Protein Fitness Prediction is Impacted by the Interplay of Language Models, Ensemble Learning, and Sampling Methods

Mehrsa Mardikoraem[1,2], Daniel Woldring[1,2*]

Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI, USA
Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI, USA
*Correspondence: woldring@msu.edu

Abstract: Advances in machine learning (ML) and the availability of protein sequences via high-throughput sequencing techniques have transformed our ability to design novel diagnostic and therapeutic proteins. ML allows protein engineers to capture complex trends hidden within protein sequences that would otherwise be difficult to identify in the context of the immense and rugged protein fitness landscape. Despite this potential, there persists a need for guidance during the training and evaluation of ML methods over sequencing data. Two key challenges for training discriminative models and evaluating their performance include handling severely imbalanced datasets (e.g., few high-fitness proteins among an abundance of non-functional proteins) and selecting appropriate protein sequence representations. Here, we present a framework for applying ML over assay-labeled datasets to elucidate the capacity of sampling methods and protein representations to improve model performance in two different datasets with binding affinity and thermal stability prediction tasks. For protein sequence representations, we incorporate two widely used methods (One-Hot encoding, physiochemical encoding) and two language-based methods (next-token prediction, UniRep; masked-token prediction, ESM). Elaboration on performance is provided over protein fitness, length, data size, and sampling methods. In addition, an ensemble of representation methods is generated to discover the contribution of distinct representations to the final prediction score. Within the context of these datasets, the synthetic minority oversampling technique (SMOTE) outperformed undersampling while encoding sequences with One-Hot, UniRep, and ESM representations. In addition, ensemble learning increased the predictive performance of the affinity-based dataset by 4% compared to the best single encoding candidate (F1-score = 97%), while ESM alone was rigorous enough in stability prediction (F1-score = 92%).

Keywords: Machine Learning, Protein Fitness Prediction, Embeddings, Sequence Representation, Imbalanced Assay-Labeled Datasets, Sampling Methods, Ensemble Learning

1. Introduction

Proteins are biological machines involved in almost all biological processes [1–4]. These molecules are made of amino acids that fold into 3-dimensional structures and perform life-sustaining biological functions [5]. Protein engineering practices aim to modify proteins to redirect what has already evolved in nature and address the industrial and medical needs of modern society [6,7]. This has been a challenging task due to the astronomical number of possible mutations and the complex sequence-function relationship of the proteins (i.e., fitness landscape) [8]. Therefore, various experimental and computational techniques have been developed to overcome the challenge of finding high-fitness proteins among mostly non-functional mutants. Recently, machine learning (ML) has shown promise as a tool to supplement already established techniques, such as rational design and directed evolution [9–12]. Unlike directed evolution, ML models can learn from non-functional mutants instead of simply discarding them during enrichment for functional clones. ML-assisted protein engineering, therefore, has potential as a time-efficient and cost-effective approach to searching for desired protein functionality. This provides a unique opportunity to create smart protein libraries, elevate and accelerate directed evolution strategies, and enhance the probability of finding unexplored high-fitness variants in the protein fitness landscape [13–15]. Machine learning methods have attained a high success rate in predicting essential protein properties including secondary structure, solubility, binding affinity, flexibility, and specificity [16–20]. Despite these recent milestones, generalizability and robustness of models will require further explorations in different protein fitness prediction tasks and training details.

Solving these challenges will require us to view proteins from a new perspective that supplements our biochemical knowledge with lessons from written languages. Recent advances in ML and artificial intelligence have applied natural language processing (NLP) methods to identify context-specific patterns from written or spoken text. NLP tasks learn how words function grammatically (syntax) and how they deliver meaning within themselves and in surrounding words (semantics) [21,22]. This has given rise to virtual assistants with voice recognition and sentiment analysis of text from diverse

languages [23,24]. Similarly, protein engineering can leverage these NLP tools – treating a string of amino acids as if they were letters on a page – to understand the language of proteins, providing a promising route to capture nuances (e.g., epistatic relationships, functional motifs) in complex sequence-function mappings [25,26]. The rapid expansion of publicly available protein sequence data (e.g., Uniprot [27], SRA [28]) further supports the use of big data and language models in the domain of protein engineering[29]. Self-supervised language models learn the context of the provided text by reconstructing the masked tokens/linguistic units of the text string using the unmasked parts. For the context of protein engineering, pre-trained protein language models – carrying valuable information about the epistasis/interaction of amino acids – can be applied to downstream tasks by extracting the optimized weight functions as a fixed-size vector (embedding) [25,30,31]. Among early embedding developments, Alley et al. introduced UniRep, a deep learning model that was trained on 24 million unique protein sequences to perform the next amino acid prediction tasks for extracting information about the global fitness landscape of proteins. Rives et al. trained ESM, a language model for masked amino acid prediction tasks, on over 250 million protein sequences [32]. The learned representations – including UniRep [33], ESM [32], TAPE [34], and ProteinBERT [35] - have generated promising results in diverse areas such as predicting protein fitness, protein localization, protein-protein interaction, and disease risk of mutations in terms of improved prediction scores, increased generalizability, and mediated data requirements [36–40]. Using embeddings for sequence representations (transfer learning) enables knowledge transfer between protein domains and future prediction tasks by further optimizing the already-learned weights. For example, Min et al. obtained a 20% increase in the F1-score (the harmonic mean of precision and recall) for a heat shock protein identification task when training their NLP-based model, DeepHSP [41], on top of pre-trained representations.

In this study, we perform protein sequence fitness prediction with ML techniques to demonstrate how model performance varies given the choice of protein representation, protein size, and the biological attribute (e.g., binding affinity and thermal stability) to be predicted. This work provides actionable insights for effectively building discriminative models and improving their prediction scores via sampling techniques and ensemble

learning. As efficient use of embedding methods on experimental datasets is in its infancy, rigorous studies are needed to gain new insights into the performance of the pre-trained models given various training conditions and distinct biological function predictions. To this end, we two large datasets that were representative of common protein engineering tasks. First, we leverage a highly imbalanced dataset (93% non-functional; Table 1), consisting of our previously described affinity-evolved affibody sequences [42] to explore NLP-driven practices. We then expand our analysis to include thousands of protein sequences labeled with their experimentally measured stabilities (melting temperatures, Tm) obtained from the Novozymes Enzyme Stability Prediction (NESP) dataset [43]. With the two datasets having unique attributes, we were well positioned to address multiple questions: i) How do different representation methods perform in predicting distinct fitness attributes such as stability or affinity? ii) How do sampling methods perform in the imbalanced protein dataset? iii) Is ensemble learning over different protein representations helpful in boosting the performance of discriminative models? By addressing these challenges, we also gain direct insights for model interpretation and reveal the features that are most important for discriminating between fit and non-fit sequences (Figure S1, Figure S3-4). We discovered that oversampling generally outperformed the undersampling techniques (Figure 4). In addition, ensemble over representations greatly improved the predictive performance in the affibody data (Figure 5). For protein representations, UniRep and One-Hot outperformed other methods in the affibody (affinity) dataset while ESM achieved the best score in stability prediction in NESP (Figure S6). Finally, it was observed that the performance of various protein representation methods is strongly impacted by protein sequence length (Figure 7).

2. Materials and Methods

2-1 Obtaining Experimentally Labeled Sequence Data

Two different datasets with varying data characteristics were explored. The first is our experimental data of affibody sequences that previously were iteratively evolved for binding affinity and specificity against a panel of diverse targets [44]. The second collection of labeled protein sequences was obtained from the recently released Kaggle

dataset wherein numerous proteins (n= 18,190) of various lengths are labeled according to their thermal stability (Tm). The NESP dataset was filtered to only include sequences characterized at pH=7. For the affibody dataset, raw sequence data were cleaned by removing any sequences that contained stop codons or invalid characters. Afterwards, the frequency of each unique sequence in the experimental steps was tabulated. Infrequent sequences appearing fewer than ten and four times (within magnetic activated cell sorting (MACS) and fluorescent activated cell sorting (FACS), respectively) were treated as background and removed from the analysis. Note that more stringent frequency removal for MACS was mainly due to the experiment type and higher probability to introduce noise in the dataset. After removing the background, sequences from MACS and FACS were combined to form the final high-fitness population of binders. The non-binding population included the initial affibody sequence pool which did not appear in the enriched population of binder sequences. The Initial affibody sequences that were within one hamming distance (i.e., a single amino acid mutation) of any enriched sequence were removed as well to account for potential errors encountered during deep sequencing. All affibody sequences were exactly 58 amino acids in length with mutations present at up to 17 of these positions.

