Deep Learning for Protein Thermostability Engineering

Rafael Espinheira Alves
rafael.e.alves@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

January 2021

Abstract

Recent advances in natural language processing show that modern neural networks are capable of learning context and semantics of language with unsupervised pre-training. Using such language models, researchers are now introducing new ways to represent protein sequences as continuous vectors (embeddings) that capture biophysical properties directly from unlabelled sequences. In this work, the embeddings generated by *SeqVec*, a deep learning model based on the *ELMo* language model and trained on the *UniRef50* data set, were studied for their capacity to capture protein thermostability.

Three thermostability data sets were prepared and used to train and evaluate several machine learning models for their capacity to predict protein thermostability properties using only the *SeqVec* embeddings as features. Although far from perfect, experiments on wild-type proteins show that such models produce meaningful predictions of protein melting temperature, and can isolate proteins with high thermostability. Additionally, models trained to predict the effect of mutations on the protein thermostability were capable of achieving Matthews correlation coefficients as high as 0.354 on independent testing data, a competitive value compared with recent literature.

Using transfer-learning for protein stability prediction opens up a new form of sequence-based tools that do not rely on biophysical features and do not require protein structure information. With this work it was shown that this approach to protein thermostability prediction has a lot of potential, but the lack of data is still a large limitation.

Keywords: Machine learning; Deep learning; Language models; Protein engineering; Protein thermostability prediction.

1. Introduction

Protein engineering aims to obtain proteins with useful properties for technology, science and medicine. As the amino acid sequence determines the protein's properties [1], specific amino acid modifications have already resulted in new protein designs and in the optimization of existing enzymes [2]. However, traditional protein engineering processes face overwhelming amounts of possible mutations to model [3], from which most are not functional or can produce unaccounted effects in stability [4], or are limited to an iterative approach of trial and error with expensive and time-consuming screening procedures [3].

One of the protein properties with industrial interest is the thermostability of enzymes. Increasing their thermostability is useful to facilitate purification steps based on heat treatments, higher associated stability to destabilizing agents, and also allows the use of higher reaction temperatures making for a faster and more sterile process [3], [5]. Numerous protein thermostability models and machine learning (ML) predictors have been developed, but protein stability modelling is a very difficult task and for which there is still limited data [6].

With the exponential increase in protein sequence databases [7], already we are seeing some efforts in making the connection between sequence and function, with natural language processing models that find high-level protein representations, called embeddings, that are closely associated with the protein function and properties. The hypothesis that these models can be applied to model protein sequences and learn the biological rules that dictate protein properties directly from the amino acids is seeing a lot of support [8], [9], [10], [11], [12], [13].

In this work, the *SeqVec* protein sequence model [9], based on the *ELMo* natural language model [14], was studied for its capacity to capture protein thermostability information directly from sequence data, to determine its potential use in protein thermostability prediction for protein engineering applications.

2. Validation of the SeqVec model

Aiming to explore the biological meaning of the *SeqVec* embeddings and to validate the original paper's results, a data set was prepared for visualization of biological properties in the embedding space and for the implementation of a secondary structure prediction algorithm.

2.1. Materials and Methods

A protein secondary structure data set was prepared from the *Protein Data Bank* [15], after removing sequences with over 50% identity using *CD-HIT* [16]. The Structure Integration with Function, Taxonomy and Sequence database [17] was used to obtain the Enzyme Commission (EC) numbers of 26999 proteins. *SeqVec* was then used to embed the protein sequences. The proteins were represented by the sequence average of the sum of the outputs of *SeqVec*, and the 23969 amino acids of 100 of these proteins were represented by the output of the middle layer of the model.

The previously mentioned amino acid embeddings were split into a training set of 20 655 residues and a testing set of 3314 residues, used to implement a k-Nearest Neighbours (k-NN) classifier, trained to predict the secondary structure label of the embeddings. A Principal Component Analysis (PCA) was used to reduce the dimensionality of the embeddings to 100 components (40.2% explained training data variance), and the value of k was chosen by cross-validation as 25.

2.2. Results

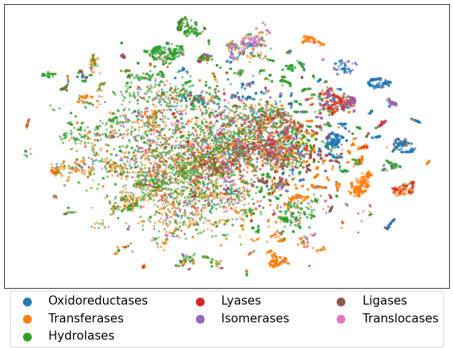

Figure 1: t-SNE visualization of the protein embeddings from the data set, coloured by EC number of the proteins (x-axis: t-SNE 1; y-axis: t-SNE 2. The projection into two dimensions shows small isolated clusters that are representative of protein function.

Projecting the protein embeddings to two dimensions using t-Distributed Stochastic Neighbour Embedding (t-SNE) (figure 1), the representation obtained shows several small isolated clus-

ters that are representative of enzyme class, with a mixed and uninformative central cluster.

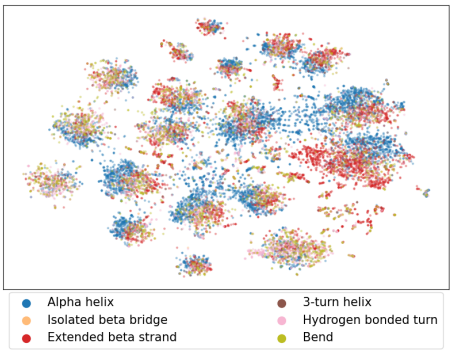

Figure 2: t-SNE visualization of the amino acid embeddings from the data set (x-axis: t-SNE 1; y-axis: t-SNE 2). Colouring by the secondary structure label of the amino acids shows some separation between different structures.

