Information Fusion 102 (2024) 102035

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Addressing data scarcity in protein fitness landscape analysis: A study on semi-supervised and deep transfer learning techniques

José A. Barbero-Aparicio *, Alicia Olivares-Gil, Juan J. Rodríguez, César García-Osorio, José F. Díez-Pastor

Departamento de Ingeniería Informática, Universidad de Burgos, Burgos, 09001, Spain

ARTICLE INFO

Keywords:
Bioinformatics
Machine learning
Transfer learning
Semi-supervised learning
Protein fitness prediction
Small datasets

ABSTRACT

This paper presents a comprehensive analysis of deep transfer learning methods, supervised methods, and semi-supervised methods in the context of protein fitness prediction, with a focus on small datasets. The analysis includes the exploration of the combination of different data sources to enhance the performance of the models. While deep learning and deep transfer learning methods have shown remarkable performance in situations with abundant data, this study aims to address the more realistic scenario faced by wet lab researchers, where labeled data is often limited.

The novelty of this work lies in its examination of deep transfer learning in the context of small datasets and its consideration of semi-supervised methods and multi-view strategies. While previous research has extensively explored deep transfer learning in large dataset scenarios, little attention has been given to its efficacy in small dataset settings or its comparison with semi-supervised approaches.

Our findings suggest that deep transfer learning, exemplified by ProteinBERT, shows promising performance in this context compared to the rest of the methods across various evaluation metrics, not only in small dataset contexts but also in large dataset scenarios. This highlights the robustness and versatility of deep transfer learning in protein fitness prediction tasks, even with limited labeled data.

The results of this study shed light on the potential of deep transfer learning as a state-of-the-art approach in the field of protein fitness prediction. By leveraging pre-trained models and fine-tuning them on small datasets, researchers can achieve competitive performance surpassing traditional supervised and semi-supervised methods. These findings provide valuable insights for wet lab researchers who face the challenge of limited labeled data, enabling them to make informed decisions when selecting the most effective methodology for their specific protein fitness prediction tasks.

Additionally, the study investigated the combination of two different sources of information (encodings) through our enhanced semi-supervised methods, yielding noteworthy results improving their base model and providing valuable insights for further research.

The presented analysis contributes to a better understanding of the capabilities and limitations of different learning approaches in small dataset scenarios, ultimately aiding in the development of improved protein fitness prediction methods.

0. Introduction

Proteins, as the fundamental building blocks of life, hold immense significance in various scientific disciplines, and protein engineering has emerged as a powerful field aimed at harnessing the potential of these versatile molecules for a wide range of applications [1]. A protein fitness landscape represents the connection between protein sequences and their functional characteristics, associating a specific fitness value to its corresponding sequence. It illustrates how changes in protein sequences impact their behavior and performance. Moreover, it can be a very useful tool in protein engineering, helping us understand the intricate relationships between protein sequence, structure and function, enabling us to decipher the underlying principles behind protein behavior and facilitating the design of novel proteins with tailored properties [2]. The utilization of protein fitness landscapes has demonstrated its effectiveness in unraveling the mechanisms by which directed evolution exerts strong selection pressures, driving proteins to evolve rapidly towards desired characteristics [3].

* Corresponding author.
E-mail address: jabarbero@ubu.es (J.A. Barbero-Aparicio).

https://doi.org/10.1016/j.inffus.2023.102035
Received 21 July 2023; Received in revised form 7 September 2023; Accepted 19 September 2023
Available online 22 September 2023

The efficiency of protein engineering is hampered by the vastness of the protein sequence space [4,5], as well as the inherent limitations in screening capabilities and analysis capacity. To navigate this huge space of possibilities, the integration of machine learning and directed evolution has emerged as a promising approach [6]. By using these techniques, it becomes possible to predict the fitness associated with specific protein sequences, thereby drastically reducing the need of time-consuming manual screening required for this task.

Studying protein fitness landscapes poses significant challenges [7], particularly in situations where data is scarce due to limitations in experimental resources or the availability of annotated data. The scarcity of data restricts the comprehensive exploration of the vast protein sequence space, making it difficult to gain a complete understanding of the relationships between sequence variations and protein phenotypes. Additionally, limited availability of annotated data hampers the development and validation of accurate predictive models. These challenges need the exploration of innovative approaches that can effectively extract the maximum amount of information from the available data and address the scarcity issue, enabling more robust analysis and modeling of protein fitness landscapes.

While several studies on supervised machine learning (where prediction models are obtained from fully labeled datasets) applied to protein fitness prediction have emerged [6,8–13], there remains a gap in the literature regarding the application of these techniques to small dataset situations and the exploration of specific methods, such as semi-supervised learning (where the models are obtained using the aggregation of information from both labeled and unlabeled data), that can address the challenges associated with limited data availability. Small datasets pose unique difficulties due to the limited number of labeled samples, making it challenging to train accurate predictive models. Moreover, traditional supervised approaches may struggle to generalize well in these scenarios, which has led to the rise of alternative methodologies like semi-supervised learning. By harnessing both labeled and unlabeled data, and with the capacity of taking advantage of learning from multiple views or encodings [14] (an encoding is a numerical representation of a protein sequence that allows machine learning models to process and analyze the sequence information effectively), semi-supervised learning has the potential to improve model performance, overcome data scarcity, and enhance the accuracy of protein fitness predictions in small dataset situations, as has been demonstrated in various domains [15]. Addressing these gaps in the literature is crucial for advancing the field and developing robust methods that can effectively handle the challenges posed by scarce data in protein fitness prediction.

One of the most promising areas of semi-supervised learning is deep transfer learning, which harnesses the power of pre-trained models to leverage knowledge from large datasets in related domains [16]. By transferring learned representations and weights, deep transfer learning enables the effective utilization of unlabeled labeled data in the target domain, thereby improving the performance of predictive models. This approach has shown great potential in various domains associated with small datasets like emotion [17] and micro-expression [18] recognition, detection of diabetic retinopathy [19] or machinery fault diagnosis [20] and has the capacity to also address the challenges in protein landscape analysis.

Specifically, in the field of protein engineering, which involves the design and modification of proteins for specific applications or functions, several methods have emerged that apply deep transfer learning techniques to a wide range of protein-related tasks [12,21–25]. This technique consists of applying deep learning models that, although trained to solve a task, are used to solve a different task. The idea is to be able to reuse the knowledge learned from feature extraction in the first layers, which is assumed to be common to both tasks, so it is only necessary to slightly adjust the weights of the layers to solve the new task. These approaches make use of pre-trained models, often based on transformer architectures or BERT-like models, which have been trained on large-scale protein sequence or structure datasets. By fine-tuning these pre-trained models on task-specific data, such as protein fitness prediction, deep transfer learning enables the efficient transfer of learned representations and knowledge to the target protein engineering task. These methods have demonstrated improved performance, enhanced generalization, and reduced reliance on large labeled datasets, making them promising tools for protein engineering applications in small data contexts.

In scenarios with limited data, it has been observed that simple supervised methods often exhibit superior performance compared to deep learning approaches [26]. This can be attributed to the inherent complexity and high-dimensional nature of deep learning models, which require large amounts of labeled data to generalize effectively. However, the utilization of semi-supervised learning techniques offers a promising alternative, surpassing the performance of traditional supervised learning methods in situations with sparse data.