2-2 Obtaining the Sequence Representations

We obtained four different numerical representations for our sequence data: One-Hot and physiochemical encoding, UniRep, and ESM embeddings. One-Hot encoding refers to building a matrix (amino acids × protein length) and filling it with one when there is a specific amino acid in the given position, filling the rest with zeros. For physiochemical encoding, we used the modlamp [45] package in python which is used for extracting the physical features from protein sequences. There were two types of features represented in modlamp package (global and peptide descriptors). All the global (e.g., sequence length, molecular weight, aliphatic index, etc.) and local physiochemical features based on Eisenberg scale were extracted for this analysis (twenty in total).

Embedding refers continuous representation of the protein sequence in a fixed-size vector, and it should contain meaningful information about proteins [46]. For example, in

the embedding visualization of amino acids in low dimensions for both UniRep and ESM, similar amino acids (in terms of size, charge, hydrophobicity, etc.) were close to each other. For UniRep representation, we used the 1900 dimension and mean representation over layers. We used Jax_UniRep since for obtaining the UniRep embeddings, https://github.com/ElArkk/jax-unirep. UniRep uses mLSTM structure for performing next-token prediction, and it was trained on 24 million sequences in the Uniref50 dataset with 18M parameters. For ESM, we chose ESM2 [47] with 1280 vector dimensions and 650M parameters and means over layer representations. GitHub for ESM is https://github.com/facebookresearch/esm.

2-3 Sampling & Splitting

Sampling refers to choosing a random subset of data to represent the underlying population. Three different sampling methods were tested for our severely imbalanced affibody dataset: undersampling, random oversampling, and synthetic minority oversampling technique (SMOTE) [48]. Due to the sparse and rugged nature of the protein fitness landscape, it is common for experimental data obtained in the protein domain to be highly imbalanced. One practical approach for resolving the imbalanced dataset issue is using sampling techniques when training the dataset. Oversampling is randomly repeating the minority class examples; thus, it could be prone to overfitting in comparison to undersampling. However, undersampling may discard the useful information especially in severely imbalanced datasets as it is removing many samples from the majority class. SMOTE is a more recent addition to sampling methods, and it is oversampling the minor population by synthetically generating more instances that are highly similar to the minority class. While SMOTE has shown promising results in increasing the prediction performance for various imbalanced datasets [49–51], there are also studies indicating undersampling superior performance compared to oversampling methods [52,53]. As a result, we examined the performance of all three sampling techniques to validate which sampling method performs well within our wet-lab protein dataset over different encoding methods.

For splitting the datapoints within the test set in an imbalanced dataset, sampling equally from each class may lead to an overestimation of the model performance[54]. As a result, we made sure that the test set distribution follows the initial data distribution (93% naïve vs. 7% Enriched). The classification performance was examined with F1-score. Note that hyperparameter optimization was implemented when necessary with OPTUNA [55] and the objective was set to maximize the F1-score in validation set.

2-4 Algorithm Selection & Training Details

For classification, logistic regression (LR) was chosen and L2 penalization (Ridge) was used to reduce the likelihood of overfitting. We reasoned that a simple logistic regression enables a fair comparison between cases. One regression task was also implemented over the NESP dataset with random forest regressor (RFR). We used regression to observe how models perform with increasing the prediction challenge, from binary prediction to actual label prediction. The rationale for using RFR was that linear regression model was not viable to meet the prediction task complexity. For a fair comparison between protein encoding performances in regression, the RFR hyperparameters, max number of estimators and max_depth, were optimized with OPTUNA.

2-5 Ensemble Learning

To improve the predictive performance of protein encoding predictions, we developed a framework that combines various encoding methods. We experimented with two approaches: **concatenation and voting**. In concatenation, the encodings were combined by adding them together and used the resulting representation as input for our predictive model. In voting, separate predictive models for each encoding method were trained. The final prediction is then calculated with the majority-voted label over a fixed test set.

2-6 Metrics & Statistical Analysis

The metric used for analyzing classification performance is the F1-score which quantifies the prediction power even in imbalanced datasets as it takes both precision

and recall into account. For regression, we used mean squared error (MSE) and R^2 to indicate how the models perform. MSE is the mean of the square of differences between the actual labels and the predicted values in the test set while R^2 represents the variation explained by the independent variables. For sensitivity analysis, experiments were implemented with multiple random seeds (20 in affibody and 30 in NESP dataset) and p-value has been calculated to analyze the null hypothesis. The null hypothesis assumes that the performances of the methods are similar and when rejected, we consider the methods to be statistically significant in their obtained output. The results for multiple seeds are shown with violin plots where the white dots represent the mean value.

The techniques outlined in this section are summarized in Figure 1.