Projecting the amino acid embeddings in a similar procedure shows that these group each amino acid type in mostly separate clusters, indicating that the model learns how to identify each amino acid and to take into account its context. Amino acids with similar physicochemical properties were also found to be closer in the embedding space. Additionally, the clusters also exhibit some clustering within themselves, representative of the secondary structure labels of the amino acids (figure 2).

Using these embeddings to train a k-NN secondary structure predictor produced a positive performance in the testing set, with an accuracy score of 57.8% and an F1 score (weighted average based on class support) of 54.9%. Analysis of the model's Confusion Matrix (CM) and Receiver Operating Characteristic (ROC) curves show a good performance of the model for the most common labels and that it can still maintain a better than random performance on the most difficult classes (figures 3 and 4.

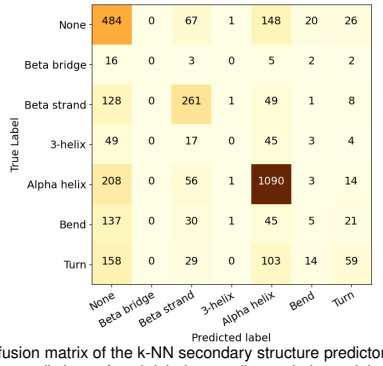

Figure 3: Confusion matrix of the k-NN secondary structure predictor on the testing set, detailing the predictions of each label, according to their true labels.

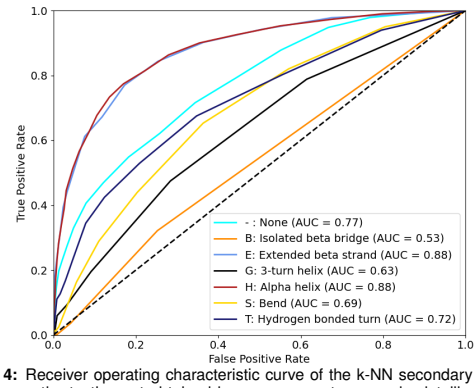

Figure 4: Receiver operating characteristic curve of the k-NN secondary structure predictor on the testing set obtained by a one-vs-rest approach, detailing the true positive rate as a function of the false positive rate, obtained at different decision thresholds.

2.3. Discussion

As was observed by the authors of the *SeqVec* model [9], the unsupervised embeddings learned by the *ELMo* model trained on protein sequences contain biological information which can be used to model aspects of protein biochemistry. Such conclusions were also obtained by other authors of relevant deep learning protein embedders such as *UniRep* [10], *D-SPACE* [8], the bidirectional transformer as published by [11], and the transfer-learning repository published by [13].

The obtained t-SNE projection of the protein embeddings is similar to the authors' results with the same procedure on the *SCOPe* data set. By obtaining equivalent results in this small experiment, the generalizability of the results to different data sets is confirmed.

We also observed that the *SeqVec* amino acid embeddings show the capacity to learn the physico-chemical properties of the amino acids, as well as some indication that they can be used for secondary structure prediction tasks with clusters representative of the secondary structure labels. However, due to time limitations this experiment used a reduced data set with only 100 proteins, and although it was processed to remove protein sequences with at least 50% sequence similarity, it may not be representative of the diverse range of proteins that are found in nature.

For secondary structure prediction, we obtained an accuracy score of 57.8% that is inferior to the best application of *SeqVec* by the authors, which implements a deep learning model with evolutionary profiles together with the amino acid embeddings, and obtained an accuracy of 64.1%, and even this method was inferior to the state of the art secondary structure prediction method *NetSurfP-2.0*, which was applied by the *SeqVec* model's authors and obtained an accuracy score of 71.1%. Another of the previously mentioned transfer-learning effort used for secondary structure prediction is the bidirectional transformer, which shows similar performances, with an accuracy score of 60.8%.

This can be due to the small training set, the implementation of k-NN, one of the most simple ML algorithms, and the severe class imbalance of this prediction task. State of the art models use deep neural networks for this problem that have been trained on more extensive data. Given the difficulty of this prediction task, our implementation is considered a success.

Additionally, more effort could have been performed to explore individual protein sequences and their secondary structure annotation, because certain patterns might have arisen that could explain why this amino acid prediction task was not considered optimal by the original authors, facilitating further efforts to improve this method for protein annotation.

3. Thermostability prediction with the ProTherm wild-type data set

The first effort to develop a model of protein thermostability directly from protein sequences was attempted with the *ProTherm* database. Using its wild-type data set, a ML regression model was implemented, using only the *SeqVec* protein embeddings as features, to test whether these encode thermostability information.

3.1. Materials and Methods

The *ProTherm* database [18] was used to prepare a data set of protein sequences and their free Gibbs energy of unfolding (ΔG) annotation. With several unusable records and multiple records per protein in the database, only a total of 794 experimental records were collected, coming from 119 different protein sequences. In order to use only one label per sequence, all ΔG records were converted to $kcal/mol$, and the mean ΔG value of each protein across all available experimental conditions was calculated. The protein sequences were represented by the sequence average of the amino acid embeddings obtained as the output of the middle layer of *SeqVec*.

