Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy

Yunxin Xu[1,2], Di Liu[1,2], and Haipeng Gong[1,2,*]

[1]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, 100084, China

[2]Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, 100084, China

[*]To whom correspondence should be addressed: hgong@tsinghua.edu.cn

ABSTRACT

Accurate prediction of the fitness and stability of a protein upon mutations is of high importance in protein engineering and design. Despite the rapid development of deep learning techniques and accumulation of experimental data, the multi-labeled nature of fitness data hinders the training of robust deep-learning-based models for the fitness and stability prediction tasks. Here, we propose three geometric-learning-based models, GeoFitness, GeoDDG and GeoDTm, for the prediction of the fitness score, $\Delta\Delta G$ and ΔT_m of a protein upon mutations, respectively. In the optimization of GeoFitness, we designed a novel loss function to allow supervised training of a unified model using the large amount of multi-labeled fitness data in the deep mutational scanning (DMS) database. By this means, GeoFitness efficiently learns the general functional effects of protein mutations and achieves better performance over the other state-of-the-art methods. To further improve the downstream tasks of $\Delta\Delta G$/ΔT_m prediction, we re-utilized the encoder of GeoFitness as a pre-trained module in GeoDDG and GeoDTm to overcome the challenge of lack of sufficient amount of specifically labeled data. This pre-training strategy in combination with data expansion remarkably improves model performance and generalizability. When evaluated on the benchmark test sets (S669 for $\Delta\Delta G$ prediction and a newly collected set S571 for ΔT_m prediction), GeoDDG and GeoDTm outperform the other state-of-the-art methods by at least 35% and 70%, respectively, in terms of the Spearman correlation coefficient between predicted and experimental values. An online server for the suite of these three predictors, GeoStab-suite, is available at http://structpred.life.tsinghua.edu.cn/server_geostab.html.

Keywords protein stability · fitness prediction · mutation effects · geometric learning · pre-training

Introduction

Protein fitness is defined as the ability of a protein to perform a specific function, but is frequently quantified by different indicators (*e.g.*, enzyme activity, peptide binding affinity and protein stability)[1] in various experimental cases. One of the main goals of protein design and engineering is to improve the protein fitness so as to enhance protein performance in biotechnological and biopharmaceutical processes[2,3]. Among the various indicators of protein fitness, the protein stability[4] is of high interest, which is commonly evaluated by two metrics, ΔG and T_m. ΔG denotes the unfolding free energy change at room temperature and describes the thermodynamic stability of a protein, while T_m stands for the protein melting temperature and reflects the ability of a protein to maintain its folded state against temperature fluctuations[5,6]. Understanding protein stability not only facilitates the acquisition of engineered proteins with sustained activities in harsh bioprocesses and/or under manufacturing conditions[7], but also aids in the elucidation of the role of human genetic variants in the etiology of diverse diseases in the field of biomedicine[8]. Indeed, much attention has been paid to learn the changes of fitness (and particularly the stability) of proteins upon mutations. For instance, the deep mutational scanning (DMS) database[9] contains over 3,000,000 pieces of protein fitness data, which were measured using various experimental indicators upon mutations. The experimental strategy is, however, highly costly and labor-intensive, considering the excessively large number of possible protein sequences even for single-point mutations[10–12]. Therefore, developing computational tools for fast and accurate prediction of the mutation effects on protein fitness is in urgent demand.

The protein fitness prediction methods can be developed and optimized based on the DMS database. However, the multi-labeled nature of DMS data hinders the training of a unified prediction model (Figure 1). Existing models[1,3,4,12–30] mainly take two strategies to circumvent this obstacle. One strategy taken by methods like ECNet[16] and SESNet[15] is to train a specific model based on each definition of fitness in the concerned issue. Therefore, additional fitness data with the same corresponding meaning to the interested protein property are required to retrain these models before the practical prediction. This strategy leads to satisfying performance but sacrifices the model usability. Another strategy is to train the model with the masked language modeling objectives, like in $\mathrm{RF}_{\mathrm{joint}}$[22,23], MSA transformer[18], ESM-1b[24], ESM-1v[17] and ESM-2[25], which completely bypasses the need of any fitness data. This strategy can produce unified models for fast prediction but at the cost of model performance. Presently, there still lacks a robust method to generate a unified fitness prediction model with satisfying performance.

Different from the multi-labeled fitness data, the protein stability changes upon mutations are clearly defined by two metrics, $\Delta\Delta G$ and ΔT_m, and the accumulation of experimental data allows the development of corresponding prediction algorithms. Recently, great efforts have been exerted in $\Delta\Delta G$ prediction[31–49]. Current methods can be mainly classified as mechanistic predictors, machine learning predictors, and deep learning predictors[50]. Mechanistic predictors are based on protein energy or statistical potentials, like DDGun and DDGun3D[46], while machine learning predictors, including INPS-Seq[39], I-Mutant3.0-Seq[40], PremPS[41], *etc.*, feed features extracted from protein sequences or structures into various machine learning models for prediction. With the rapid development of deep learning techniques, deep learning models have shown great potential in the accurate prediction of protein stability, exemplified by ACDC-NN-Seq[38], ACDC-NN[35] and ThermoNet[37]. Unfortunately, these methods have not demonstrated absolute advantages over traditional methods when evaluated on a newly released benchmark dataset, S669[51]. The critical issue lies in the limited amount of training data ($\sim 10,000$), which easily leads to overfitting for deep learning models with millions of parameters. Moreover, available experimental data is biased toward several hotspot protein families, specific target amino acid mutations and destabilizing mutations[6,52]. One manifestation of such biases is that most current predictors are unable to reproduce the anti-symmetry property of $\Delta\Delta G$, *i.e.* inversion of the value upon the swapping of the statuses of mutant and wild type[52].

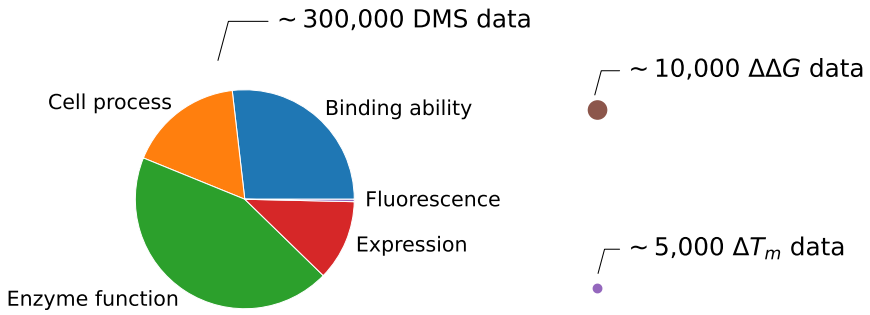


Figure 1. Summary on the DMS, $\Delta\Delta G$, and ΔT_m data. DMS data are multi-labeled and could be roughly divided into 5 categories. More detailed classification of DMS data utilized in the model training of this work could be found in Table S1. Unlike the DMS data, $\Delta\Delta G$, and ΔT_m data are single-labeled but have significantly reduced scales, with their data sizes approximately indicated by the radii of circles. In general, although the DMS data adopted in this work have no overlap with the $\Delta\Delta G$/ΔT_m data, the size of the former is about 20 times of the size of latter ones in combination.

ΔT_m prediction is relatively less explored when compared with $\Delta\Delta G$ prediction. Existing models[10,53–58], like AUTOMUTE[54,58] and HoTMuSiC[55], utilize physicochemical properties or statistical potentials as features and integrate with machine-learning-based methods for prediction. To our knowledge, deep-learning-based methods are absent in this field, possibly due to the significantly reduced amount of available ΔT_m data in comparison with $\Delta\Delta G$, despite the higher potential impact in practice for the prediction of ΔT_m[6]. Besides the same challenges confronted in the $\Delta\Delta G$ prediction, another dilemma in the ΔT_m prediction arises from the lack of established open-source benchmark datasets, which hinders the fair comparison of model performance.

In this work, we first developed a geometric-learning-based model, GeoFitness, for the prediction of protein fitness. Specifically, we designed a novel loss function to allow the training of a unified model with the multi-labeled fitness data in the DMS database. The model derived by this means avoids the prior limitation of model retraining before practical usage and simultaneously achieves a slightly improved performance over the other state-of-the-art methods like ECNet. Furthermore, by reutilizing the geometric encoder of GeoFitness, we developed two additional downstream models, GeoDDG and GeoDTm, to predict $\Delta\Delta G$ and ΔT_m of a protein upon mutations, respectively, with the model architecture specifically designed to ensure the anti-symmetry of prediction results. Notably, we addressed the challenge of limited data in $\Delta\Delta G$ and ΔT_m predictions by two strategies: expanding the training data by data collection as well as inheriting the geometric encoder of the GeoFitness model pre-trained on the DMS database. Particularly, the latter strategy remarkably improved the model performance and generalizability, considering the fact that fitness data of protein variants outnumber those of $\Delta\Delta G$ and ΔT_m by at least one order of magnitude as well as the relevance between protein fitness and stability in biology (Figure 1). When evaluated on the benchmark test sets, S669[51] for $\Delta\Delta G$ and S571 (a newly collected set in this work) for ΔT_m predictions, GeoDDG and GeoDTm outperform the other state-of-the-art methods by at least 35% and 70%, respectively, in terms of the Spearman correlation coefficient between predicted and experimental values.

Methods

Overview of the prediction models

The GeoStab-suite developed in this work comprises three distinct software programs, namely GeoFitness, GeoDDG, and GeoDTm, all of which aggregate the information from both protein sequence and structure into a geometric-learning-based encoder for prediction. The Geometric Encoder adopts a graph attention (GAT) neural network architecture, where the nodes (1D) represent amino acid residues and edges (2D) reflect inter-residue interactions (Figure 2a). GeoFitness is a unified model capable of predicting fitness landscape of protein variants of single mutations (Figure 2b). GeoDDG and GeoDTm reutilize the pre-trained information extractor of GeoFitness to predict the $\Delta\Delta G$ and ΔT_m values of proteins upon arbitrary mutations, respectively (Figure 2c). The protein structure information could be either derived from the Protein Data Bank (PDB) or predicted purely based on sequence using AlphaFold2[59]. Consequently, we have trained two versions of GeoDDG and GeoDTm, with the suffixes of "-3D" and "-Seq" to annotate the version that relies on experimental structures and that only needs sequence information in practical usage, respectively.

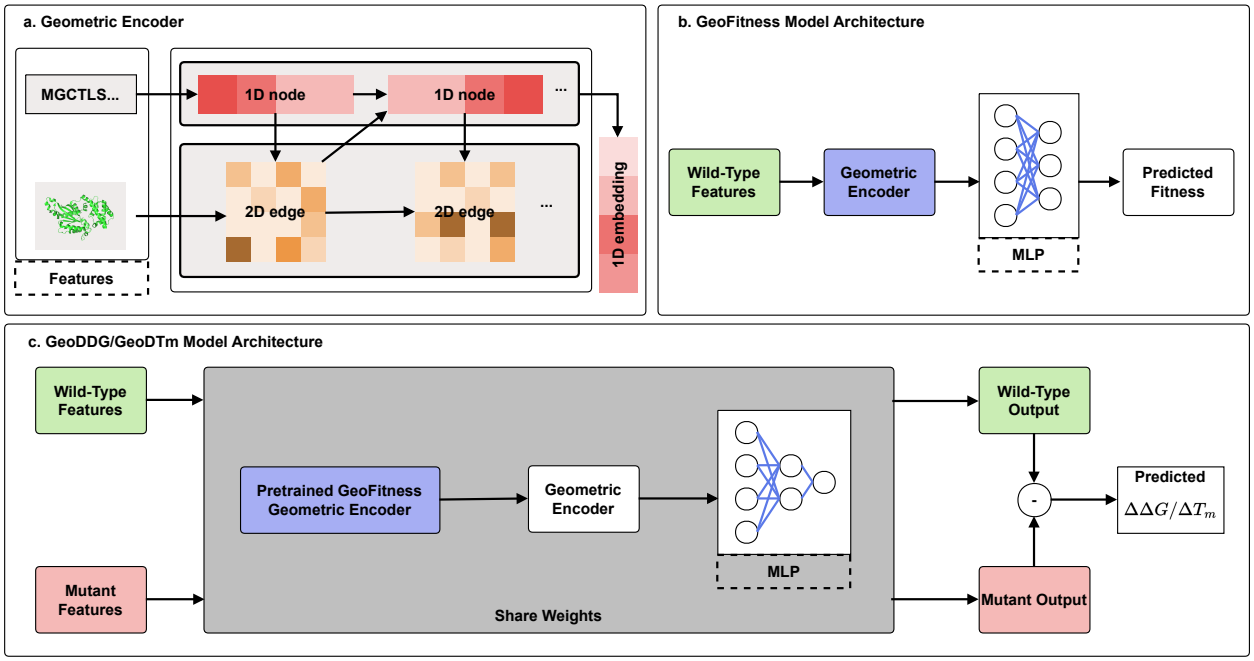


Figure 2. Schematic overview of the model architecture. a) The input of Geometric Encoder consists of both sequence and structural information, which initialize the node and edge embeddings, respectively. After alternate updating of nodes and edges in an N-layer GAT network, the final embedding is generated. **b)** The GeoFitness model takes the features of wild-type protein sequence and structure as input, and outputs the predicted fitness value through the Geometric Encoder and multi-layer-perceptron (MLP) layers. **c)** The model architectures of GeoDDG and GeoDTm are similiar. Features of wild-type and mutant are processed using the same encoder (with shared weights) and their differences are taken as the prediction results for $\Delta\Delta G/\Delta T_m$. Notably, the encoder reutilizes the pre-trained network module of GeoFitness to facilitate information processing, by inheriting the network architecture and parameter values: the modules colored violet in (**b**) and (**c**) are identical.