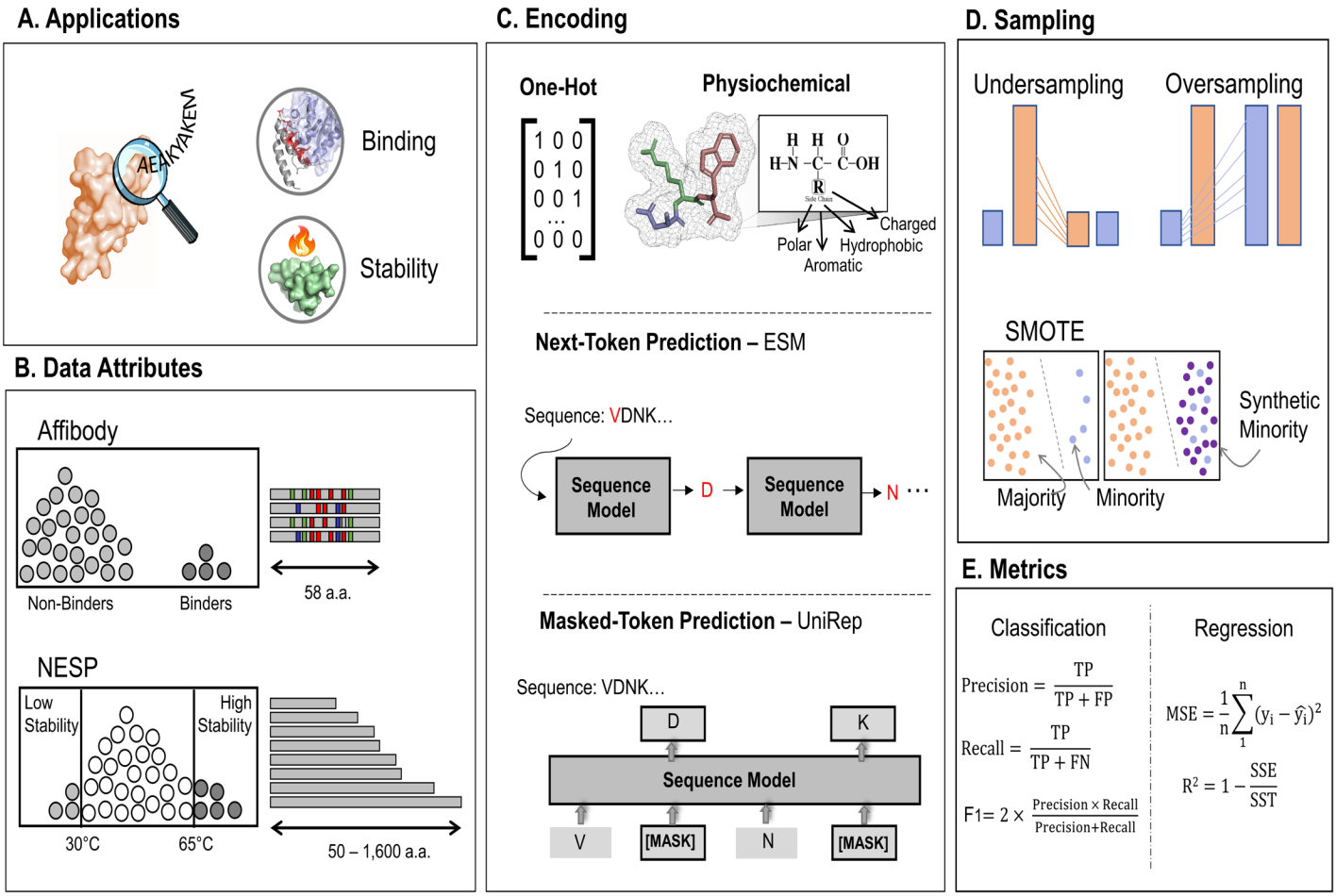

Figure 1. Overview of the implemented techniques, data attributes, and evaluation metrics. A. Illustrates the use of sequence-function mapping to identify protein sequence functionality (e.g., therapeutics, diagnostics, enzymatic function). B. Data attributes for the two datasets used in this study. The first dataset includes high-fitness protein binders among a pool of non-binder affibody sequences with up to 17 mutation sites. The other dataset includes a wide array of proteins with their associated melting point. C. One-hot encoding, physicochemical encoding, and pre-trained models were used to encode the protein sequences present in our datasets. All present protein amino acid information in a machine-readable format, but in different ways. One-hot encoding converts each amino acid to a binary vector of all 0s but 1 where it belongs to its position in the matrix. In physicochemical encoding, each amino acid is represented by its physiochemical characteristics, such as polarity, charge, size, etc. Pretrained models are trained over a large corpus of unlabeled data capturing the syntax and semantics of protein language via NLP-driven models, such as next-token prediction (e.g., UniRep) and masked token prediction (e.g., ESM). D. The sampling methods to be discovered in this literature are undersampling, oversampling, and synthetic minority oversampling techniques (SMOTE). E: The metrics used for evaluating the performance of prediction tasks (classification and regression) are defined.

3. Results

3.1. Sequence-Function Mapping Obtained from High-Throughput Selection Methods & Deep Sequencing Affibody Dataset

To investigate the impact of feature representation, ensemble learning, and sampling methods, several prediction tasks were leveraged. We performed a classification task on the obtained sequences to predict the scarce high affinity binder class among the pool of non-binder class in the affibody data. For NESP dataset, in the classification task, we simplified the data by choosing two classes of low- (Tm ≤35°C) and high-stability (Tm ≥60°C). In addittopm, regression was implemented to increase the prediction difficulty and to observe how protein encodings perform releatively. The models were tasked with predicting the stability (Tm) value and all the sequences with measured pH=7 were included. The details of obtained sequences after cleaning, and the type of prediction tasks are reported in Table1**. Note that the NESP results will be overviewed in section 3-5 and supplementary information.**

Table 1. Dataset attributes and prediction tasks

Dataset	Task	Fitness	Model	Attributes
Affibody	Classification	Binding Affinity	Logistic Regression	82,663 non-binders 6,077 binders
NESP	Classification	Stability	Logistic Regression	3,743 high-stability 1,311 low-stability
NESP	Regression	Stability	Random Forest Regressor	18,190 total

3.2. Physiochemical Feature Encoding, Interpretable yet Lower Predictive Capacity.

The classification results in physiochemical encodings are shown in Figures 2 and 3. We ranked the leading features in discriminating non-binder and binder classes and listed the encoding method's F1-score in different sampling methods. The physiochemical encoding performance was not among the lead encoding methods, yet it achieved a high F1-score with only 20 features. It also provided insights on how physiochemical features correlate with each other in the given data (Figure S1).

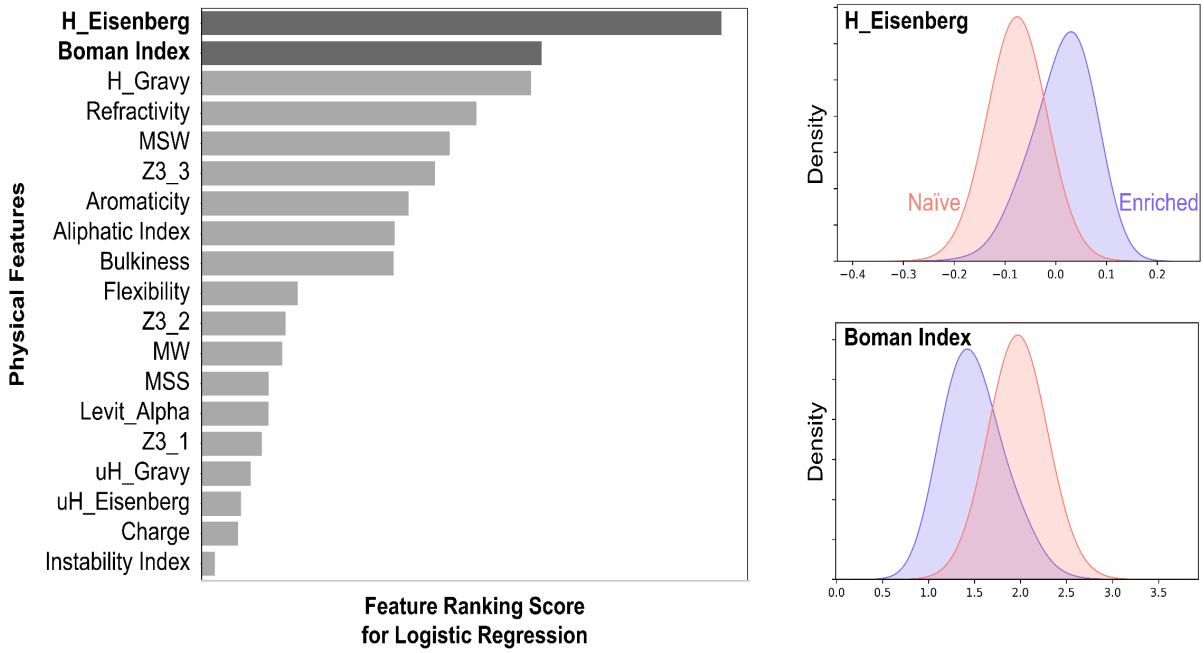

Figure 2. The lead physiochemical features in naïve and enriched class discriminations were H_Eisenberg, Boman Index, and H_Gravy. Gravy and Eisenberg capture hydrophobicity scales. Boman Index is a measure of the protein's ability to interact with its environment based on the solubility of individual residues. The enriched proteins in our library have gone through negative screening and are specific to their target. Therefore, there is a shift to a lower Boman index for this population. Note that the plot is the result of oversampling, SMOTE, in the logistic regression task.