This data set was discretized into 5 bins of equal ΔG intervals, and a random, stratified stratified split was performed,

where 85% of the data was used for the training of a *Lasso* linear regression model, and the remaining 15% for an independent evaluation of the model on unseen data. PCA was applied to reduce the dimensionality of the data to 50 dimensions (explained data variance of 94.33%). This model was also applied with polynomial features of degrees 2, 3 and 4, and all implementations used an α regularization strength manually chosen as 0.1.

Considering the poor performances obtained with this implementation, different approaches to incorporate the experimental conditions were implemented, such as calculating a weighted mean ΔG value that takes into account the experimental conditions and gives more weight to records closer to physiological conditions [19], removing the effect of the experimental conditions by fitting a linear regression to the data and removing the residuals (individually for each protein, and also to the entire data set at once), and also incorporating these as additional features. Two baselines were also developed. The first is a naïve baseline that always predicts the average ΔG value of the data set, and the second one is a linear regression that predicts the ΔG values based on temperature and pH only.

3.2. Results

Neither the PCA nor the t-SNE two-dimensional projections of the protein embeddings revealed a separation between proteins with different ΔG values, suggesting that the *SeqVec* embeddings are not capable of capturing this information directly from the sequence.

Using polynomial features of degree 3 and the mean ΔG values, the model could achieve a test Root Mean Squared Error (RMSE) of 4.474, but the test set r^2 score of -0.036 and an Explained Variance Score (EVS) of -0.027 are very poor. The negative predictive power of this model was compared to the first baseline model, where a constant prediction of the data set's average ΔG produced a test RMSE of 4.088. This indicates that using the *SeqVec* protein embeddings to train a thermostability prediction model directly from sequence is not directly possible. The inclusion of the temperature and pH as additional features for the model produced the best results, with the lowest test RMSE and the highest r^2 and EVS scores of all the experiments, as expected since it uses all 794 records in the data set and includes large amounts of repeated protein sequences. The second baseline indicates that the temperature and pH by themselves do not include any relevant information for the prediction of protein thermostability.

Table 1: Root Mean Squared Error (RMSE), r^2 correlation coefficient and Explained Variance Score (EVS) performances of the regression models using different preprocessing approaches to take into account the experimental conditions. Calculating a weighted mean of ΔG that gives more weights to experiments closer to physiological conditions was considered the best approach, because it was the only approach with positive r^2 and *EVS* that did not used repeated records.

Model	Features degree	RMSE test	r^2 test	EVS test
Average ΔG baseline	1	4.088	-0.015	0.000
T and pH only baseline	1	4.009	0.024	0.049
Mean ΔG	3	4.474	-0.036	-0.027
Weighted mean ΔG	2	4.592	0.035	0.048
Remove T and pH effect from the entire data set	1	7.860	-0.229	0.000
Remove T and pH effect from each protein	1	8.455	-0.175	-0.171
Including T and pH as additional features	3	3.643	0.461	0.463

It is also noteworthy that the experiment where the effect of the experimental conditions was globally removed from the data set produced very similar ΔG values for all records, producing an almost constant prediction of the ΔG values, inaccurate in the testing set. In addition, removing the effect of the experimental conditions from each protein separately produced the worst model in this experiment, as a result of records with extreme experimental conditions being assigned unexpected ΔG values during the data processing. We can also observe that using the weighted average of each protein's ΔG values produced better results than a simple average calculation of this value, with similar RMSE values, but positive r^2 and EVS values in the testing set, which were only achieved by the second baseline and the model which included the experimental conditions.

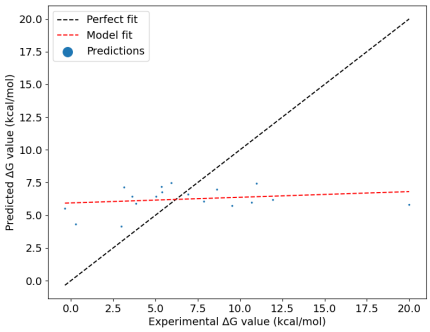

Figure 5: Scatter plot of the predicted free Gibbs energy of unfolding values (y axis) and their true values (x axis) in the testing set. The trend is almost horizontal and uninformative.

In a final effort to extract conclusions from this model, scatter plots of the predictions of this model in the training set and in the testing set were produced. An overall positive correlation with positive slope could be found in the training set, but the predictions were almost horizontal in the testing set (figure 5), which further proves the difficulty of developing such a machine learning protein thermostability predictor from this data set.

3.3. Discussion

Prediction of the free Gibbs energy of unfolding of wild-type proteins is not usually performed directly from sequence. This procedure is usually based on additional structural information, or based on physicochemical models of amino acid interactions. In this experiment, the *ProTherm* database of wild-type proteins proved to be unusable for the development of a machine learning model to predict protein thermostability directly from protein sequence, using the *SeqVec* embeddings as features.

This can be a result of three steps of the process: inadequate data processing, inadequate features or inadequate regression models. However, not much could have been done with only 119 different protein sequences, given the issues related to the *ProTherm* database. The visualization of the features suggested instead that these were not capable of capturing elements of thermostability, which was further confirmed by the inaccurate prediction models.

However, the correlation obtained by the polynomial regression with the *SeqVec* protein embeddings was positive, and although mostly horizontal, indicates that perhaps with more records a more accurate predictor could be developed. Additionally, using the experimental conditions as additional features could, in fact, be more adequate for visualization of the predictions (instead of the method using the weighted ΔG average of each protein). This method allowed the use of more records, and should not have been discarded because of this, as the use of experimental conditions as predictive features is also frequently used in literature on thermostability engineering.