This paper aims to explore the performance of various learning methodologies in protein fitness prediction, including supervised, semi-supervised and deep transfer learning methods in the context of limited data availability. Through a comparative analysis, we seek to identify the optimal approach for researchers seeking to apply these techniques in protein engineering. By examining their strengths, limitations, and effectiveness in dealing with scarce data, we found that deep transfer learning can outperform classical supervised models and semi-supervised methods, even in situations where labeled data is very limited. Despite the potential of multi-view learning to exploit the mixture of multiple sources of information and unlabeled data, our findings highlight the significant impact of transfer learning in enhancing the performance of deep learning models, enabling them to surpass the capabilities of traditional supervised and semi-supervised approaches in small dataset situations. The insights gained from this study provide valuable guidance to researchers in selecting the most effective methodology for protein engineering tasks with limited data, furthering our understanding and advancement in the field.

1. Material and methods

1.1. Data

All the models used in this paper were trained and tested using the green fluorescent protein (GFP) dataset from [27]. This dataset comprises 32 610 protein sequences, each associated with a corresponding fitness value. In the context of this dataset, the fitness value indicates the intensity of fluorescence exhibited by the protein. The fluorescence intensity serves as a quantitative measure of the protein's performance or activity in terms of its ability to emit green light when excited by an external light source. The protein sequences in this dataset exhibit varying numbers of variants in comparison to the wild type sequence. The distribution of sequences by number of variants can be found in Table 1.

1.2. Models

1.2.1. Encoding

Protein encodings play a fundamental role in bioinformatics, enabling researchers to represent and analyze the rich information embedded within protein sequences and structures. Additionally, in order to process protein data using machine learning models, it is often necessary to convert text sequences into numerical representations. One of the most widely used strategies for this purpose is one-hot encoding, which represents protein sequences using zeros and ones. In this encoding scheme, each amino acid in a sequence is assigned a binary vector where only one element is set to 1, indicating its presence at that position. Although one-hot encoding is a well-established approach, alternative methods are also employed in this study.

Table 1
Number of sequences by amount of amino acid substitutions in the GFP dataset from [27].

# variants	# sequences
1	970
2	9 686
3	8 206
4	5 653
5	3 814
6	2 164
7	1 172
8	568
9	237
10	95
11	28
12	9
13	5
14	3
Total	32 610

In our experiments, we have utilized one-hot encoding as a standard method due to its widespread usage and ease of implementation. However, it is worth noting that some of the methods employed in this paper are capable of employing different encoding strategies. For instance, deep transfer learning methods such as ProteinBERT, which we utilized in our experimentation, possess their own encoders that capture the underlying patterns and relationships within protein sequences, representing them in a latent space. This encoding strategy allows the model to make effective use of its unique encoding capabilities to extract meaningful features.

Furthermore, certain semi-supervised methods have the ability to train using diverse data representations. They typically employ various subsets of features, providing different perspectives and complementing each other in the training process. In our study, we have taken advantage of this approach to train a model using a combination of information from two different encodings, thereby exploring the optimal mixture of encodings that yield improved performance. The analysis of this optimal encoding combination can be found in Section 2.

1.2.2. Semi-supervised models

When working with small datasets, where labeled examples are limited, getting additional information from unlabeled instances becomes crucial for improving model performance. Some semi-supervised learning methods offer a valuable approach in such scenarios, for example, by effectively utilizing unlabeled data through pseudo-labeling techniques. Unlike traditional supervised learning that relies solely on labeled data, semi-supervised learning takes advantage of the vast amount of unlabeled data available, which is often more easily accessible. This allows the model to learn from a larger pool of information and potentially discover hidden patterns and structures in the unlabeled data. A general description of the pseudo-labeling process in semi-supervised learning models is represented in Fig. 1. By incorporating both labeled and unlabeled data, semi-supervised learning can enhance the generalization ability of the model and improve its performance on tasks with limited labeled examples. Given the nature of the problem we are trying to solve, where unlabeled instances are readily available while the process of labeling them is both expensive and time-consuming, we have found it highly promising to explore the application of semi-supervised methods in this particular context.

However, it is important to acknowledge the challenges and limitations associated with semi-supervised learning. One major challenge is the reliability of the pseudo-labels assigned to the unlabeled data. Since these labels are generated by the model itself, there is the potential for misclassification and noise, which can impact the quality of the training process. Another limitation is the assumption that the distribution of labeled and unlabeled data is similar, which may not always hold true in practice. If the distribution mismatch is significant, the model's performance may be compromised. In this context, our experimental design involved randomly selecting and excluding a certain percentage of instances from the training dataset, effectively creating a set of unlabeled data. Subsequently, we utilized the remaining labeled instances to generate pseudo-labels for the previously unlabeled data. This approach ensured a balanced distribution between labeled and unlabeled data, facilitating the exploration of semi-supervised learning methods.

Moreover, in the context of semi-supervised learning, it is crucial to perform thorough hyperparameter tuning and employ suitable regularization techniques to mitigate the risk of overfitting to the unlabeled data. To ensure optimal model performance, we conducted a grid search cross-validation process to select the most suitable hyperparameters for our experiments. This systematic approach allowed us to fine-tune the models and maximize their generalization capabilities.

Despite the challenges involved, semi-supervised learning presents valuable opportunities to get useful information from unlabeled data and combine that information with the labeled data to improve the model performance in scenarios with limited labeled examples. By effectively utilizing both labeled and unlabeled data and addressing the associated challenges, semi-supervised learning emerges as a powerful tool to deal with small dataset problems in diverse domains, including the field of protein engineering.

Although the number of articles that study classification in the field of semi-supervised learning is abundant in the literature [15], there are hardly any semi-supervised methods for regression [28]. Additionally, it is worth noting that, during the time of conducting this study, the availability of readily accessible implementations for semi-supervised regression methods was extremely limited. In fact, the only available method we found was the single-view version of the Co-Regression model, available in LAMDA-SSL [29], which served as a foundation for our improved multi-view version. Further details on our Co-Regression model will be provided in the subsequent paragraphs.

In this context, the initial approach could involve transforming the regression problem into a classification task. The primary goal is to establish a reliable ranking of protein fitness, and thus, the fitness values can be partitioned into categories above and below the median. Subsequently, a classifier can be trained to predict whether a given sequence belongs to the category above or below the median (or any other specified percentile split). Instead of utilizing the actual class label, the probability obtained from the classifier can serve as a surrogate measure for ordering the samples based on their fitness. This alternative has been discarded based on the results obtained in a preliminary experiment, which revealed that the performance of classification methods was significantly inferior to that of native regression methods.

Therefore, we have decided to focus exclusively on using the few existing semi-supervised regression methods, namely TriTraining Regressor and Multi-view Co-Regression. These methods have shown potential for yielding favorable outcomes in addressing the challenges posed by limited data availability.

Multi-view co-regression. Co-Training [30] is one of the most studied paradigms of semi-supervised classification.

