Fitness translocation: improving variant effect prediction with biologically-grounded data augmentation

Adrien Mialland
Artificial Intelligence Research Center, AIST [1]
mialland.a@gmail.com

Shuzo Fukunaga
University of Toronto
sfukunaga.academia@gmail.com

Riku Katsuki
The University of Electro-Communications
k2343004@gl.cc.uec.ac.jp

Yunfei Dong
The University of Tokyo
dong_yunfei_23@stu-cbms.k.u-tokyo.ac.jp

Hideki Yamaguchi
The University of Tokyo
yamaguchi.h.et.al@gmail.com

Yutaka Saito
Artificial Intelligence Research Center, AIST [1]
The University of Tokyo
School of Frontier Engineering, Kitasato University
saito.yutaka@kitasato-u.ac.jp

Abstract

Navigating the protein fitness landscape is critical for understanding sequence-function relationships and improving variant effect prediction. However, the limited availability of experimentally measured functional data poses a significant bottleneck. To address this, we present a novel data augmentation strategy called fitness translocation, which leverages fitness landscapes from related proteins to enhance the performance of variant effect predictors on a target protein. Using embeddings from protein language models and by translocating the features within the sequence space, we transfer the fitness information from homologous protein datasets to a target protein to augment its dataset. Our approach was evaluated across diverse protein species, including IGPS orthologs, GFP orthologs, and SARS-CoV-2 spike proteins strains for cell entry and ACE2 binding. The results demonstrate consistent and substantial improvements in predictive performances, particularly for datasets with limited training data. Furthermore, we introduce a systematic selection framework for identifying the most beneficial protein datasets for augmentation and optimizing predictive gains. This study highlights the potential of related protein fitness translocation to advance protein engineering and variant effect prediction. The implementation of the method is available at https://github.com/adrienmialland/ProtFitTrans.

***Keywords*—** machine learning, protein engineering, directed evolution, library design, training data, sequence space exploration

[1] National Institute of Advanced Industrial Science and Technology

1 Introduction

An effective mapping of protein sequence-function relationship has the potential to address a wide range of challenges, from the design of new proteins with useful properties to the acceleration of drug discovery or the prediction of mutation impacts [1]. The function of a protein is essentially its biological role, and a fundamental principle of biology is that the amino acid sequence determines the function. Therefore, the creation of protein variants through the mutation of one or multiple amino acids would impact the protein's function, and the possibility to map each variant to its function would give access to the fitness landscape [2]. So, navigating the fitness landscape would allow for the identification of higher regions and therefore improved proteins.

However, given a number k of critical amino acids to mutate, 20^k variants would need to be created and evaluated to get a comprehensive understanding of the fitness landscape, which rapidly results in a prohibitively large sequence space. While recent advancements in genetic engineering allows for the creation of extremely large libraries of sequences [3–5], measuring the functional data for each individual sequence remains costly and several orders of magnitude slower than the sequence generation [4]. So, this bottleneck prevents effective fitness landscape exploration and has recently motivated the use of machine learning (ML) methods to learn the sequence-function mapping [1, 6, 7, 4, 8, 9].

These computational approaches enable the construction of fitness landscape models from experimental data, allowing for the prediction of unobserved protein sequences and to extrapolate into unexplored and distant region of the sequence space [10]. However, ML models are often trained with limited data, and their effectiveness significantly depends on the diversity of the training dataset. Therefore, even though different datasets or ML models may support the exploration of distinct yet desirable subspaces within the sequence space [1, 11], achieving sufficient diversity would still require an extensive number of experiments to gather a comprehensive range of sequence-function relationships. To address these limitations, the field of ML has introduced various data augmentation methods across different applications [12, 13], but an effective method is still lacking for protein variant effect prediction tasks. However, a popular data augmentation approach consists in the addition of closely related but easily accessible data. Recent investigations on the evolutionary constraints imposed by sequence and structural requirements, have highlighted the significant correlations in the fitness landscape that distant orthologues may have [15]. This suggest that fitness landscapes of protein species from a given protein family can be translocated in the sequence spaces, which would allow for the construction of extended datasets.

In this study, we propose a data augmentation method for variant effect prediction based on *fitness translocation*. It aims to enhance the predictive performances of a target protein variants by leveraging data from related proteins, where their fitness landscape is translocating to the target protein within the sequence space. To evaluate our approach, we applied deep representation learning [14] using Transformer models on datasets from IGPS [15], GFP [16], and SARS-CoV-2 spike proteins for both Cell Entry and ACE2 Binding [17], where training data are available for variant effect prediction on respectively three, three, tow, and two protein species. For each family, we trained a variant effect predictor using data from one protein species as the target and compared its performances to a predictor trained with the addition of data from the remaining proteins. It showed that augmented training data significantly improved the performances. Also, we developed a method for selecting the most beneficial additional protein for data augmentation. These results indicate that data from previous protein engineering efforts can be leveraged to predict variant effects in proteins with limited available data.

2 Methods

2.1 Variant Effect Prediction Task

To capture intricate structural and functional features from amino acids sequence data, transformer-based deep representation learning methods with protein language models (pLM) were employed [14]. Given a target protein and additional protein datasets, this allowed us to obtain the corresponding embeddings of each sequence data along with the associated fitness values.

Then, a downstream variant effect prediction task was developed based on the embeddings datasets, to assess the impact of our data augmentation method through fitness translocation. Specifically, for a given target protein and a set of additional proteins, the prediction task considered various dataset combinations, including both augmented target protein data through fitness translocation and original target protein data with no translocation.

The variant effect prediction models were evaluated using a nested cross-validation (CV) scheme [18]. Concretely, when evaluating a non-augmented dataset, for each outer loop of the nested CV, a new training set with a fixed size S was randomly sampled, and the remaining data was used as the new test set. Within each loop, an inner CV was applied to the training set to identify the best model hyper-parameters. The optimized model was then retrained on the entire training set and evaluated on the test set. This procedure was repeated over N outer loop to estimate the performances. When evaluating an augmented dataset, the nested CV procedure remained the same, except that the additional part of the dataset was entirely included in the training set for each loop,

while the test set was not changed. Each model was therefore consistently exposed to the augmented data to allow for a reliable comparison of any performances increase attributable to the translocated data.

2.2 Fitness Translocation Core Principle

Using the embeddings obtained from a pLM, with one protein type used as a target and another protein type as additional data, the translocation of the fitness landscape was performed by transferring the feature embeddings of the additional protein variants, to the target protein wildtype. More specifically, the differences between the embeddings of the additional protein variants and their corresponding wildtype are computed in the feature space. These differences are then applied to the embedding of the target wildtype, allowing for the use of new variants that preserve their biologically relevant characteristics. Therefore, the procedure can be repeated for any number of additional protein datasets.