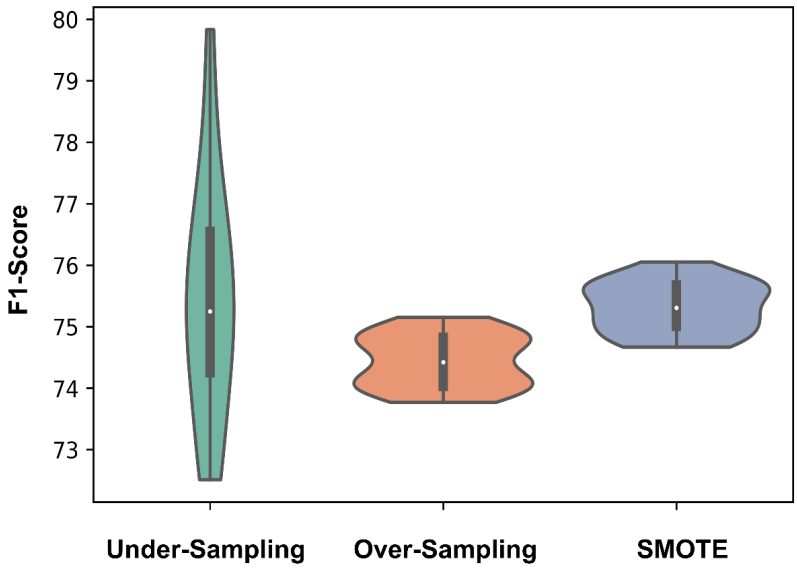

Figure 3. SMOTE with mean F1-score= 75.5% was found to be the most effective sampling method when encoding the affibody sequences with physiochemical features. A) Sensitivity analysis of the predictive performances with 20 random seeds. The results from the analysis indicate that undersampling might perform as powerfully as SMOTE since the p-value>0.05. However, incorporating all the p-values, SMOTE has shown a highly significant difference in F1-score compared to oversampling.

3.3. Comparison Over All the Encoding and Sampling Methods

Once the lead physical features for high-affinity binders were determined, we demonstrated the performance of different protein representations within our selected sampling techniques. The prediction performance indicates that each encoding method performed differently in predicting the fitness of proteins, and One-Hot and UniRep were the top performers. In addition, among the samplings, SMOTE boosted the F1-score in almost all cases. Figure 4 exhibits the F1-score distributions within 20 different random seeds. A complete T-Test analysis is reported in Table S1-3.

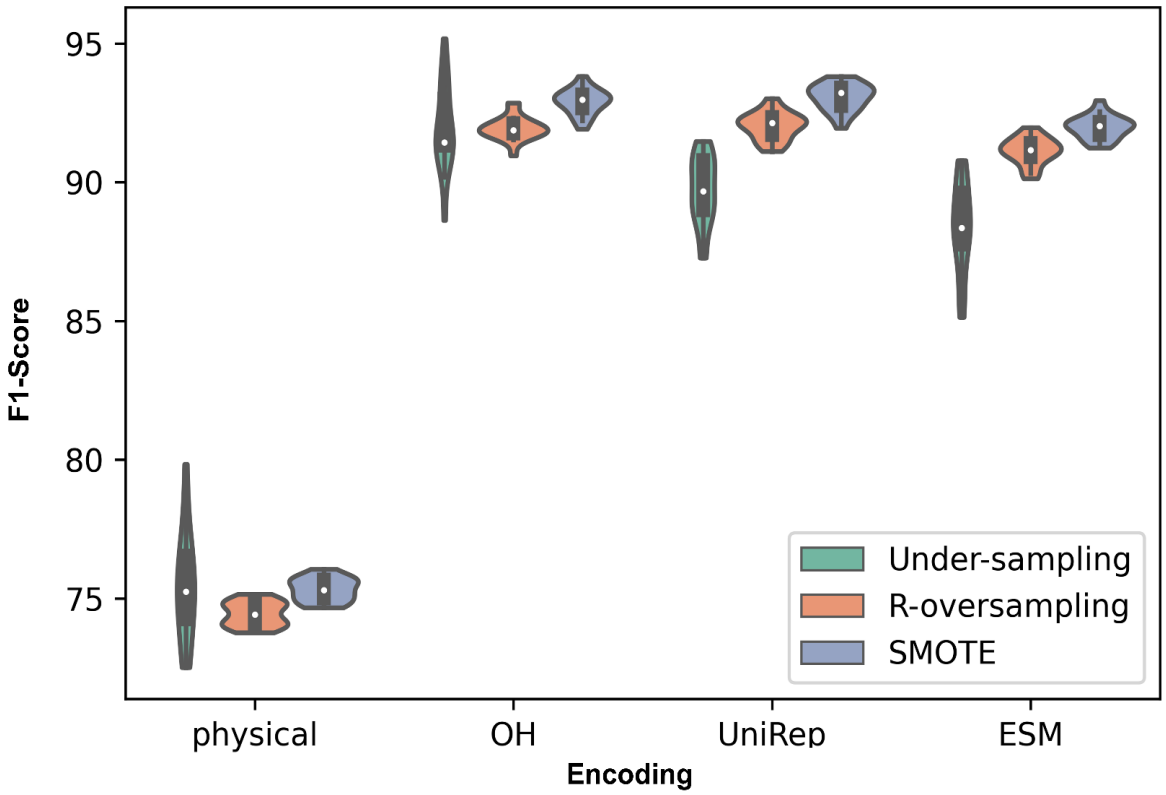

Figure 4. Performance analysis of encoding methods highlights the shortcomings of physical features and strength of the SMOTE sampling method. Protein sequences encoded using physical features, one-hot, UniRep, and ESM were used to perform classification tasks among the affibody dataset. Within each encoding method, under-sampling, random over-sampling, and SMOTE sampling methods were evaluated for each encoding. The resulting F1 scores over 20 random seeds are shown here as violin plots. Detailed statistical analysis is provided in Table S1.

3.4. Increased Generalizability & Predictive Performance via Ensemble Learning

Due to the varying performances of the protein encodings, we postulated that ensemble learning increases the models' predictive performance. As oversampling performed better than undersampling in three out of four encoding methods, we exclusively analyzed the ensemble learning for the two oversampling types (i.e., R-oversampling, and SMOTE). The physicochemical encoding for this analysis was discarded as its performance was not as potent as the other encodings.

Figure 5 represents ensemble technique, voting, remarkably enhanced the performance with respect to all the methods with a mean F1-score=97% over the 20 random seeds. This represented voting method enhanced the prediction score by combining the predictions of multiple models based on single encodings. We concluded that as different encodings might capture distance and relationship of the datapoints differently, combining their predictions boosted the final model performance. The encoding methods used for voting technique in the dataset are visualized in Figure 6 in a UMAP (Uniform Manifold Approximation and Projection) [56] plot