Protein fitness database and dataset

MaveDB

MaveDB is an open-source database designed to store data of large-scale mutation effects, which was first published in 2019 and had been updated to version 2 in 2021[60,61]. MaveDB includes over 300 datasets related to protein fitness, from which we chose 52 DMS studies on single-point mutation effects. The details of the datasets are summarized in Table S1.

DeepSequence Dataset

We also filtered DMS data from the DeepSequence dataset[1] by collecting single-point mutation data but excluding the portions redundant with MaveDB. In this way, we obtained 22 DMS datasets. The details of the datasets are summarized in Table S1.

Division of training, validation and test sets

The DMS data used in this work contain approximately 300,000 entries from 74 proteins in total (Table S1). For each protein, all available DMS data were randomly assigned as training, validation and testing with a ratio of 7:1:2. In order to reproduce the results of ECNet and SESNet, a specific model was optimized using the corresponding training data for each individual protein. The training data of all proteins were further combined as a comprehensive training set for GeoFitness. Notably, the DMS datasets used here have no overlap in labels with the databsets of protein stability, particularly with the benchmark test sets of $\Delta\Delta G$/ΔT_m (see Figure S1). Consequently, model pre-training using the plethora of DMS data brings no information leak to the downstream prediction tasks of $\Delta\Delta G$ and ΔT_m as well as the performance evaluation.

Protein stability database and dataset

ProThermDB

ProThermDB, a thermodynamic database for protein mutants, is an upgraded version of its predecessor ProTherm database that was first released in 1999 and then continuously updated until 2006[62–67]. As a refined version, ProThermDB[68] contains approximately 31,500 pieces of protein stability data with annotation errors in ProTherm corrected.

ThermoMutDB

ThermoMutDB is a manually curated thermodynamic database for protein mutants, consisting of two parts of data. The first part is derived from the 1,902 literatures related to the ProTherm database and all data have been double-checked and validated. The second part contains new thermodynamic mutation data collected from approximately 34,000 literatures. The database contains about 15,000 pieces of protein stability data[69].

S2648

The S2648 dataset[70] is a $\Delta\Delta G$ dataset consisting of 2,648 single-point mutations of 131 proteins collected from ProTherm database. It has been widely utilized as a training set in prior softwares for $\Delta\Delta G$ prediction and is considered the most extensively employed dataset of its kind.

S669

The S669 dataset[51] is made up with 669 single-point mutations of 94 proteins, which have < 25% sequence similarity with those in S2648 and VariBench[71]. As many softwares for $\Delta\Delta G$ prediction have previously utilized S2648 or VariBench for model training, S669 is generally considered as a fair benchmark test set for the performance evaluation of $\Delta\Delta G$ prediction methods.

S783

S783 is a subset of S971, which was first introduced by Chen et al.[41] and has long been a recognized test dataset for $\Delta\Delta G$ prediction models trained on S2648. Performance of GeoDDG is also tested on this dataset to fairly compare with the existing models. We filtered redundant entries from S971 and got S783 as the test set for several reasons. First, there are indeed overlapped entries between S921 and S2648, which needs exclusion. Second, as GeoDDG is designed to treat the wild-type and mutant proteins simultaneously at the same position, the reverse mutation involved in S971 is actually the same entry with its direct mutation and thus can be filtered out.

S8754

The S8754 dataset, composed of 8,754 single-point mutations of 301 proteins, is generated in this work as the training set for the $\Delta\Delta G$ prediction. We collected data from two $\Delta\Delta G$ databases, ProThermDB and ThermoMutDB, cautiously cleaned each individual piece of data, combined non-overlapping parts, and then manually verified the overlapping parts to guarantee the uniqueness and high level of credibility for each entry. Notably, we replenished the critical experimental conditions including pH and temperature for each entry of $\Delta\Delta G$ data. After removing data redundant with the test set (*i.e.* sequence identity > 25% with the S669 dataset), we finally constructed the largest $\Delta\Delta G$ dataset at present for model training.

S1626

The S1626 dataset[5] is a ΔT_m dataset consisting of 1,626 single-point mutations of 95 proteins collected from the ProTherm database. It has been widely employed as the training set in prior softwares for ΔT_m prediction and is considered the most extensively employed dataset of its kind.

S571

The S571 dataset, composed of 571 single-point mutations of 37 proteins, is generated in this work as the benchmark test set for the evaluation of ΔT_m prediction methods. We collected data from two ΔT_m databases, ProThermDB and ThermoMutDB, and performed data cleaning in an approach similar to S8754. For each entry, we also replenished experimental pH information. After eliminating data redundant with the S1626 dataset (*i.e.* sequence identity > 25%), this dataset finally becomes a benchmark test set for the fair evaluation of ΔT_m prediction methods, since most prior ΔT_m prediction softwares utilized S1626 as the training set.

S4346

The S4346 dataset, composed of 4346 single-point mutations of 349 proteins, is generated in this work as the training set for the ΔT_m prediction. The data were collected and checked just as in the construction of S571. After removing data redundant with the test set (*i.e.* sequence identity > 25% with the S571 dataset), we finally constructed the largest ΔT_m dataset at present for model training.

Information processing in the geometric-learning-based model architecture

As shown in Figure 2a, the Geometric Encoder adopts the GAT architecture, using the nodes and edges to process input features from the sequence and structure of a protein, respectively. The sequence features are derived from the large-scale protein language model ESM-2[25], which has been improved in terms of model architecture, training parameters and training data compared to its predecessor, ESM-1b[24], in order to capture global context information. Hence, each node in the GAT is initialized by the embedding of the corresponding residue derived from the ESM-2 output. Unlike the traditional approaches that engage inter-residue distances/contacts as edges, in our geometric learning model, each edge is initialized by the relative geometric relationship between a pair of residues derived from the protein 3D structure. Specifically, the position and orientation of each residue could be delicately quantified by a local coordinate system using the positions of its C_α, N and carboxyl carbon (C) atoms. When the local coordinate system of a given residue is taken as the reference, the global coordinates of all other residues can be transformed into relative translations (as vectors) and rotations (as matrices) to this reference frame, which jointly compose the edge features. Here, we follow the AlphaFold2[59] definition on the residue-level local coordinate system. Thus, the bases of the local coordinate system of the ith residue, R_i, are calculated as

$$\begin{aligned} v_{i1} &= p_i^N - t_i \\ v_{i2} &= p_i^C - t_i \\ u_{i1}, u_{i2} &= (v_{i1}, v_{i2}) \\ R_i &= \left[\frac{u_{i1}}{\|u_{i1}\|} \quad \frac{u_{i2}}{\|u_{i2}\|} \quad \frac{u_{i1}}{\|u_{i1}\|} \times \frac{u_{i2}}{\|u_{i2}\|} \right], \end{aligned} \tag{1}$$

where t_i, p_i^N and p_i^C denote positions (as vectors) of the C_α, N and C atoms in the global coordinate system, respectively, while the notation $(\cdot\,,\cdot)$ refers to the Gram-Schmidt orthogonalization operation, $\|\cdot\|$ stands for the metric of vector length and $\times$ denotes the cross product of vectors.

After initialized by sequence and structural features, the node embedding x_i and the edge embedding z_{ij} are further fused in each layer of the Geometric Encoder by

$$\begin{aligned} a_{ij} &= Linear(z_{ij}) \\ b_{ij} &= \frac{1}{w_{ij}} \|(R_i Linear(x_i) + t_i) - (R_j Linear(x_j) + t_j)\| \\ \alpha &= Softmax(\sigma([A, B])), \end{aligned} \tag{2}$$

where A and B denote the matrices composed of a_{ij} and b_{ij}, respectively, σ is a transformation of $\sigma(x) = LeakyRelu(Conv2d(InstanceNorm(x)))$, and α is the computed attention matrix.

Finally, the node embedding x_i and the edge embedding z_{ij} are updated through neighboring layers by

$$\begin{aligned} Z^k &= \sigma([Z^{k-1}, \alpha]) \\ X^k &= LayerNorm(f(\alpha, R, t, p^{C_\beta}, X^{k-1}) + X^{k-1}), \end{aligned} \tag{3}$$

where k denotes the kth layer of the Geometric Encoder, X and Z stand for the vector and matrix composed of x_i and z_{ij}, respectively, and f is an abstract function referring to a series of operations for the self-attention mechanism. X generated from the last layer represents the 1D feature extracted by the Geometric Encoder, which can be used for downstream predictions, *e.g.*, by the MLP module in GeoFitness (Figure 2b).

In the model architecture of GeoDDG and GeoDTm (Figure 2c), wide-type features and mutant features are processed by a geometric encoder with shared weights. Hence, difference between the output values of the wide-type and mutant proteins is supposed to change sign when their input features are switched, a design that naturally guarantees the permutational anti-symmetry of prediction results.

Training details

We utilized S2648 and S8754 as the training set of GeoDDG and S669 as the test set for $\Delta\Delta G$ prediction. Similarly, we utilized S1626 and S4346 as the training set of GeoDTm and S571 as the test set for ΔT_m prediction. Notably, S8754 and S4336 are the enriched training sets collected in this work, and S571 constructed in this work is the first benchmark test set for the fair evaluation on ΔT_m prediction. As mentioned previously, redundancy between training sets and test sets have been carefully removed, using a stringent criterion: no proteins in the training set have sequence identity > 25% to those in the test set.

Thanks to the success of AlphaFold2 in the field of protein structure prediction, it is possible to generate accurately predicted wild-type structures from protein sequences. It was reported that models utilizing structures predicted by AlphaFold2 (pLDDT > 90) and real experimental structures could achieve similar prediction accuracies in the prediction of mutation effects[72]. Nevertheless, we trained two versions of GeoDDG/GeoDTm, utilizing experimental wild-type structures (GeoDDG-3D/GeoDTm-3D) and AlphaFold2-predicted wild-type structures (GeoDDG-Seq/GeoDTm-Seq) for training, respectively. The details of input features are briefly stated below and summarized in Table S2.

We selected the 33-layer 650M-parameter model of ESM-2[25] to generate the "word" embedding for each residue. What's more, FoldX[33] software was utilized to predict the structures of mutants based on the wild-type structure. We only selected the 32 residues closest to the mutation location in the Cartesian space as the considered residues for structural feature extraction.

In the optimization of GeoFitness, in order to utilize the multi-labeled DMS data for model training, we converted all fitness data into the relative ranks for all possible mutations at each individual position, regardless of the original meaning of each fitness indicator. This pre-processing precedure unifies labels in the DMS database and simultaneously retains key information (*i.e.* the relative ranking between mutations) required in practical protein engineering. We took the Spearman correlation coefficient, a metric relying on ranks, as the loss function for model training. Particularly, we adopted the soft Spearman correlation coefficient developed by Blondel *et al.*, which unlike the traditional metric, is differentiable and thus allows the gradient update in the neural network[73]. This strategy enables the training of a unified GeoFitness model utilizing multi-labeled DMS data in this work. In the model training of GeoDDG and GeoDTm, parameters of the pre-trained geometric encoder inherited from GeoFitness were initially frozen to allow the rapid optimization of the other parameters. Subsequently, all parameters were allowed to change slightly in the fine-tuning stage. Furthermore, we adopted a conjugated loss function here for the $\Delta\Delta G$ and ΔT_m predictions, which includes both the soft Spearman correlation coefficient and the mean square error (MSE) between predicted and experimental values. If not mentioned otherwise, the Adam optimizer was chosen to train the model for 200 iterations. The initial learning rate was set as 10^{-3} and was declined by half if the validation loss stopped dropping for 10 consecutive iterations.

Evaluation metrics

Pearson correlation coefficient (indicated by r) is defined as

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}, \tag{4}$$

where Cov represents the covariance and σ stands for the standard deviation.

Spearman correlation coefficient (indicated by ρ) is calculated similarly to Pearson correlation coefficient r but is applied on the rank

$$\rho_{X,Y} = r_{R(X),R(Y)} = \frac{Cov(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}, \tag{5}$$

where R denotes the rank of a variable.

In this work, we mainly used the Spearman correlation coefficient, mean absolute error (MAE) and root mean square error (RMSE) between the predicted and experimental $\Delta\Delta G/\Delta T_m$ values to evaluate the performance of the methods. Additionally, we adopted two evaluation indicators to assess the ability of the predictors to reproduce the anti-symmetric property of $\Delta\Delta G$ and ΔT_m, namely r_{d-i} and $\langle\delta\rangle$, defined as

$$r_{d-i} = \frac{Cov(\Delta\Delta G^{dir}, \Delta\Delta G^{inv})}{\sigma_{dir}\sigma_{inv}} \tag{6}$$

$$\langle\delta\rangle = \frac{\sum_{i=1}^{N}(\left|\Delta\Delta G_i^{dir} + \Delta\Delta G_i^{inv}\right|)}{2N}. \tag{7}$$

r_{d-i} is the Pearson correlation coefficient between the direct (*i.e.* wild-type-to-mutant) and the corresponding inverse (*i.e.* mutant-to-wild-type) sequence variations, while $\langle\delta\rangle$ quantifies the bias between direct and inverse prediction values. A method that perfectly conforms to anti-symmetry should have r_{d-i} equal to -1 and $\langle\delta\rangle$ equal to 0.