2.3 Fitness Translocation Dataset Selection

The decision to include an additional dataset in the final augmented dataset was evaluated based on the resulting improvement in the protein variant effect prediction task. Different criteria were considered to assess the improvement: (i) All additional datasets were translocated at once. The resulting dataset was kept if the target protein performances was improved. (ii) The additional datasets were translocated and assessed individually. The ones that improved the performances were then selected and included in the final augmented dataset. (iii) The additional datasets selected through (ii) were used in a greedy algorithm [19], which consists in choosing the dataset that seems best at the moment. Specifically, the selected datasets were successively included from the best performing to the lowest performing. The dataset currently included were then kept only if it improved the performances of the augmented dataset. The final augmented dataset should therefore contain only the additional datasets that consistently increase the performances. (iv) The significance of the performances improvement in (ii) were evaluated statistically with a paired t-test [20]. Then, the significant ones underwent the greedy selection in (iii). Finally, for the sake of clarity, the remaining of the paper will be referring to each of the four criteria as, respectively: No-Selection, Individual-Select, Individual-Greedy, Statistical-Greedy.

2.4 Fitness Translocation Method

The protein variant effect prediction task was used to express the dataset augmentation and selection process of the fitness translocation method. Therefore, for a given target protein with additional datasets and a selection criterion, all possible combinations of fitness translocation specified by the selection criterion were considered, including no-translocation. The task was then run iteratively to seek for the final augmented dataset that yielded the best improvement in performances.

3 Experiments

3.1 Datasets

Indole-3-Glycerol Phosphate Synthase (IGPS) [15]**.** Three orthologous TIM barrel fold of the IGPS family from thermophilic species S. solfataricus (SsIGPS), T. maritima (TmIGPS), and T. thermophilus (TtIGPS), underwent targeted mutations across eight 10-residue segments, covering β-barrel cores and adjacent α-β loops, using EMPIRIC deep mutational scanning [21], which yielded a 30-40% identity. The fitness of each mutant was assessed in a yeast strain lacking endogenous IGPS, with selection coefficients quantifying the relative growth of each mutant compared to the wildtype over time. A total of 1680 (10 AA $\times$ 8 $\times$ 21 mutation types) mutations were created per ortholog. So, excluding stop codon and MmeI, the resultant library contained 1497, 1502, and 1489 variant for SsIGPS, TmIGPS, and TtIGPS species respectively.

Green Fluorescent Protein (GFP) [16]**.** Three orthologs of the GFP family from Aequorea macrodactyla (amacGFP), Clytia gregaria (cgreGFP), and Pontellina plumata (ppluGFP2) were used to generate mutant libraries through random mutagenesis by error-prone PCR, introducing an average of four mutations per variant. amacGFP and cgreGFP had 17% and 19% identity to ppluGFP, respectively, while amacGFP and cgreGFP shared 43% identity. Each GFP variant was fused to the red fluorescent protein mKate2 for expression normalization. To measure fitness, constructs were integrated into the E. coli genome, and cells were sorted by fluorescence intensity within a narrow red fluorescence gate to control for gene expression level and other errors. The resultant library contained 35500, 26165, and 32260 variants for amacGFP, cgreGFP, and ppluGFP, species, respectively.

SARS-CoV-2 Spike Protein [17]**.** Two strains of the SARS-CoV-2 spike protein family from XBB.1.5 and BA.2 were used to generate a dataset using pseudovirus-based deep mutational scanning [22], designed to assess each variants impact on spike-mediated cell entry and ACE2 binding. Spike mutants were created

to include mutations frequently observed in natural SARS-CoV-2 evolution, as well as targeted changes at functionally significant sites, covering over 7,000 unique amino acid changes across the spike protein. Each variant contains, on average, two mutations. Cell entry efficiency was measured in a 293T cell line expressing the ACE2 receptor, where pseudovirus entry into cells was quantified. ACE2 binding was assessed by leveraging neutralization assays with soluble ACE2, where the neutralization potency provided a proxy for receptor-binding affinity. The final library contained 7129 BA.2 and 7029 XBB.1.5 variants for cell entry, and 6257 BA.2 and 6106 XBB.1.5 variants for ACE2 bindings.

3.2 Configurations Evaluated

Deep Representation Learning Models The embeddings for each wildtype protein and their variants were generated with both ESM-1v [23] and ESM2 [24] pre-trained models, taking the mean representation of their final layers. This resulted in 1280-dimensional embeddings for both models.

Variant Effect Prediction Task For a given embeddings dataset, training size S, and untrained predictive model, the prediction task was built using the nested CV as follow. First off, the dataset splits of each outer loops were obtained by shuffling the dataset and extracting non-overlapping samples of size S as the train sets, while using the remaining data as the test sets. The dataset was then shuffled again, and additional train/test splits were extracted until all the splits for the N outer loops were obtained. In addition, for each outer loop, a 5 folds CV was performed for the inner loop, coupled with a grid-search to optimize the model's hyper-parameters. Then, the best performing model was retrained on the entire training set, and its performances were evaluated using Spearman's correlation between the true fitness values and the predicted fitness values on the entire test set.

Fitness Translocation Evaluation. Applying the fitness translocation method should only select the set of additional protein datasets that improve the performances of a protein variant effect prediction task. Therefore, both the selection criterion and the ability of the method to generalize to new data should be evaluated. To do so, the evaluation was performed in two steps.

First, the fitness translocation method was applied without the dataset selection step to compute every possible combination of translocation, including no-translocation, which allows to assess the core principle of the method. This procedure is repeated within different configuration for a comprehensive assessment of the fitness translocation method. That is, for a given protein family, all species were successively considered as the target, while the remaining species were considered as additional protein datasets for translocation. The same is then conducted for all protein family included in this study. In addition, for a given target protein, three prediction models were considered: Support Vector Regressor (SVR), Random Forest Regressor (RF), and Lasso. Finally, this entire experiment was performed with a total of 26 different target training sizes S, which resulted in training sets that contained from 45 up to 1125 variants. The remaining large amount of data was used as the test sets.

second, the fitness translocation method was applied with no modification, but in conditions that reproduces realistic scenarios, which often come with limited datasets. So, the same 26 target training sizes than the first evaluation step were considered, and the same train/test splits from the nested CV outer loops were obtained. However, for each outer loop, the test set was ignored and a new train/validation split (80%/20%) was obtained from the current training set. The variant effect prediction task was then built without modification, using the new split. This makes the resulting performances directly comparable to the first evaluation step. Finally, this second evaluation step was repeated by considering all dataset selection criteria successively.