The idea is to simultaneously train two base classifiers on different views of the data, views that are assumed to be conditionally independent of each other. The predictions for which each of the classifiers is most confident are used to extend the other classifier's dataset with new pseudo-labeled instances. With this new dataset, the classifiers are retrained, to obtain additional pseudo-labeled instances and so on, until a stopping criterion is reached.

CoReg [31] is an adaptation of the Co-Training paradigm for regression. This adaptation tries to reduce the re-training times of the base estimators by employing k-nearest neighbors regressors. While this also eliminates the need to use independent views of the data by using different distance metrics on the calculations made by each regressor, for our current problem using two different encodings of the same

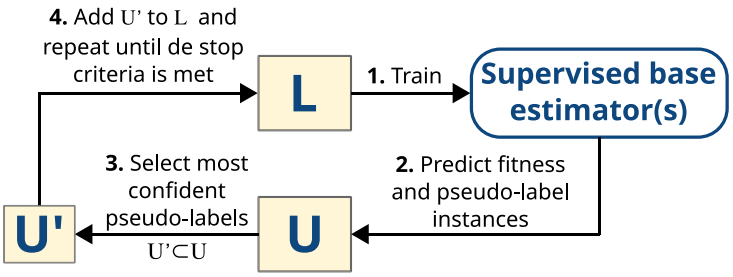

Fig. 1. Pseudo-labeling process in semi-supervised learning. The set of labeled instances in the dataset is represented by L, and the unlabeled instances are represented by U.

sequence as independent views could actually lead to an improvement in the performance of the model. For this reason, we use an adaptation of the implementation available in the LAMDA-SSL library [29] that allows the use of independent views.

For the pseudo-labeled instance selection criterion, heuristically, the error of the regressor on the labeled example set should decrease the most if the most confidently labeled example is utilized. CoReg employs a computationally less costly approximation of the mean squared error (MSE) calculated using only the k-nearest neighboring labeled examples of the pseudo-labeled instance. Let Δ_u denote the result of subtracting the latter MSE from the former MSE, only the pseudo-labeled instance with the higher Δ_u is added to the training set on each iteration.

In the case of Multi-view Co-Regression, the only parameters that required adjustment were the orders of the Minkowski distance, $D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$, utilized by each base k-nearest neighbors regressor internally. We tested the values $p = 2$, $p = 3$, $p = 4$, and $p = 5$. Ultimately, we achieved the best results by using an order of $p = 2$ for one model and an order of $p = 5$ for the other.

TriTraining regressor. The TriTraining algorithm [32] is an evolution of the Co-Training paradigm that uses three base classifiers instead of two, and also eliminates the need to introduce multiple views of the data. Each base classifier is first trained with a subset of labeled instances. Then, those classifiers are used to predict the label of some unlabeled instances. In TriTraining, the set of labeled instances of each classifier is extended with new pseudo-labeled instances when the label it assigns to an unlabeled instance differs from the label the other two classifiers agree on. In other words, if two classifiers agree on the label they assign to an unlabeled instance and this label differs from the one assigned by the third classifier, this instance is added to the dataset of the latter with the label assigned by the first two. As before, the process is repeated until some stop criteria is reached.

In this study, we evaluated various base models and ultimately selected a Support Vector Regressor (SVR) and Ridge Regressor due to their superior performance.

As we have not found any existing implementations of TriTraining for regression, we have developed our own. In order to adapt this algorithm to a regression task, a new threshold parameter needs to be established. If two predicted values are closer than this threshold, it is considered that they agree. We introduce the consequent changes taking as a basis the implementation available in the sslearn library [33].

The TriTraining method and its base models underwent a grid search to explore different parameter combinations. For the Ridge base model, the grid search considered a range of values for the alpha parameter, using the values 0, 0.05, and 0.85. The tolerance parameter was varied using the values 0.0001, 0.01, 0.1, 1, and 10. The selected values for Ridge were 0.85 for the alpha and 1 for the tolerance. Similarly, for the SVR model, the grid search examined two kernel options: radial basis function kernel and linear kernel. The tolerance parameter was again tested using the same set of values. The chosen kernel for SVR was the radial basis function and the tolerance was set to 1.

1.2.3. Supervised models

In light of the limited data availability, we sought to assess the performance of the base regressors used in the semi-supervised models in this study, considering their reduced number of parameters compared to deep learning models. This examination is particularly important in low-data contexts where overfitting can pose significant challenges. This analysis will provide a deeper understanding of the impact of the strategies employed by the semi-supervised methods and allow us to assess their ultimate effectiveness. Since the base models (SVR, Ridge, and k-nearest neighbors regressor) were already optimized through a grid search in the previous section, those optimized values are the ones used for these models. In our experiments, we employed these supervised models both independently and as base models for the semi-supervised learning approaches. By conducting this evaluation, we can gain insights into the efficacy of the semi-supervised approaches and verify their performance in our specific scenario, while being able to compare them to their supervised base models.

1.2.4. Deep transfer learning

Within the field of protein engineering, numerous models have been developed to address the protein fitness prediction challenge. Specifically, deep learning models based on latent space representations have been proven to accurately capture intrinsic properties of proteins and their relationships between sequences [34]. These models range from relatively small architectures based on long short-term memory (LSTM) neural networks [12,21,35], to large-scale architectures based on large language models (LLMs) originally developed in the natural language processing (NLP) field [22,24].

We did a preliminary evaluation of several deep learning methods for potential use in transfer learning, including options such as ProtT5-XL [22], ProtBert-BFD [22], TAPE Transformer [23], ESM-1b [35]. Despite the typically high hardware demands associated with deep learning methods, one of our goals was to use one that could be executed with moderate hardware requirements, which is why we ultimately discarded these methods. When we compare the size of ProteinBERT [25], our chosen model, with ≈ 16M parameters with the mentioned methods (TAPE Transformer: ≈ 38M parameters; ProtBert-BFD: ≈ 430M parameters; ESM-1b: ≈ 650M parameters; ProtT5-XL: ≈ 3B parameters). ProteinBERT is much lighter, thus reducing execution times and alleviating hardware requirements. Moreover, the TCR-BERT [36] and EpiBERTope [37] methods were discarded because their adaptation for this particular task seemed complex and would have been time-consuming.

These approaches offer diverse perspectives on protein engineering, but most of the methods require expensive graphics cards to train the models. Interestingly, we only found one deep transfer learning model [21] that takes into account small training set scenarios.

ProteinBERT [25], the deep learning method finally selected for our study, is a general model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture [38], which is one of the most commonly used architectures in natural language processing (NLP) tasks. BERT and ProteinBERT are based on the transformer architecture, which utilizes self-attention mechanisms to capture contextual dependencies in a sequence of words or tokens.

The key idea behind ProteinBERT is to pre-train the model on a large corpus of unlabeled protein sequences in a self-supervised manner. During pre-training, ProteinBERT learns to predict masked amino acids in a protein, where a certain percentage of amino acids are randomly masked out, and the model needs to predict the missing amino acids based on the context provided by the surrounding ones. This masked language modeling objective allows ProteinBERT to learn contextualized representations of proteins, capturing the relationships of amino acids inside them.

ProteinBERT also employs another pre-training task called Gene Ontology (GO) [39] annotation prediction. In this task, the model is trained to predict a label based on diverse protein functions. This objective helps ProteinBERT understand protein-level relationships and dependencies.