Results

Evaluation of protein fitness prediction by GeoFitness

As mentioned in **Methods**, we unified the multi-labeled fitness data into the relative rankings of available mutations at each individual position for each protein to facilitate the model training of GeoFitness. Therefore, the final GeoFitness is a universal fitness prediction model, which does not need to be retrained using extra fitness data of the target protein before practical prediction, *i.e.* allowing zero-shot prediction.

Here, we evaluated the performance of GeoFitness using the DMS database (DeepSequence + MaveDB). Specifically, for each protein in the DMS database, we predicted the fitness scores of all candidate single mutations in the test set, calculated the Spearman correlation coefficient between predicted and experimental values at each individual position, and averaged this value for the whole protein.

We first compared GeoFitness with other unified models that are completely independent of fitness data, including $\mathrm{RF}_{\mathrm{joint}}$, MSA Transformer, ESM-1b, ESM-1v and ESM-2. As shown in Figure 3, GeoFitness remarkably outperforms these unsupervised models, approaching higher Spearman correlation coefficient for almost all proteins in the DMS database, which indicates that the use of fitness data in model training significantly improves the model performance. Subsequently, we continued the evaluation against ECNet and SESNet, two state-of-the-art methods that need to train a specific model for each individual protein. Despite the unfairness of such a comparison to our method, GeoFitness exhibits better performance on more than half of the proteins tested (Figure 3 and Figure 4a). When averaged over all

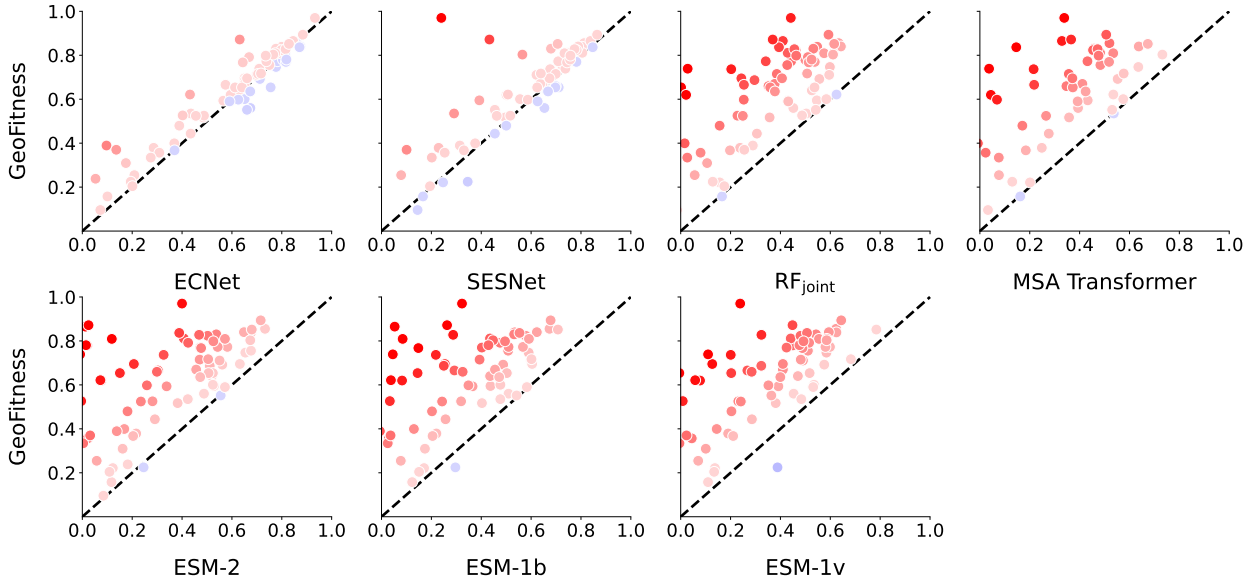


Figure 3. Pairwise comparisons of GeoFitness with other methods for the prediction of mutation effects on protein fitness. Each point in the figure represents the evaluation results of one protein in the DMS database using the metric of Spearman correlation coefficient. The point is colored red if GeoFitness outperforms the other method and blue otherwise, with the degree of opaqueness reflecting the magnitude of difference. Detailed data could be found in Table S3.

proteins in the DMS database, the universal GeoFitness model reaches a slightly higher Spearman correlation coefficient than the series of specifically optimized ECNet and SESNet models (0.62 vs. 0.59 and 0.57). Therefore, although information is lost by the transformation of absolute fitness data into relative rankings in the optimization of GeoFitness, such a negative impact is compensated by the large-scale enrichment of multi-labeled training data.

Furthermore, we conducted an ablation study of GeoFitness to analyze the contributions of model architecture and data enrichment, where the baseline model was trained using MLP as the network structure and DeepSequence as the training set. As shown in Figure 4b, the geometric-learning-based architecture slightly enhances the performance, whereas enrichment of data from the MaveDB database makes a more significant contribution, increasing the mean Spearman correlation coefficient by 0.16.

In summary, through effective utilization of the large scale of multi-labeled fitness data, GeoFitness successfully learns the effects of single mutations without sacrificing model usability. Unlike the fitness prediction, the amount of labeled data is typically reduced by at least one order of magnitude in downstream tasks like the protein stability prediction. Considering the biological relevance between protein fitness and stability, we reused the pre-trained geometric encoder of GeoFitness as information extractor (with inherited network architecture and parameter values) in the GeoDDG/GeoDTm models for $\Delta\Delta G/\Delta T_m$ predictions (Figure 2b and Figure 2c). Notably, the lack of label overlap between the DMS fitness data and protein stability data safely eliminates the possibility of information leaking by this reutilization.

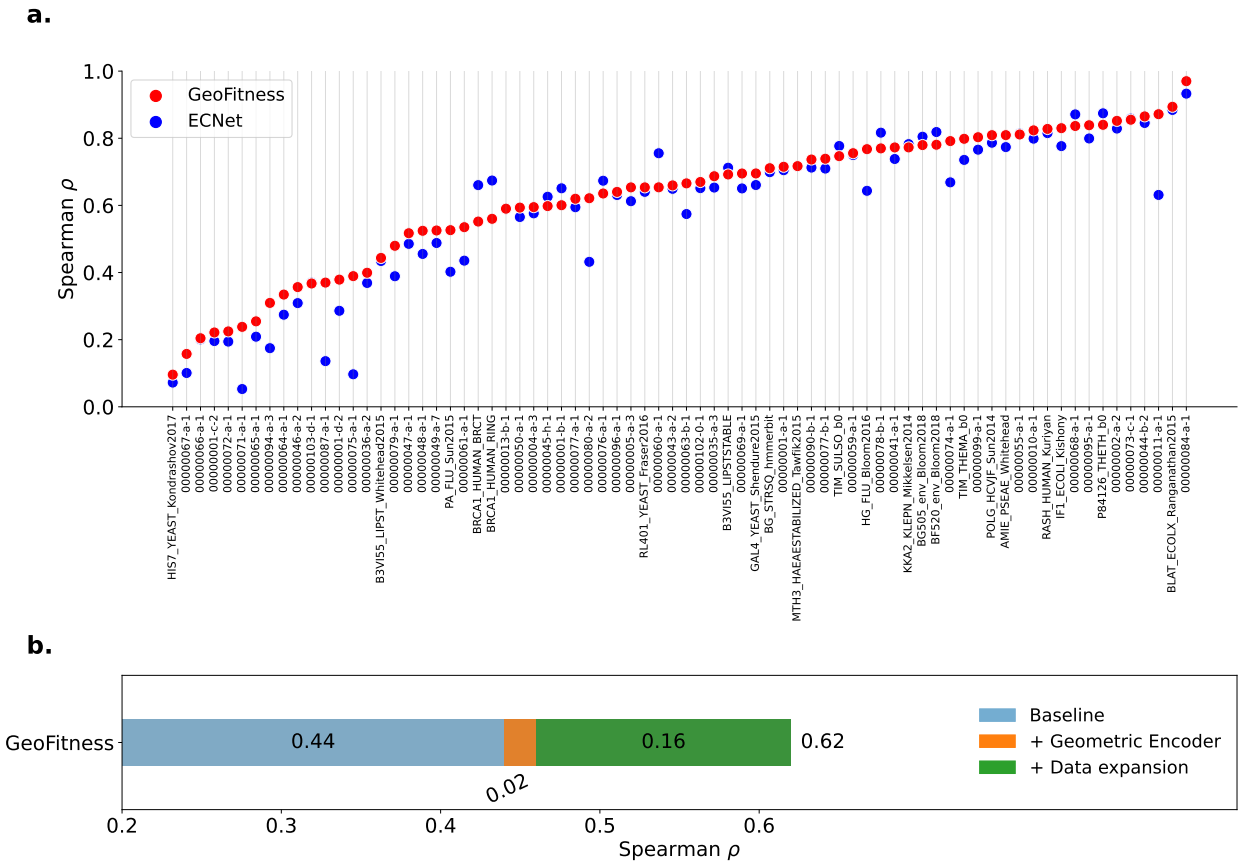


Figure 4. Detailed analysis for GeoFitness. a) Comparison between GeoFitness and ECNet on the DMS database. GeoFitness has higher Spearman correlation coefficient on 60 out of the 74 proteins. Detailed data could be found in Table S3. **b)** Ablation study of GeoFitness.

Evaluation of $\Delta\Delta G$ prediction by GeoDDG

GeoDDG uses a module with shared weights to process the sequence and structural information of both wild-type and mutant proteins and takes their difference for prediction. Unlike GeoFitness that is only applicable for single-mutation effects, GeoDDG is, in principle, able to predict the change of ΔG caused by an arbitrary number of mutations.

Available $\Delta\Delta G$ prediction methods can be classified into sequence-based and structure-based, depending on whether the experimental structural information is required for inference. Here, we trained two versions of GeoDDG, namely GeoDDG-Seq and GeoDDG-3D, which retrieve information from AlphaFold2-predicted structures and experimental structures, respectively, and evaluated their performance against mainstream state-of-the-art methods on the S669 benchmark test set. Notably, in this work, redundancies in training sets of GeoDDG have been strictly removed against the test set S669 to guarantee fair evaluation (see **Methods** for details), and more importantly, the zero overlap between the training set of the pre-trained GeoFitness model and S669 (Figure S1) precludes the possibility of information leaking during performance evaluation.

As shown in Table 1, GeoDDG-Seq and GeoDDG-3D significantly outperform the other methods of their respective categories in all metrics. Particularly, our methods exhibit an advantage of a large margin in terms of the Spearman correlation coefficient for the direct mutational effects, with relative improvements of 36.4% (0.60 vs. 0.44, see Figure

Table 1. Comparison of GeoDDG with exsiting models on the S669 dataset.

	Total			Direct			Inverse			Antisymmetry	
Method	ρ	**RMSE**	**MAE**	ρ	**RMSE**	**MAE**	ρ	**RMSE**	**MAE**	r_{d-i}	$\langle\delta\rangle$
Sequence-based											
INPS-Seq	0.64	1.52	1.10	0.44	1.52	1.09	0.44	1.53	1.10	-0.99	0.07
ACDC-NN-Seq	0.61	1.53	1.08	0.43	1.53	1.08	0.43	1.53	1.08	**-1.00**	**0.00**
DDGun	0.59	1.76	1.27	0.43	1.75	1.27	0.41	1.77	1.27	-0.98	0.12
I-Mutant3.0-Seq	0.36	1.92	1.47	0.35	1.56	1.16	0.23	2.22	1.79	-0.45	1.52
MuPro	0.31	2.03	1.58	0.26	1.61	1.21	0.22	2.38	1.96	-0.29	1.90
SAAFEC-Seq	0.19	2.01	1.53	0.35	1.54	1.12	-0.01	2.40	1.94	-0.03	1.66
GeoDDG-LE-Seq	0.69	1.41	1.02	0.55	1.40	1.02	0.54	1.41	1.02	-1.00	0.01
GeoDDG-Seq	**0.73**	**1.37**	**0.99**	**0.60**	**1.38**	**0.99**	**0.60**	**1.37**	**0.99**	**-1.00**	0.01
Structure-based											
ACDC-NN	0.63	1.50	1.05	0.45	1.49	1.05	0.44	1.50	1.06	-0.99	0.09
PremPS	0.63	1.50	1.07	0.42	1.51	1.09	0.42	1.49	1.05	-0.82	0.38
INPS3D	0.57	1.64	1.19	0.44	1.50	1.07	0.36	1.76	1.31	-0.46	1.02
DDGun3D	0.55	1.61	1.13	0.43	1.60	1.11	0.41	1.62	1.14	-0.97	0.18
Dynamut	0.48	1.65	1.21	0.37	1.60	1.19	0.36	1.69	1.24	-0.55	0.62
ThermoNet	0.46	1.64	1.20	0.37	1.62	1.17	0.34	1.66	1.23	-0.83	0.31
PopMusic	0.43	1.82	1.36	0.41	1.51	1.09	0.22	2.09	1.64	-0.23	1.42
FoldX	0.41	2.40	1.53	0.28	2.32	1.57	0.34	2.48	1.50	-0.33	1.51
DUET	0.39	1.86	1.39	0.42	1.52	1.10	0.23	2.14	1.68	-0.07	1.48
MAESTRO	0.39	1.80	1.35	0.46	1.44	1.06	0.19	2.10	1.65	-0.17	1.32
mCSM	0.34	1.96	1.49	0.36	1.54	1.13	0.21	2.30	1.86	-0.02	1.75
I-Mutant3.0	0.28	1.96	1.49	0.35	1.54	1.12	0.15	2.32	1.87	-0.01	1.62
SDM	0.28	1.93	1.45	0.39	1.67	1.26	0.14	2.16	1.64	-0.42	0.97
GeoDDG-LE-3D	0.72	1.38	0.99	0.57	1.38	0.99	0.57	1.38	0.99	-1.00	0.00
GeoDDG-3D	**0.74**	**1.37**	**0.99**	**0.62**	**1.37**	**0.99**	**0.62**	**1.37**	**0.99**	**-1.00**	**0.00**

The winner in each metric is highlighted in bold. "Direct" refers to the mutations from the wild type to mutant, "Inverse" means the corresponding mutations in the reverse direction, and "Total" stands for the combination of the above categories. Here, ρ denotes Spearman correlation coefficient and details of all metrics are described in **Methods**. The raw results of the other models are taken from a prior work[51].