So, the following results report the performances for all possible translocation, including no-translocation, obtained from the first evaluation step that used the full dataset. Then, the results from the selection criteria are shown as red lines to express what translocated combination has resulted from each selection criterion.

4 Results

We developed a new data augmentation method, called fitness translocation, that aims to enhance performances of a protein variant effect prediction task. It enables the transfer of fitness landscapes from multiple additional protein variant datasets to that of a target protein, while seeking to identify the translocated combination that maximizes the predictive performances on the target. The method was applied and evaluated on three protein family, namely IGPS orthologs [15], GFP orthologs [16], and SARS-CoV-2 spike proteins both for cell entry and ACE2 binding [17]. For a given target protein, the variant effect prediction task was built with a nested CV of $N = 50$ outer loops, and the inner loop hyper-parameter grids were optimized prior to grid-searches.

4.1 Fitness translocation Core Principal

The assessment of all possible combination of fitness translocation, including no-translocation, reveals that incorporating additional protein data in a target protein dataset significantly improves the performances off a variant effect prediction task. As representative examples, Figures 1, 2, 3, 4 respectively show the results for SARS-CoV-2 Spike protein for cell entry and ACE2 bindings, IGPS, and GFP proteins using ESM2 pLM and SVR predictor. The additional results are reported in the supplemental file. The effect of fitness translocation was substantial for most target training sizes and was especially marked for smaller sizes. It was the biggest for SARS-CoV-2 cell entry (S1, S5), followed by SARS-CoV-2 ACE2 binding (S2, S6), IGPS (S3, S7) and GFP (S4, S8) families. In addition, the results of fitness translocation were essentially dependent on the type of predictor and the combination being translocated and showed consistency across various configurations. However, improvement on GFP family was limited and occurred essentially for smaller target training sizes in most cases, although some configurations yielded substantial improvements.

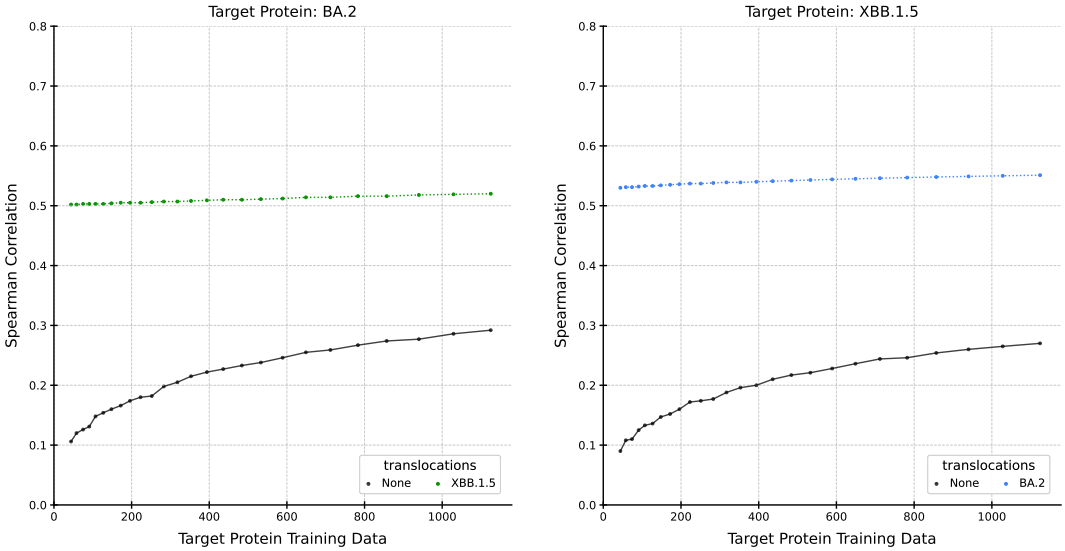

Figure 1: Results for SARS-CoV-2 spike protein Cell Entry, ESM2 pLM, and SVR predictor.

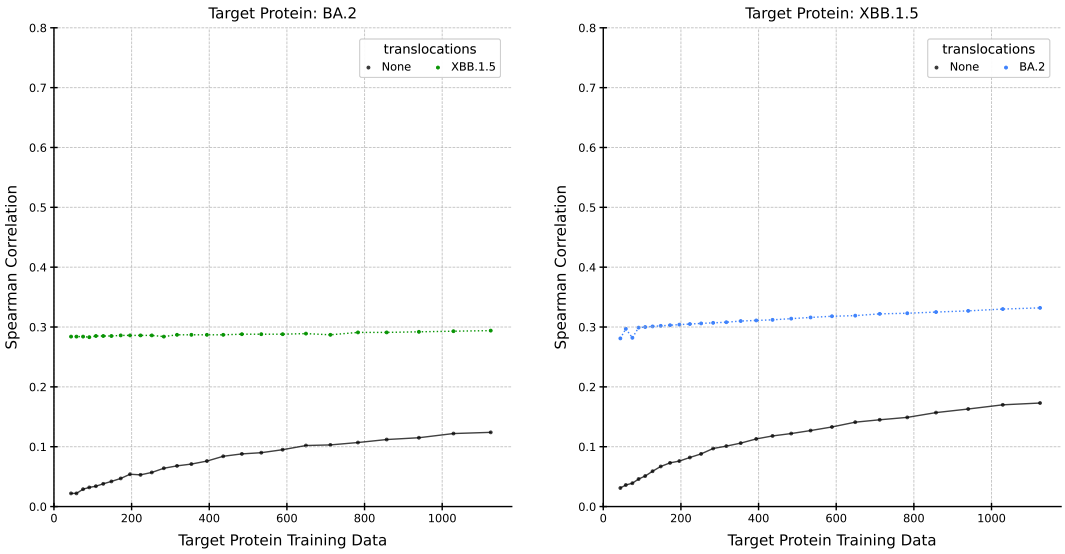

Figure 2: Results for SARS-CoV-2 spike protein ACE2 Bindings, ESM2 pLM, and SVR predictor.