4. Prediction of thermostability changes with the ProTherm single-mutants data set

Instead of modelling protein thermostability directly from sequence, the prediction of changes to protein thermostability as a result of point mutations (described by the $\Delta\Delta G$ value) is more frequently used in protein engineering. To assess the usefulness of the *SeqVec* embeddings in the development of such a model, a thermostability data set was compiled. The amino acid embeddings of the obtained protein sequences were explored for the development of predictive features for several ML models, from which the most promising were studied in detail.

4.1. Materials and Methods

Considering the issues with the *ProTherm* database found in the previous experiment, and looking to use as much data as possible, we used the protein thermodynamic data made available by the *iStable 2.0* [6] and the *PremPS* [20] prediction models. However, there is a large overlap between these data sets, as all originate from *ProTherm* data. Duplicate records were removed, and a unique $\Delta\Delta G$ value per record was calculated. After this processing step we obtained a data set with 3706 unique records from 305 different proteins, without redundancy from experimental conditions. This data set was named S3706, and was split in a partition with 3272 mutation records, called S3272, for the training of sequence-based protein thermostability machine learning models, and another with 434 records, called S434, used for an unbiased evaluation of the models (table 2). This split was performed manually to avoid a random split, so that the models could be evaluated in entirely different protein sequences.

Table 2: Description of the data partitions of the data set S3706 of protein thermostability changes upon single mutations used for the separate training and evaluation of the machine learning algorithms, evidencing the imbalanced representation of the positive and negative classes of records.

	S3706	S3272	S434
Total number of records	3706	3272	434
Records with positive label	855	648	171
Records with negative label	2851	2588	263
Total number of proteins	305	155	150

The *SeqVec* model was used to process the wild-type and the mutant sequences of the records in this data set, from which the amino acid residue embeddings were obtained as the output of the middle layer of the model. 10 different feature sets were attempted to describe the mutation records in the embedding space, from which it was observed that by describing each wild-type and mutant sequences by the average of the residue embeddings in a window of 5 residues to each side of the mutation, and then subtracting the mutant representation from the wild-type representation, the best results could be observed in both visualization and prediction experiments. For the development of ML models, PCA was used to reduce the features to 250 dimensions, with 83.35% explained data variance.

Several ML classifiers were attempted, from which the linear Support Vector Machine (SVM) produced the best results based on the Matthews Correlation Coefficient (MCC) and the precision score. Hyperperparameter tuning was performed by cross-validation, in which the primal formulation and the dual formulation of the problem were both attempted as well as the standard SVM loss function and its squared formulation, and the l_1 and l_2 regularization types, applied with different regularization strengths C. A squared loss function with a l_1 penalty and a C value of 50 provided the best mean cross-validation MCC of 0.133.

A baseline model was also developed, which uses a Decision Tree classifier based on simple features to describe the protein mutations: a one-hot-encoding label of the amino acid types and of the physicochemical properties (aliphatic, aromatic, polar neutral, acidic, basic or unique), their molecular weights and hydrophobicity values, and the *BLOSUM62* value for the substitution were used.

4.2. Results

The feature set prepared to describe the mutation records was projected to two-dimensions by t-SNE (figure 6), where some separation between mutations with a positive and a negative effect in the $\Delta\Delta G$ can be observed, suggesting that the features capture some thermostability information directly from the protein sequence.

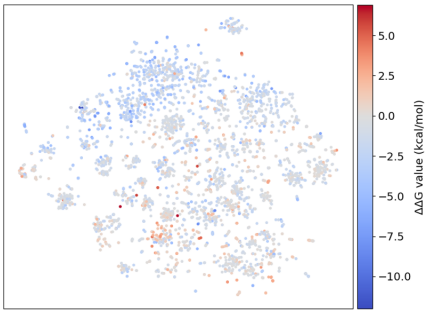

Figure 6: Projection to two dimensions of the mutation records, represented by the difference between the wild-type features and the mutant features, where each protein is described by the average of the embeddings of the amino acids in a window of 5 residues in each direction of the mutation by t-SNE (x-axis: t-SNE 1; y-axis: t-SNE 2). The projection shows separation between mutations with a positive and a negative effect on thermostability.

Using these features, a linear SVM with hyperparameters tuned by cross-validation in the training set was used to predict the $\Delta\Delta G$ value of the records, achieving a MCC of 0.318 and precision score od 0.774 in the testing set. The model's CM and Precision-Recall Curve (PRC) are shown in figures 7 and 8, respectively, showing that although severely biased to the over-represented negative class, this model is capable of correctly predicting a large number of positive labels. The baseline model was only capable of achieving a MCC of 0.200 and precision of 0.73, with a severly worse PRC.

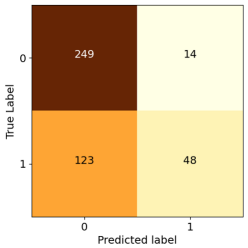

Figure 7: Confusion Matrix (CM) of the tuned linear Support Vector Machine (SVM), evaluated on the testing set S434 to predict the protein thermostability changes upon single mutations.

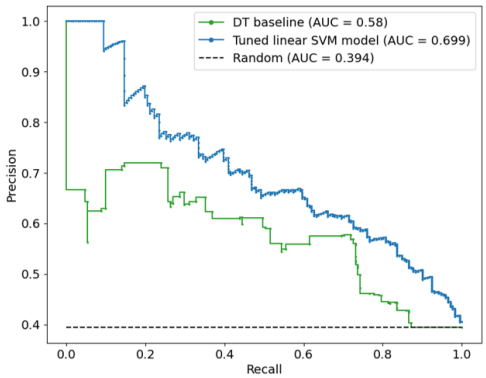

Figure 8: Precision-Recall curve (PRC) of the tuned linear Support Vector Machine (SVM), evaluated on the testing set S434 to predict the protein thermostability changes upon single mutations.