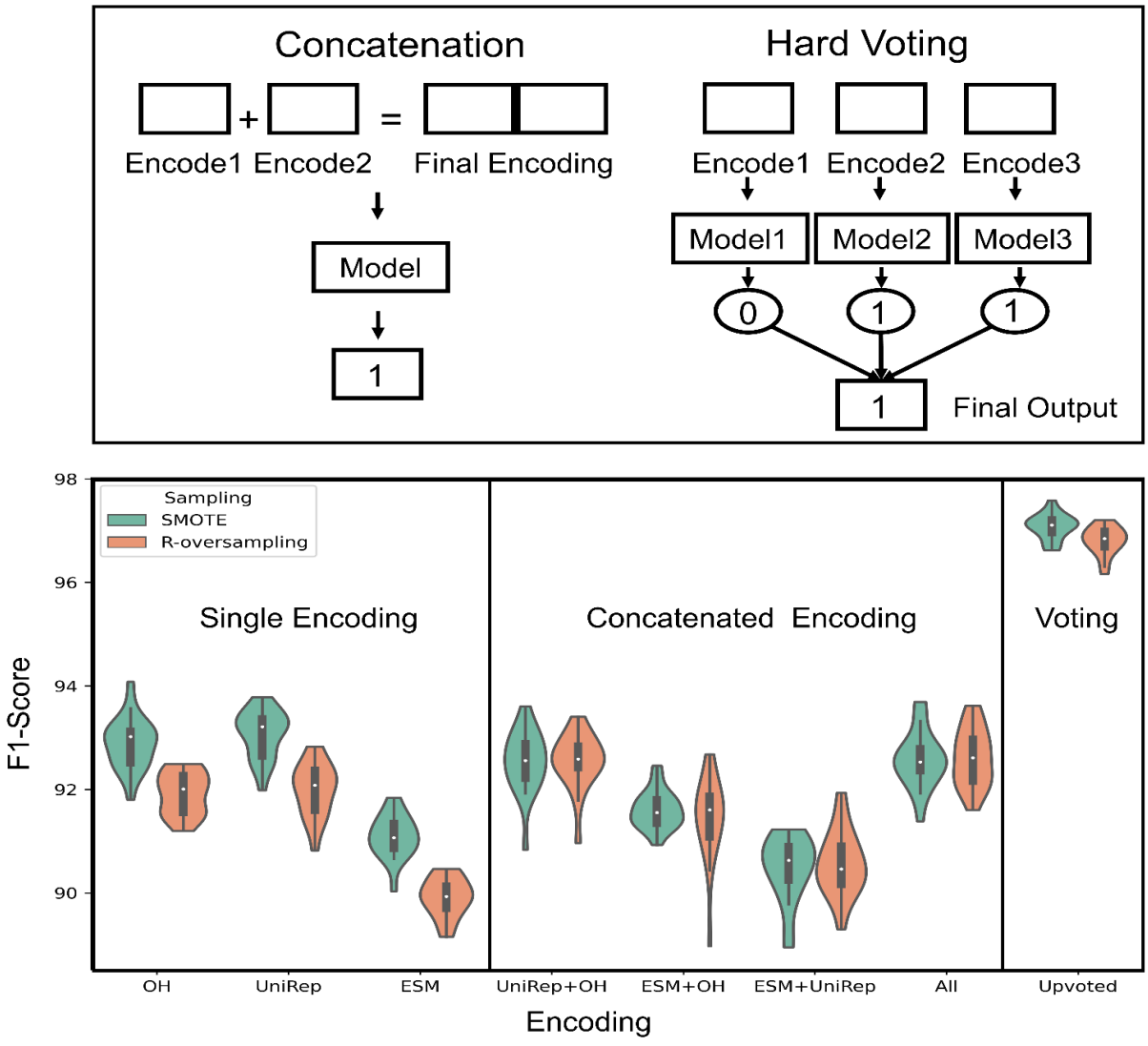

Figure 5. Voting Substantially Improved the predictive performance in All Random Initializations over Different Encoding Methods. The plot above has three regions from left respectively; it includes single encoding methods, concatenation of encodings, and voting of predictions. The vote was performed such that each encoding went through a predictive model over the same dataset. Then, the final prediction is obtained by majority voting. It is insightful how voting increases the models' robustness and generalizability. The concatenation performed similarly or worse than the best model in single encodings. For the statistical analysis, please refer to the supplementary. The best model among all predictions was Upvote with SMOTE sampling, Mean-F1-score=97%. Refer to the supplementary for a summary of statistics (Table S2).

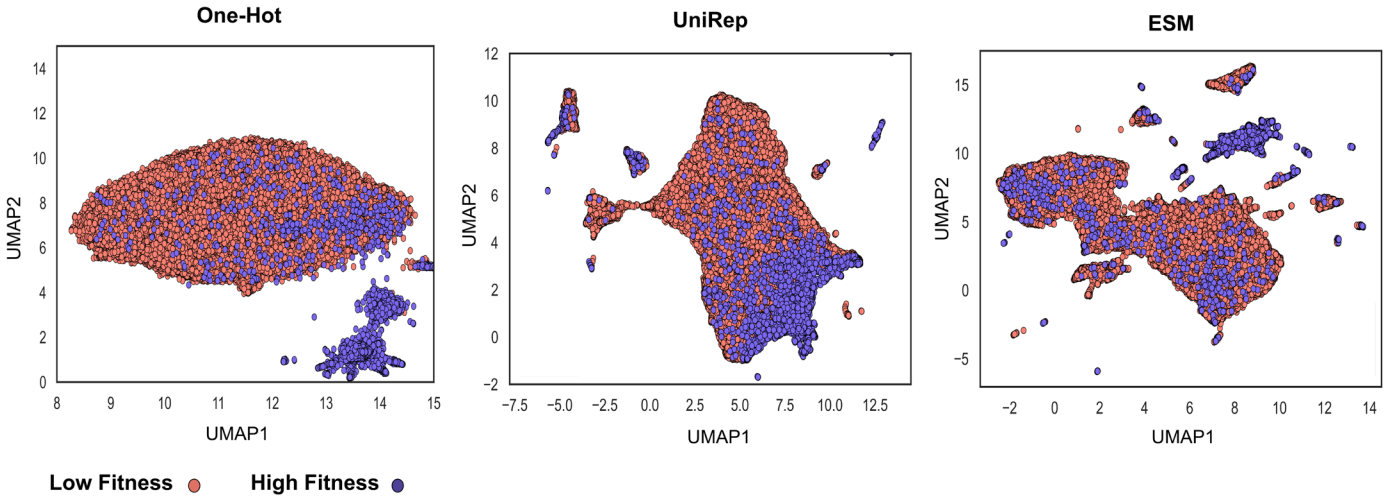

Figure 6: **Different protein encodings potentially capture distinct functional aspects of the proteins.** A 2D visualization of the encoding techniques which resulted in improved prediction in the voting method in UMAP. This method is a dimensionality reduction technique such as PCA with unique advantages such as preserving the local structure of the data and capturing non-linear relationships between data points. In observing the sequence-function relationship in proteins, one can conclude that each protein sequence representation/encoding has the potential to capture different aspects of the fitness to be predicted.

Note that while F1-score is used as an overall measure of model's performance, individual analysis of confusion matrix values of a classification algorithm affords clarity on how the model performs predicting the labels for classes (Figure S2).

3.5. How Protein Encodings Perform Considering the Data Attributes.

The hypotheses were tested over affibody datasets that had notable attributes such as severe imbalance, multiple mutation sites, affinity and specificity enrichment, and small molecular protein length. The obtained results indicated voting and oversampling were highly effective methods to boost the fitness prediction performance. However, individual protein-encoding performance comparisons need more convincing explanation and thorough exploration. Specifically, we wondered why ESM underperformed one-hot and UniRep despite more powerful setup in pretraining and a being showcased in studies for high prediction potential [57]. While the performance could be due to the datatype (e.g., small protein, complex fitness, etc.), we decided to further analyze the encoding prediction scores in a completely different dataset and bring insights on embedding performances in various conditions (e.g., data size in training, protein length, prediction task difficulty). The curated data contains 18,190 sequences with varying lengths and provides melting points which indicate the protein stability. Figure 7 is the performance comparison in stability prediction of embeddings, their concatenation, and voting using different data sizes. Despite down performing in the affibody affinity data, ESM performed best for stability prediction when including proteins with max length=500 (Figure S6).