Fitness Translocation Absolute Amplitudes trend. In all family − target − pLM − predictor configurations, the effect of fitness translocation eventually reached a saturation point as the target training sizes increased (S3, S4, S7, S8), or indicated a trend toward saturation (S1, S2, S5, S6) for target training sizes

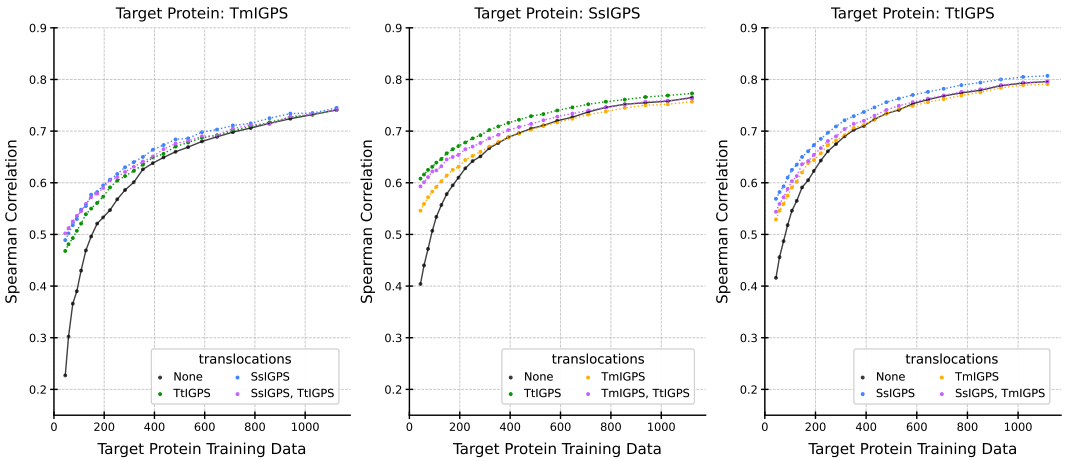

Figure 3: Results for IGPS orthologs, ESM2 pLM, and SVR predictor.

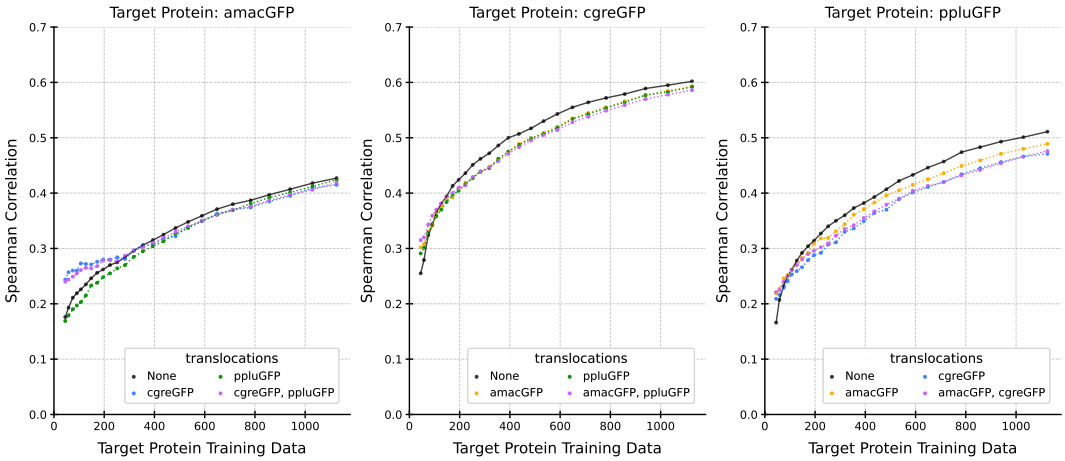

Figure 4: Results for GFP orthologs, ESM2 pLM, and SVR predictor.

beyond the maximum size assessed in this study. In other words, the absolute difference between the target without translocation and with translocation ultimately stopped increasing as the target training size increased. The SARS-COV-2 spike proteins for cell entry had 12 configurations tested (2 targets $\times$ 2 pLMs $\times$ 3 predictors, S1, S5) and had the biggest improvement, directly followed by SARS-CoV-2 spike proteins for ACE2 bindings that also had 12 configurations tested (S2, S6). In both cases, the results were consistent across all configurations and was substantial for all 26 target training sizes. The IGPS family had 18 configurations tested (3 targets $\times$ 2 pLMs $\times$ 3 predictors, S3, S7) and had the second biggest improvement. It was consistent across target $-$ pLM configurations, and reaches saturation toward the higher target training sizes, but was dependent on the predictor. Specifically, the RF predictor resulted in decreased performances for the higher target training sizes (S3.3, S7.3). The GFP family had 18 configurations tested (S5, S8) and had the lowest performances. It was consistent across pLM models, but showed varying trend over the target training sizes, and was dependent on both the target and the predictor. So, no clear trend emerged from the GFP family, but some configurations still provided substantial improvement for all 26 target training sizes (S4.2, S8.2).

Besides, as an overall trend, the final performances of the fitness translocation method were mainly dependent on the content of the translocated combination and the type of predictor, where specific combinations could either increase or decrease the original performances. In comparison, the effects of the pLMs were mostly marginal and essentially acted on the magnitude of the fitness translocation results, rather than the inherent capabilities of the additional datasets. Finally, the SARS-CoV-2 Spike proteins without translocation had the lowest performances but showed the biggest improvement, while the IGPS proteins had the highest performances and showed reduced improvement in comparison. This trend suggests an additional saturation mechanism, where the amplitude of the improvement is inversely proportional to the initial performances of the target protein. However, although the GFP family results seats in between SARS-COV-2 Spike proteins and IGPS families, it actually did not exhibit a clear pattern and had varying performances depending of the choice of species used as the target (S4, S8).

Fitness Translocation Relative Amplitudes trend. IGPS and GFP families having more protein species than SARS-COV-2 Spike proteins, they allow for the assessment of the relative amplitude between pairs of translocated combinations. For a given target, the relative amplitude between each pair showed a repeatable trend over each pLMs and each predictor. In other word, the translocated combinations could reliably be classified from the highest to the lowest performing. This was especially visible when the relative amplitude was more pronounced, and a representative example can be seen with IGPS family and SsIGPS species used as the target (Figure 3). The best improvement over both ESM model and all 3 predictors was consistently observed with the translocation of TtIGPS alone, followed by TmIGPS + TtIGPS together, and then by TmIGPS alone (S3, S7). Even for the RF classifier, that performed significantly poorer than the SVR and Lasso classifier, no significant disruption of the overall trend could be noticed. In addition, it should be specified that when the relative amplitude between pairs of combinations is not substantial (e.g S3.3, S7.3), the overall trend can probably be assessed as a whole as any impact on the trend, over different configurations, may be considered as statistical variations. Besides, and more interestingly, the relative amplitudes tended to be conserved over different targets. For instance, in IGPS family the translocation of TmIGPS consistently performed the worst over all targets that used it, and over all ESM and predictor configurations (S3, S7). Finally, despite the reduced improvement that fitness translocation provided on GFP family, it still exhibited similar trends in relative amplitude between the translocated combinations (S4, S8).

