) A visual snapshot of the Pythia webapp interface.

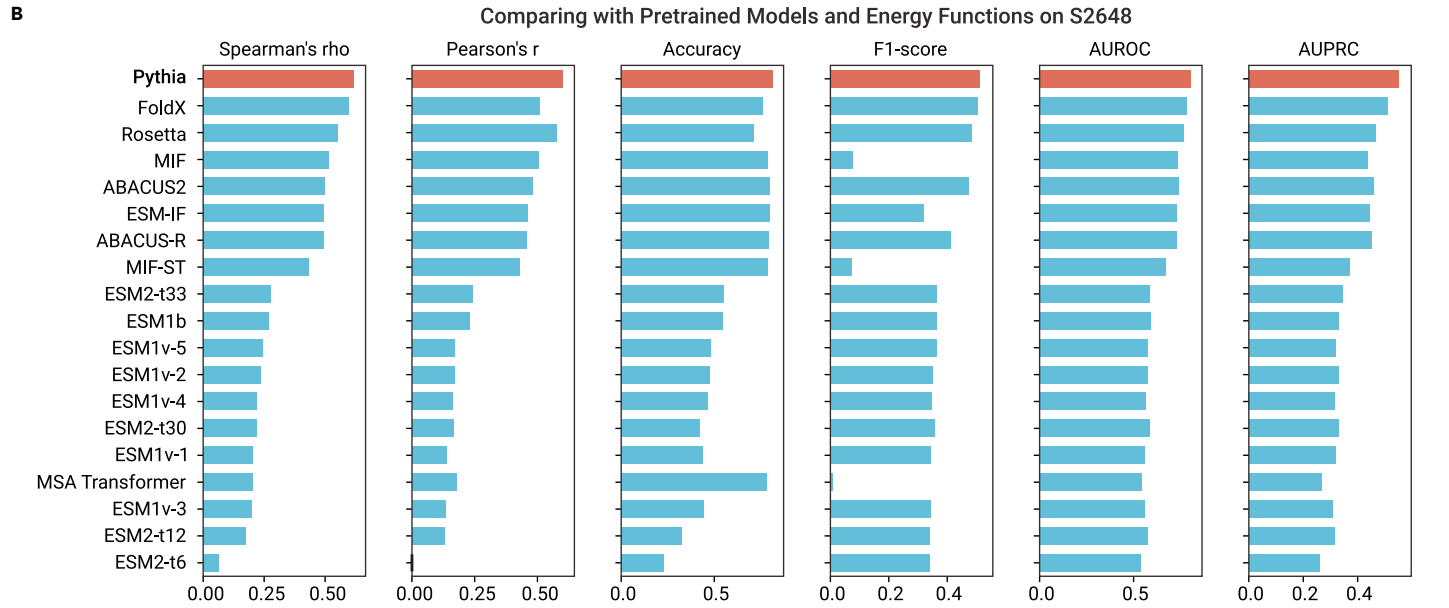
of alternative methods by a factor ranging from 625 to 50,000 on the same hardware (Figure 2F).

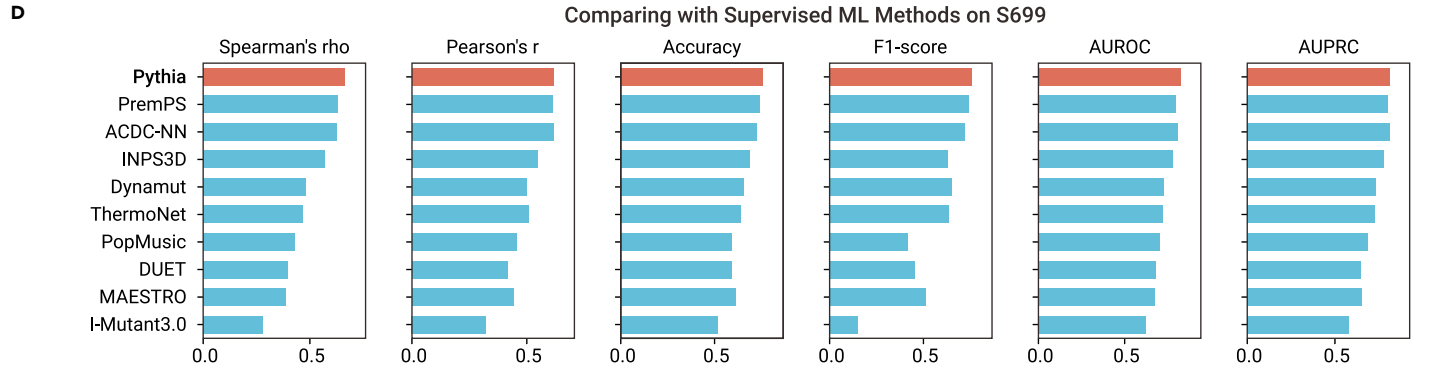
Evaluation of Pythia on a mega-scale dataset

Expanding the scope of our investigation, we applied predictive analytics to a mega-scale dataset of approximately 1 million mutations across 600 proteins including natural, redesigned, and hallucinated domains.[39] Performance was evaluated on 177,315 mutations within 181 well-characterized natural protein domains. The overall performance metrics indicated a Spearman's rho value of 0.602 and a Pearson's correlation coefficient (r) of 0.633 (Figure 3A), while the AUROC reached 0.83 (Figure 3B), and the AUPRC reached 0.88 in predicting the stabilizing potential of a mutation (Figure 3C). These results align closely with the performance metrics reported in S2648 and S669. Notably, of the 181 evaluated natural domains, 127 domains (approximately 70%) exhibited a Spearman's rho surpassing 0.6 (Figure 3D), indicating a relatively robust correlation.[64]

This compelling observation prompted a more granular exploration of domain-specific correlations. Unlike a holistic assessment across all point mutations, analyzing the correlation values for individual domains provides insights that are particularly beneficial for applications in protein engineering