5) and 34.9% (0.62 vs. 0.46, see Figure 6) over the second best methods in the sequence-based and structure-based categories, respectively. Moreover, GeoDDG is the first method that can reduce the absolute prediction error (*i.e.* MAE) of $\Delta\Delta G$ to be less than 1 kcal/mol. Additionally, both versions of GeoDDG successfully ensure the anti-symmetric $\Delta\Delta G$ values between the direct and inverse mutations, as demonstrated by r_{d-i} and $\langle\delta\rangle$ values of -1 and 0, respectively.

We further conducted ablation studies on the two versions of GeoDDG. As expected, the baseline models that were trained using the architecture of MLP on the smaller S2648 dataset show similar performances to INPS-Seq (Figure 5) and MAESTRO (Figure 6), the second best methods in the sequenced-based and structure-based categories, respectively. In both cases, the introduction of geometric encoder brings an increase of 0.04 to the Spearman correlation coefficient, while the engage of soft rank loss and data expansion further enhance the performance slightly. Interestingly, the

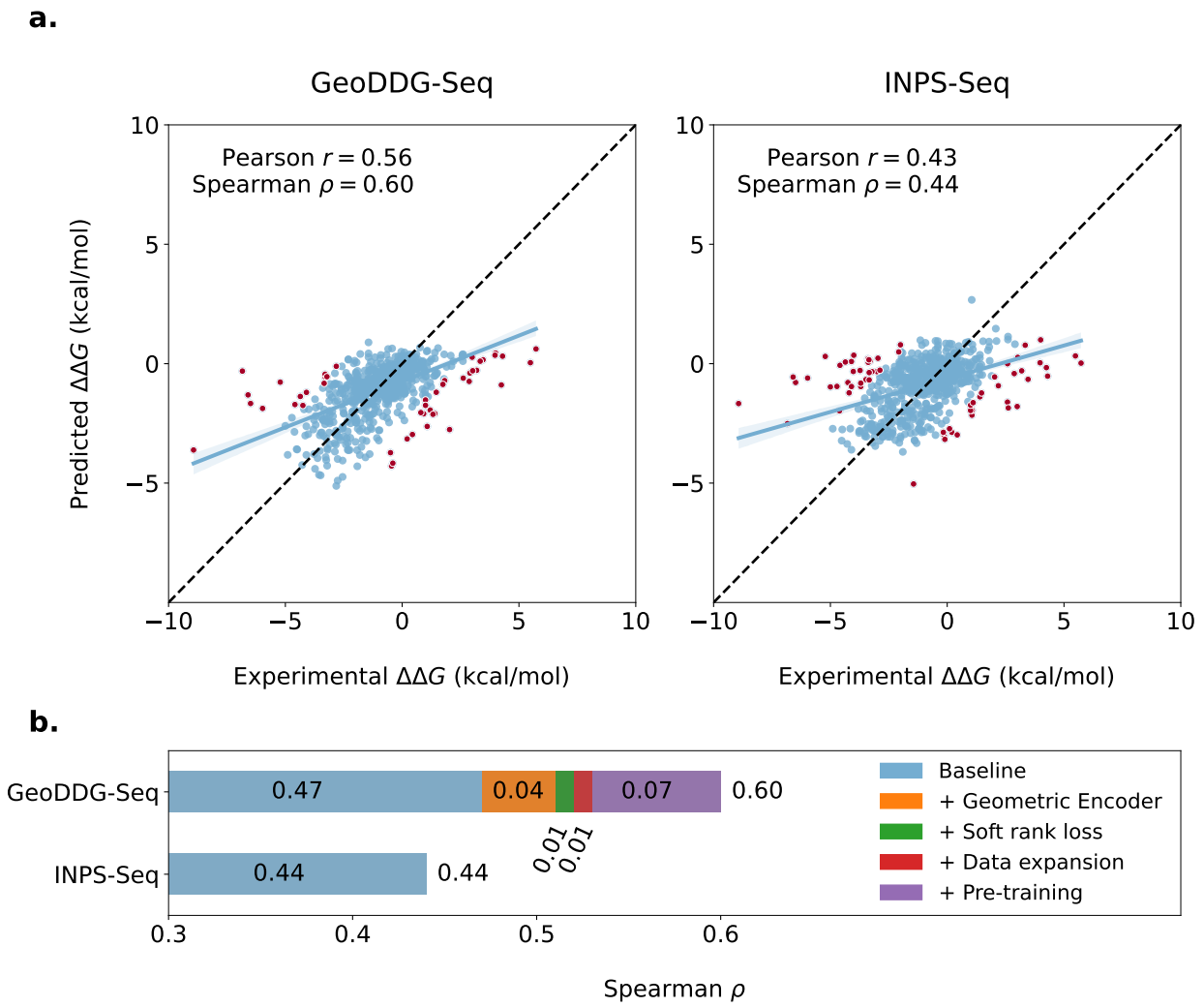


Figure 5. Detailed analysis for GeoDDG-Seq. a) Side-by-side comparison between GeoDDG-Seq and the second best sequence-based predictor, INPS-Seq, on S669. Predictions with error > 2.5 kcal/mol are identified as outliers (46 in GeoDDG-Seq vs. 64 in INPS-Seq) and are colored red in the figures. **b)** Ablation study of GeoDDG-Seq.

pre-training strategy introduces the largest improvement (0.07 in both GeoDDG-Seq and GeoDDG-3D), supporting that model pre-training using the large amount of protein fitness data can effectively overcome the challenge of limited data that have perplexed the $\Delta\Delta G$ prediction in the past years.

To further validate the importance of the innovation points of our GeoDDG model, principally the pre-training strategy and the design of network architecture, we retrained GeoDDG using limited training data, *i.e.* on the basic S2648 dataset, from which many of the previous methods were developed, and named this model as GeoDDG-LE. As shown in Table 1, when evaluated on the S669 test set, the model performance of GeoDDG-LE declines slightly in comparison with the full version GeoDDG, but still prevails all of the other methods by a large margin (*e.g.*, > 10 percentage points in the Spearman correlation coefficient for direct mutations), for both sequence-based and structure-based predictors. We also tested GeoDDG-LE against previous methods developed purely from the basic S2648 training set. Notably, the S971 testing set that was traditionally used to evaluated these models instead has overlaps with the S2648 training set. After removing these redundant data, we composed a subset of S971, namely the S783 dataset (see **Methods** for details), for

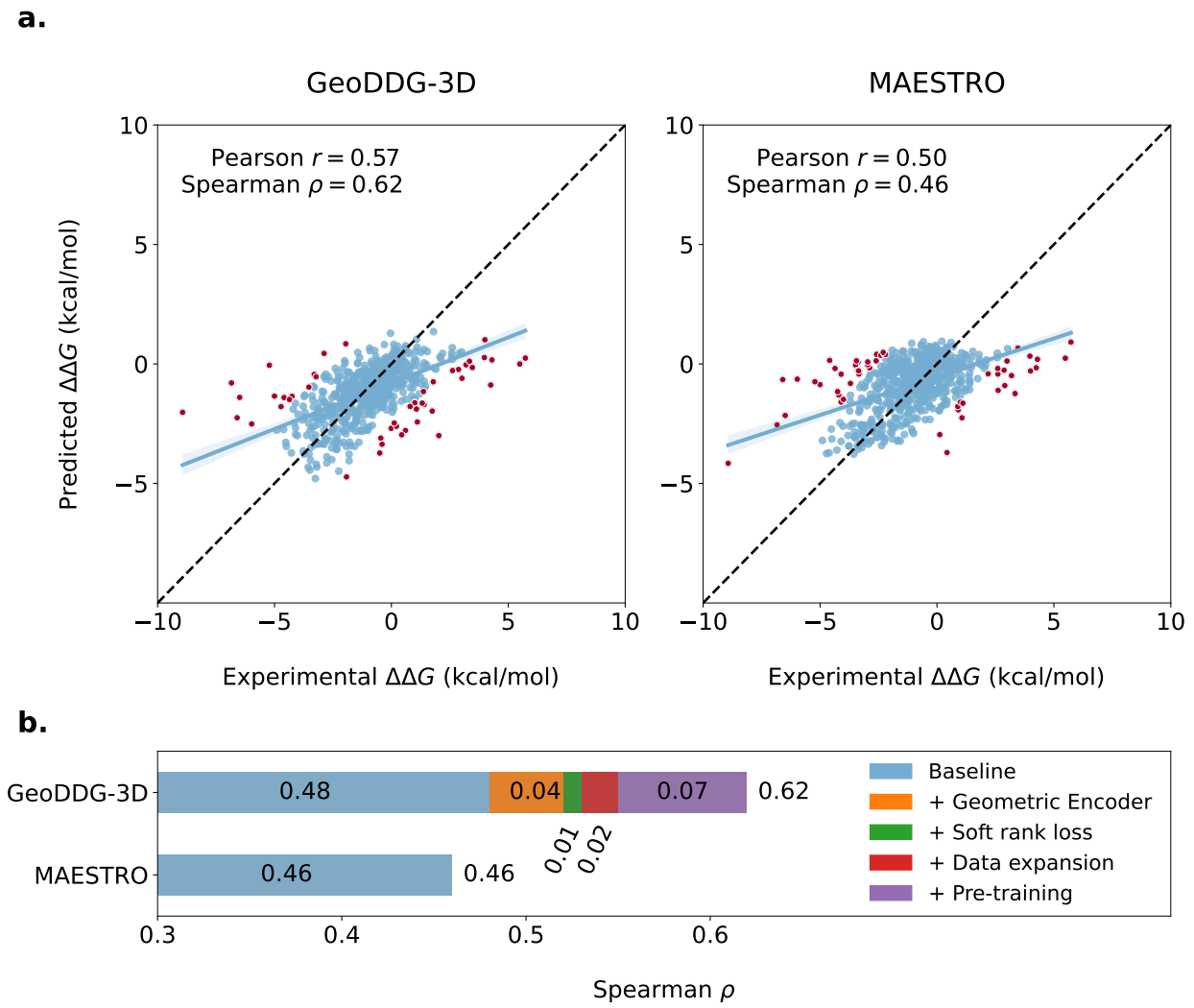


Figure 6. Detailed analysis for GeoDDG-3D. a) Side-by-side comparison between GeoDDG-3D and the second best structure-based predictor, MAESTRO, on S669. Predictions with error > 2.5 kcal/mol are identified as outliers (48 in GeoDDG-3D vs. 54 in MAESTRO) and are colored red in the figures. **b)** Ablation study of GeoDDG-3D.

the model performance evaluation. According to the testing results, our GeoDDG-LE models outperform all of the other methods remarkably on the S783 test set (Table 2). Moreover, Figure S2 shows the performance of GeoDDG-LE models by 5-fold and 10-fold cross validations within the S2648 training set. The fairly balanced performance of GeoDDG-LE models among cross validations and benchmark tests (on S669/S783 sets) laterally supports good generalizability and robustness of our method. Given the above systematic tests, we believe that our method is capable of generating reliable prediction models for the mutation effects on the protein stability.

In conclusion, through the effective use of data by the pre-training strategy and the adequate processing of protein structural information by geometric learning, GeoDDG enormously improves the prediction of mutation effects on the ΔG of proteins.

Table 2. Comparison of GeoDDG-LE with exsiting models on the S783 dataset.

	Total			Direct			Inverse			Antisymmetry	
Method	ρ	**RMSE**	**MAE**	ρ	**RMSE**	**MAE**	ρ	**RMSE**	**MAE**	r_{d-i}	$\langle\delta\rangle$
Sequence-based											
INPS-Seq				0.56	1.68	1.15					
GeoDDG-LE-Seq	**0.80**	**1.44**	**1.00**	**0.70**	**1.44**	**1.00**	**0.70**	**1.44**	**1.00**	**-1.00**	**0.00**
Structure-based											
PremPS				0.66	1.51	1.05					
PoPMuSiC				0.60	1.51	1.05					
INPS3D				0.59	1.58	1.07					
FoldX				0.60	1.99	1.26					
mCSM				0.50	1.65	1.17					
GeoDDG-LE-3D	**0.81**	**1.49**	**1.01**	**0.71**	**1.49**	**1.01**	**0.71**	**1.49**	**1.01**	**-1.00**	**0.00**

The winner in each metric is highlighted in bold. "Direct" refers to the mutations from the wild type to mutant, "Inverse" means the corresponding mutations in the reverse direction, and "Total" stands for the combination of the above categories. Here, ρ denotes Spearman correlation coefficient and details of all metrics are described in **Methods**. The raw results of the other models are taken from a prior work[41].