The key innovation of BERT lies in its bidirectional nature. Unlike previous models that process token sequences in a left-to-right or right-to-left manner, BERT uses a bidirectional transformer architecture, allowing ProteinBERT to consider the entire protein sequence during both pre-training and fine-tuning. This enables ProteinBERT to capture dependencies from both preceding and succeeding amino acids, leading to a deeper understanding of their context and relationships.

We selected ProteinBERT as our specific model of choice due to several compelling reasons. Firstly, ProteinBERT offers exceptional adaptability to a wide range of protein-related problems, making it a versatile tool for our research. Additionally, the fine-tuning process of ProteinBERT is straightforward, allowing us to easily adapt the model to our specific task. Moreover, ProteinBERT exhibits a relatively lightweight structure compared to other models while maintaining comparable performance, making it efficient in terms of computational resources. Lastly, ProteinBERT aligns with the current state-of-the-art in the field by applying the transformer-based architecture [40], which is widely recognized and utilized in the protein engineering and NLP domains.

After pre-training, ProteinBERT can be fine-tuned on specific downstream protein-related tasks such as secondary structure prediction, remote homology prediction or protein fitness prediction. During fine-tuning, ProteinBERT is trained on task-specific labeled data, where the model's parameters are adjusted to optimize performance on the target task. In this case, all the parameters have been selected after a preliminary cross-validated grid search.

ProteinBERT was trained using a grid search approach to optimize its hyperparameters. The following values were tested for each parameter: batch size values of 8, 32, and 64 (with 8 selected as the best value), maximum epochs per stage of 50, 200, and 1000 (with 1000 selected), learning rate of 1e−03, 1e−04, and 1e−05 (with 1e−04 selected), learning rate with frozen pre-trained layers of 1e−03, 1e−04, and 1e−05 (with 1e−05 selected), number of final epochs of 50, 200, and 500 (with 200 selected), dropout rate of 0.5 and 0.75 (with 0.5 selected), factor for the learning rate schedule of 0.5, 0.25, and 0.1 (with 0.1 selected), patience for learning rate reduction of 5, 10, and 50 (with 50 selected), minimum learning rate of 1e−08 and 1e−09 (with 1e−09 selected), patience for early stopping of 10, 20, and 60 (with 60 selected), and beginning with frozen or unfrozen pre-trained layers (with frozen selected). These hyperparameters were carefully selected to optimize the training of ProteinBERT and ensure its effective performance for protein-related tasks.

1.3. Model training methodology

The incorporation of additional information extracted from the unlabeled instances of the dataset is especially relevant in the training process of the models studied in this work. Fig. 2 shows the training processes of semi-supervised models and those obtained by transfer learning (in our study, from ProteinBERT). The upper part of the diagram shows the process implemented in this study, which involves progressively eliminating labeled instances. To simulate situations with fewer instances, we have randomly selected some of the available labeled instances, removed their label, and moved them to the unlabeled set. This approach allows us to simulate situations with far fewer labeled instances from our original dataset of approximately 30,000 instances.

Once the instances have been divided into labeled and unlabeled, the process utilized by the semi-supervised models is shown in the lower left and the process followed by ProteinBERT in the lower right. It should be noted that the training process of the semi-supervised models involves the pseudo-labeling process, which was already illustrated in Fig. 1. On the other hand, the unlabeled instances used by ProteinBERT are from the UniRef90 database, and these are used in the pre-training process, after which the labeled instances are used for the fine-tuning process.

1.4. Evaluation metrics

We have employed three metrics to evaluate the performance of the models analyzes in this paper. First, we utilize the mean squared error (MSE) as a baseline metric for assessing the performance of the models in a regression problem. The MSE provides a quantitative measure of the average squared difference between the predicted and actual fitness values.

However, our primary focus, and the basis for our conclusions, is the Spearman's rank correlation coefficient [41], denoted as Spearman's ρ. This metric is widely used in the literature on protein engineering and it has been used in a reference benchmark paper by Brandes et al. [23]. Spearman's ρ measures the relationship between the predicted and the actual rankings of the protein fitness values. It accounts for the relative order of the values rather than their specific numerical differences. By using Spearman's ρ, we can effectively assess the models' performance in capturing the relative ordering of the proteins according to their fitness values, which is crucial in many bioinformatics applications.

Spearman's rank correlation coefficient is defined by the following equation:

$$\rho = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)} \tag{1}$$

In this context, d_i represents the difference in rank for each instance i between the two rankings and n represents the number of ranked instances.

The values for Spearman's ρ can range from -1 to $+1$. A value of $+1$ indicates a perfect positive monotonic relationship, where higher ranks in one variable correspond to higher ranks in the other variable. A value of -1 indicates a perfect negative monotonic relationship, where higher ranks in one variable correspond to lower ranks in the other variable. A value of 0 indicates no monotonic relationship between the variables. Intermediate values between -1 and $+1$ represent varying degrees of monotonicity, with higher absolute values indicating stronger correlations.

In the context of selecting mutations for comprehensive analysis in the laboratory, it is of greater interest to ensure the accuracy of the top positions in the obtained ranking, even if the lower positions may not be as reliable. Hence, it is pertinent to consider a novel metric that emphasizes the prioritization of the top-ranking positions. A weighted version [42] of the Kendall's rank correlation coefficient [43] has the potential to achieve such results by adding a higher weight to the most important positions in the ranking that gradually decreases for the less relevant ranks. In this regard, we have employed this metric, referred to as weighted τ hereafter. In its non-weighted version Kendall's rank correlation coefficient is defined by Eq. (2), where C is the number of concordant pairs and D is the number of discordant pairs.

$$\tau = \frac{C - D}{C + D} \tag{2}$$

In the weighted version, that we used in this paper, the assignment of weights relies on a rank array, which allocates nonnegative ranks to individual elements. These ranks, inversely proportional to the values

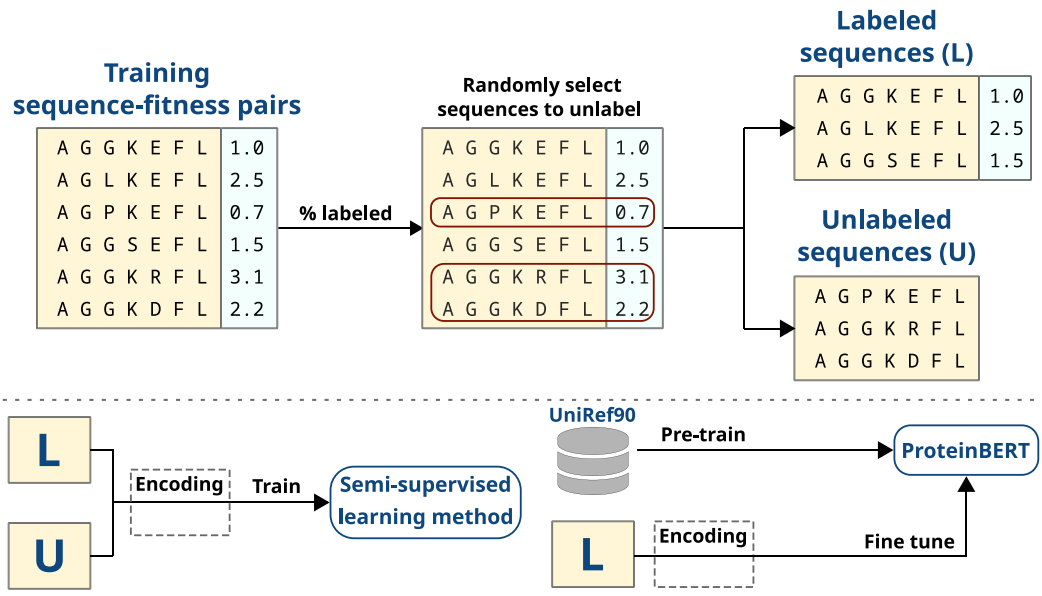

Fig. 2. Training process for semi-supervised and transfer learning methods.