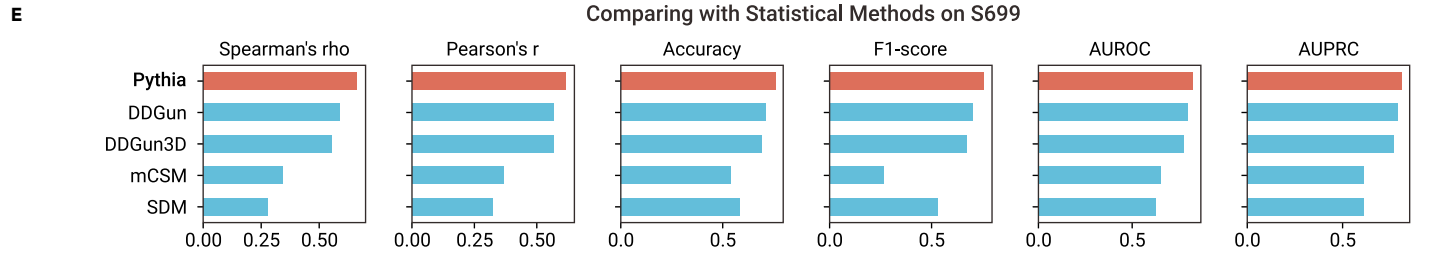


Figure 2. Evaluation of Pythia in predicting ΔΔG compared with current state-of-the-art methods (A) The density scatterplot for Pythia's predictions. (B) Parameters of the top 5 deep learning methods and their Spearman's rho on S2648. (C) Inference speeds of the top 6 methods ranked by Spearman's rho. (D) Comparisons of Pythia against pretrained models and energy functions. The correlation of the predicted values with experimental ΔΔG is indicated by Spearman's rho and Pearson's r, revealing the ranking and linear correlation. The metrics for classification tasks (accuracy, F1 score, AUROC, and AUPRC) categorize the stabilizing factor ($\Delta\Delta G_{folding} > 0$) using the S2648 dataset. (E) Comparisons of Pythia against nine supervised ML methods using the S669 dataset. (F) Comparisons of Pythia against four knowledge-based statistical methods using the S669 dataset. The top-performing method is highlighted red, and the remaining methods are highlighted blue.

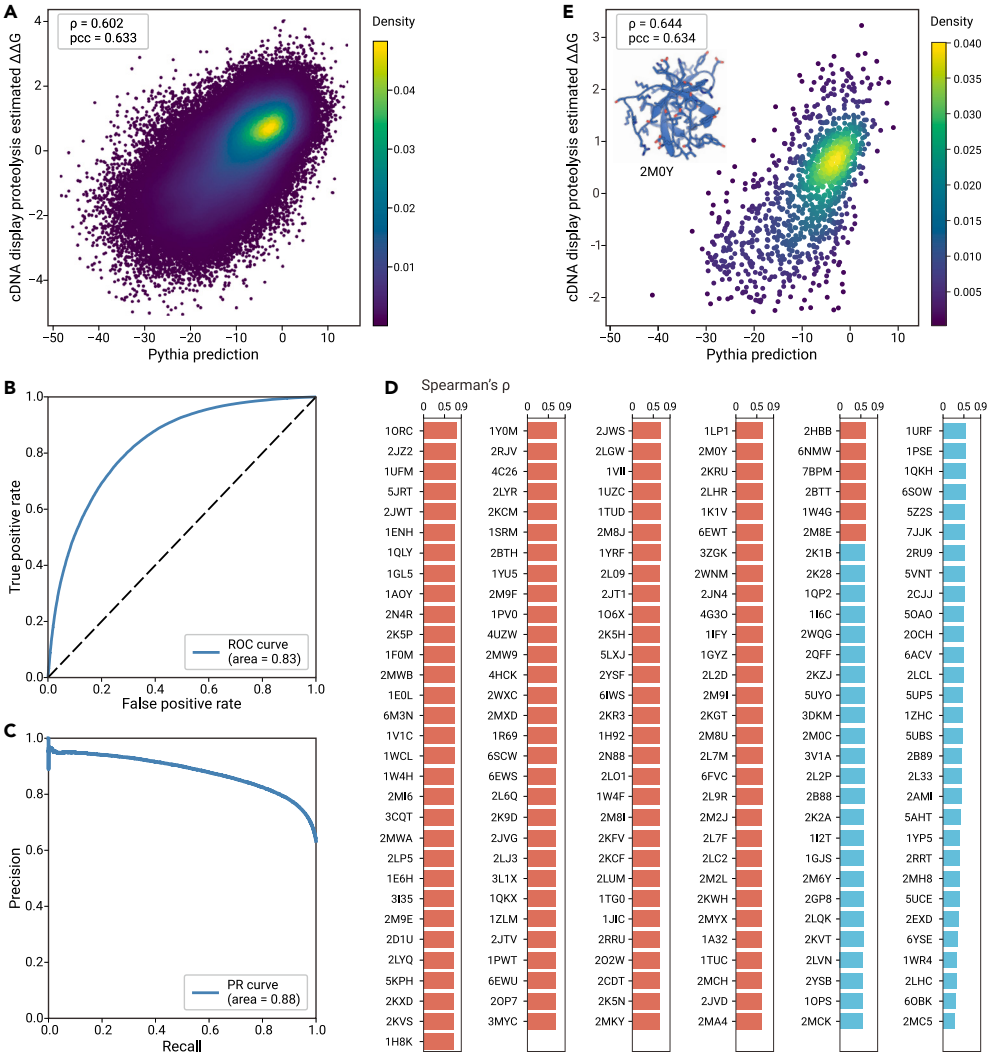


Figure 3. Validation of Pythia on the mega-scale dataset (A) Density scatterplot showcasing Pythia's predictions on the mega-scale dataset. (B) ROC curve illustrating Pythia's ability to classify stabilizing mutations. (C) PR curve highlighting Pythia's prediction accuracy in classifying stabilizing mutations. (D) The correlation between Pythia's predictions and the cDNA display proteolysis estimated ΔΔG is represented by Spearman's rho across all 181 domains. The Spearman's rho of prediction and measured values higher than 0.6 is colored red, otherwise colored blue. (E) Density scatterplot of Pythia's predictions for the structural domain of SH3 domain of DOCK180.

and mutation prediction. Through this domain-specific analysis, we attained a higher average Spearman's rho of 0.620. A noteworthy case emerged from our examination of the SH3 structural domain of DOCK180 in *Mus musculus* (PDB: 2M0Y), where the correlation between scores predicted by Pythia and the ΔΔG values derived from cDNA display proteolysis yielded a Spearman's rho of 0.644 (Figure 3E), positioning this result within the median range across the 181 tested domains.