Evaluation of ΔT_m prediction by GeoDTm

Similar to GeoDDG, GeoDTm can predict the influence on T_m of a protein by multiple mutations, with "Seq" and "3D" versions denoting the models trained using AlphaFold2-predicted structures and experimental structures, respectively.

Table 3. Comparison of GeoDTm with existing models on the S571 dataset.

	Total			Direct			Inverse			Antisymmetry	
Method	ρ	**RMSE**	**MAE**	ρ	**RMSE**	**MAE**	ρ	**RMSE**	**MAE**	r_{d-i}	$\langle\delta\rangle$
Sequence-based											
GeoDTm-Seq	**0.57**	**7.99**	**5.39**	**0.51**	**7.99**	**5.39**	**0.51**	**7.99**	**5.39**	**-1.00**	**0.00**
Structure-based											
HoTMuSiC				0.30	8.41	5.70					
AUTO-MUTE				0.25	8.50	5.79					
GeoDTm-3D	**0.56**	**8.16**	**5.45**	**0.51**	**8.16**	**5.45**	**0.51**	**8.16**	**5.45**	**-1.00**	**0.00**

The winner in each metric is highlighted in bold. "Direct" refers to the mutations from the wild type to mutant, "Inverse" means the corresponding mutations in the reverse direction, and "Total" stands for the combination of the above categories. Here, ρ denotes Spearman correlation coefficient and details of all metrics are described in **Methods**. Blank region means that the corresponding result is unavailable.

However, unlike the $\Delta\Delta G$ prediction, there is no previously established benchmark test set for the ΔT_m prediction, thus disallowing the fair evaluation of all methods. To address this problem, we curated a test set of 571 samples for ΔT_m, named S571, which is non-redundant with currently available training datasets (see **Methods** for details) and has zero overlap with the training set of the pre-trained GeoFitness model at the same time (Figure S1). Subsequently, we evaluated the performance of GeoDTm-Seq and GeoDTm-3D on the S571 set against two available ΔT_m predictors,

AUTO-MUTE and HoTMuSiC, which are, unfortunately, both structured-based models and unable to provide the prediction for the inverse mutations. As shown in Table 3, GeoDTm-3D prevails the other structure-based methods in all metrics. Notably, our method tremendously enhances the Spearman correlation coefficient from 0.30 to 0.51, with a relative improvement of more than 70%. Interestingly, GeoDTm-Seq achieves a slightly lower RMSE and MAE than GeoDTm-3D, which laterally demonstrates the assistance of accurate structural prediction by AlphaFold2 in functional predictions. Additionally, like the case in GeoDDT, both versions of GeoDTm perfectly conform to the anti-symmetric requirement of prediction results.

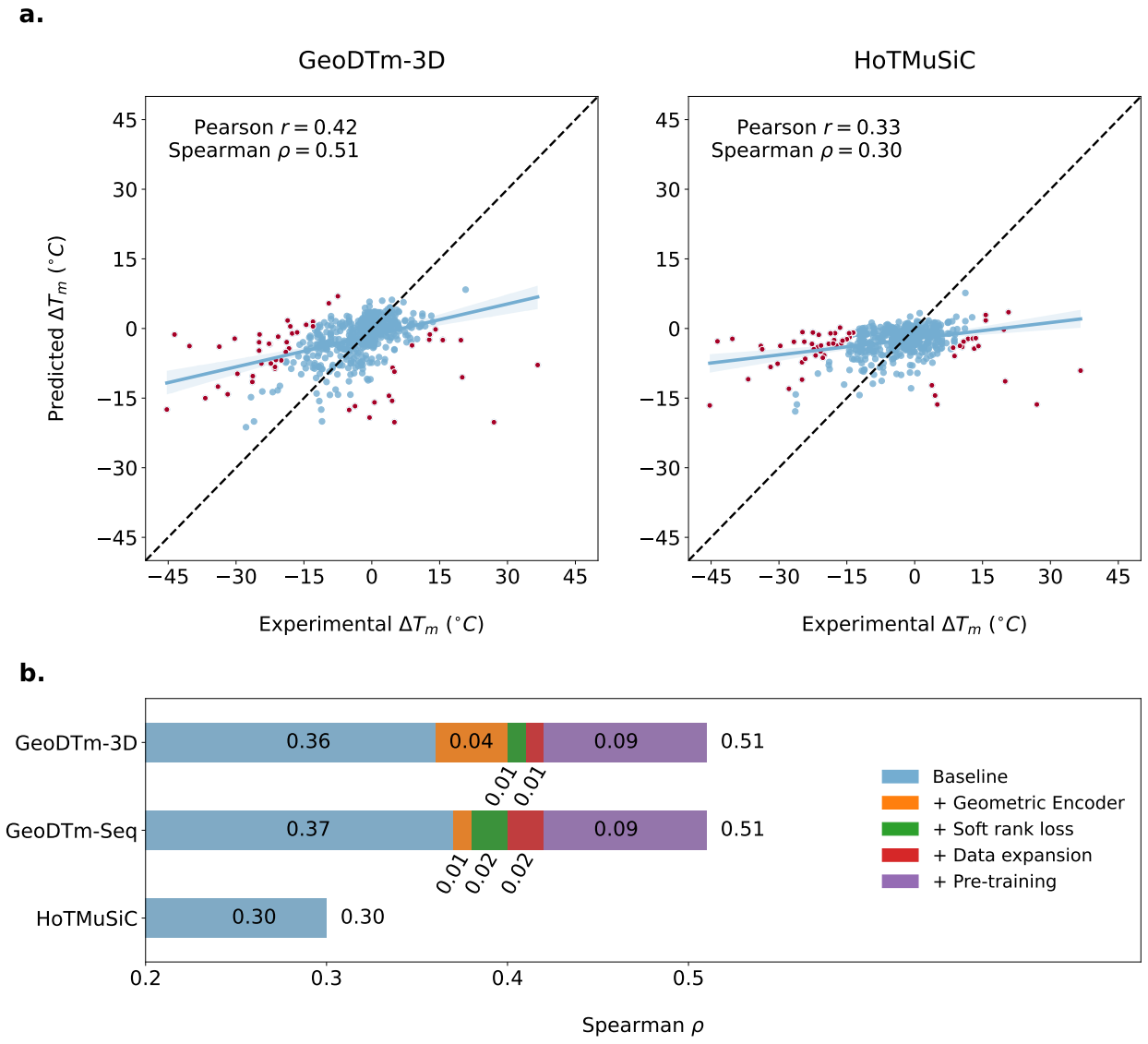


Figure 7. Detailed analysis for GeoDTm. a) Side-by-side comparison between GeoDTm-3D and the second best structure-based predictor, HotMuSiC, on S571. Predictions with error > 12.5 $^\circ C$ are identified as outliers (54 in GeoDDG-3D vs. 62 in HotMuSiC) and are colored red in the figures. **b)** Ablation study of GeoDTm-3D and GeoDTm-Seq.

We also performed ablation studies for the two versions of GeoDTm .As shown in Figure 7, the baseline models that were trained using the MLP architecture on the smaller S1624 training set already prevail the other methods

(HoTMuSiC here), which were likely to be over-trained considering the significantly reduced performance on the S571 set compared with the reported cross-validation values in their original papers. The geometric encoder, soft rank loss and data expansion jointly improve the Spearman correlation coefficient by 0.06 and 0.05 for GeoDTm-3D and GeDTm-Seq, respectively. In both versions of GeoDTm, the pre-training strategy makes the largest contribution, increasing the Spearman correlation coefficient by 0.09. This significant enhancement further demonstrates the role of pre-training using large amount of related data in eliminating overfitting and improving model generalizability.

Hence, similar to GeoDDG, GeoDTm tremendously improves the prediction of mutation effects on the T_m of proteins.

Discussion

Understanding the mutation effects is an important issue in practical protein design and engineering, and hence extensive attention has been paid to developing computational methods to learn and predict these effects accurately and efficiently. Despite the continual efforts in the model architecture design, insufficient and ineffective data utilization gradually becomes the bottleneck that prohibits the further improvement of model performance. In the fitness prediction field, although the large scale of DMS data have paved the way for the training of deep learning models, existing models have to sacrifice either the prediction power or the model usability due to the improper treatment of the multi-labeled data. In the $\Delta\Delta G$ and ΔT_m prediction field, though the problem definition is clear and the data are single-labeled, the limited size and significant bias of data prohibit the training of deep learning models without overfitting. A systematic review[52] even claimed that the average accuracy of $\Delta\Delta G$ prediction models had not improved in the past 15 years given the evaluation results on the S669 test set[51]. The situation is even worse for the ΔT_m prediction, since objective test sets (like S669 in $\Delta\Delta G$ prediction) are absent, which disallows fair evaluation for the ΔT_m prediction methods.

In this work, we adopted the differentiable ranking algorithm[73] to construct a soft Spearman correlation coefficient loss, which enables training of a unified model with all DMS data regardless of their multi-labeled nature. This novel loss function allows us to make full use of the DMS dataset in the supervised training of GeoFitness, which could predict the fitness landscape of the interested protein with an accuracy higher than the state-of-the-art models without the need of model retraining using sufficient data of the same kind before practical use.

Since protein stability is closedly related with protein fitness in biology, we utilized the geometric encoder of GeoFitness that was pre-trained on the fitness data to guide the prediction of $\Delta\Delta G$ and ΔT_m. As suggested by the ablation studies, the pre-training strategy enhances the Spearman correlation coefficient by 0.07 and 0.09 for $\Delta\Delta G$ and ΔT_m prediction, respectively, indicating its marked role on alleviating the negative impact of limited data and on improving the model generalizability. Such remarkable improvements mainly arise from the correlation between the prediction targets, fitness and stability here, as well as the overwhelming data scale of the former over the latter. Moreover, the improvement introduced by the pre-training strategy is more significant in the ΔT_m prediction task that has less available training data than the $\Delta\Delta G$ prediction. This observation implies the possible aids of our pre-training strategy in the methodology development of other downstream tasks such as predicting the luminescence intensity and expression level of proteins upon mutations, application tasks that are of great importance but have even less available experimental data.

To further improve the prediction powers on protein stability, we collected, cleaned and manually checked the data from ProThermDB[68] and ThermoMutDB[69] and curated the largest open-source protein stability datasets: S8754 for $\Delta\Delta G$ and S4346 for ΔT_m. Ablation studies on GeoDDG and GeoDTm support the positive impact of data expansion on model performance. Before the introduction of S669, in the $\Delta\Delta G$ prediction field, cross validation is usually used to demonstrate the effectiveness of the model, a protocol that frequently causes bias[52] and hinders the fair evaluation of the model performance. To address the same challenge confronted in the ΔT_m prediction field, we constructed a

new benchmark test set S571 in this work to enable the fair model evaluation on ΔT_m prediction. The three newly collected datasets, S8754, S4346 and S571, are freely downloadable and may further benefit the field of protein stability prediction.

The models proposed in this work, GeoFitness, GeoDDG and GeoDTm, are all based on the geometric encoder, a geometric-learning-based network architecture that are designed to simultaneously process the information of protein sequences and structures and to ensure the anti-symmetry of prediction results. Based on the evaluation of prior models (Table 1 and Table 3), utilization of structural information clearly benefits the stability prediction (see the better performance of structure-based models vs. sequence-based ones in general), but introduces additional limitations, since experimental structures are frequently unavailable for the concerned protein target. Therefore, we trained two versions of protein stability predictors, GeoDDG/GeoDTm-3D and GeoDDG/GeoDTm-Seq, which retrieve structural information from the experimental structures and AlphaFold2-predicted ones[59], respectively. In practical predictions, the "3D" version is suggested for better performance when the reliable experimental structure is available, while the "Seq" version is an alternative when only sequence information is known. On the other hand, whether AlphaFold2 prediction benefits the downstream protein functional prediction tasks is contraversial[72,74–78]. As shown in Figure S3, correlation between the predicted ΔpLDDT by AlphaFold2 and $\Delta\Delta G$ is poor and is completely lost if we focus on data points of higher significance (*i.e.* absolute $\Delta\Delta G$ value > 5 kcal/mol). This observation indicates that the naive use of AlphaFold2 indicators like the pLDDT is unlikely to improve the functional prediction. In contrast, among the stability predictors proposed in this work, our "Seq" version can reach a prediction power very close to that of the "3D" version (Table 1 and Table 3), suggesting that proper utilization of the accurately predicted structures from AlphaFold2 as spatial constraints (*e.g.*, the edge embedding in this work) can effectively assist the downstream tasks like the stability prediction, which agrees with previous findings[72].

We have established a web server for GeoStab-suite, a suite of the three predictors GeoFitness, GeoDDG and GeoDTm. We expect the GeoStab-suite to be a useful tool for researchers in the field of protein science.

Author Contribution

Y. X. and H. G. proposed the methodology and designed the experiment. Y. X. implemented the experiment. Y. X. and D. L. analyzed the results. Y. X., D. L. and H. G. wrote the manuscript. All authors agreed with the final manuscript.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (#32171243) and by the Beijing Frontier Research Center for Biological Structure.