they represent, prioritize elements with lower values, with the highest rank assigned to the value of 0. Additionally, a weight function is employed to derive weights from the assigned ranks. The weight assigned to an exchange is then determined by aggregating the weights associated with the ranks of the exchanged elements, either through summation or multiplication. The resulting values of weighted τ can vary from -1 to $+1$, as previously explained for Spearman's ρ.

2. Results

2.1. Encodings multi-view analysis

One of the most relevant aspects to consider in the field of semi-supervised learning is its ability to leverage information from different data representations or encodings. In this case, given the numerous options available for encoding protein sequences, we can provide valuable information to the models by utilizing two different encodings. To study the effect of combining different encodings, a preliminary experiment was conducted using a reduced dataset, from which the best combination of encodings for subsequent experiments could be determined. This dataset consists of a 10% subset of the previously explained general dataset, which is formed by 2600 train instances (of which 500 are labeled) and 660 test instances. The objective of this dataset size is to simulate a situation with a realistically small amount of labeled data. The Multi-view Co-Regression model was used in this experiment since it is the only one, among the models studied in this paper, that supports multiple views or encodings. To obtain a more stable estimation of the model performance, the training process was repeated five times due to the randomness of the method for each pair of encodings using different random seeds. This approach allows for a more reliable and consistent measure of the average model performance across multiple iterations.

Fig. 3 displays the results obtained by Multi-view Co-Regression when combining one encoding with another. The values on the diagonal represent the results obtained using the same encoding twice (which strictly speaking does not correspond to a correct application of the multi-view approximation, but the values obtained have been included as a reference).

In this figure we can observe that binary-based encodings do not give the best performance. Although one-hot encoding is the most commonly used representation in the field, it is not the most suitable encoding for KNN. This is due to the high dimensionality and sparsity of the one-hot encoded features, which can lead to challenges in accurately measuring similarity and consequently have a negative impact on KNN performance.

The utilization of other encodings based on continuous numerical terms is enough to improve the results. Particularly, representations based on physicochemical, evolutionary, or structural features of amino acids provide a substantial enhancement. However, these values are still far from those obtained by combining different encodings. The figure clearly demonstrates that for each diagonal value, there is always another value in the corresponding row or column that outperforms it. This observation strongly indicates that the multi-view model can extract valuable information from the utilization of two complementary encodings, and is by itself already an interesting result of the experiments presented in this article. From this figure, we must also highlight that the best performance was achieved by combining a physicochemical property (Acthely factors) and an evolutionary property (BLOSUM62), which is logical as the model can benefit from information of different aspects of the amino acids.

2.2. Protein fitness landscape results

In the protein fitness prediction experiments, we analyzed the performance of supervised, semi-supervised, and transfer learning methods using the complete dataset defined in Section 1.1. Since the main focus of this study is to evaluate performance in low-data contexts, we conducted a series of experiments with different percentages of labeled data which were run 5 times each to obtain more statistically meaningful results. Firstly, we partitioned the dataset into training, validation, and test sets with percentages of 65%, 15%, and 20% respectively. For ProteinBERT, the validation set was used for model validation, following the common practice in deep learning. For the other methods, the validation set was combined with the training set to ensure equal data availability during the training process. Secondly, we performed experiments for each method where the number of labeled instances was limited to specific percentages. These percentages were 100%, 75%, 50%, 25%, 10%, 5%, 1%, 0.5%, 0.25%, and 0.01%. Although some of these labeling percentages may be too small to obtain reliable models, we included a progression to sufficiently low percentages to visualize the progress of each model with respect to dataset size. The case of semi-supervised models is special as they were trained on both the labeled dataset and the corresponding percentage of remaining unlabeled instances.

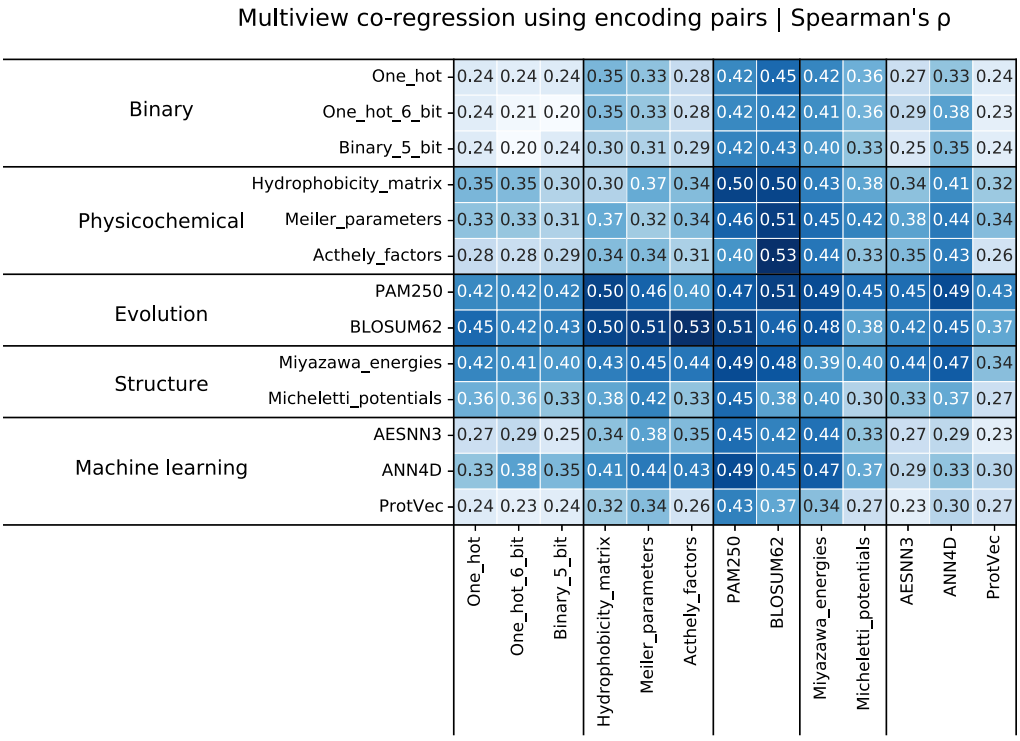

Fig. 3. Multi-view Co-regression results using a combination of two different views (encodings). In the labels we can see the different encoding names grouped by the kind of property group they are part of according to the classification in [44].