The achieved MCC values are quite behind the state of the art *iStable 2.0* prediction model, which achieved a value of 0.708 for the same metric, but are better than the *PoPMuSiC* model with a MCC of 0.291 and the *MUpro* model, with a MCC of 0.248 [6].

Although a logistic regression and a MLP classifier were capable of achieving slightly better MCC values than the linear SVM shown here, these models produced worse precision metrics, and were considered inadequate for a protein engineering application. To further study the precision of this model, different subsets of the testing set were created to evaluate the

performance of the model on protein sequences of decreasing sequence similarity to the training set, where it was found that the model achieves a better precision on sequences similar to those it was trained on.

4.3. Discussion

Overall, the *SeqVec* embeddings provided predictive features to develop ML models of $\Delta\Delta G$ changes upon single mutations that provide better performances than some well established models but still fall behind state of the art.

However, this comparison is not straightforward, as the data sets used in this experiment for the training and the testing of the machine learning models were different from those used by other models. Since different models frequently use different data sets, only the review papers that train and test each model on the same data present a valid comparison, and this was not implemented for the models developed in this thesis.

In protein engineering, it is desirable to perform as least mutations to a protein as possible in order to avoid altering its fitness, and since each protein can be mutated in a wide number of different ways, from which only a few will result in real positive stability changes, a thermostability predictor does not need to accurately predict a lot of the records correctly (which would translate into a high MCC), as long as the ones it predicts as positive are correct (high precision score). If the most confidently predicted positive records are correct, such a model can still be interesting for application in a protein thermostability engineering procedure. In addition, the study of the PRC in testing subsets of different sequence identities was also useful for this evaluation, where although it would have been interesting to see that the model can predict the positive class correctly independently of sequence identity to the training set, observing a correlation between sequence similarity and higher precision indicates that the model is learning biologically significant features that describe the effect of the mutations in the proteins, and can accurately generalize this knowledge to similar proteins.

It is noteworthy that the models developed in this experiment used solely the *SeqVec* embeddings as features, with no additional structural features, nor even the explicit amino acid sequence of the protein. This approach, very different from the sequence or structure-based models seen in literature, skips the difficult step of generation of complex protein features of those models, which sometimes rely on other physicochemical models themselves. Capable of achieving MCC values that are relevant for recent literature, this opens a completely new approach to protein thermostability prediction with the helpful advantages associated with transfer-learning, namely the possibility to compare different proteins or mutations in the high-dimensional feature space. A way to conclude about the usefulness of the embedding space for direct comparison of mutations could have been based on the two-dimensional projections generated for the features, as the t-SNE plot showed some clustering that could be representative of specific amino acid substitutions, which could be analyzed as positive or negative according to their context in the entire protein sequence.

A trade-off between increasing the MCC at a decrease on the precision score, and vice versa, was also observed. In no attempt could the models maintain both of these scores elevated, indicating that, when using the *SeqVec* features for machine learning prediction of protein thermostability upon point mutations, a choice needs to be made on a very precise model that misses a lot of potential positive mutations, or a model that can overall separate the two classes of mutations but might suggest some negative mutations as positives.

In general, it was expected that this experiment would find better results, as the prediction of protein thermostability directly from sequence is quite difficult due to the complexity of the protein folding process, but the prediction of the effect that a mutation can have in a protein can more easily be modelled.

5. The effect of mutations in the embeddings

Seeing that the effect of a mutation in the thermostability of a protein is better modelled by *SeqVec* when only the embeddings of amino acids close to the mutation are used to generate the protein embedding, it became interesting to study the effect that mutations have in the embeddings of the entire amino acid sequence.

5.1. Materials and Methods

To describe the effect of a mutation in the amino acid embeddings, the euclidean distance between the wild-type and the mutant sequence amino acid embeddings of each protein was used. This was used to study the mutation records in the S3706 data set, and an additional mutagenesis data set prepared from a subset of 84 proteins from the S3706 data set. These proteins were chosen due to the availability of binding-site information and for having a sequence length smaller than 250 amino acids. For each protein, a single-mutant sequence was generated by changing each amino acid in the protein individually, according to the *BLOSUM62* matrix [21], where each amino acid was mutated according to the highest scoring substitution, or randomly between the highest scoring substitutions. This was performed under the hypothesis that this would make the mutations as equivalent as possible, and allow us to focus entirely on the mutation location and its effect on the protein sequence as a function of the position.

For each wild-type and mutant sequences, the amino acid embeddings were obtained as the sum of the output of the three layers of *SeqVec*. The euclidean distance effect was studied both in terms of linear distance in the amino acid sequence and also in terms of 3D distance between the amino acids in the wild-type protein conformation. The Wilcoxon rank-sum statistical test was used to verify the statistical relevance of differences in the euclidean distances between the embeddings.

5.2. Results

The mutation records in the S3706 data set were explored, from which a few arbitary records were plotted in terms of euclidean distance between the wild-type and the mutant sequence embeddings, showing that the highest difference was usually concentrated in and close to the mutation location, but that in some cases there was also an effect in distant amino acids (figure 9). Performing a similar visualization of all records showed that quite frequently there is also a strong effect in amino acids distant from the mutation. Note that except for the mutated amino acid in the center of the x axis, these amino acids are of the same type in the wild-type and in the mutant sequences, showing the capacity of the *SeqVec* model to capture the different contexts, as a consequence of a mutation in another amino acid of the protein.