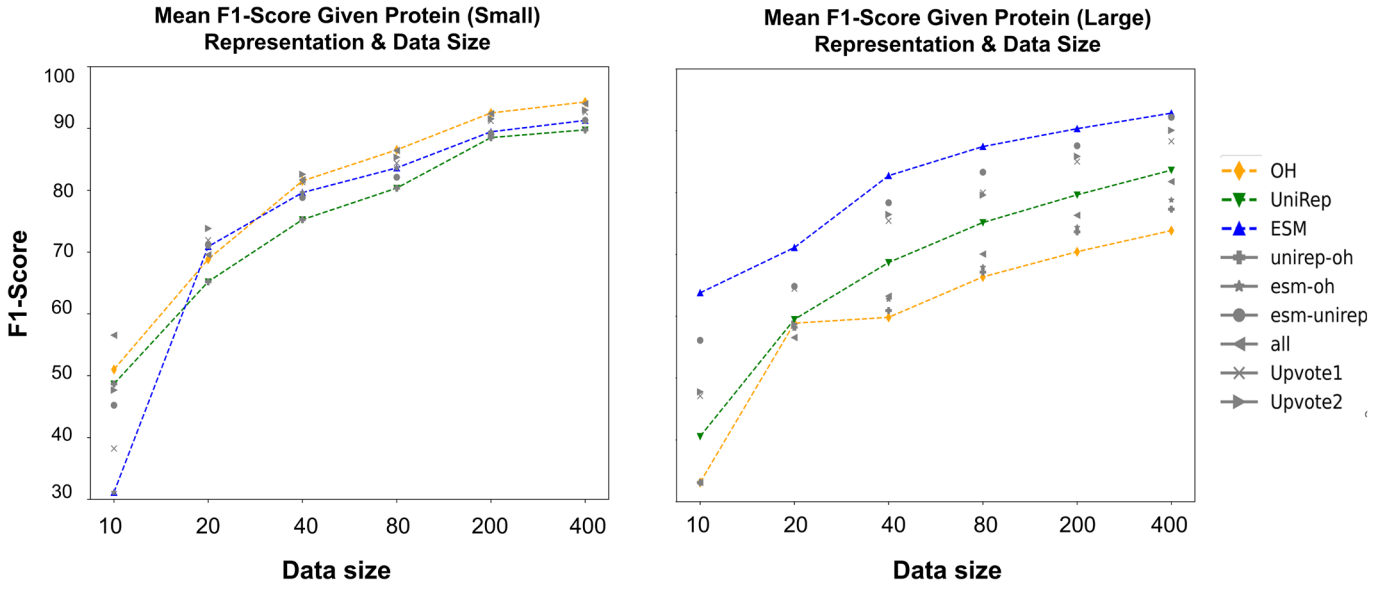

Figure 7. **Protein representation performances over different data sizes for small proteins (length<=120) vs. large(400<=length<=1500).** The obtained results are largely different with respect to the protein size. **Highlights: For small proteins,** upon comparing the violin plots and statistical test results, protein sequence encoding methods were performed distinctively with respect to the initial dataset (protein max length=500). One-Hot encoding had a more significant contribution in boosting the classification metrics for small proteins. As an example, when N=400, both One-Hot and All-Encoding concatenation with a mean F1-score of 94% outperformed the other encoding methods**. One-Hot tends to be problematic for large proteins** as it results in a highly sparse encoding vector. This has been shown in this plot when One-Hot encoding performance is not satisfactory enough in comparison with ESM and UniRep. When N=400, Based on both the violin plots and Welch t-test, either ESM or ESM_UniRep with 92% mean F1-score achieved the highest performance. One-Hot with *a* 73% mean F1-score was the lowest score among all the encodings. Refer to the supplementary information for all one-by-one comparisons of the statistics and classification scores.

The last analysis is a regression task for predicting the melting point value. We wondered how different encoding methods perform if we use all the data and increase the prediction challenge (Tm prediction rather than stability class prediction). MSE and R^2 are shown in predicting the Tm values of a dataset of 18,190 sequences with 0.3 test size. There was a significant difference in the performance of encoding methods which was not the case in classification task. ESM was the best encoding method in predicting stability (R^2 score= 0.65). Note that we used all the data and did not use sampling methods for training. Therefore, sensitivity analysis was not critical to perform for performance estimations. The regression metrics are reported in Table2.

Table2: Regression metrics for encoding methods in validation & test set.

Encoding	Validation		Test	
	R^2	MSE	R^2	MSE
One-Hot	0.21	141	0.24	130
UniRep	0.49	108	0.4	102
ESM	0.65	63	0.65	60

4. Discussion

In this study, we shed light on two key challenges of applying discriminative models over amino acid sequence data for protein engineering applications: 1) handling

imbalanced data and 2) choosing an appropriate protein representation (i.e., encoding). Assay-labeled sequence data in this domain is often severely imbalanced (due to the rugged and sparse nature of the protein fitness landscape) and requires careful consideration in data sampling, splitting, and choice in data representation for model training. To capture this common occurrence of imbalanced data, we trained discriminative ML models over our cytometry-sorted deep-sequenced small protein (affibody) data to distinguish between functional sequences (n=6,077) among a large collection of non-functional protein sequences (n=82,663). We then explored the impact of encoding protein sequences using two simplistic approaches (One-Hot encoding, physiochemical encoding) and two language-based methods (UniRep, ESM). We hypothesized that, as each protein representation may capture distinct information, combining representations via embedding concatenation and ensemble learning increases overall performance and generalizability.

To address the issue of imbalanced data, we implemented multiple sampling techniques – undersampling, random oversampling, and SMOTE – and compared performances via F1 scores. Our results indicate that implementing oversampling techniques over imbalanced datasets improves predictive performance relative to undersampling or the exclusion of sampling methods. Among the sequence representation methods, embeddings are the answer to improved fitness prediction and data requirements. However, it is essential to consider the choice of protein representation, its benefits, and its drawbacks. For example, the choice of fitness to be predicted (e.g., thermal stability, binding affinity, target specificity) and the language model pretraining procedure affect the model's predictive performance and need further discussion. Therefore, we analyzed an additional dataset (i.e., the NESP dataset which included a variety of protein sequences with their Tm) to discuss the effect of protein representations over the variables such as protein length, protein fitness, and prediction type (i.e., classification vs. regression). For ensemble learning, we used majority voting to combine the prediction of each representation over the same ML model which significantly improved the F1-score in Affibody dataset.

As only a very small fraction of protein sequences is experimentally annotated with properties, the primary goal of embeddings is to distill valuable information from unlabeled data and use them for property/fitness prediction. Previous reports have observed that there are sequence motifs, conserved regions, and evolutionary information in the protein databases that can be learned by language models [33,34,58]. This has been tested with different NLP techniques, varying model parameters, and clustering sizes for databases used and resulted in a wide array of language-based protein representations [30,57,59,60]. These promising embeddings (e.g., ESM, UniRep) have been evaluated in many studies and have improved the fitness prediction scores and alleviated the assay-labeled data requirements [59,61]. However, there are also studies that report minor improvements to predictions by using solely embedding methods. In some cases, prediction scores were improved by simpler representations such as one-hot or physiochemical encoding [62,63]. Similarly, Rao et al. pointed out a different performance of embeddings in TAPE [34] with 38 million parameters based on different protein engineering tasks. Their model performed outstandingly in fluorescence and stability prediction while it did not perform as well as hand-engineered features in contact prediction.

The current capabilities and limitations of language models motivate the need for optimizing the pretraining task and improving the methodology for supervising the pre-trained models. Consider ESM2, one of the largest language models used for protein sequences which has shown significant improvement in protein structure prediction compared to previous embeddings. In our study, protein representations obtained via ESM2 significantly outperformed UniRep or One-Hot in stability prediction. However, in the context of predicting binding functionality among small protein affibody variants, its performance was exceeded by UniRep and One-Hot (Figure 4). This motivates looking into what knowledge is transferred by pretraining models and how useful they are for specific fitness predictions, with or without further supervision. Here we covered the core challenges and considerations in supervising the models in fitness prediction, yet additional downstream analysis and posing insightful questions will give us more understanding and directions in discriminating the protein sequences based on their

fitness. In order to improve the pretraining step, we might adopt techniques such as adjusting the masking rate [64], adding biological priors [60,65], increasing the model parameters [57], and building specialized language models for the desired fitness [66] , given the growing data availability and computational resources. Additional studies are required for improved downstream fitness predictions, such as fine-tuning with a reduced chance of overfitting [67], incorporating the effect of post-translational modifications, and characterizing the performance of embeddings in different data setups [68] with varying protein types and finesses for supporting the development of novel proteins in diagnostics and therapeutics.

5. Conclusions

We integrated machine learning and protein engineering knowledge to identify high-fitness protein sequences. We quantified model performance while varying the choice of feature representation, ensemble learning, and sampling methods. Analysis across a broad range of protein chain lengths revealed the ESM language model to be most beneficial for encoding large protein sequences (Figure 7). Yet, in the context of small protein sequences, comparable performance was observed between one-hot encoding and the language models (ESM and UniRep). In our analysis, oversampling proved to be an effective technique to improve performance when dealing with severely imbalanced datasets (Figure 4). Finally, ensemble learning was a promising method for boosting the prediction scores when using unique, competitive encoding methods (Figure 5).