4.2 Fitness Translocation Dataset Selection

The assessment of the dataset selection methods revealed that the Statistical-Greedy approach allows the fitness translocation method to reliably select among the best performing translocated combinations, over the 60 configurations ((4 SARS-CoV-2 + 3 IGPS + 3 GFP) targets $\times$ 2 pLMs $\times$ 3 predictors) considered, and the 26 target training sizes. As a representative example, Figures 5, 6 7, 8 respectively show the results for SARS-CoV-2 Spike, IGPS, and GFP families using ESM2 pLM and SVR predictor. The results from each target training size are reported as the average performance over the N=50 outer loops of the nested CV and the corresponding train/test splits. This allowed to repeatedly samples Spearman's correlation coefficients of a given dataset, and a one-sided paired t-test was then used to evaluate whether the mean difference between two sample of Spearman's coefficients indicated an increase in performances. It estimated the distribution of paired differences and compared the resulting p-value to a significance threshold of $\alpha = 0.05$. Then, the corresponding minimal value that a paired t-test can reliably detect is given by the following formula [20]:

$$\Delta_{\text{min}} = t_{\alpha,N-1} \cdot \frac{s_d}{\sqrt{N}} \tag{1}$$

Where α is the significance threshold, N is the number of outer loops in the nested CV, $t_{\alpha,N-1}$ is the critical value of the Student-distribution for $N-1$ degrees of freedom at the significance level α, and s_d is the standard deviation of the paired differences. Consequently, the minimum detectable improvement Δ_{min} decreases as N increases and depends on the ability of a predictor to provide stable performances over multiple train/test splits.

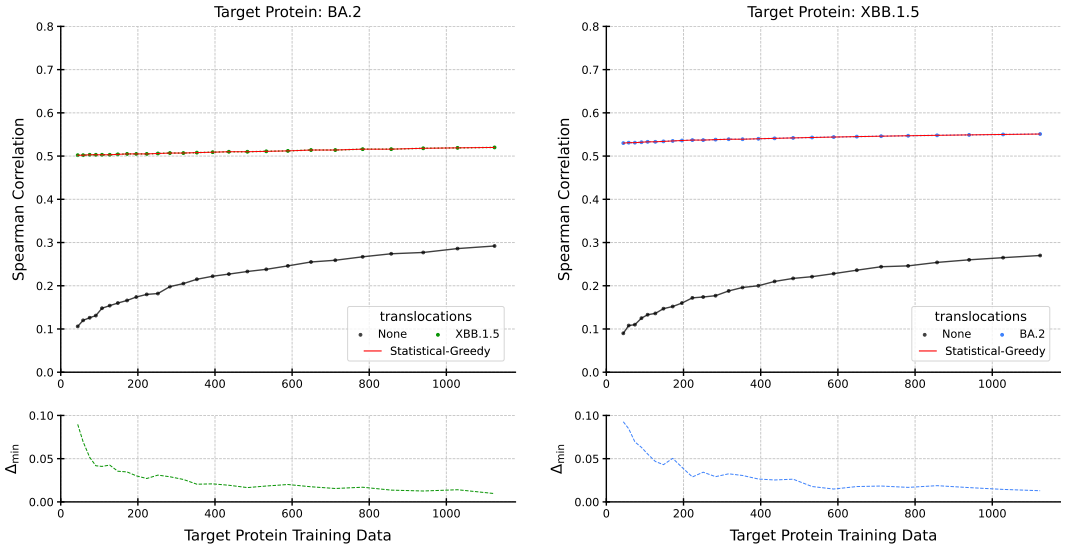

Figure 5: Results for SARS-CoV-2 spike protein Cell Entry, ESM2 pLM, and SVR predictor.

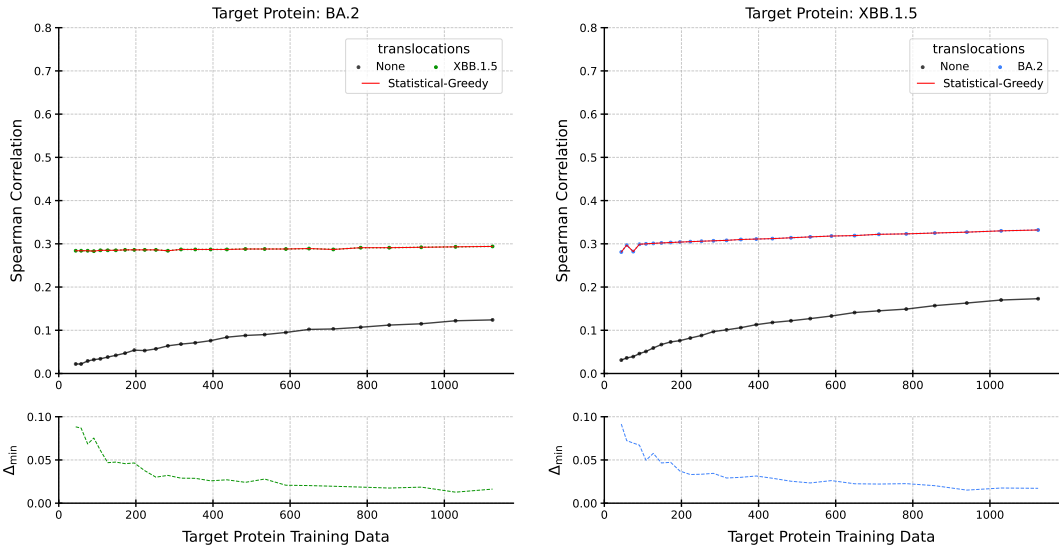

Figure 6: Results for SARS-CoV-2 spike protein ACE2 Binding, ESM2 pLM, and SVR predictor.

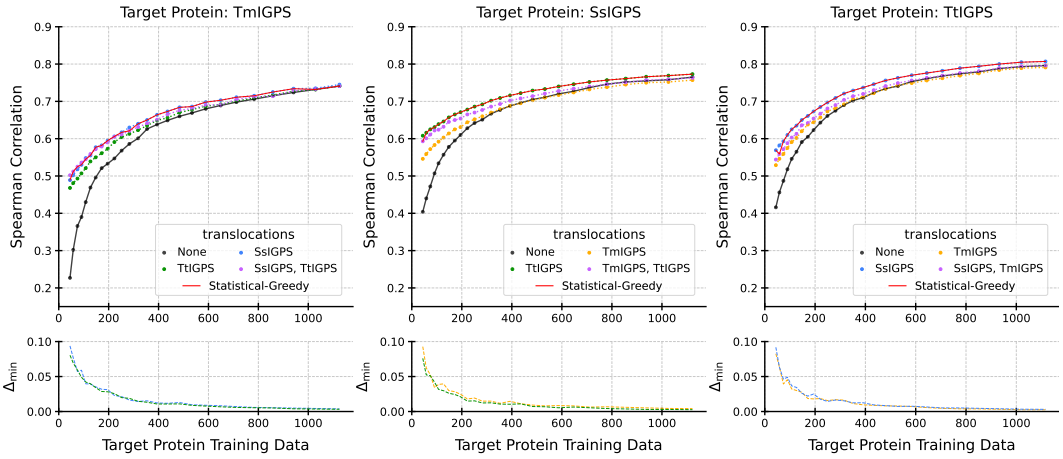

Figure 7: Results for IGPS, ESM2 pLM, and SVR predictor.