In this dataset, all sequences, regardless of the number of substitutions they contain, are part of the same dataset. This approach allows us to study the models' ability to predict the protein fitness landscape from a varied yet reduced dataset. In Fig. 4, we observe that, for any given labeling percentage, ProteinBERT consistently outperforms the other models. It is remarkable that the performance of ProteinBERT remains consistently superior to the other models, despite the reduction in the training dataset. One might expect that simpler models, compared to the complex architectures of deep learning, would give better results, since having fewer parameters would make their generalization capabilities less affected by the limited availability of labeled training data. However, the ability to transfer knowledge from the pre-training phase enables ProteinBERT to achieve significantly better results than the other models, even with limited data.

Fig. 4 provides an interesting overview of the evolution of the performances of the different methods as the number of labeled instances is reduced. But it is also especially interesting to focus on what happens at two of the extremes, in which the number of labeled instances are 5% and 75%.

Firstly, it is worth examining the case where there is 5% of labeled instances, which corresponds to a scenario with approximately 1300 instances, a common situation in many real-world problems. In Table 2, we observe that ProteinBERT achieves the highest performance, significantly outperforming other methods. This is evident across multiple evaluation metrics, including the classic regression metric MSE and the two ranking metrics. The supervised methods, SVR and Ridge, obtain results that come closest to ProteinBERT's performance. Additionally, the semi-supervised method TriTrainingRegressor exhibits a slight improvement over Ridge, suggesting that semi-supervised methods can get some useful information from unlabeled instances. These findings highlight the effectiveness of ProteinBERT in scenarios with limited labeled data, showcasing its superior performance compared to alternative approaches.

Secondly, we have chosen to examine the scenario with 75% labeled instances because it represents a sufficiently large dataset (approximately 20 000 instances) and allows for the inclusion of an additional 25% of unlabeled instances, enabling us to analyze the performance of semi-supervised methods. In Table 3, we observe that once again ProteinBERT achieves the best results across all metrics, followed by Ridge and SVR, as well as their semi-supervised counterparts utilizing TriTrainingRegressor. The consistent superior performance of ProteinBERT underscores its effectiveness across a wide range of dataset sizes, including both small and big scenarios.

A notable conclusion that can be drawn from this experimentation is that semi-supervised models do not achieve the expected performance. In fact, the results obtained are in some cases slightly worse than the supervised models they are wrapping. From this experimentation, we can infer that the insights that semi-supervised methods can extract from unlabeled data not only fail to provide useful information to the models but also introduces noise that worsens the final results. There is one notable exception, which is the case of Multi-view Co-regression. In situations with limited data but sufficient for maintaining acceptable performance (between 5% and 0.5%), Co-regression significantly improves the results of its base classifier, specially when it is able of getting advantage of two different encodings applying the multi-view approach. However, the issue lies with its base classifier, KNeighborsRegressor, which performs significantly worse than the others even after the parameter tuning process, as explained in Section 2.1.

3. Discussion

Based on the results presented in the previous section, it can be concluded that deep transfer learning methods have established a new state-of-the-art in protein fitness prediction. The inherent language understanding capabilities of these methods enable them to adapt effectively to both high-data and low-data scenarios. Remarkably, deep transfer learning surpasses the performance of methods specifically designed for low-data situations. Neither the simplicity of classical regression methods nor the ability of semi-supervised methods to extract information from unlabeled instances come close to matching the performance of deep transfer learning methods.

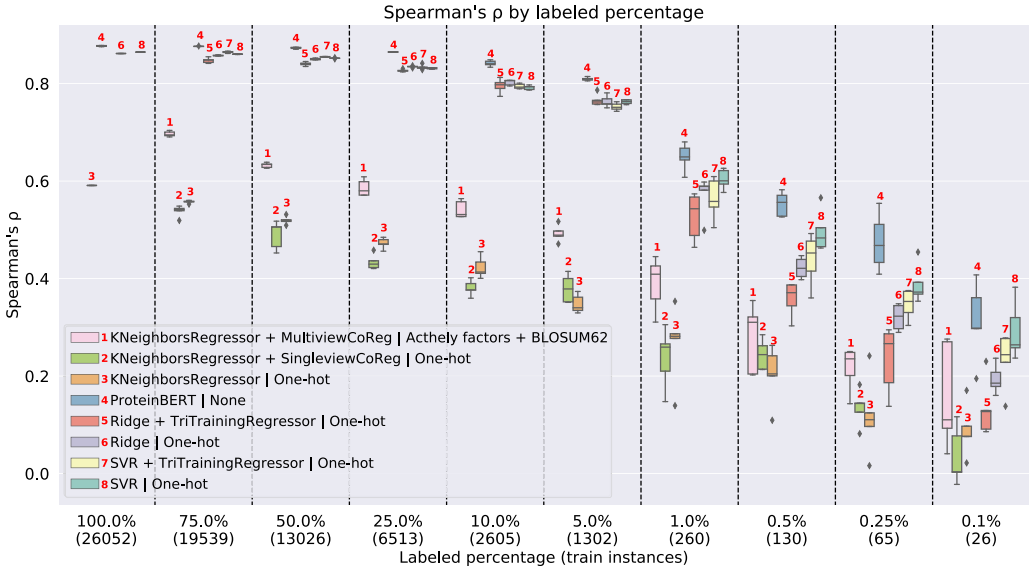

Fig. 4. Spearman's ρ results across a selection of labeled training sizes. Please note that semi-supervised methods are not included in the 100% step as they require a certain quantity of unlabeled instances to operate effectively.

Table 2
Protein fitness prediction results training with ≈ 1300 labeled instances from the training set. (MSE: Mean squared error; ρ: Spearman's correlation coefficient; τ: weighted version of Kendall's τ).

Model	Semi-supervised	MSE	ρ	τ
Ridge	–	0.5146	0.7631	0.5935
SVR	–	0.9233	0.7625	0.6399
KNeighborsRegressor	–	1.4481	0.3478	0.3397
Ridge	TriTrainingRegressor	0.5228	0.7640	0.6096
SVR	TriTrainingRegressor	0.9202	0.7575	0.6224
KNeighborsRegressor	SingleviewCoReg	1.4639	0.3793	0.3629
KNeighborsRegressor	MultiviewCoReg	1.0512	0.4923	0.4115
ProteinBERT	Transfer learning	**0.2663**	**0.8094**	**0.7034**

Note: The best values are highlighted in bold.

Table 3
Protein fitness prediction results training with 75% of the dataset ≈ 20 000 labeled instances from the training set. (MSE: Mean squared error; ρ: Spearman's correlation coefficient; τ: weighted version of Kendall's τ).

Model	Semi-supervised	MSE	ρ	τ
Ridge	–	0.3603	0.8575	0.6873
SVR	–	0.3890	0.8601	0.7742
KNeighborsRegressor	–	1.2902	0.5573	0.4929
Ridge	TriTrainingRegressor	0.3665	0.8524	0.7027
SVR	TriTrainingRegressor	0.3846	0.8622	0.7570
KNeighborsRegressor	SingleviewCoReg	1.2517	0.5384	0.4570
KNeighborsRegressor	MultiviewCoReg	0.8848	0.6965	0.5503
ProteinBERT	Transfer learning	**0.0701**	**0.8763**	**0.8022**

Note: The best values are highlighted in bold.