Data and Code Availability: The NESP data is available at https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/data.

The Affibody dataset is available upon request. The source code for this project can be found On the GitHub repository: https://github.com/WoldringLabMSU/Sequence_Fitness_Prediction.

Acknowledgments: We would like to thank the department of chemical engineering and material science at Michigan State University for funding this project. For the support in computational resources, we thank MSU's High Performance Computing Center (HPCC). Finally, we would like to express our gratitude to Alex Golinski for his invaluable comments and insights.

Conflicts of Interest: The authors declare no conflict of interest.

Supplementary Materials: The following supporting information can be downloaded at the provided link.

Figure S1: Physicochemical feature correlation plot affibody dataset.

Table S1: Figure 4 t-test results.

Table S2: Figure 5 t-test results.

Table S3: T-test results for R-Oversampling vs. SMOTE.

Figure S2: Violin plot-based confusion matrix for Figure 5 results.

Table S4: T-test for Figure S2.

Figure S3: Physicochemical feature correlation plot NESP dataset.

Figure S4: Physicochemical feature ranking for NESP dataset.

Figure S5: Physical Feature representation while using maximum N=1000, performed poorly and have not got selected for the main figure.

Figure S6: Mean F1-score comparison between protein representations including proteins with max length=500.

References

1. Liebermeister, W.; Noor, E.; Flamholz, A.; Davidi, D.; Bernhardt, J.; Milo, R. Visual Account of Protein Investment in Cellular Functions. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 8488–8493, doi:10.1073/pnas.1314810111.
2. Schlessinger, J. Cell Signaling by Receptor Tyrosine Kinases. *Cell* **2000**, *103*, 211–225, doi:10.1016/s0092-8674(00)00114-8.
3. Hogan, B.L. Bone Morphogenetic Proteins: Multifunctional Regulators of Vertebrate Development. *Genes Dev.* **1996**, *10*, 1580–1594, doi:10.1101/gad.10.13.1580.
4. Andrianantoandro, E.; Basu, S.; Karig, D.K.; Weiss, R. Synthetic Biology: New Engineering Rules for an Emerging Discipline. *Mol. Syst. Biol.* **2006**, *2*, 1–14, doi:10.1038/msb4100073.
5. Heim, M.; Römer, L.; Scheibel, T. Hierarchical Structures Made of Proteins. The Complex Architecture of Spider Webs and Their Constituent Silk Proteins. *Chem. Soc. Rev.* **2010**, *39*, 156–164, doi:10.1039/b813273a.
6. Kolmar, H. Biological Diversity and Therapeutic Potential of Natural and Engineered Cystine Knot Miniproteins. *Curr. Opin. Pharmacol.* **2009**, *9*, 608–614, doi:10.1016/j.coph.2009.05.004.
7. Krasniqi, A.; D'Huyvetter, M.; Devoogdt, N.; Frejd, F.Y.; Sörensen, J.; Orlova, A.; Keyaerts, M.; Tolmachev, V. Same-Day Imaging Using Small Proteins: Clinical Experience and Translational Prospects in Oncology. *J. Nucl. Med.* **2018**, *59*, 885–891, doi:10.2967/jnumed.117.199901.
8. Romero, P.A.; Arnold, F.H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866–876, doi:10.1038/nrm2805.
9. Hellinga, H.W. Rational Protein Design: Combining Theory and Experiment. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 10015–10017, doi:10.1073/pnas.94.19.10015.
10. Jäckel, C.; Kast, P.; Hilvert, D. Protein Design by Directed Evolution. *Annu. Rev. Biophys.* **2008**, *37*, 153–173, doi:10.1146/annurev.biophys.37.032807.125832.
11. Li, G.; Dong, Y.; Reetz, M.T. Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes? *Adv. Synth. Catal.* **2019**, *361*, 2377–2386, doi:10.1002/adsc.201900149.
12. Anand, N.; Eguchi, R.; Mathews, I.I.; Perez, C.P.; Derry, A.; Altman, R.B.; Huang, P.S. Protein Sequence Design with a Learned Potential. *Nat. Commun.* **2022**, *13*, 1–11, doi:10.1038/s41467-022-28313-9.
13. Wu, Z.; Jennifer Kan, S.B.; Lewis, R.D.; Wittmann, B.J.; Arnold, F.H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 8852–8858, doi:10.1073/pnas.1901979116.
14. Saito, Y.; Oikawa, M.; Sato, T.; Nakazawa, H.; Ito, T.; Kameda, T.; Tsuda, K.; Umetsu, M. Machine-Learning-Guided Library Design Cycle for Directed Evolution of Enzymes: The Effects of Training Data Composition on Sequence Space

Exploration. *ACS Catal.* **2021**, *11*, 14615–14624, doi:10.1021/acscatal.1c03753.

15. Golinski, A.W.; Mischler, K.M.; Laxminarayan, S.; Neurock, N.L.; Fossing, M.; Pichman, H.; Martiniani, S.; Hackel, B.J. High-Throughput Developability Assays Enable Library-Scale Identification of Producible Protein Scaffold Variants. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, 1–11, doi:10.1073/pnas.2026658118.
16. Chen, J.; Zheng, S.; Zhao, H.; Yang, Y. Structure-Aware Protein Solubility Prediction from Sequence through Graph Convolutional Network and Predicted Contact Map. *J. Cheminform.* **2021**, *13*, 1–10, doi:10.1186/s13321-021-00488-1.
17. Wang, S.; Liu, D.; Ding, M.; Du, Z.; Zhong, Y.; Song, T.; Zhu, J.; Zhao, R. SE-OnionNet: A Convolution Neural Network for Protein–Ligand Binding Affinity Prediction. *Front. Genet.* **2021**, *11*, 1–9, doi:10.3389/fgene.2020.607824.
18. Kuzmin, K.; Adeniyi, A.E.; DaSouza, A.K.; Lim, D.; Nguyen, H.; Molina, N.R.; Xiong, L.; Weber, I.T.; Harrison, R.W. Machine Learning Methods Accurately Predict Host Specificity of Coronaviruses Based on Spike Sequences Alone. *Biochem. Biophys. Res. Commun.* **2020**, *533*, 553–558, doi:10.1016/j.bbrc.2020.09.010.
19. Das, S.; Chakrabarti, S. Classification and Prediction of Protein–Protein Interaction Interface Using Machine Learning Algorithm. *Sci. Rep.* **2021**, *11*, 1–12, doi:10.1038/s41598-020-80900-2.
20. Vander Meersche, Y.; Cretin, G.; de Brevern, A.G.; Gelly, J.C.; Galochkina, T. MEDUSA: Prediction of Protein Flexibility from Sequence. *J. Mol. Biol.* **2021**, *433*, 166882, doi:10.1016/j.jmb.2021.166882.
21. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75, doi:10.1109/MCI.2018.2840738.
22. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
23. Mnasri, M. Recent Advances in Conversational NLP : Towards the Standardization of Chatbot Building. **2019**.
24. Campagna, G.; Xu, S.; Moradshahi, M.; Socher, R.; Lam, M.S. Genie: A Generator of Natural Language Semantic Parsers for Virtual Assistant Commands. *Proc. ACM SIGPLAN Conf. Program. Lang. Des. Implement.* **2019**, 394–410, doi:10.1145/3314221.3314594.
25. Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling Aspects of the Language of Life through Transfer-Learning Protein Sequences. *BMC Bioinformatics* **2019**, *20*, doi:10.1186/s12859-019-3220-8.
26. Ofer, D.; Brandes, N.; Linial, M. The Language of Proteins: NLP, Machine Learning & Protein Sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1750–1758, doi:10.1016/j.csbj.2021.03.022.
27. Bateman, A.; Martin, M.J.; O'Donovan, C.; Magrane, M.; Apweiler, R.; Alpi, E.;

Antunes, R.; Arganiska, J.; Bely, B.; Bingley, M.; et al. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43*, D204–D212, doi:10.1093/nar/gku989.