In addition, we probed the influence of both AlphaFold2 models and ESMFold models on our predictions (Figure S3). The analysis revealed that models exhibiting higher predicted local distance difference test (pLDDT) scores are concomitantly more likely to produce elevated Spearman's rho values. Conversely, certain models with erroneous predictions were identified as having low pLDDT scores, accompanied by discrepancies between AlphaFold2 and ESMFold model outputs (Figures S4 and S5). Consistent with prior findings, our predictions generated from AlphaFold2 models either matched or surpassed those obtained from experimentally determined structural data.[65]

Identification of stabilizing mutations for a LEH

Encouraged by the superior generalization in predicting ΔΔG, we experimentally verified Pythia's predictions using the LEH from *Rhodococcus erythropolis* DCL14. This enzyme has been used widely in organic chemistry and has undergone extensive protein engineering, allowing direct comparisons between different strategies.[66] Generally, the current computational enzyme stabilizing process employs various predictors to nominate putative stabilizing mutations, followed by wet lab characterization (Section S5). We selected mutations with scores below −2 predicted by Pythia, prioritizing those with the lowest scores when multiple mutations were possible at a given position. This process led to 35 single-point mutations; 31 mutants yielded soluble expression in *Escherichia coli*, among which 17 mutations increased the protein's apparent melting temperature (T_m) (Figures 4A and 4B). Hybrid strategies employ visual inspections or molecular dynamic (MD) simulations to filter out unreasonable candidates predicted by energy function methods such as FoldX, thus improving the median ΔT_m from −1.80°C to −0.15°C. However, this requires a high level of technical expertise and hinders the widespread adoption of such a strategy. By comparison, Pythia improved the median ΔT_m to 0.80°C of all expressed mutants without any knowledge-based mutation selection (Figure 4C). Notably, the proportion of mutations with a T_m increase exceeding 1°C was significantly higher in Pythia's predictions compared with energy function methods, even with MD filtration (Figure 4D). Among these beneficial mutations, the P57A mutation, which is typically regarded as destabilizing in force field-based methods, exhibited the highest T_m increase of 8.8°C. Moreover, only 4 of the 17 beneficial mutations had been previously reported, highlighting Pythia's unique capability to identify stabilizing mutations that conventional methods may have overlooked. In light of this, Pythia is a promising tool to enhance the advancement of hybrid strategies, such as FireProt,[67] FRESCO, and GRAPE,[15] that integrate information from diverse complementary approaches to provide more options for the subsequent accumulation paths.

Structural interpretability of Pythia

Since Pythia employs an attention mechanism, we can leverage the attention scores learned by the model to explore whether it has effectively captured the intricate interactions within proteins. We visualized the attention scores for functional residues in molecular graphs from two distinct categories (Figure 5). The first instance examines the π-π interactions involving F52 and its neighboring residues within the GB1 domain (PDB: 1PGA). Nearby F52, four aromatic amino acids—Y3, F30, W43, and Y45—have the potential to form π-π interactions that stabilize the hydrophobic core of the domain (Figure 5A). Notably, Pythia assigns higher attention scores to these four amino acids along with the crucial F52, indicating the model's ability to recognize the significance of these interactions in comparison with other neighboring residues (Figure 5B).

In our analysis of pre- and postmutation structures, we focused on DuraPETase,[15] a more stable plastic-degrading enzyme engineered from *Is*PETase. Previous studies have highlighted the synergistic effect of D186 along with several stability-enhancing point mutations. We compared the attention scores assigned with the mutated residues surrounding D186 with those of their wild-type counterparts (Figure 5C). The results indicated that Pythia assigns higher attention scores to the mutated interactions, suggesting that the model is attuned to the structural implications of mutations and effectively captures the consequential relationships between mutated residues and their environments (Figures 5D and 5E).