References

1. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15,** 816–822. https://doi.org/10.1038/s41592-018-0138-4 (2018).
2. Coluzza, I. Computational protein design: a review. *Journal of Physics: Condensed Matter* **29,** 143001. https://doi.org/10.1088/1361-648x/aa5c76 (2017).
3. Dahiyat, B. I. & Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **278,** 82–87. https://doi.org/10.1126/science.278.5335.82 (1997).

4. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19,** 596–604. https://doi.org/10.1016/j.sbi.2009.08.003 (2009).
5. Pucci, F., Bourgeas, R. & Rooman, M. High-quality Thermodynamic Data on the Stability Changes of Proteins Upon Single-site Mutations. *Journal of Physical and Chemical Reference Data* **45,** 023104. https://doi.org/10.1063/1.4947493 (2016).
6. Louis, B. B. V. & Abriata, L. A. Reviewing Challenges of Predicting Protein Melting Temperature Change Upon Mutation Through the Full Analysis of a Highly Detailed Dataset with High-Resolution Structures. *Molecular Biotechnology* **63,** 863–884. https://doi.org/10.1007/s12033-021-00349-0 (2021).
7. Yeoman, C. J., Han, Y., Dodd, D., *et al.* in *Advances in Applied Microbiology* 1–55 (Elsevier, 2010). https://doi.org/10.1016/s0065-2164(10)70001-0.
8. Kopanos, C., Tsiolkas, V., Kouris, A., *et al.* VarSome: the human genomic variant search engine. *Bioinformatics* **35** (ed Wren, J.) 1978–1980. https://doi.org/10.1093/bioinformatics/bty897 (2018).
9. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature Methods* **11,** 801–807. https://doi.org/10.1038/nmeth.3027 (2014).
10. McGuinness, K. N., Pan, W., Sheridan, R. P., *et al.* Role of simple descriptors and applicability domain in predicting change in protein thermostability. *PLOS ONE* **13** (ed Permyakov, E. A.) e0203819. https://doi.org/10.1371/journal.pone.0203819 (2018).
11. Wu, Z., Kan, S. B. J., Lewis, R. D., *et al.* Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences* **116,** 8852–8858. https://doi.org/10.1073/pnas.1901979116 (2019).
12. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature Methods* **16,** 687–694. https://doi.org/10.1038/s41592-019-0496-6 (2019).
13. Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., *et al.* Mutation effects predicted from sequence co-variation. *Nature Biotechnology* **35,** 128–135. https://doi.org/10.1038/nbt.3769 (2017).
14. Khersonsky, O., Lipsh, R., Avizemer, Z., *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Molecular Cell* **72,** 178–186.e5. https://doi.org/10.1016/j.molcel.2018.08.033 (2018).
15. Li, M., Kang, L., Xiong, Y., *et al.* SESNet: sequence-structure feature-integrated deep learning method for data-efficient protein engineering. *Journal of Cheminformatics* **15.** https://doi.org/10.1186/s13321-023-00688-x (2023).
16. Luo, Y., Jiang, G., Yu, T., *et al.* ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nature Communications* **12.** https://doi.org/10.1038/s41467-021-25976-8 (2021).
17. Meier, J., Rao, R., Verkuil, R., *et al. Language models enable zero-shot prediction of the effects of mutations on protein function* in *Advances in Neural Information Processing Systems* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y., *et al.*) **34** (Curran Associates, Inc., 2021), 29287–29303. https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.
18. Rao, R., Liu, J., Verkuil, R., *et al.* MSA Transformer. *bioRxiv.* https://doi.org/10.1101/2021.02.12.430858 (2021).
19. Rao, R., Bhattacharya, N., Thomas, N., *et al. Evaluating Protein Transfer Learning with TAPE* in *Advances in Neural Information Processing Systems* (eds Wallach, H., Larochelle, H., Beygelzimer, A., *et al.*) **32** (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/37f65c068b7723cd7809ee2d31d7861c-Paper.pdf.

20. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences* **110.** https://doi.org/10.1073/pnas.1215251110 (2012).

21. Russ, W. P., Figliuzzi, M., Stocker, C., *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* **369,** 440–445. https://doi.org/10.1126/science.aba3304 (2020).

22. Wang, J., Lisanza, S., Juergens, D., *et al.* Scaffolding protein functional sites using deep learning. *Science* **377,** 387–394. https://doi.org/10.1126/science.abn2100 (2022).

23. Mansoor, S., Baek, M., Juergens, D., *et al.* Accurate Mutation Effect Prediction using RoseTTAFold. *bioRxiv*. https://doi.org/10.1101/2022.11.04.515218 (2022).

24. Rives, A., Meier, J., Sercu, T., *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118.** https://doi.org/10.1073/pnas.2016239118 (2021).

25. Lin, Z., Akin, H., Rao, R., *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130. https://doi.org/10.1126/science.ade2574 (2023).

26. Alley, E. C., Khimulya, G., Biswas, S., *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16,** 1315–1322. https://doi.org/10.1038/s41592-019-0598-1 (2019).

27. Amrein, B. A., Steffen-Munsberg, F., Szeler, I., *et al.* *CADEE*: Computer-Aided Directed Evolution of Enzymes. *IUCrJ* **4,** 50–64. https://doi.org/10.1107/s2052252516018017 (2017).

28. Bedbrook, C. N., Yang, K. K., Rice, A. J., *et al.* Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Computational Biology* **13** (ed Maranas, C. D.) e1005786. https://doi.org/10.1371/journal.pcbi.1005786 (2017).

29. Bedbrook, C. N., Yang, K. K., Robinson, J. E., *et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature Methods* **16,** 1176–1184. https://doi.org/10.1038/s41592-019-0583-8 (2019).

30. Biswas, S., Khimulya, G., Alley, E. C., *et al.* Low-N protein engineering with data-efficient deep learning. *Nature Methods* **18,** 389–396. https://doi.org/10.1038/s41592-021-01100-y (2021).

31. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* **33,** W306–W310. https://doi.org/10.1093/nar/gki375 (2005).

32. Rodrigues, C. H., Pires, D. E. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research* **46,** W350–W355. https://doi.org/10.1093/nar/gky300 (2018).

33. Schymkowitz, J., Borg, J., Stricher, F., *et al.* The FoldX web server: an online force field. *Nucleic Acids Research* **33,** W382–W388. https://doi.org/10.1093/nar/gki387 (2005).

34. Worth, C. L., Preissner, R. & Blundell, T. L. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research* **39,** W215–W222. https://doi.org/10.1093/nar/gkr363 (2011).

35. Benevenuta, S., Pancotti, C., Fariselli, P., *et al.* An antisymmetric neural network to predict free energy changes in protein variants. *Journal of Physics D: Applied Physics* **54,** 245403. https://doi.org/10.1088/1361-6463/abedfb (2021).

36. Cao, H., Wang, J., He, L., *et al.* DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *Journal of Chemical Information and Modeling* **59,** 1508–1514. https://doi.org/10.1021/acs.jcim.8b00697 (2019).

37. Li, B., Yang, Y. T., Capra, J. A., *et al.* Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLOS Computational Biology* **16** (ed Fariselli, P.) e1008291. https://doi.org/10.1371/journal.pcbi.1008291 (2020).

38. Pancotti, C., Benevenuta, S., Repetto, V., *et al.* A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes Upon Genetic Variations. *Genes* **12,** 911. https://doi.org/10.3390/genes12060911 (2021).

39. Fariselli, P., Martelli, P. L., Savojardo, C., *et al.* INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* **31,** 2816–2821. https://doi.org/10.1093/bioinformatics/btv291 (2015).

40. Capriotti, E., Fariselli, P., Rossi, I., *et al.* A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9.** https://doi.org/10.1186/1471-2105-9-s2-s6 (2008).

41. Chen, Y., Lu, H., Zhang, N., *et al.* PremPS: Predicting the impact of missense mutations on protein stability. *PLOS Computational Biology* **16** (ed Keskin, O.) e1008543. https://doi.org/10.1371/journal.pcbi.1008543 (2020).

42. Chen, C.-W., Lin, J. & Chu, Y.-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. *BMC Bioinformatics* **14.** https://doi.org/10.1186/1471-2105-14-s2-s5 (2013).

43. Dehouck, Y., Grosfils, A., Folch, B., *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25,** 2537–2543. https://doi.org/10.1093/bioinformatics/btp445 (2009).

44. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics* **79,** 830–838. https://doi.org/10.1002/prot.22921 (2010).

45. Li, G., Panday, S. K. & Alexov, E. SAAFEC-SEQ: A Sequence-Based Method for Predicting the Effect of Single Point Mutations on Protein Thermodynamic Stability. *International Journal of Molecular Sciences* **22,** 606. https://doi.org/10.3390/ijms22020606 (2021).

46. Montanucci, L., Capriotti, E., Frank, Y., *et al.* DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC Bioinformatics* **20.** https://doi.org/10.1186/s12859-019-2923-1 (2019).

47. Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., *et al.* SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Research* **45,** W229–W235. https://doi.org/10.1093/nar/gkx439 (2017).

48. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30,** 335–342. https://doi.org/10.1093/bioinformatics/btt691 (2013).

49. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research* **42,** W314–W319. https://doi.org/10.1093/nar/gku411 (2014).

50. Iqbal, S., Li, F., Akutsu, T., *et al.* Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. *Briefings in Bioinformatics* **22.** https://doi.org/10.1093/bib/bbab184 (2021).

51. Pancotti, C., Benevenuta, S., Birolo, G., *et al.* Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Briefings in Bioinformatics* **23.** https://doi.org/10.1093/bib/bbab555 (2022).

52. Pucci, F., Schwersensky, M. & Rooman, M. Artificial intelligence challenges for predicting the impact of mutations on protein stability. *Current Opinion in Structural Biology* **72,** 161–168. https://doi.org/10.1016/j.sbi.2021.11.001 (2022).

53. Jia, L., Yarlagadda, R. & Reed, C. C. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. *PLOS ONE* **10** (ed Zhang, Y.) e0138022. https://doi.org/10.1371/journal.pone.0138022 (2015).

54. Masso, M. & Vaisman, I. I. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics* **24,** 2002–2009. https://doi.org/10.1093/bioinformatics/btn353 (2008).

55. Pucci, F., Bourgeas, R. & Rooman, M. Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Scientific Reports* **6.** https://doi.org/10.1038/srep23257 (2016).

56. Saraboji, K., Gromiha, M. M. & Ponnuswamy, M. N. Average assignment method for predicting the stability of protein mutants. *Biopolymers* **82,** 80–92. https://doi.org/10.1002/bip.20462 (2006).

57. Topham, C. M., Srinivasan, N. & Blundell, T. L. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Engineering Design and Selection* **10,** 7–21. https://doi.org/10.1093/protein/10.1.7 (1997).

58. Masso, M. & Vaisman, I. I. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. *Advances in Bioinformatics* **2014,** 1–7. https://doi.org/10.1155/2014/278385 (2014).

59. Jumper, J., Evans, R., Pritzel, A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. https://doi.org/10.1038/s41586-021-03819-2 (2021).

60. Esposito, D., Weile, J., Shendure, J., *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology* **20.** https://doi.org/10.1186/s13059-019-1845-6 (2019).

61. Rubin, A. F., Min, J. K., Rollins, N. J., *et al.* MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. *bioRxiv*. https://doi.org/10.1101/2021.11.29.470445 (2021).

62. Bava, K. A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research* **32,** 120D–121. https://doi.org/10.1093/nar/gkh082 (2004).

63. Gromiha, M. M., An, J., Kono, H., *et al.* ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Research* **27,** 286–288. https://doi.org/10.1093/nar/27.1.286 (1999).

64. Gromiha, M. M. ProTherm, version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research* **28,** 283–285. https://doi.org/10.1093/nar/28.1.283 (2000).

65. Gromiha, M. M. , Thermodynamic Database for Proteins and MProThermutants: developments in version 3.0. *Nucleic Acids Research* **30,** 301–302. https://doi.org/10.1093/nar/30.1.301 (2002).

66. Kumar, M. D. S. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research* **34,** D204–D206. https://doi.org/10.1093/nar/gkj103 (2006).

67. Sarai, A., Gromiha, M. M., An, J., *et al.* Thermodynamic databases for proteins and protein-nucleic acid interactions. *Biopolymers* **61,** 121–126. https://doi.org/10.1002/1097-0282(2002)61:2%3C121::aid-bip10077%3E3.0.co;2-1 (2001).

68. Nikam, R., Kulandaisamy, A., Harini, K., *et al.* ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Research* **49,** D420–D424. https://doi.org/10.1093/nar/gkaa1035 (2020).

69. Xavier, J. S., Nguyen, T.-B., Karmarkar, M., *et al.* ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Research* **49,** D475–D479. https://doi.org/10.1093/nar/gkaa925 (2020).

70. Dehouck, Y., Kwasigroch, J. M., Gilis, D., *et al.* PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC bioinformatics* **12,** 1–12 (2011).

71. Nair, P. S. & Vihinen, M. VariBench: A Benchmark Database for Variations. *Human Mutation* **34,** 42–49. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22204. https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.22204 (2013).

72. Akdel, M., Pires, D. E. V., Pardo, E. P., *et al.* A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology* **29,** 1056–1067. https://doi.org/10.1038/s41594-022-00849-w (2022).