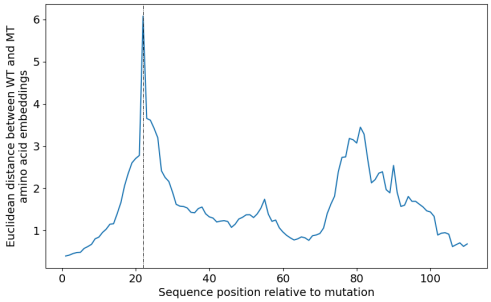

Figure 9: The euclidean distance between the wild-type sequence amino acid embeddings and the mutant sequence amino acid embeddings (y axis) as a function of the sequence position of the mutation (x axis) of the mutation W22F in protein 1AJ3.

The same procedure was performed with the mutagenesis data set prepared above. Assuming that this mutagenesis approach does not induce bias towards the possible mutations, a mutation to a binding site is expected to have a stronger effect in the embeddings, and is also expected to affect amino acids throughout the sequence more strongly. By splitting the

results in segments of linear distance to the mutation location, the statistical relevance of the difference between the euclidean distances of embeddings from sequences with binding site mutations can be compared to the euclidean distances of embeddings from sequences with mutations outside of binding sites (figure 10).

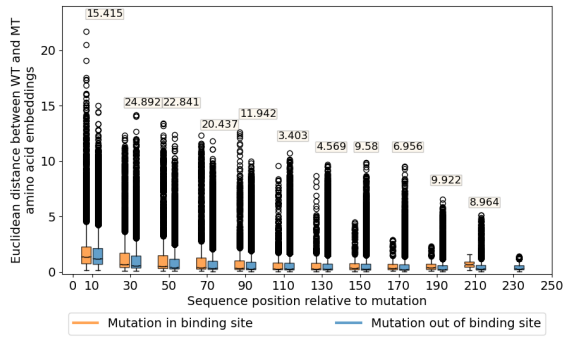

Figure 10: Boxplot representation of the euclidean distance value between wild-type sequence amino acid embeddings and mutant sequence amino acid embeddings (y axis) in segments of 20 amino acids of sequence distance to the mutation (x axis). For each segment, the rank-sum test statistic is shown, indicating that the effect of a mutation in a binding site is stronger than that of mutations out of the binding sites.

A widespread effect is can be seen throughout the amino acid embeddings, and this effect is clearly stronger in amino acids closer to the mutation, and gradually weaker with an increasing distance except for certain cases were the mutation caused a disruption throughout the sequence. A rank-sum (RS) test shows that mutations in binding sites cause a higher euclidean distance between embeddings throughout the sequence, as evidenced by the positive RS test statistic in every segment, and the p-values very close to zero obtained in all segments. By performing a collective RS test to compare the distributions of the euclidean distance between the embeddings of the binding site mutations and the non-binding site mutations, a factor of 53.237 with a p-value close to zero is obtained, further confirming the hypothesis that binding site mutations cause a significantly stronger effect in the embeddings.

Performing the same procedure in terms of 3D distance between the amino acids, we can now see that the stronger effect of the binding site mutations is only noticed by embeddings of amino acids close in space to the mutation. This is evidenced by the fact that the RS test statistic becomes negative for larger 3D distances, with p-values close to zero in all segments.

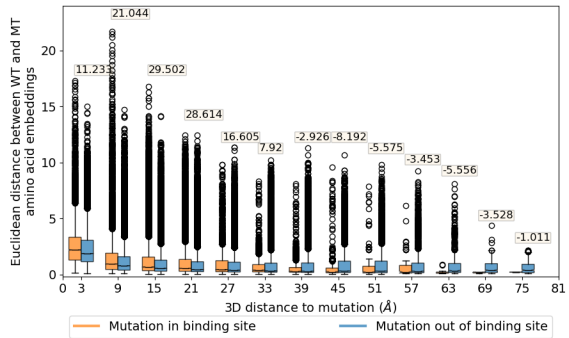

Figure 11: Boxplot representation of the euclidean distance value between wild-type sequence amino acid embeddings and mutant sequence amino acid embeddings (y axis) in segments of 6 angstrom of distance in space to the mutation (x axis). For each segment, the rank-sum test statistic is shown, indicating that the effect of a mutation in a binding site is stronger than that of mutations out of the binding sites only if the amino acid is close to the mutation.

Since amino acids of the protein binding site are usually close in space in the final conformation of the protein, and since it was previously concluded that a mutation to a binding site causes a stronger effect in the embeddings, seeing this effect concentrated in amino acids that are close together suggests that the *SeqVec* embeddings capture the protein conformation and the three-dimensional distance between amino acids.

5.3. Discussion

In this experiment we studied the effect of mutations in the *SeqVec* embeddings. Under the hypothesis that the simulated mutations were equivalent, and did not induce bias to specific amino acid changes, the euclidean distance between the amino acid residue embeddings of the wild-type and the mutant sequences suggests that the *SeqVec* embeddings can both capture long-distance effects of the mutation in the protein sequence, but also the three-dimensional conformation of the protein and the interactions between amino acids in the protein, as evidenced by the difference observed when mutations were performed to binding sites.

However, given the diversity of three-dimensional protein structures observed in nature, the assumption that the simulated mutations are equivalent in amino acid type changes is far from ideal. Although simulating the amino acid substitutions based on the *BLOSUM* concept for biologically significant amino acid changes was the best solution for this experiment, the use of a curated deep mutational scanning data set would have been more suitable. This was, however, out of the scope of this thesis, and the experiment was performed with the limitation that performing mutations to very similar amino acids might not result in large effects in the context of the other amino acids.