Statistical-Greedy Performances. For all configuration that we evaluated (S9, S10, S11, S12, S13, S14, S15, S16), the minimum detectable improvement Δ_{min} followed the expected trend expressed by Equation 1. For the smallest target training size, the $N = 50$ outer loops resulted in Δ_{min} between 0.1 and 0.07, which then rapidly decreased as the target training size increased, to reach between 0.017 and 0.005 for the biggest target training sizes. In other words, the fitness translocation method naturally allows to detect smaller improvements as the target training size increases. Besides, it should be reminded that the selection method was applied within conditions that reproduced a realistic scenario, by using limited data. The final selected combinations are then reported as a red line on the results from the full dataset, for all target training sizes (e.g Figure 8). It shows that the statistical-Greedy selection method was able to generalize to the full dataset results by reliably selecting among the best performing translocated combinations. Especially, when the relative improvement between candidate combinations were small, the lowest scoring combinations were systematically excluded (e.g Figure 7, S11.1). In addition, in the case where no improvement could be achieved with fitness translocation, the original target dataset was consistently selected as the best one (e.g Figure 8, S12.3). This result stem from the use to the one-sided version of the paired t-test. It naturally establishes an upper threshold and minimizes the influences of the statistical fluctuations from non-effective combinations, which could otherwise be selected accidentally. Therefore, the initial selection before the greedy algorithm is improved. Besides, in cases where the improvement was too small to reach significance, increasing N may benefit the selection process, to decrease the minimum detectable improvement Δ_{min}. In particular, the variations in selection observed for smaller target training sizes in GFP family results (S12, S16), arised from the limited improvements it provided, which could not consistently reach statistical significance, while still showcasing reduced but noticeable improvements.

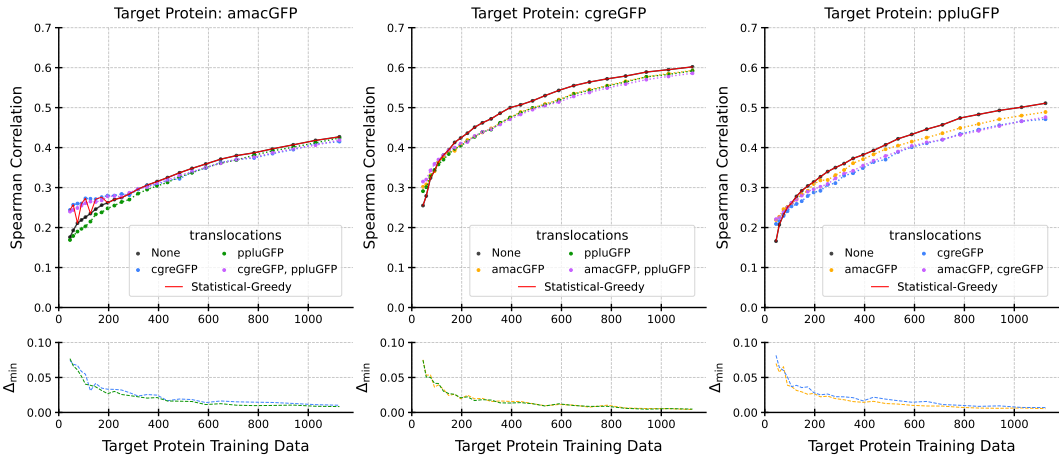

Figure 8: Results for GFP, ESM2 pLM, and SVR predictor.

Other Selection Method Performances. The effectiveness of the remaining selection method was also evaluated. First off, the Individual-Greedy method is the closest from the Statistical-Greedy. It removes the statistical testing and uses the performances of the original target dataset as an initial selection threshold, while still sorting the resulting candidate combinations in a greedy manner. Therefore, in cases where all translocated combinations resulted in a substantial improvement, both approaches were providing the greedy algorithm with the same set of combinations, resulting in similar final selection (e.g S23, S35). However, when a combination achieved poor performances, the lack of paired t-test did not allow to effectively manage the statistical fluctuations (e.g S27.2, S38.2, S25.2). The Individual-Greedy method was therefore selecting inappropriate combinations, making it unable to generalize properly. Second, the Individual-Select approach removes both the statistical testing and the greedy algorithm. So, similar consideration arises, with the additional downside that the lack of greedy algorithm could not allow for the initial selection of candidate combinations to be sorted out (e.g S24.3, S35.3, S38.3), resulting in the worst results from the three approaches.

4.3 Fitness Translocation Method

The fitness translocation method using the Statistical-Greedy selection approach is reported in algorithm 1, and its implementation is available at `https://github.com/adrienmialland/ProtFitTrans`. It takes as input the target data, the additional protein data, the number N of nested CV outer loops, and the significance threshold α, and outputs the best performing combination. It is composed of three main steps. The first one computes the paired differences in performance, as measured by the Spearman's correlation coefficient, between each individual additional dataset and the target dataset. The second one applies the initial selection using the paired t-test, which only keeps the individual dataset that resulted in a significant improvement. The third step applies the greedy algorithm based on the selected datasets and iteratively build the best performing combination.

5 Discussion

This study demonstrates the potential of fitness translocation to address challenges in protein variant effect prediction tasks when available data is limited. By leveraging fitness landscapes from related proteins, this method enables the construction of augmented datasets that improve the prediction performance of a target protein. These findings highlight the benefit of fitness translocation as an effective approach to expand sequence-function annotations based on prior engineering efforts, and to improve the predictive performances for proteins with limited experimental data.