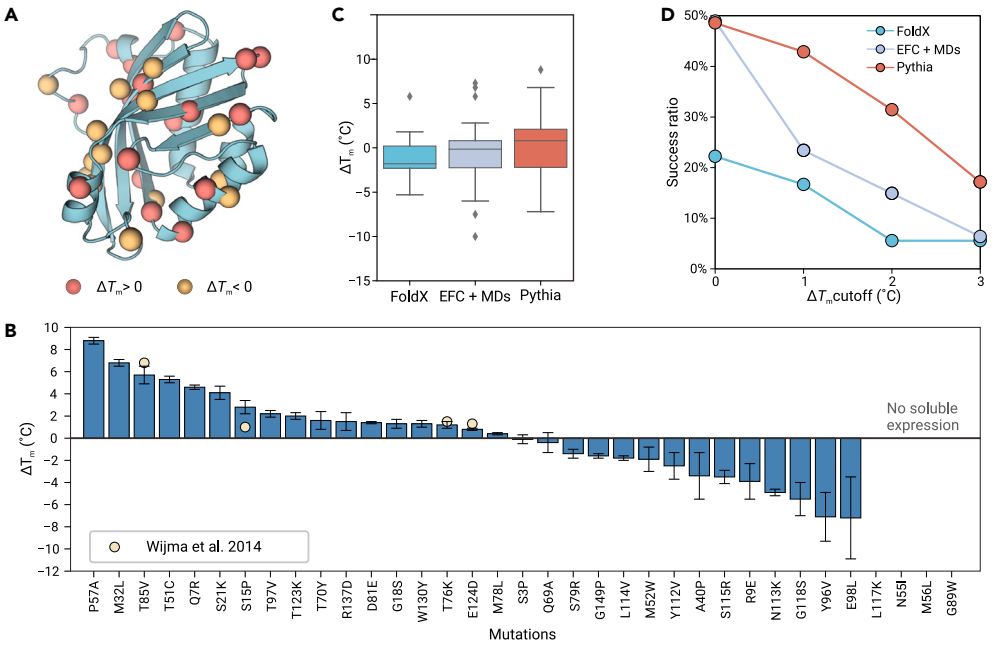


Figure 4. Experimental validation of Pythia's predictions (A) The monomer structure of LEH is rendered in cartoon form, with the C-alpha atom locations of mutations shown as spheres. Stabilizing mutations are represented with red spheres, while destabilizing mutations are shown in yellow. The visualization of protein structures was prepared using PyMOL. (B) The bar plot represents the measured ΔT_m of mutants characterized in this study (error bars represent standard deviation, n = 3 technical replicates). Bars accentuated with light blue dots present the stabilizing mutations, as reported by Wijma et al.[68] The Pearson's r between our measurement and the previous report is 0.895, and the RMSE is 1.094. (C) A boxplot comparing three different mutation prediction strategies. The central line within each box represents the median values of ΔT_m for that strategy. The top and bottom boundaries of the box represent the first and third quartiles, respectively. The height of the box represents the interquartile range (IQR). Data points outside of the 1.5 × IQR range are considered outliers and are plotted individually. (D) The success ratio of characterized mutations versus different ΔT_m cutoff values across three strategies. The blue curve depicts results using only FoldX. The gray curve represents the outcomes of energy function calculations supplemented with further MD filtration (EFC + MD simulations), corresponding to the FRESCO strategy's single-point mutation prediction component. The red curve demonstrates results achieved using only Pythia, in the absence of other selection criteria.

ΔΔG prediction at the protein universe scale

Several previous studies have established exemplary approaches for conducting large-scale mutation analyses across proteomes, yielding valuable insights into the potential of mutations that cause diseases,[66] influence fitness,[68] and predict ΔΔG values.[25] We further examined the prediction speed of Pythia at three different scales: (1) proteome scale, (2) annotated proteins, and (3) the protein universe. For the proteome-scale assessment, we utilized the proteome of *E. coli* K-12 as a representative example (Figure 6A). Pythia efficiently predicted all 25,189,782 mutations across 4,214 structures (with an averaged C-alpha pLDDT > 70) in only 3 min using a single NVIDIA GeForce RTX 4090.

Next, we expanded our analysis to encompass all mutation predictions for the 134,276 structures in SwissProt with a pLDDT score above 95. Remarkably, this extensive computational task was completed in approximately 2 h, scanning a total of 770,105,473 mutations. Finally, we processed all possible mutations for 26 million high-quality AlphaFold2 structures. Pythia completed the entire computation in 3 days using a machine equipped with 8 NVIDIA GeForce RTX 4090 GPUs. This clearly demonstrates Pythia's immense computational efficiency for large-scale mutation prediction.

In our preliminary analysis of millions of mutations derived from a uniform distribution, we observed that the average scores of amino acid substitutions correlated with the substitution scores in the BLOSUM62 matrix (Figure 6B). These findings align with previous research suggesting that stability plays a significant role in protein evolution; however, factors such as function, solubility, and aggregation also contribute to the evolutionary process.[44,69,70]

Moreover, we identified a significantly higher average mutation score in thermophilic proteins compared with nonthermophilic proteins, with a p value of 0.0 from the Mann-Whitney U test (Figure 6C). Although this difference is marginal, it suggests that sourcing stabilizing mutations from a thermostable scaffold may be more challenging, indicating a more constrained sequence space for thermophilic proteins.

Drawing upon a comprehensive dataset of mutational variations, we undertook an analysis into the role of residue type in influencing protein stability by comparing thermophilic and nonthermophilic proteins. A clear pattern emerged in the predicted mutations, indicating that smaller substituents (A or C) tend to be generally favorable. Conversely, substitutions involving aromatic rings (F, Y, and W) appear to be disadvantageous in thermophilic proteins (Figure 6D).