28. Katz, K.; Shutov, O.; Lapoint, R.; Kimelman, M.; Rodney Brister, J.; O'Sullivan, C. The Sequence Read Archive: A Decade More of Explosive Growth. *Nucleic Acids Res.* **2022**, *50*, D387–D390, doi:10.1093/nar/gkab1053.

29. Torrisi, M.; Pollastri, G.; Le, Q. Deep Learning Methods in Protein Structure Prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1301–1310, doi:10.1016/j.csbj.2019.12.011.

30. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Yu, W.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *14*, 1–16, doi:10.1109/TPAMI.2021.3095381.

31. Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. *Nat. Commun.* **2022**, *13*, 4348, doi:10.1038/s41467-022-32007-7.

32. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; et al. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, doi:10.1073/pnas.2016239118.

33. Alley, E.C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G.M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16*, 1315–1322, doi:10.1038/s41592-019-0598-1.

34. Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y.S. Evaluating Protein Transfer Learning with Tape. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9689.

35. Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; Linial, M. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *Bioinformatics* **2022**, *38*, 2102–2110, doi:10.1093/bioinformatics/btac020.

36. Hsu, C.; Nisonoff, H.; Fannjiang, C.; Listgarten, J. Combining Evolutionary and Assay-Labelled Data for Protein Fitness Prediction. *bioRxiv* **2021**, 2021.03.28.437402.

37. Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. *Adv. Neural Inf. Process. Syst.* **2021**, *35*, 29287–29303.

38. Chu, S.K.S.; Siegel, J. Predicting Single-Point Mutational Effect on Protein Stability. **2021**.

39. Lv, Z.; Wang, P.; Zou, Q.; Jiang, Q. Identification of Sub-Golgi Protein Localization by Use of Deep Representation Learning Features. *Bioinformatics* **2020**, *36*, 5600–5609, doi:10.1093/bioinformatics/btaa1074.

40. Li, K.; Zhong, Y.; Lin, X.; Quan, Z. Predicting the Disease Risk of Protein Mutation

Sequences With Pre-Training Model. *Front. Genet.* **2020**, *11*, 1–10, doi:10.3389/fgene.2020.605620.

41. Min, S.; Kim, H.G.; Lee, B.; Yoon, S. Protein Transfer Learning Improves Identification of Heat Shock Protein Families. *PLoS One* **2021**, *16*, 1–14, doi:10.1371/journal.pone.0251865.

42. Woldring, D.R.; Holec, P. V.; Stern, L.A.; Du, Y.; Hackel, B.J. A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold. *Biochemistry* **2017**, *56*, 1656–1671, doi:10.1021/acs.biochem.6b01142.

43. Pultz, D.; Friis, E.; Inversion; Salomon, J.; Maggie; Fischer Hallin, P.; Baagøe Jørgensen, S. Novozymes Enzyme Stability Prediction. Kaggle. **2022**.

44. Woldring, D.R.; Holec, P. V.; Stern, L.A.; Du, Y.; Hackel, B.J. A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold. *Biochemistry* **2017**, *56*, 1656–1671, doi:10.1021/acs.biochem.6b01142.

45. Müller, A.T.; Gabernet, G.; Hiss, J.A.; Schneider, G. modlAMP: Python for Antimicrobial Peptides. *Bioinformatics* **2017**, *33*, 2753–2755, doi:10.1093/bioinformatics/btx285.

46. Yang, K.K.; Wu, Z.; Bedbrook, C.N.; Arnold, F.H. Learned Protein Embeddings for Machine Learning. *Bioinformatics* **2018**, *34*, 2642–2648, doi:10.1093/bioinformatics/bty178.

47. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model. **2022**.

48. Kovács, B.; Tinya, F.; Németh, C.; Ódor, P. SMOTE: Synthetic Minority Over-Sampling Technique Nitesh. *Ecol. Appl.* **2020**, *30*, 321–357.

49. Fernández, A.; García, S.; Herrera, F.; Chawla, N. V. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905, doi:10.1613/jair.1.11192.

50. Mohammed, A.J. Improving Classification Performance for a Novel Imbalanced Medical Dataset Using SMOTE Method. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 3161–3172, doi:10.30534/ijatcse/2020/104932020.

51. Rupapara, V.; Rustam, F.; Shahzad, H.F.; Mehmood, A.; Ashraf, I.; Choi, G.S. Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model. *IEEE Access* **2021**, *9*, 78621–78634, doi:10.1109/ACCESS.2021.3083638.

52. Hasanin, T.; Khoshgoftaar, T.M.; Leevy, J.L.; Bauder, R.A. Severely Imbalanced Big Data Challenges: Investigating Data Sampling Approaches. *J. Big Data* **2019**, *6*, doi:10.1186/s40537-019-0274-4.

53. Blagus, R.; Lusa, L. Evaluation of SMOTE for High-Dimensional Class-Imbalanced Microarray Data. *Proc. - 2012 11th Int. Conf. Mach. Learn. Appl. ICMLA 2012* **2012**,

2, 89–94, doi:10.1109/ICMLA.2012.183.

54. van den Goorbergh, R.; van Smeden, M.; Timmerman, D.; Van Calster, B. The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1525–1534, doi:10.1093/jamia/ocac093.

55. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna. **2019**, 2623–2631, doi:10.1145/3292500.3330701.

56. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

57. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Costa, A. dos S.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction. *bioRxiv* **2022**, 2022.07.20.500902.

58. Marquet, C.; Heinzinger, M.; Olenyi, T.; Dallago, C.; Erckert, K.; Bernhofer, M.; Nechaev, D.; Rost, B. Embeddings from Protein Language Models Predict Conservation and Variant Effects. *Hum. Genet.* **2021**, 1–35, doi:10.1007/s00439-021-02411-y.

59. Biswas, S. Low-N Protein Engineering with Data-Efficient Deep Learning A Paradigm for Low-N Protein Engineering. **2020**, 1–39.

60. Rao, R.M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. In Proceedings of the Proceedings of the 38th International Conference on Machine Learning; Meila, M., Zhang, T., Eds.; PMLR, 2021; Vol. 139, pp. 8844–8856.

61. Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. *bioRxiv* **2021**, 2021.07.09.450648.

62. Shanehsazzadeh, A.; Belanger, D.; Dohan, D. Is Transfer Learning Necessary for Protein Landscape Prediction? **2020**, 1–10.

63. Wittmann, B.J.; Yue, Y.; Arnold, F.H. Informed Training Set Design Enables Efficient Machine Learning-Assisted Directed Protein Evolution. *Cell Syst.* **2021**, *12*, 1026-1045.e7, doi:10.1016/j.cels.2021.07.008.

64. Wettig, A.; Gao, T.; Zhong, Z.; Chen, D. Should You Mask 15% in Masked Language Modeling? **2022**.

65. Lupo, U.; Sgarbossa, D.; Bitbol, A.F. Protein Language Models Trained on Multiple Sequence Alignments Learn Phylogenetic Relationships. *Nat. Commun.* **2022**, *13*, 1–11, doi:10.1038/s41467-022-34032-y.

66. Nourani, E.; Asgari, E.; Mc Hardy, A.; Mofrad, M. TripletProt: Deep Representation Learning of Proteins Based on Siamese Networks. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2021**, 1–1, doi:10.1109/tcbb.2021.3108718.

67. Hua, H.; Li, X.; Dou, D.; Xu, C.-Z.; Luo, J. Fine-Tuning Pre-Trained Language

Models with Noise Stability Regularization. **2022**, *14*, 1–15.

68. Wang, B.; Member, S.; Wang, A.; Chen, F.; Member, S.; Wang, Y.; Kuo, C.J. Evaluating Word Embedding Models : Methods and Experimental Results. 1–13.