Using the euclidean distance between embeddings to represent their differences could also induce some wrong conclusions, as it does not capture the notion of closeness between two high-dimensional vectors. The cosine similarity could have been used instead but, even with this measure, capturing distances in high-dimensional data is difficult [22]. On the other hand, the secondary structure prediction experiment applied a k-NN algorithm using the euclidean distance with success, so this distance metric is expected to be accurate.

Although the *SeqVec* model was not used for protein engineering nor was it studied for the prediction of mutational effects, the *UniRep* model [10] was applied to predict the stability of naturally occurring proteins and of *de novo* designed proteins using deep mutational scanning data sets, achieving better results than well established methods such as *Rosetta* [23], and was also able to predict the functional effects of mutations to proteins, as well as modelling the fitness landscapes of diverse proteins. This model was also capable of predicting mutations that increase certain properties such as protein fluorescence, which was also observed with the *D-SPACE* model [8]. Additionally, the bidirectional transformer from [11] was used to successfully predict amino acid residue contact points as well as to predict enzyme activity changes upon mutations. Our results with the *SeqVec* model are in agreement with the observations that such unsupervised language models can compete with state of the art models of protein biology.

6. Thermostability prediction with the Meltome Atlas wild-type data set

With the publication of the *Meltome Atlas* [24], an extensive and uniform data set of protein thermostability is made available. Considering the obstacles faced in the prediction of thermostability directly from sequence with the *ProTherm* database, another attempt to develop such a model was performed, using the *SeqVec* models to represent the protein sequences of the *Meltome Atlas*, and studying them for their capacity to capture the melting temperature values.

6.1. Materials and Methods

From the *Meltome Atlas* data set, we used all proteomes except those coming from human cell lines, choosing to use only proteins that were clearly identified with a *UniProt* database entry code [7], with a total of 34501 unique protein sequences. We used only proteins with a melting temperature annotation coming from cell lysate experiments, resulting in 27354 unique protein sequences. For repeated sequences across different

tissues, strains and organisms, the mean melting temperature value was used. The protein sequences were processed by *SeqVec*, from which the protein-level embeddings were generated as the sequence average of the amino acid residue embeddings of each protein, obtained as the output of the middle-layer of the model.

This data set was split in two random, shuffled and stratified partitions (obtained by dividing the original data set in 10 quantiles, and performing a random, shuffled split on each), where 85% of the data was used for training of ML models that use the *SeqVec* protein embeddings to predict the melting temperature values, and the remaining 15% was left out for an independent testing of the performance of the models on previously unseen data. PCA was used to reduce the dimensionality of the features to 100 principal components, fit to the training set with an explained variance of 86.4%, and used to reduce both data set partitions.

From several regression models, the Pearson Correlation Coefficient (PCC) in the testing set was used to choose the Multilayer Perceptron (MLP) as the best model. An architecture with 2 hidden layers of 256 and 20 neurons was chosen by cross-validation based on the Mean Squared Error (MSE). A learning rate of 0.001 and 23 training epochs were implemented, chosen by experiments with an early stopping callback.

A baseline feature set was also developed, which describes each protein sequence by a vector with the frequencies of each amino acid type in the protein. The performance of the previously mentioned MLP with the *SeqVec* embeddings was compared with its performance with this baseline feature set.

6.2. Results

The protein embeddings were projected to two dimensions by t-SNE, revealing that embeddings from thermophile organisms are different from the mesophile embeddings and are grouped together in well-defined clusters, while the mesophile proteins are dispersed throughout the embedding space (figure 12). Colouring this same projection by the organism of each protein, we found that the thermophile organisms are the only ones that are in well-defined clusters in the embedding space, while all other organisms are spread throughout the feature space, suggesting that *SeqVec* can identify thermophile proteins from among a varied data set of protein sequences from multiple organisms.

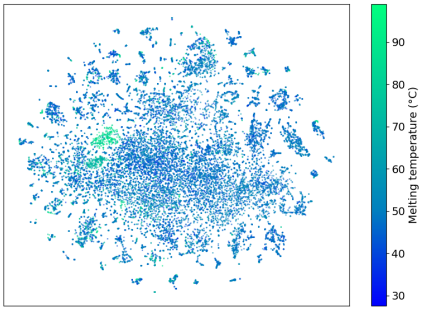

Figure 12: t-SNE projection to two dimensions of the protein embeddings, coloured by melting temperature (x-axis: t-SNE 1; y-axis: t-SNE 2).

Although the MLP model implemented to predict the melting temperature of the proteins showed some signs of overfitting, with a decrease in performance between the training set and the testing set evaluation metrics, this model showed the highest PCC value of 0.74 in the testing set, the most widely used parameter to evaluate protein thermostability regression models. With a close to best test RMSE value of 7.04, this model was considered the best model obtained in this experiment.

For this model, the test r^2 and EVS scores of 0.52 and 0.53, respectively, show a positive predictive power, and their similar values suggest that the model is unbiased. The performance of this model was studied further, by visualization of the predictions of the testing set (figure 13). From this figure we can observe that the *SeqVec* protein embeddings can be used to train a MLP to predict the melting temperature directly from sequence that accurately models the stability of wild-type proteins, as this model has the capacity to predict the melting temperature of the most stable proteins in the data as significantly different from the melting temperature of the more frequent mesophile proteins, while also achieving a good PCC.