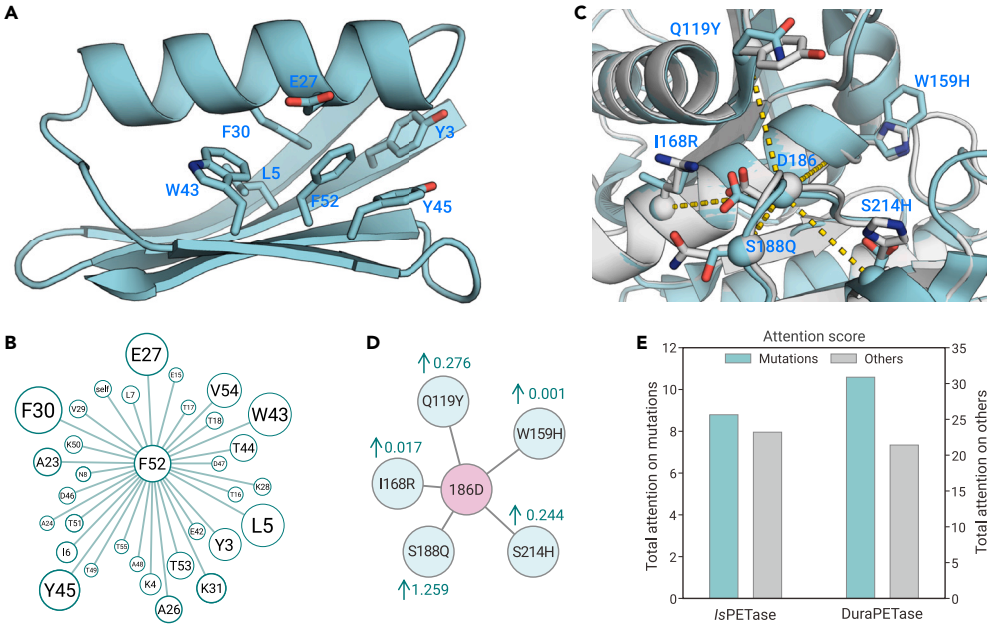


Figure 5. Interpretability of Pythia The attention score can be interpreted as a measure of the impact of amino acids in the environment on the central amino acid distribution (A) The π−π interactions of F52 with its neighboring residues in the GB1 domain (PDB: 1PGA). Possible π-π interactions are highlighted with blue dashed lines. (B) In the k-NN graph of the F52 in the GB1, the weights are assigned based on the attention score in the final AMPL. Bigger dots have higher attention with the central node. (C) A comparison between *Is*PETase (PDB: 5XH3) with the DuraPETase (PDB: 6KY5). *Is*PETase is green, and the DuraPETase is white. The side chain of mutation positions and the central node D186 are shown in sticks, and their C-alphas are displayed as spheres. (D) The change in attention scores between mutated residues surrounding D186 with their wild-type counterparts. (E) The sum of total attention scores at the five mutated positions and positions remains unchanged. Protein structure visualizations were prepared with PyMOL.

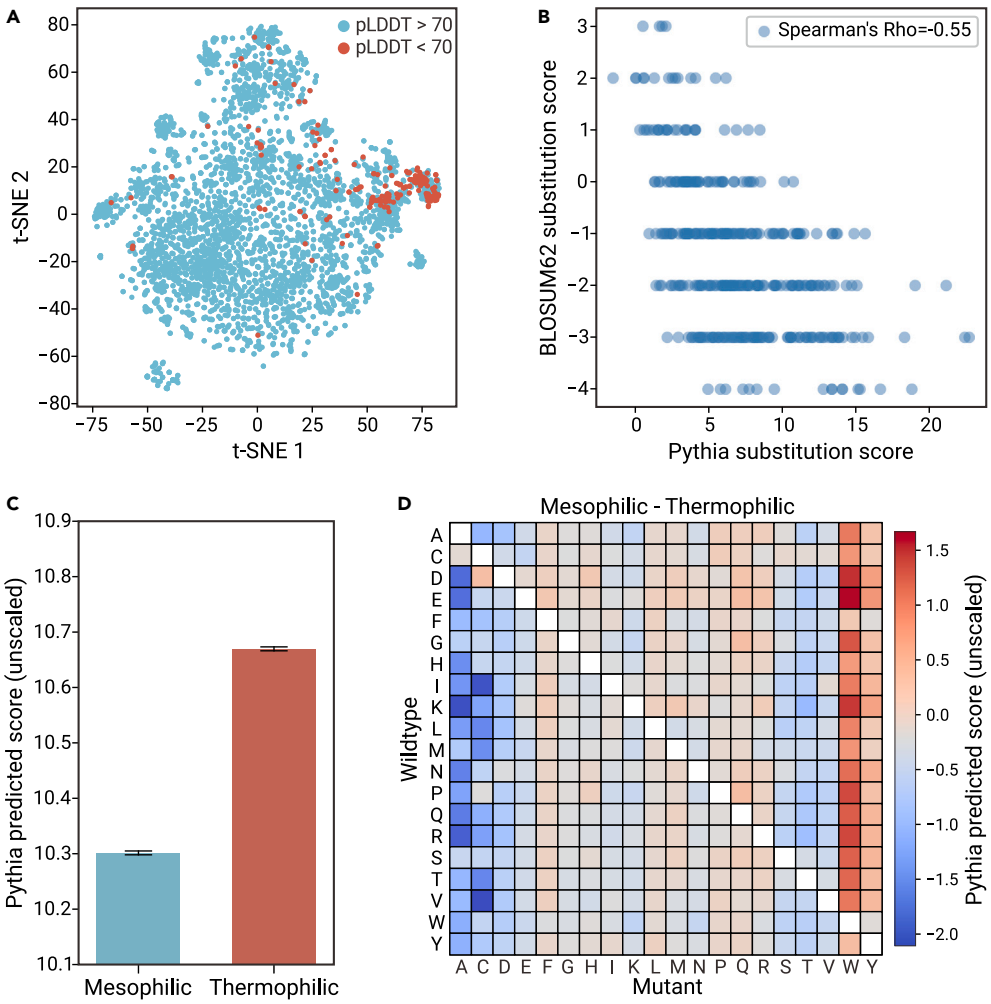


Figure 6. Insights gained from large-scale mutagenesis predictions (A) Visualization of the protein space with a t-SNE embedding of the *E. coli* K-12 proteome. Blue dots represent proteins with averaged C-alpha pLDDTs ≥ 70, while red dots represent proteins with averaged C-alpha pLDDTs < 70. (B) Scatterplot comparing amino acid substitution scores of Pythia and BLOSUM62. (C) Bar plot depicting the averages of all mutations. Mutations in thermotolerant proteins exhibit significantly higher Pythia scores (p = 0.0 in Mann-Whitney U test) compared with mutations in randomly sampled proteins (most likely non-thermostable). (D) The comparison of energetic effects caused by substitution between proteins derived from the mesophile and thermophile. Heatmap illustrating energy differences caused by various mutation types, with 380 mutation types color coded based on their average energy difference.

Utilizing the benefits of SSL, our investigation into large-scale protein mutations revealed intricate details that are often overlooked in isolated protein mutation studies.