Comparison with Traditional Data Augmentation. Traditional data augmentation methods have been commonly used in various modern machine learning approaches, demonstrating their effectiveness in fields such as computer vision [25, 13] or natural language processing (NLP) [26, 27]. These methods range from basic manipulations, such as rotation or cropping, to the generation of synthetic data using advanced methods. However, these approaches are difficult to apply on protein data due to their unique characteristics. Proteins are subject to complex sequence-function relationships, where even a single residue mutation can significantly alter the properties of a given protein and must be carefully considered [28]. Moreover, recent investigations explored the advantages of multiple image and text related data augmentation methods for protein sequence prediction [29]. The authors suggest that while improvement can be achieved, the performance gains may be

Algorithm 1 Protein Fitness translocation

```
Require: target, homologs, N, α
  differences ← []
  significant ← []

  for n ← 1 to N do                                   ▷ Compute paired differences
      train, test ← split_dataset(target)
      t_result ← evaluate_model(train, test)
      for all hom ∈ homologs do
          trans ← translocate(target, hom)
          result ← evaluate_model(train||trans, test)
          append (hom, result − t_results) to differences
      end for
  end for

  for all hom ∈ homologs do                           ▷ Compute differences significance
      σ_e ← standard_error(differences, hom)
      Δμ ← mean_difference(differences, hom)
      if p-value(Δμ, σ_e, N) < α then
          append (hom, Δμ) to significant
      end if
  end for

  best_result ← extract_best(significant)
  candidates ← sort_all_but_best(significant)

  for all hom ∈ candidates do                         ▷ Sort out candidate combinations
      new_results ← []
      for n ← 1 to N do
          trans ← translocate(target, hom, best_trans)
          result ← evaluate_model(train||trans, test)
          append result to new_results
      end for

      if mean(new_results) > best_trans then
          add hom to best_result homologs
          set best_result result to new_results
      end if
  end for

  return best_result
```

suboptimal and that innovative methods tailored to protein data are required. In that regard, our proposed fitness translocation method provides a simple framework for protein variant effect prediction that enables the use of protein data and fitness landscapes from past experiments. It allows to augment a given dataset without relying on synthetic data and with no alteration of the sequence-function pairs.

Fitness Translocation Evaluation. The generalizability of the fitness translocation method across multiple configurations, as shown by our results, makes it a reliable and easy to apply data augmentation approach for enhancing target protein datasets. Even for the GFP protein family, where improvements were not consistent across all configurations, significant gains were observed under specific conditions (S4.2, S8.2). This later point suggests that certain protein species may only benefit from subsets of additional datasets, highlighting the potential advantage of increasing the number of datasets considered for fitness translocation. However, this may result in a significant increase in processing time if all possible combinations are evaluated. Typically, with a number of m candidate datasets, each one of them being either included or not in the final augmented dataset, this would require a number of 2^m combination to be evaluated. The introduction of the greedy algorithm allows to effectively address this issue, by prioritizing the most promising combinations and reducing the number of evaluations to a maximum of $2m$. This upper limit is only maximized if all datasets are selected during the initial selection process. The use of a paired t-test can further improve both the processing times

and the final performances, by reducing the impact of statistical variations and only considering the meaningful improvements. In that regard, Equation 1 provides a simple solution to set a relevant minimum detectable improvement Δ_{min}, with the downside that increasing N does not scale linearly with Δ_{min}, as Δ_{min} is proportional to $1/\sqrt{N}$. However, the use of $N = 50$ nested CV outer loops in this study was still sufficient to achieve a Δ_{min} below 0.1 for the smallest target training size, with Δ_{min} rapidly decreasing as the training size increased (Figures 5, 6, 7, 8). In addition, Δ_{min} also depends on s_d, which is fixed by a given configuration. So, for a specific configuration, using the paired t-test coupled with the greedy algorithm and a reasonably high number N of outer loops can both provide reliable and meaningful results in a reduced processing time.

Biological Validity of Fitness Translocation. Recent investigations into phylogenetically divergent IGPS proteins with the TIM barrel fold, showing only 30-40% sequence identity, have explored the evolutionary constraints imposed by sequence and structural requirements [15]. The study revealed unexpected long-range allosteric pathways linking distal residues to the active site, which likely play a critical role in maintaining function. Also, a significant correlation was observed between the fitness landscapes of the distant orthologs. The authors suggested that amino-acid preferences at a given position in the structure are mostly conserved during evolution. These findings propose a general mechanism by which evolutionary pressures contribute to the conservation of fitness landscapes, where such pressures maintain critical biophysical and functional features of the protein, despite sequence divergence. Additionally, previous studies on closely related homologs (influenza virus nucleoproteins), with 94% sequence identity, reported similar results on amino-acid preferences at specific positions [30], further supporting the notion that fitness landscapes are shaped by conserved structural and functional properties, even for low sequence identity. So, the results from the study on IGPS led the authors to interpret landscapes conservation as the translocation of fitness landscapes in the sequence space.

Building on these findings, our study employed pre-trained pLMs to generate latent representations of protein sequences, leveraging their ability to capture intricate structural and functional features encoded in amino acids [14]. These models are pre-trained on large-scale protein sequence datasets containing billions of amino acids and learn internal representations by predicting masked amino acids. This allows the model to infer context-dependent features in a self-supervised manner, producing embeddings that represent the biological properties of protein sequences, while preserving the sequence-function relationship [31]. Our findings show that the translocation of the fitness landscape effectively take advantage of the conserved properties between protein species, with the possibility to generalize beyond the particular TIM barrel fold.

Implications for Protein Engineering. By introducing sequence diversity in datasets for machine learning models, fitness translocation has the potential to improve model generalization capabilities across diverse contexts, facilitating applications such as enzyme engineering or therapeutic development. In the case of directed evolution, the properties of a protein sequence are iteratively improved by mimicking the process of natural selection. It involves several rounds of mutagenesis, expression, and selection, where each round builds upon the previous one to construct a library of sequences that contains high quality variants. The effectiveness of directed evolution is therefore dependent on the selection process, where a machine learning model should predict the best set of sequences to include in subsequent round [8]. The cost and time required for directed evolution make the fitness translocation method particularly suited to increase the model capabilities, select high quality variants, and limit the number of rounds. More generally, many applications that rely on sequence-function mappings of protein variants would likely benefit from fitness translocation. As an example, modern methods for designing artificial proteins with specific functions, or *de novo* protein design, often involves generative models trained on embeddings from a pLM [32]. These methods may therefore benefit from an augmented dataset with fitness translocation.

6 Conclusion

In this study, we introduced *fitness translocation*, a data augmentation technique for protein variant effect prediction which leverage evolutionary conservation of fitness landscapes. This method enables the translocation of sequence space embedding features from additional protein datasets to a target protein, enriching the training data with biologically relevant variations from previous protein engineering efforts. We also presented a systematic selection framework that identifies the most beneficial datasets for data augmentation. This approach significantly improved the predictive performances and demonstrates the utility of fitness translocation, particularly in scenarios with limited data. The implementation of the method is available at `https://github.com/adrienmialland/ProtFitTrans`.