73. Blondel, M., Teboul, O., Berthet, Q., *et al. Fast Differentiable Sorting and Ranking* in *INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 119* (eds Daume, H. & Singh, A.) **119.** International Conference on Machine Learning (ICML), ELECTR NETWORK, JUL 13-18, 2020 (2020).

74. Buel, G. R. & Walters, K. J. Can AlphaFold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology* **29,** 1–2. https://doi.org/10.1038/s41594-021-00714-2 (2022).

75. Hu, M., Yuan, F., Yang, K. K., *et al. Exploring evolution-aware & -free protein language models as protein function predictors* 2022. arXiv: 2206.06583 [q-bio.QM].

76. McBride, J. M., Polev, K., Reinharz, V., *et al.* AlphaFold2 can predict single-mutation effects on structure and phenotype. *bioRxiv*. https://doi.org/10.1101/2022.04.14.488301 (2022).

77. Pak, M. A., Markhieva, K. A., Novikova, M. S., *et al.* Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLOS ONE* **18** (ed Budisa, N.) e0282689. https://doi.org/10.1371/journal.pone.0282689 (2023).

78. Zhang, Y., Li, P., Pan, F., *et al.* Applications of AlphaFold beyond Protein Structure Prediction. *bioRxiv*. eprint: https://www.biorxiv.org/content/early/2021/11/04/2021.11.03.467194.full.pdf. https://www.biorxiv.org/content/early/2021/11/04/2021.11.03.467194 (2021).

Supplementary Materials

Improving the prediction of protein stability changes upon mutations by geometric learning and a pre-training strategy

Yunxin Xu[1,2], Di Liu[1,2], and Haipeng Gong[1,2,*]

[1]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, 100084, China

[2]Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, 100084, China

[*]To whom correspondence should be addressed: hgong@tsinghua.edu.cn

Table S1. The details of DMS datasets for GeoFitness training.

Dataset name	Source	Type	Detailed Type	Number
BRCA1_HUMAN_BRCT	DeepSequence	Binding ability	Protein interaction	1,262
B3VI55_LIPSTSTABLE	DeepSequence	Enzyme function	Transferase activity	7,890
RL401_YEAST_Fraser2016	DeepSequence	Cell process	Protein modifer	1,253
BG_STRSQ_hmmerbit	DeepSequence	Enzyme function	Hydrolase activity	2,999
BF520_env_Bloom2018	DeepSequence	Binding ability	Protein interaction	12,577
B3VI55_LIPST_Whitehead2015	DeepSequence	Enzyme function	Transferase activity	7,631
GAL4_YEAST_Shendure2015	DeepSequence	Binding ability	DNA binding	1,195
TIM_SULSO_b0	DeepSequence	Enzyme function	Lyase activity	1,519
POLG_HCVJF_Sun2014	DeepSequence	Binding ability	drug binding	1,630
HIS7_YEAST_Kondrashov2017	DeepSequence	Enzyme function	Lyase activity	168
MTH3_HAEAESTABILIZED_Tawfik2015	DeepSequence	Enzyme function	Transferase activity	1,777
AMIE_PSEAE_Whitehead	DeepSequence	Enzyme function	Hydrolase activity	6,227
P84126_THETH_b0	DeepSequence	Enzyme function	Lyase activity	1,519
IF1_ECOLI_Kishony	DeepSequence	Binding ability	DNA binding	1,367
PA_FLU_Sun2015	DeepSequence	Enzyme function	Transferase activity	1,820
KKA2_KLEPN_Mikkelsen2014	DeepSequence	Enzyme function	Transferase activity	4,960
TIM_THEMA_b0	DeepSequence	Enzyme function	Lyase activity	1,519
BG505_env_Bloom2018	DeepSequence	Binding ability	Protein interaction	12,729
HG_FLU_Bloom2016	DeepSequence	Binding ability	Protein interaction	10,715
BRCA1_HUMAN_RING	DeepSequence	Enzyme function	Transferase activity	575
BLAT_ECOLX_Ranganathan2015	DeepSequence	Enzyme function	Hydrolase activity	4,996
RASH_HUMAN_Kuriyan	DeepSequence	Binding ability	Protein interaction	3,134
00000001-a-1	MaveDB	Enzyme function	Transferase activity	3,021
00000001-b-1	MaveDB	Cell process	Protein modifer	1,919
00000001-c-2	MaveDB	Cell process	Calmodulin function	2,831
00000001-d-2	MaveDB	Enzyme function	Transferase activity	4,617
00000002-a-2	MaveDB	Binding ability	Protein interaction	513
00000004-a-3	MaveDB	Enzyme function	Transferase activity	940

Table S1. Continued from previous page

Dataset name	Source	Type	Detailed Type	Number
00000005-a-3	MaveDB	Enzyme function	Lyase activity	10,450
00000010-a-1	MaveDB	Binding ability	RNA binding	1,188
00000011-a-1	MaveDB	Cell process	Chaperon function	171
00000013-b-1	MaveDB	Expression	Protein abundancy	3,689
00000035-a-3	MaveDB	Enzyme function	Oxidoreductase activity	16,853
00000036-a-2	MaveDB	Binding ability	Protein interaction	5,814
00000041-a-1	MaveDB	Enzyme function	Transferase activity	3,714
00000043-a-2	MaveDB	Cell process	Receptor activation	562
00000044-b-2	MaveDB	Expression	Protein abundancy	3,798
00000045-h-1	MaveDB	Cell process	Cellular toxicity	2,599
00000046-a-2	MaveDB	Cell process	Resistence to ubiquitination	467
00000047-a-1	MaveDB	Expression	Cell surface protein abundancy	6,639
00000048-a-1	MaveDB	Expression	Cell surface protein abundancy	6,645
00000049-a-7	MaveDB	Enzyme function	Oxidoreductase activity	11,357
00000050-a-1	MaveDB	Cell process	DNA repair	16,749
00000055-a-1	MaveDB	Expression	Protein abundancy	2,922
00000059-a-1	MaveDB	Binding ability	DNA binding	2,967
00000060-a-1	MaveDB	Cell process	Cellular toxicity	1,342
00000061-a-1	MaveDB	Binding ability	Protein interaction	297
00000063-b-1	MaveDB	Enzyme function	Oxidoreductase activity	2,527
00000064-a-1	MaveDB	Binding ability	Protein interaction	1,870
00000065-a-1	MaveDB	Cell process	GDP-GTP exchange	2,997
00000066-a-1	MaveDB	Cell process	Synaptic transmission	4,365
00000067-a-1	MaveDB	Enzyme function	Oxidoreductase activity	3,756
00000068-a-1	MaveDB	Cell process	Cell cycle (p53)	7,467
00000069-a-1	MaveDB	Binding ability	Protein interaction	2,223
00000071-a-1	MaveDB	Enzyme function	Hydrolase activity	2,404
00000072-a-1	MaveDB	Binding ability	Protein interaction	666

Table S1. Continued from previous page

Dataset name	Source	Type	Detailed Type	Number
00000073-c-1	MaveDB	Binding ability	Substrate binding	5,023
00000074-a-1	MaveDB	Cell process	Chaperon function	4,323
00000075-a-1	MaveDB	Binding ability	Protein interaction	9,576
00000076-a-1	MaveDB	Cell process	Subcellular localization	507
00000077-a-1	MaveDB	Binding ability	Protein interaction	2,986
00000077-b-1	MaveDB	Binding ability	Protein interaction	2,777
00000078-b-1	MaveDB	Expression	Protein abundancy	2,695
00000079-a-1	MaveDB	Expression	Protein abundancy	4,840
00000080-a-2	MaveDB	Fluorescence	Fluorescence intensity	1,084
00000084-a-1	MaveDB	Cell process	Cellular toxicity	1,176
00000087-a-1	MaveDB	Enzyme function	Transferase activity	9,462
00000090-b-1	MaveDB	Binding ability	Protein interaction	1,576
00000094-a-3	MaveDB	Cell process	Channel function	3,319
00000095-a-1	MaveDB	Enzyme function	Oxidoreductase activity	6,142
00000096-a-1	MaveDB	Enzyme function	Transferase activity	8,570
00000099-a-1	MaveDB	Expression	Cell surface protein abundancy	165
00000102-0-1	MaveDB	Expression	Protein abundancy	5,083
00000103-d-1	MaveDB	Enzyme function	Transferase activity	6,810

Table S2. Features adopted by the GeoStab-suite.

Feature	Shape	Source or meaning
1D information	[L, 1280]	Calculated by ESM-2 (esm2_t33_650M_UR50D)
Physicochemical properties	[L, 7]	Including steric parameters, hydrophobicity, volume, polarizability, isoelectric point, helix probability and sheet probability
pLDDT	[L, 1]	Reported by AlphaFold2
2D information	[L, L, 7]	Calculated from coordinates
Coordinate (x, y, z)	[L, 14, 3]	Obtained from the PDB file
esm1v-logits	5	Predicted by five ESM-1v models.
pH&temp	2	pH and temperature

L refers to the number of residues of the target protein.

Table S3. Side-by-side comparison of fitness prediction models on the DMS database.

Dataset name	GeoFitness	ECNet	SESNet	RF_{joint}	MSA Transformer	ESM-2	ESM-1b	ESM-1v
BRCA1_HUMAN_BRCT	0.55	0.66	0.45	0.54	0.53	0.55	0.54	0.36
B3VI55_LIPSTSTABLE	0.69	0.71	0.67	0.56	0.55	0.56	0.49	0.58
RL401_YEAST_Fraser2016	0.65	0.64	0.70	0.36	0.37	0.51	0.14	0.20
BG_STRSQ_hmmerbit	0.71	0.70	0.62	0.60		0.55	0.60	0.49
BF520_env_Bloom2018	0.78	0.82	0.76	0.52		0.01	0.43	0.48
B3VI55_LIPST_Whitehead2015	0.44	0.43	0.45	0.31	0.28	0.29	0.26	0.29
GAL4_YEAST_Shendure2015	0.70	0.66	0.67	0.41	0.37	0.63	0.60	0.41
TIM_SULSO_b0	0.75	0.78	0.74	0.60	0.64	0.65	0.59	0.58
POLG_HCVJF_Sun2014	0.81	0.79	0.75	0.58		0.12	0.08	0.55
HIS7_YEAST_Kondrashov2017	0.10	0.07	0.14	-0.02	0.03	0.08	-0.05	-0.04
MTH3_HAEAESTABILIZED_Tawfik2015	0.72	0.72	0.64	0.51	0.58	0.50	0.60	0.68
AMIE_PSEAE_Whitehead	0.81	0.77	0.73	0.61	0.53	0.57	0.59	0.59
P84126_THETH_b0	0.84	0.87	0.79	0.64	0.64	0.65	0.62	0.59
IF1_ECOLI_Kishony	0.83	0.78	0.68	0.53	0.46	0.54	0.53	0.54
PA_FLU_Sun2015	0.53	0.40		0.49		-0.00	0.03	0.01
KKA2_KLEPN_Mikkelsen2014	0.77	0.78	0.77	0.44	0.43	0.58	0.57	0.60
TIM_THEMA_b0	0.80	0.74	0.76	0.46	0.48	0.53	0.50	0.49
BG505_env_Bloom2018	0.78	0.80		0.50		0.02	0.43	0.45
HG_FLU_Bloom2016	0.77	0.64	0.74	0.51		0.48	0.15	0.49
BRCA1_HUMAN_RING	0.56	0.67	0.65	0.45	0.39	0.49	0.51	0.45
BLAT_ECOLX_Ranganathan2015	0.89	0.88	0.86	0.59	0.51	0.72	0.68	0.64
RASH_HUMAN_Kuriyan	0.83	0.82	0.77	0.46	0.47	0.47	0.29	0.32
00000001-a-1	0.72	0.70	0.70	0.39	0.36	0.47	0.39	0.48
00000001-b-1	0.60	0.65	0.56	0.58	0.57	0.52	0.48	0.53
00000001-c-2	0.22	0.20	0.25	0.16	0.20	0.12	0.17	0.14
00000001-d-2	0.38	0.29	0.23	0.24	0.25	0.22	0.22	0.19
00000002-a-2	0.85	0.83	0.71	0.62	0.67	0.68	0.71	0.78
00000004-a-3	0.59	0.58	0.39	0.55	0.43	0.53	0.44	0.53