Previously, implementing and training deep learning models could pose challenges in terms of difficulty and computational requirements for wet lab researchers. However, the simplicity of fine-tuning pre-trained models like ProteinBERT means that even non-specialized researchers in machine learning will be able to train a powerful model without issue. Additionally, as demonstrated in this study, these models do not require a large amount of data to outperform other methods in common scenarios encountered by laboratory researchers. This accessibility and efficiency make deep transfer learning models, such as ProteinBERT, a highly attractive option for researchers seeking accurate predictions with limited data and minimal expertise in machine learning.

The advantage of deep transfer learning methods over other options studied in this work can be attributed to their ability to understand the language of proteins through their self-training phase. This self-training step is related to the techniques used by the semi-supervised models, where the model is able to extract information from unlabeled data. Even some authors would classify this deep transfer learning models into the semi-supervised category [45]. Nevertheless, ProteinBERT has a much more efficient capacity of extracting that information after seeing the results of the semi-supervised methods. The incorporation of the encoding phase within the model architecture also ensures that less information is lost in the sequences and enables the discovery of an optimal encoding for the problem at hand. This combination of language comprehension and effective encoding could explain the superior performance of deep transfer learning models in protein fitness prediction.

Based on the findings, we can conclude that further research is needed on semi-supervised methods, especially in the context of regression. While it appears that these methods have the potential to improve upon supervised approaches, it is challenging for them to surpass the performance of transfer learning, even in scenarios with limited labeled data available. Nevertheless, the fusion of information from different encodings in multi-view semi-supervised methods is worth investigating further, as they have shown to be highly effective in enhancing the performance of the base model. Therefore, more exploration and experimentation are required to fully understand the capabilities and limitations of semi-supervised methods in protein-related tasks.

In conclusion, this study highlights the emergence of deep transfer learning methods as a new state-of-the-art in protein fitness prediction. Through their ability to comprehend the language of amino acids and their efficient fine-tuning process, these models demonstrate superior performance across a wide range of dataset sizes, surpassing both traditional regression approaches and semi-supervised methods. The integration of self-training and encoding within the model architecture ensures optimal information retention and sequence representation for the problem at hand. These findings showcase the potential of deep transfer learning methods to revolutionize protein-related research, providing accessible and effective tools for researchers, including those with limited machine learning expertise. With further advancements and research in this field, deep transfer learning holds promise for accelerating discoveries and breakthroughs in protein analysis and bioinformatics.

In addition to the findings presented in this study, there are several promising directions for future research. One area of investigation is exploring alternative deep learning models that incorporate inputs combining multiple encodings. This is motivated by the observed improvements achieved by multi-view methods that make use of diverse encodings. Additionally, investigating the performance of shallow networks using concatenated encodings as input is another interesting research question. Examining whether such networks outperform those working solely on individual encodings would provide valuable insights. These future research endeavors have the potential to advance our understanding of protein fitness prediction and contribute to the development of more effective modeling approaches.

CRediT authorship contribution statement

José A. Barbero-Aparicio: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Alicia Olivares-Gil:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Juan J. Rodríguez:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **César García-Osorio:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **José F. Díez-Pastor:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this paper is already publicly available.

Acknowledgments

This work is supported by the Junta de Castilla Leon, Spain under project BU055P20 (JCyL/FEDER, UE), and the Ministry of Science and Innovation, Spain under project PID2020- 119894 GB-I00 co-financed through European Union FEDER funds. José A. Barbero-Aparicio is funded through a pre-doctoral grant by the University of Burgos and Alicia Olivares-Gil is funded by the predoctoral grant from the Department of Education of Junta de Castilla y León (VA) (ORDEN EDU/875/2021) (Spain).

References

[1] J.A. Brannigan, A.J. Wilkinson, Protein engineering 20 years on, Nat. Rev. Mol. Cell Biol. 3 (12) (2002) 964–970, http://dx.doi.org/10.1038/nrm975.

[2] P.A. Romero, A. Krause, F.H. Arnold, Navigating the protein fitness landscape with Gaussian processes, Proc. Natl. Acad. Sci. 110 (3) (2013) E193–E201, http://dx.doi.org/10.1073/pnas.1215251110, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1215251110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1215251110.

[3] P.A. Romero, F.H. Arnold, Exploring protein fitness landscapes by directed evolution, Nat. Rev. Mol. Cell Biol. 10 (12) (2009) 866–876, http://dx.doi.org/10.1038/nrm2805.

[4] J. Maynard Smith, Natural selection and the concept of a protein space, Nature 225 (5232) (1970) 563–564, http://dx.doi.org/10.1038/225563a0.

[5] D.T. Dryden, A.R. Thomson, J.H. White, How much of protein sequence space has been explored by life on Earth? J. R. Soc. Interface 5 (25) (2008) 953–956, http://dx.doi.org/10.1098/rsif.2008.0085, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2008.0085. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2008.0085.

[6] C.R. Freschlin, S.A. Fahlberg, P.A. Romero, Machine learning to navigate fitness landscapes for protein engineering, Curr. Opin. Biotechnol. 75 (2022) 102713, http://dx.doi.org/10.1016/j.copbio.2022.102713, URL https://www.sciencedirect.com/science/article/pii/S0958166922000465.

[7] D.L. Hartl, What can we learn from fitness landscapes? Curr. Opin. Microbiol. 21 (2014) 51–57, http://dx.doi.org/10.1016/j.mib.2014.08.001, Antimicrobials. URL https://www.sciencedirect.com/science/article/pii/S1369527414001209.

[8] C. Hsu, H. Nisonoff, C. Fannjiang, J. Listgarten, Learning protein fitness models from evolutionary and assay-labeled data, Nature Biotechnol. 40 (7) (2022) 1114–1122, http://dx.doi.org/10.1038/s41587-021-01146-5.

[9] T.A. Hopf, J.B. Ingraham, F.J. Poelwijk, C.P.I. Schärfe, M. Springer, C. Sander, D.S. Marks, Mutation effects predicted from sequence co-variation, Nature Biotechnol. 35 (2) (2017) 128–135, http://dx.doi.org/10.1038/nbt.3769.

[10] A.-M. Illig, N.E. Siedhoff, U. Schwaneberg, M.D. Davari, A hybrid model combining evolutionary probability and machine learning leverages data-driven protein engineering, bioRxiv (2022) http://dx.doi.org/10.1101/2022.06.07.495081, arXiv:https://www.biorxiv.org/content/early/2022/06/07/2022.06.07.495081.full.pdf. URL https://www.biorxiv.org/content/early/2022/06/07/2022.06.07.495081.

[11] Y. Luo, G. Jiang, T. Yu, Y. Liu, L. Vo, H. Ding, Y. Su, W.W. Qian, H. Zhao, J. Peng, ECNet is an evolutionary context-integrated deep learning framework for protein engineering, Nature Commun. 12 (1) (2021) 5743, http://dx.doi.org/10.1038/s41467-021-25976-8.

[12] E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G.M. Church, Unified rational protein engineering with sequence-based deep representation learning, Nature Methods 16 (12) (2019) 1315–1322, http://dx.doi.org/10.1038/s41592-019-0598-1.