DISCUSSION

The prediction of ΔΔG following mutations plays a crucial role in elucidating the effects of genetic variations on protein function and stability. Due to the limited availability of labeled ΔΔG data needed for deep learning, we introduce Pythia, an efficient approach tailored for zero-shot predictions that leverages the capabilities of SSL. Pythia's architecture enables the integration of both sequence and structural data, with a focus on the interactions between residues. It has learned to infer how the spatial arrangement of neighboring residues affects the probability of the central masked residue being a specific amino acid, thereby improving the accuracy of stability predictions. In addition, its attention weights provide valuable biological insights (Section S6), making interaction patterns interpretable and improving the model's explainability. The dual capability of assessing the likelihoods of amino acids at the central residue and explaining inter-residue interactions contributes to a deeper understanding of genomic variation and its implications for protein functions.

Comparative assessments demonstrate that Pythia outperforms other self-supervised pretraining models in correlating predictions with experimental ΔΔG values, achieving superior accuracy with the fewest parameters. In comparison with conventional energy calculation methods, Pythia not only delivers a modest improvement in prediction correlation but also boasts an extraordinary computational speed increase of up to 10^5-fold. This remarkable efficiency makes Pythia particularly well-suited for large-scale, high-throughput studies across extensive protein datasets. Comprehensive *in silico* benchmarks and *in vivo* experiments further validate Pythia as a robust and versatile tool for protein engineering.

A recent advancement in the field of protein mutation prediction is the introduction of the mega-scale dataset,[40] which has not only deepened our understanding of protein stability but has also provided valuable data for model development. Blaabjerg et al. evaluated Rasp[25] on a curated subset of the mega-scale dataset containing 164,524 mutants across 164 protein domains and achieved Pearson's r of 0.62. This result is comparable with the Pearson's r of 0.63 that Pythia obtained with the same dataset. Very recently, some models used mega-scale datasets for training and outperformed previous training methods in terms of various benchmarks[42,43] This vast amount of experimental data can be used to fine-tune Pythia for better prediction of protein domains that are currently underpredicted.

One notable limitation of Pythia is that the predicted values are not expressed in the physical unit of kcal/mol. This may restrict its application in situations where a physical unit is essential. We have partially addressed this limitation by calibrating the predictions with ΔΔG using the S2648 dataset (Section S3). Another constraint of Pythia is its dependence on predicted structures as the starting point. However, with 152 billion genetic variations predicted in this study, it appears feasible to integrate pretrained protein language models with the probabilities generated by Pythia. Such integration could enhance the accuracy and impartiality of sequence-based ΔΔG prediction of mutations.

MATERIALS AND METHODS

See supplemental information for details.

DATA AND CODE AVAILABILITY

The source code is available on GitHub (https://github.com/Wublab/Pythia.git) under the Apache-2.0 license. The source code of the web server is available upon request at https://pythia.wulab.xyz/. The ColabPythia is available at: https://colab.research.google.com/gist/JinyuanSun/83ff4323ff751dc665f96381a02df18a/colabpythia.ipynb. The structures that

Pythia used to make predictions and the predicted results involved in the benchmark are available along with the source code for both training and prediction at https://github.com/Wublab/Pythia. Preprocessed data required to train the Pythia from scratch are also included in the GitHub repository. The pdb_utils.py script in the GitHub repository can convert untreated PDB files to training data. For large-scale analysis, all computed mutations of the *E. coli* proteome, high-quality SwissProt structures, and thermophilic proteins used in the analysis can be found at https://zenodo.org/records/8231999.