Table S3. Continued from previous page

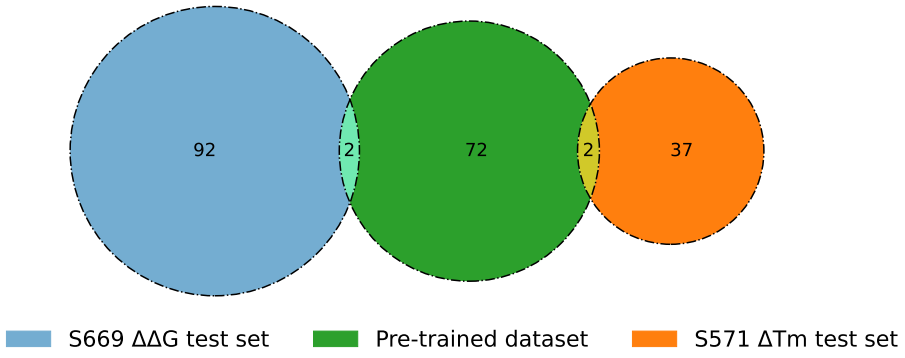
Dataset name	GeoFitness	ECNet	SESNet	RF_{joint}	MSA Transformer	ESM-2	ESM-1b	ESM-1v
00000005-a-3	0.65	0.61	0.62	0.54		0.52	0.50	0.52
00000010-a-1	0.82	0.80	0.78	0.46	0.45	0.50	0.58	0.48
00000011-a-1	0.87	0.63	0.43	0.37	0.37	0.02	0.26	0.45
00000013-b-1	0.59	0.59	0.63	0.43	0.39	0.57	0.58	0.53
00000035-a-3	0.69	0.65		0.30		0.51	0.25	0.32
00000036-a-2	0.40	0.37	0.38	0.02	-0.01	0.17	0.13	0.14
00000041-a-1	0.77	0.74	0.71	0.34		0.54	0.51	0.58
00000043-a-2	0.66	0.65	0.70	0.35	0.35	0.30	0.33	0.28
00000044-b-2	0.87	0.85	0.84	0.41	0.33	0.01	0.05	-0.03
00000045-h-1	0.60	0.63	0.59	0.25	0.07	0.26	0.35	0.35
00000046-a-2	0.36	0.31	0.25	0.08	0.02	0.02	0.03	0.04
00000047-a-1	0.52	0.49	0.50	0.36	0.36	0.38	0.40	0.38
00000048-a-1	0.52	0.46	0.46	0.25	0.27	0.23	0.24	0.24
00000049-a-7	0.53	0.49	0.52	0.23		0.28	0.32	0.23
00000050-a-1	0.59	0.57		0.43		0.34	0.37	0.39
00000055-a-1	0.81	0.81	0.81	0.43	0.48	0.41	0.44	0.44
00000059-a-1	0.76	0.75	0.72	0.43	0.48	0.68	0.51	0.51
00000060-a-1	0.65	0.76	0.71	0.00	-0.17	0.15	-0.05	-0.01
00000061-a-1	0.54	0.44	0.29	0.50	0.54	0.42	0.48	0.48
00000063-b-1	0.67	0.57	0.65	0.25	0.22	0.30	0.29	0.27
00000064-a-1	0.33	0.27	0.20	0.03	0.08	0.00	0.02	-0.00
00000065-a-1	0.25	0.21	0.08	0.06	0.08	0.06	0.08	0.07
00000066-a-1	0.20	0.20	0.19	0.17		0.11	0.15	0.13
00000067-a-1	0.16	0.10	0.17	0.17	0.16	0.12	0.12	0.11
00000068-a-1	0.84	0.87	0.85	0.39	0.15	0.39	0.55	0.51
00000069-a-1	0.70	0.65		0.24		0.21	0.25	0.13
00000071-a-1	0.24	0.05				0.18		
00000072-a-1	0.22	0.19	0.35	0.13	0.13	0.25	0.30	0.39

Table S3. Continued from previous page

Dataset name	GeoFitness	ECNet	SESNet	RF_{joint}	MSA Transformer	ESM-2	ESM-1b	ESM-1v
00000073-c-1	0.86	0.86	0.81	0.63	0.52	0.73	0.67	0.62
00000074-a-1	0.79	0.67		0.51		0.42	0.48	0.52
00000075-a-1	0.39	0.10	0.31	0.29		0.14	-0.01	-0.03
00000076-a-1	0.64	0.67	0.67	0.38	0.42	0.47	0.47	0.40
00000077-a-1	0.62	0.59	0.50	0.02	0.04	0.08	0.08	0.08
00000077-b-1	0.74	0.71	0.74	0.03	0.04	-0.01	0.05	0.11
00000078-b-1	0.77	0.82	0.78	0.54	0.52	0.47	0.40	0.42
00000079-a-1	0.48	0.39	0.50	0.16	0.17	0.18	0.20	0.20
00000080-a-2	0.62	0.43	0.51	0.62	0.54	0.07	0.04	0.06
00000084-a-1	0.97	0.93	0.24	0.44	0.34	0.40	0.32	0.24
00000087-a-1	0.37	0.14	0.10	0.26		0.03	0.04	0.02
00000090-b-1	0.74	0.71	0.68	0.20	0.22	0.33	0.22	0.20
00000094-a-3	0.31	0.17		0.11		0.16	0.17	0.10
00000095-a-1	0.84	0.80	0.80	0.55		0.68	0.53	0.62
00000096-a-1	0.64	0.63	0.66	0.53		0.52	0.44	0.47
00000099-a-1	0.80	0.77	0.57	0.61	0.73	0.67	0.45	0.59
00000102-0-1	0.67	0.65	0.64	0.38	0.41	0.46	0.46	0.41
00000103-d-1	0.37	0.37	0.33	0.20	0.18	0.20	0.20	0.21

Data listed in the table represent the Spearman correlation coefficient between predicted and experimental values.

a.



b.

Protein name (pre-trained dataset)	Label (pre-trained dataset)	Protein name (test dataset)	Label (test dataset)
00000090-b-1	PSD95^{pdz3}/T-$_2$F PSD95^{pdz3}/CRIPT protein-protein interaction	rcsb_3K82_A_wt rcsb_5JXB_A_wt	ΔΔG
00000061-a-1	KRAS4b/RAF protein-protein interaction	rcsb_1GUA_B_wt	ΔΔG
00000084-a-1	CcdB mutation effect on bacterial toxicity	rcsb_3VUB_A_wt	ΔTm
RL401_YEAST_Fraser2016	Ubiquitin mutation effect on yeast growth	rcsb_1OTR_B_wt	ΔTm

Figure S1. Lack of overlap between the datasets used for model pre-training and benchmark testing. a) The DMS dataset adopted for pre-training the GeoFitness model contains data collected from 74 proteins in total, among which 2 protein entries exhibit sequence identity > 40% with the proteins adopted by S669 and S571, the test datasets for $\Delta\Delta G$ and ΔT_m predictions, respectively. **b)** The labels of the shared proteins in the pre-training dataset are, however, totally different from the metrics of protein stability, say $\Delta\Delta G$ and ΔT_m. Hence, there is no overlap between the datasets for model pre-training and benchmark testing, thereby precluding the possibility of information leaking during performance evaluation.

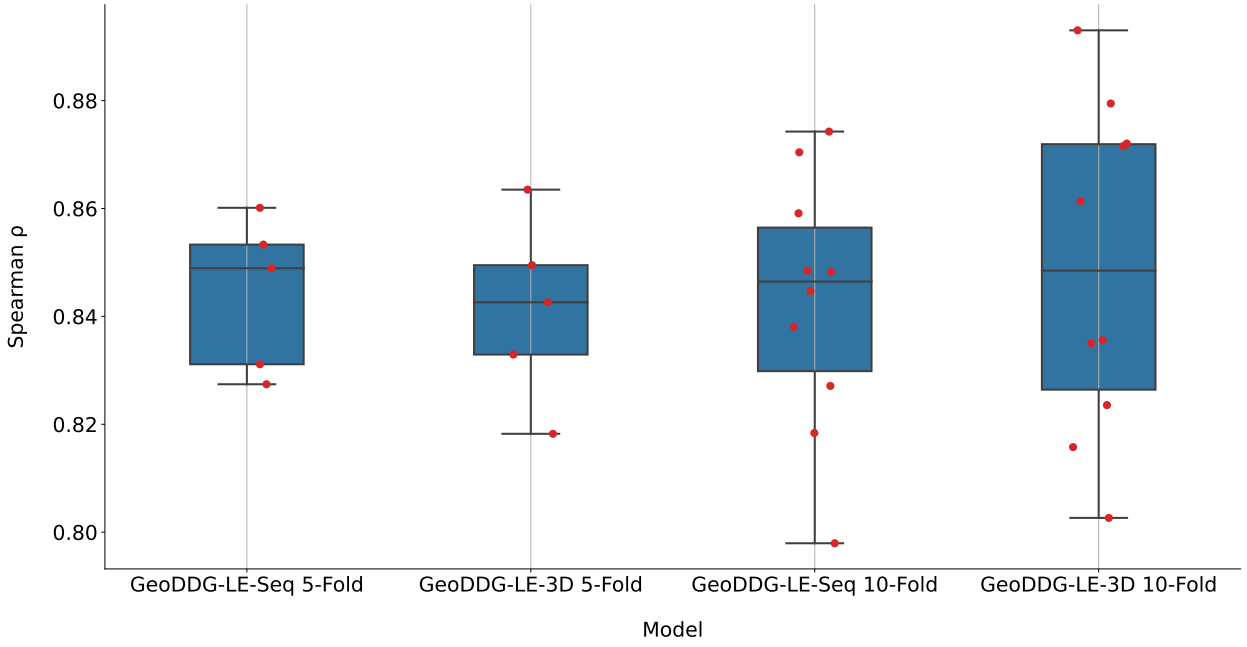


Figure S2. 5-fold and 10-fold cross validation for GeoDDG-LE on the S2648 dataset. The "Seq" and "3D" versions of GeoDDG-LE are evaluated on S2648 with both 5-fold and 10-fold cross validations. Each red dot refers to an individual result of the cross-validation and the overall results are shown by the box plot.

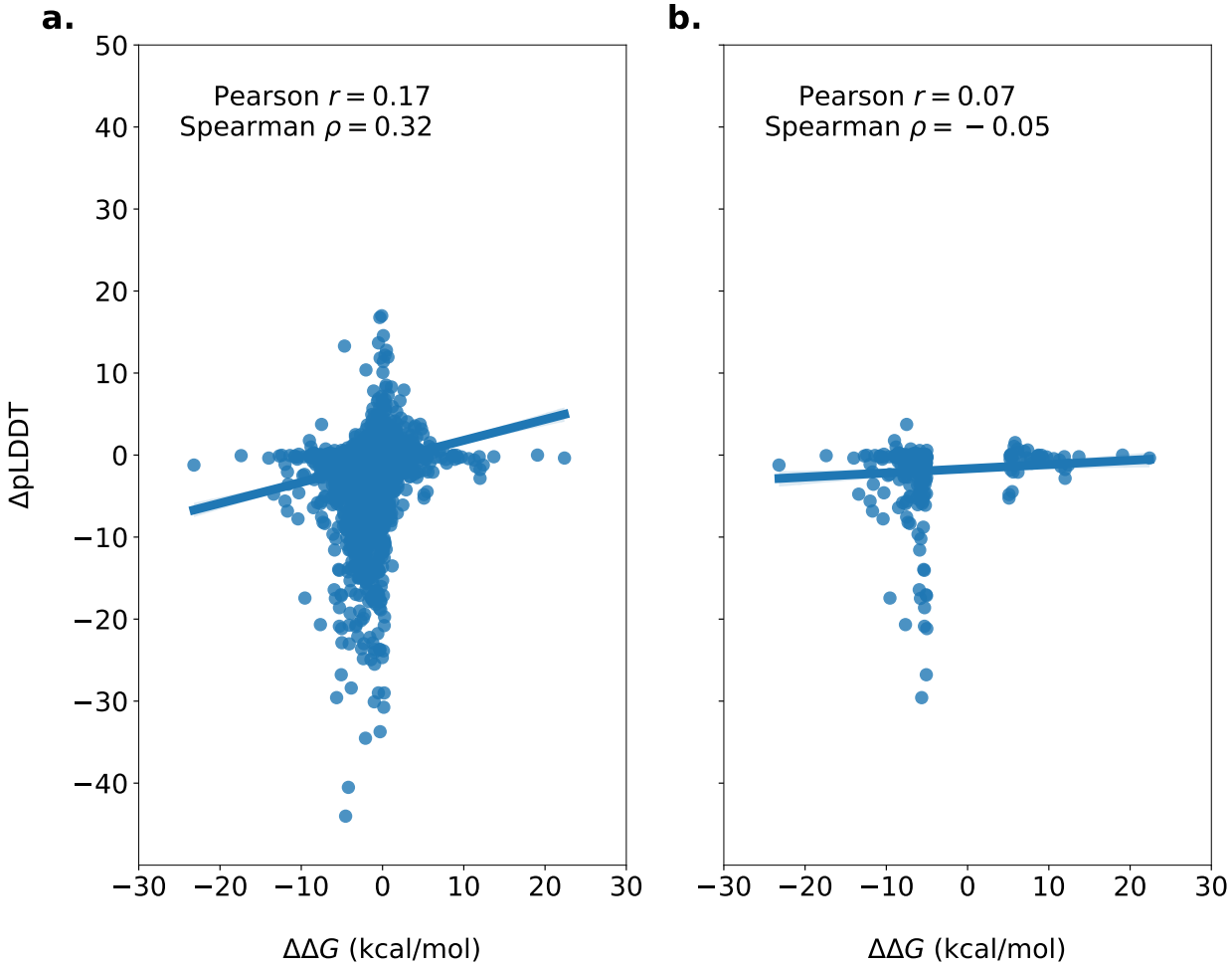


Figure S3. The correlation between $\Delta\Delta G$ from experiments and ΔpLDDT by AlphaFold2. For each individual protein variant (shown as a dot in the figures), we calculated the difference of its ΔG and pLDDT (reported by AlphaFold2) with the corresponding value of the wide-type protein to evaluate the mutational effects on protein stability and AlphaFold2 prediction. **a)** All protein variants are considered. **b)** Only protein variants of high significance (*i.e.* $|\Delta\Delta G| > 5$ kcal/mol) are considered.