[13] S. Biswas, G. Khimulya, E.C. Alley, K.M. Esvelt, G.M. Church, Low-N protein engineering with data-efficient deep learning, Nature Methods 18 (4) (2021) 389–396, http://dx.doi.org/10.1038/s41592-021-01100-y.

[14] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, 2013, http://dx.doi.org/10.48550/arXiv.1304.5634, arXiv:1304.5634. arXiv:1304.5634.

[15] J.E. van Engelen, H.H. Hoos, A survey on semi-supervised learning, Mach. Learn. 109 (2) (2020) 373–440, http://dx.doi.org/10.1007/s10994-019-05855-6.

[16] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2018, Springer International Publishing, Cham, 2018, pp. 270–279.

[17] H.-W. Ng, V.D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 443–449, http://dx.doi.org/10.1145/2818346.2830593.

[18] M. Peng, Z. Wu, Z. Zhang, T. Chen, From macro to micro expression recognition: Deep learning on small datasets using transfer learning, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 657–661, http://dx.doi.org/10.1109/FG.2018.00103.

[19] M.T. Hagos, S. Kant, Transfer learning based detection of diabetic retinopathy from small dataset, 2019, http://dx.doi.org/10.48550/arXiv.1905.07203, arXiv:1905.07203. arXiv:1905.07203.

[20] T. Han, C. Liu, R. Wu, D. Jiang, Deep transfer learning with limited data for machinery fault diagnosis, Appl. Soft Comput. 103 (2021) 107150, http://dx.doi.org/10.1016/j.asoc.2021.107150, URL https://www.sciencedirect.com/science/article/pii/S1568494621000739.

[21] N. Strodthoff, P. Wagner, M. Wenzel, W. Samek, UDSMProt: universal deep sequence models for protein classification, Bioinformatics 36 (8) (2020) 2401–2409, http://dx.doi.org/10.1093/bioinformatics/btaa003, arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/8/2401/48984849/bioinformatics_36_8_2401.pdf.

[22] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2022) 7112–7127, http://dx.doi.org/10.1109/TPAMI.2021.3095381.

[23] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, Y.S. Song, Evaluating protein transfer learning with TAPE, Adv. Neural Inf. Process. Syst. 32 (2019) 9689–9701.

[24] N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design, Nature Commun. 13 (1) (2022) 4348, http://dx.doi.org/10.1038/s41467-022-32007-7.

[25] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (2022) 2102–2110, http://dx.doi.org/10.1093/bioinformatics/btac020, arXiv:https://academic.oup.com/bioinformatics/article-pdf/38/8/2102/49009610/btac020.pdf.

[26] L. Brigato, L. Iocchi, A close look at deep learning with small data, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 2490–2497, http://dx.doi.org/10.1109/ICPR48806.2021.9412492.

[27] K.S. Sarkisyan, D.A. Bolotin, M.V. Meer, D.R. Usmanova, A.S. Mishin, G.V. Sharonov, D.N. Ivankov, N.G. Bozhanova, M.S. Baranov, O. Soylemez, N.S. Bogatyreva, P.K. Vlasov, E.S. Egorov, M.D. Logacheva, A.S. Kondrashov, D.M. Chudakov, E.V. Putintseva, I.Z. Mamedov, D.S. Tawfik, K.A. Lukyanov, F.A. Kondrashov, Local fitness landscape of the green fluorescent protein, Nature 533 (7603) (2016) 397–401, http://dx.doi.org/10.1038/nature17995.

[28] G. Kostopoulos, S. Karlos, S. Kotsiantis, O. Ragos, Semi-supervised regression: A recent review, J. Intell. Fuzzy Syst. 35 (2018) 1483–1500, http://dx.doi.org/10.3233/JIFS-169689, 2.

[29] L.-H. Jia, L.-Z. Guo, Z. Zhou, Y.-F. Li, LAMDA-SSL: Semi-supervised learning in python, 2022, http://dx.doi.org/10.48550/arXiv.2208.04610, arXiv preprint arXiv:2208.04610.

[30] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, in: COLT' 98, Association for Computing Machinery, New York, NY, USA, 1998, pp. 92–100, http://dx.doi.org/10.1145/279943.279962.

[31] Z.-H. Zhou, M. Li, et al., Semi-supervised regression with co-training, in: IJCAI, Vol. 5, 2005, pp. 908–913.

[32] Z.-H. Zhou, M. Li, Tri-training: exploiting unlabeled data using three classifiers, IEEE Trans. Knowl. Data Eng. 17 (11) (2005) 1529–1541, http://dx.doi.org/10.1109/TKDE.2005.186.

[33] J.L. Garrido-Labrador, jlgarridol/sslearn: v1.0.3.1, 2023, http://dx.doi.org/10.5281/zenodo.7781117.

[34] X. Ding, Z. Zou, C.L. Brooks III, Deciphering protein evolution and fitness landscapes with latent space models, Nature Commun. 10 (1) (2019) 5644, http://dx.doi.org/10.1038/s41467-019-13633-0.

[35] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118, http://dx.doi.org/10.1073/pnas.2016239118, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

[36] K. Wu, K.E. Yost, B. Daniel, J.A. Belk, Y. Xia, T. Egawa, A. Satpathy, H.Y. Chang, J. Zou, TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses, bioRxiv (2021) http://dx.doi.org/10.1101/2021.11.18.469186, arXiv:https://www.biorxiv.org/content/early/2021/11/20/2021.11.18.469186.full.pdf. URL https://www.biorxiv.org/content/early/2021/11/20/2021.11.18.469186.

[37] M. Park, S. woo Seo, E. Park, J. Kim, EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively, bioRxiv (2022) http://dx.doi.org/10.1101/2022.02.27.481241, arXiv:https://www.biorxiv.org/content/early/2022/03/02/2022.02.27.481241.full.pdf. URL https://www.biorxiv.org/content/early/2022/03/02/2022.02.27.481241.

[38] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, http://dx.doi.org/10.48550/arXiv.1810.04805, arXiv:1810.04805. arXiv:1810.04805.

[39] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology, Nature Genet. 25 (1) (2000) 25–29, http://dx.doi.org/10.1038/75556.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, http://dx.doi.org/10.48550/arXiv.1706.03762, arXiv:1706.03762. arXiv:1706.03762.

[41] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1) (1904) 72–101, URL http://www.jstor.org/stable/1412159.

[42] G.S. Shieh, A weighted Kendall's tau statistic, Statist. Probab. Lett. 39 (1) (1998) 17–24, http://dx.doi.org/10.1016/S0167-7152(98)00006-6, URL https://www.sciencedirect.com/science/article/pii/S0167715298000066.

[43] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.

[44] X. Jing, Q. Dong, D. Hong, R. Lu, Amino acid encoding methods for protein sequences: A comprehensive review and assessment, IEEE/ACM Trans. Comput. Biol. Bioinform. 17 (6) (2020) 1918–1931, http://dx.doi.org/10.1109/TCBB.2019.2911677.

[45] A.M. Dai, Q.V. Le, Semi-supervised sequence learning, Adv. Neural Inf. Process. Syst. 28 (2015).