REFERENCES

1. Dill, K.A. (1990). Dominant forces in protein folding. Biochemistry ***29***(31): 7133–7155. https://doi.org/10.1021/bi00483a001.
2. Bell, E.L., Finnigan, W., France, S.P., et al. (2021). Biocatalysis. Nat. Rev. Methods Primers ***1***(1): 46. https://doi.org/10.1038/s43586-021-00044-z.
3. Goldenzweig, A., and Fleishman, S.J. (2018). Principles of protein stability and their application in computational design. In Annu. Rev. Biochem., *87*, R.D. Kornberg, ed., pp. 105–129. https://doi.org/10.1146/annurev-biochem-062917-012102.
4. Grantcharova, V.P., and Baker, D. (1997). Folding dynamics of the src SH3 domain. Biochemistry ***36***(50): 15685–15692. https://doi.org/10.1021/bi971786p.
5. Liberles, D.A., Teichmann, S.A., Bahar, I., et al. (2012). The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. ***21***(6): 769–785. https://doi.org/10.1002/pro.2071.
6. Yue, P., Li, Z., and Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. J. Mol. Biol. ***353***(2): 459–473. https://doi.org/10.1016/j.jmb.2005.08.020.
7. Bloom, J.D., Labthavikul, S.T., Otey, C.R., et al. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. USA ***103***(15): 5869–5874. https://doi.org/10.1073/pnas.0510098103.
8. Pyser, J.B., Chakrabarty, S., Romero, E.O., et al. (2021). State-of-the-art biocatalysis. ACS Cent. Sci. ***7***(7): 1105–1116. https://doi.org/10.1021/acscentsci.1c00273.
9. Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., et al. (2012). Engineering the third wave of biocatalysis. Nature ***485***(7397): 185–194. https://doi.org/10.1038/nature11117.
10. Cui, Y., Sun, J., and Wu, B. (2022). Computational enzyme redesign: large jumps in function. Trends Chem. ***4***(5): 409–419. https://doi.org/10.1016/j.trechm.2022.03.001.
11. Mazurenko, S., Prokop, Z., and Damborsky, J. (2020). Machine learning in enzyme engineering. ACS Catal. ***10***(2): 1210–1223. https://doi.org/10.1021/acscatal.9b04321.
12. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., et al. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theor. Comput. ***13***(6): 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.
13. Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. J. Mol. Biol. ***320***(2): 369–387. https://doi.org/10.1016/s0022-2836(02)00442-4.
14. Xiong, P., Hu, X., Huang, B., et al. (2020). Increasing the efficiency and accuracy of the ABACUS protein sequence design method. Bioinformatics ***36***(1): 136–144. https://doi.org/10.1093/bioinformatics/btz515.
15. Cui, Y., Chen, Y., Liu, X., et al. (2021). Computational redesign of a PETase for plastic biodegradation under ambient condition by the GRAPE strategy. ACS Catal. ***11***(3): 1340–1350. https://doi.org/10.1021/acscatal.0c05126.
16. Montanucci, L., Capriotti, E., Frank, Y., et al. (2019). DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC Bioinf. ***20***(1): 335. https://doi.org/10.1186/s12859-019-2923-1.
17. Folkman, L., Stantic, B., Sattar, A., et al. (2016). EASE-MM: Sequence-based prediction of mutation-induced stability changes with feature-based multiple models. J. Mol. Biol. ***428***(6): 1394–1405. https://doi.org/10.1016/j.jmb.2016.01.012.
18. Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. ***33***: W306–W310. https://doi.org/10.1093/nar/gki375.
19. Quan, L., Lv, Q., and Zhang, Y. (2016). STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics ***32***(19): 2936–2946. https://doi.org/10.1093/bioinformatics/btw361.
20. Wang, S., Tang, H., Zhao, Y., et al. (2022). BayeStab: Predicting effects of mutations on protein stability with uncertainty quantification. Protein Sci. ***31***(11): e4467. https://doi.org/10.1002/pro.4467.
21. Li, B., Yang, Y.T., Capra, J.A., et al. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput. Biol. ***16***(11): e1008291. https://doi.org/10.1371/journal.pcbi.1008291.
22. Chen, Y., Lu, H., Zhang, N., et al. (2020). PremPS: Predicting the impact of missense mutations on protein stability. PLoS Comput. Biol. ***16***(12): e1008543. https://doi.org/10.1371/journal.pcbi.1008543.
23. Wang, S., Tang, H., Shan, P., et al. (2023). ProS-GNN: Predicting effects of mutations on protein stability using graph neural networks. Comput. Biol. Chem. ***107***: 107952. https://doi.org/10.1016/j.compbiolchem.2023.107952.
24. Zhou, Y., Pan, Q., Pires, D.E.V., et al. (2023). DDMut: predicting effects of mutations on protein stability using deep learning. Nucleic Acids Res. ***51***(W1): W122–W128. https://doi.org/10.1093/nar/gkad472.
25. Blaabjerg, L.M., Kassem, M.M., Good, L.L., et al. (2023). Rapid protein stability prediction using deep learning representations. Elife ***12***: e82593. https://doi.org/10.7554/eLife.82593.
26. Pancotti, C., Benevenuta, S., Birolo, G., et al. (2022). Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Briefings Bioinf. ***23***(2): bbab555. https://doi.org/10.1093/bib/bbab555.
27. Benevenuta, S., Pancotti, C., Fariselli, P., et al. (2021). An antisymmetric neural network to predict free energy changes in protein variants. J. Phys. D Appl. Phys. ***54***(24): 245403. https://doi.org/10.1088/1361-6463/abedfb.
28. Iqbal, S., Li, F., Akutsu, T., et al. (2021). Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. Briefings Bioinf. ***22***(6): bbab184. https://doi.org/10.1093/bib/bbab184.
29. Yu, T., Cui, H., Li, J.C., et al. (2023). Enzyme function prediction using contrastive learning. Science ***379***(6639): 1358–1363. https://doi.org/10.1126/science.adf2465.
30. Geiping, J., Garrido, Q., Fernandez, P., et al. (2023). A cookbook of self-supervised learning. Preprint at arXiv. https://doi.org/10.48550/arXiv:2304.12210.
31. Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv:1810.04805.
32. Alley, E.C., Khimulya, G., Biswas, S., et al. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods ***16***(12): 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.
33. Rao, R., Liu, J., Verkuil, R., et al. (2021). MSA Transformer. In International Conference on Machine Learning (ICML).
34. Yang, K.K., Zanichelli, N., and Yeh, H. (2023). Masked inverse folding with sequence transfer for protein representation learning. Protein Eng. Des. Sel. ***36***: gzad015. https://doi.org/10.1093/protein/gzad015.
35. Meier, J., Rao, R., Verkuil, R., et al. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. In 35th Annual Conference on Neural Information Processing Systems (NeurIPS).
36. Strokach, A., Becerra, D., Corbi-Verge, C., et al. (2020). Fast and flexible protein design using deep graph neural networks. Cell Syst. ***11***(4): 402–411.e4. https://doi.org/10.1016/j.cels.2020.08.016.
37. Liu, Y., Zhang, L., Wang, W., et al. (2022). Rotamer-free protein sequence design based on deep learning and self-consistency. Nat. Comput. Sci. ***2***(7): 451–462. https://doi.org/10.1038/s43588-022-00273-6.
38. Boomsma, W., and Frellsen, J. (2017). Spherical convolutions and their application in molecular modelling. In 31st Annual Conference on Neural Information Processing Systems (NIPS).
39. Tsuboyama, K., Dauparas, J., Chen, J., et al. (2023). Mega-scale experimental analysis of protein folding stability in biology and design. Nature ***620***(7973): 434–444. https://doi.org/10.1038/s41586-023-06328-6.
40. Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods ***15***(10): 816–822. https://doi.org/10.1038/s41592-018-0138-4.
41. Diaz, D.J., Gong, C., Ouyang-Zhang, J., et al. (2024). Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. Nat. Commun. ***15***(1): 6170. https://doi.org/10.1038/s41467-024-49780-2.
42. Dieckhaus, H., Brocidiacono, M., Randolph, N.Z., et al. (2024). Transfer learning to leverage larger datasets for improved prediction of protein stability changes. Proc. Natl. Acad. Sci. USA ***121***(6): e2314853121. https://doi.org/10.1073/pnas.2314853121.
43. Xu, Y., Liu, D., and Gong, H. (2024). Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy. Nat. Comput. Sci. ***4***(11): 840–850. https://doi.org/10.1038/s43588-024-00716-2.
44. DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: A biophysical view of protein evolution. Nat. Rev. Genet. ***6***(9): 678–687. https://doi.org/10.1038/nrg1672.
45. Varadi, M., Anyango, S., Deshpande, M., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. ***50***(D1): D439–D444. https://doi.org/10.1093/nar/gkab1061.
46. Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci. ***15***(11): 2507–2524. https://doi.org/10.1110/ps.062416606.
47. Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature ***596***(7873): 583–589. https://doi.org/10.1038/s41586-021-03819-2.
48. Dauparas, J., Anishchenko, I., Bennett, N., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science ***378***(6615): 49–56. https://doi.org/10.1126/science.add2187.
49. Gilmer, J., Schoenholz, S.S., Riley, P.F., et al. (2017). Neural Message Passing for Quantum Chemistry. In 34th International Conference on Machine Learning.
50. Orengo, C.A., Michie, A.D., Jones, S., et al. (1997). CATH - a hierarchic classification of protein domain structures. Structure ***5***(8): 1093–1108. https://doi.org/10.1016/s0969-2126(97)00260-8.
51. Burley, S.K., Berman, H.M., Bhikadiya, C., et al. (2019). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. ***47***(D1): D464–D474. https://doi.org/10.1093/nar/gky1004.
52. Dehouck, Y., Grosfils, A., Folch, B., et al. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics ***25***(19): 2537–2543. https://doi.org/10.1093/bioinformatics/btp445.

53. Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science ***379***(6637): 1123–1130. https://doi.org/10.1126/science.ade2574.
54. Weinstein, E.N., Amin, A.N., Frazer, J., et al. (2022). Non-identifiability and the blessings of misspecification in models of molecular fitness. In 36th Conference on Neural Information Processing Systems (NeurIPS).
55. Fang, J. (2023). The role of data imbalance bias in the prediction of protein stability change upon mutation. PLoS One ***18***(3): e0283727. https://doi.org/10.1371/journal.pone.0283727.
56. Savojardo, C., Fariselli, P., Martelli, P.L., et al. (2016). INPS-MD: a web server to predict stability of protein variants from sequence and structure. Bioinformatics ***32***(16): 2542–2544. https://doi.org/10.1093/bioinformatics/btw192.
57. Rodrigues, C.H., Pires, D.E., and Ascher, D.B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res. ***46***(W1): W350–W355. https://doi.org/10.1093/nar/gky300.
58. Dehouck, Y., Kwasigroch, J.M., Gilis, D., et al. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinf. ***12***: 151. https://doi.org/10.1186/1471-2105-12-151.
59. Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. ***42***(W1): W314–W319. https://doi.org/10.1093/nar/gku411.
60. Laimer, J., Hiebl-Flach, J., Lengauer, D., et al. (2016). MAESTROweb: a web server for structure-based protein stability prediction. Bioinformatics ***32***(9): 1414–1416. https://doi.org/10.1093/bioinformatics/btv769.
61. Capriotti, E., Fariselli, P., Rossi, I., et al. (2008). A three-state prediction of single point mutations on protein stability changes. BMC Bioinf. ***9***: S6. https://doi.org/10.1186/1471-2105-9-s2-s6.
62. Pires, D.E.V., Ascher, D.B., and Blundell, T.L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics ***30***(3): 335–342. https://doi.org/10.1093/bioinformatics/btt691.
63. Worth, C.L., Preissner, R., and Blundell, T.L. (2011). SDM-a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. ***39***: W215–W222. https://doi.org/10.1093/nar/gkr363.
64. Akoglu, H. (2018). User's guide to correlation coefficients. Turk. J. Emerg. Med. ***18***(3): 91–93. https://doi.org/10.1016/j.tjem.2018.08.001.
65. Akdel, M., Pires, D.E.V., Pardo, E.P., et al. (2022). A structural biology community assessment of AlphaFold2 applications. Nat. Struct. Mol. Biol. ***29***(11): 1056–1067. https://doi.org/10.1038/s41594-022-00849-w.
66. Wijma, H.J., Floor, R.J., Jekel, P.A., et al. (2014). Computationally designed libraries for rapid enzyme stabilization. Protein Eng. Des. Sel. ***27***(2): 49–58. https://doi.org/10.1093/protein/gzt061.
67. Bednar, D., Beerens, K., Sebestova, E., et al. (2015). FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants. PLoS Comput. Biol. ***11***(11): e1004556. https://doi.org/10.1371/journal.pcbi.1004556.
68. Cheng, J., Novati, G., Pan, J., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science ***381***(6664): eadg7492. https://doi.org/10.1126/science.adg7492.
69. Bastolla, U., Dehouck, Y., and Echave, J. (2017). What evolution tells us about protein physics, and protein physics tells us about evolution. Curr. Opin. Struct. Biol. ***42***: 59–66. https://doi.org/10.1016/j.sbi.2016.10.020.
70. Echave, J., and Wilke, C.O. (2017). Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. In Annu. Rev. Biophys., *46*, K.A. Dill, ed., pp. 85–103. https://doi.org/10.1146/annurev-biophys-070816-033819.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (grant no. 2023YFA0916000), the National Natural Science Foundation of China (32225002, 32170033, and 32422001), the Key Research Program of Frontier Sciences (ZDBS-LY-SM014), the Biological Resources Program (KFJ-BRP-009 and KFJ-BRP-017-58) from the Chinese Academy of Sciences, the Informatization Plan of Chinese Academy of Sciences (CAS-WX2021SF-0111), and the Youth Innovation Promotion Association CAS (2022086). The authors acknowledge the support of the Huawei MindSpore team.

AUTHOR CONTRIBUTIONS

J.S. developed the model, conducted the analysis, and wrote the web server software. T.Z. did the wet lab experiments and related data analysis. Y.C. and B.W. supervised the research. All authors discussed the results, and wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.1016/j.xinn.2024.100750.

LEAD CONTACT WEBSITE

https://www.im.cas.cn/jgsz2018/yjtx/gywswyswjsyjs/201911/t20191113_5430831.html