Acknowledgement This paper is based on results obtained from a project, JPNP14004, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Notin Pascal, Kollasch Aaron, Ritter Daniel et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, 2024.

[2] Romero Philip A et Arnold Frances H. Exploring protein fitness landscapes by directed evolution. In *Nature reviews Molecular cell biology*, 2022. 10.1038/nrm2805.

[3] Gantz Maximilian, Neun Stefanie, Medcalf Elliot J et al. Ultrahigh-throughput enzyme engineering and discovery in in vitro compartments. In *Chemical Reviews*, 2023. 10.1021/acs.chemrev.2c0091.

[4] Vanella Rosario, Kovacevic Gordana, Doffini Vanni et al. High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering. In *Chemical Communications*, 2022. 10.1039/D1CC04635G.

[5] Markel Ulrich, Essani Khalil D, Besirlioglu Volkan et al Advances in ultrahigh-throughput screening for directed enzyme evolution. In *Chemical Society Reviews*, 2020. 10.1039/C8CS00981C.

[6] Freschlin Chase R, Fahlberg Sarah A et ROMERO Philip A. Machine learning to navigate fitness landscapes for protein engineering. In *Current opinion in biotechnology*, 2022. 10.1016/j.copbio.2022.102713.

[7] Horne Jesse et Shuhla Diwakar. Recent advances in machine learning variant effect prediction tools for protein engineering. In *Industrial & engineering chemistry research*, 2022. 10.1021/acs.iecr.1c04943.

[8] Wittmann Bruce J, Johnston Kadina E, Wu Zachary et al. Advances in machine learning for directed evolution. In *Current opinion in structural biology*, 2021. 10.1016/j.sbi.2021.01.008.

[9] Yang Kevin K, Wu Zachary et Arnold Frances H. Machine-learning-guided directed evolution for protein engineering. In *Nature methods*, 2019. 10.1038/s41592-019-0496-6

[10] Freschlin Chase R, Fahlberg Sarah A, heinzelman Pete et al. Neural network extrapolation to distant regions of the protein fitness landscape. In *Nature Communications*, 2024. 10.1038/s41467-024-50712-3.

[11] Saito Yutaka, Oikawa Misaki, Sato Takumi et al. Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration. In *Acs Catalysis*, 2021. 10.1021/acscatal.1c03753.

[12] Iglesias Guillermo, Talavera Edgar, Gonzalez-Prieto Ángel et al. Data augmentation techniques in time series domain: a survey and taxonomy. In *Neural Computing and Applications*, 2023. 10.1007/s00521-023-08459-3.

[13] Shorten Connor et Khoshgoftaar Taghi M A survey on image data augmentation for deep learning. In *Journal of big data*, 2019. 10.1186/s40537-019-0197-0.

[14] Rives Alexander, Meier Joshua, Sercu Tom et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. In *Proceedings of the National Academy of Sciences*, 2021. 10.1073/pnas.2016239118.

[15] Chan Yvonne H, Venev Sergey V, Zeldovich Konstantin B et al. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. In *Nature communications*, 2017. 10.1038/ncomms14614.

[16] Somermeyer Louisa Gonzalez, Fleiss Aubin, Mishin Alexander et al. Heterogeneity of the GFP fitness landscape and data-driven protein design. In *Elife*, 2022. 10.7554/eLife.75842.

[17] Dadonaite Bernadeta, Brown Jack, Mcmahon Teagan E et al. Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. In *Nature*, 2024. 10.1038/s41586-024-07636-1.

[18] Vabalas Andrius, Gowen Emma, Poliakoff Ellen et al. Machine learning algorithm validation with a limited sample size. In *Plos one*, 2019. 10.1371/journal.pone.0224365.

[19] Jungnickel Dieter. The greedy algorithm. In *Graphs, Networks and Algorithms*, 2013. 10.1007/978-3-642-32278-5_5.

[20] Hsu Henry et Lachenbruch Peter A. Paired t test. In *Wiley StatsRef: statistics reference*, 2014. 10.1002/9781118445112.stat05929.

[21] Hietpas Ryan, Roscoe Benjamin, Jiang Li et al. Fitness analyses of all possible point mutations for regions of genes in yeast. In *Nature protocols*, 2012. 10.1038/nprot.2012.069.

[22] Dadonaite Bernadeta, Crawford Katharine HD, Radford Caelan E. et al. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. In *Cell*, 2023. 10.1016/j.cell.2023.02.001.

[23] Meier Joshua, Rao Roshan, Verkuil Robert et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in neural information processing systems*, 2022. 10.1101/2021.07.09.450648

[24] Lin Zeming, Akin Halil, Rao Roshan et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. In *bioRxiv*, 2022. 10.1101/2022.07.20.500902

[25] Takahashi, Ryo, Matsubara Takashi et Uehara Kuniaki. Data augmentation using random image cropping and patching for deep CNNs. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 10.1109/TCSVT.2019.2935128.

[26] Li Bohan, Hou Yutai et Che Wanxiang. Data augmentation approaches in natural language processing: A survey. In *Ai Open*, 2022. 10.1016/j.aiopen.2022.03.001.

[27] Feng Steven Y, Gangal Varun, Wei Jason et al. A survey of data augmentation approaches for NLP. In *arxiv preprint*, 2021. 10.48550/arXiv.2105.03075

[28] Le Hyunjung, Ozbulak Utku, Park Homin et al. Assessing the reliability of point mutation as data augmentation for deep learning with genomic data. In *BMC bioinformatics*, 2024. 10.1186/s12859-024-05787-6.

[29] Sun Rui, Wu Lirong, Lin Haitao et al. Enhancing Protein Predictive Models via Proteins Data Augmentation: A Benchmark and New Directions. In *arXiv preprint*, 2024. 10.48550/arXiv.2403.00875.

[30] Doud Michael B, Ashenberg Orr, et Bloom Jesse D. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. Molecular biology and evolution. In *Molecular biology and evolution*, 2021. 10.1093/molbev/msv167

[31] Alley Ethan C, Khimulya Grigory, Biswas Surojit et al. Unified rational protein engineering with sequence-based deep representation learning. In *Nature methods*, 2019. 10.1038/s41592-019-0598-1

[32] Tang Xiangru, Dai Howard, Knight Elizabeth et al. A survey of generative ai for de novo drug design: new frontiers in molecule and protein generation. In *Briefings in Bioinformatics*, 2024. 10.1093/bib/bbae338