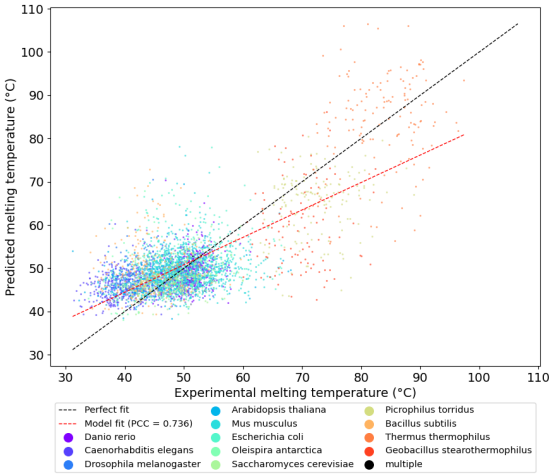

Figure 13: Scatter plots of the melting temperatures of the testing set predicted by the MLP model developed with the protein embeddings, and their true values

To compare the *SeqVec* protein features to hand-crafted features, a baseline feature set that describes each protein by a vector with the frequency of each amino acid in the sequence was developed and used to train and test the same MLP architecture. This baseline model obtained very similar results, with a test PCC value of 0.73 and a test SCC value of 0.43, as well as a similar test RMSE of 7.14, the hand-crafted features provide very similar predictive performances, which suggest that the *SeqVec* features are not an improvement to protein thermostability prediction directly from sequence.

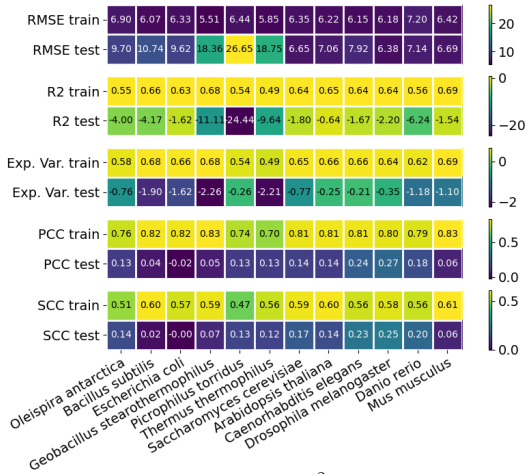

Figure 14: Root Mean Squared Error (RMSE), r^2 correlation coefficient, Explained Variance Score (EVS), Pearson Correlation Coefficient (PCC) and Spearman Correlation coefficient (SCC) performance metrics of the MLP model implemented with the protein embeddings for melting temperature prediction, upon training on all organisms except one, which was used for evaluation. The results indicate that the predictions are not accurate on organisms in which the model was not trained on.

An additional generalization experiment was performed, where the MLP model with the *SeqVec* features was trained on the proteins of all organisms except one, and evaluated on the proteins of the left-out organism.

This experiment showed that this prediction model is not accurate when applied to proteins from organisms it was not trained on (figure 14). Ideally, the model would be capable of maintaining the performance metrics when evaluated on a different organism, but this was not observed due to the nega-

tive r^2 score and EVS throughout all test organisms. The correlation coefficients were also both reduced in all cases, except for certain organisms such as *C. elegans*, *D. melanogaster* and *D. rerio* that show the best PCC and SCC values of this experiment, which are still lower than the correlations on the training data. Additionally, analysis of the RMSE, r^2 score and EVS values of the evaluation organisms shows a clear distinction between evaluating the model on thermophiles and mesophiles/psycrophiles, in which the model shows severely worse values when evaluated on the thermophiles.

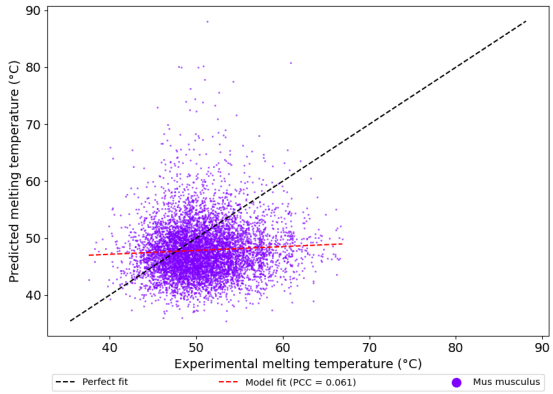

Figure 15: Scatter plots of the melting temperatures predicted for the testing set by the MLP model developed with the protein embeddings and their true values, on a training set without proteins from *M. musculus*, which were used as the testing set. The performance in the testing set indicates that the model is not accurate enough to correctly predict the melting temperatures of proteins with similar melting temperatures belonging to an organism outside of the training set.

The reduced performances in this experiment were studied further, and the prediction plots were generated, in which the reduced PCC in the left-out organisms can be clearly explained. As an example, the experiment where *M. musculus* proteins were isolated from the training and were used as the testing set are shown in figure 15. Although centered in the line of perfect prediction, the predictions on the test proteins are severely dispersed, an unexpected result, considering the good correlation in the training data. This suggests that the model is only learning how to identify the organisms.

The results of these experiments suggest that the *SeqVec* protein embeddings only superficially describe the thermostability of the proteins, because although the first regression model trained on all organisms shows very promising results, this model is incapable of generalization to different organisms. A naïve baseline was constructed, using a linear regression that simply uses the average melting temperature of the organism of the protein as a single feature. The first experiment with the entire data set produces a test RMSE of 4.50 and a test PCC of 0.895, which are better than those of the MLP with either the baseline features or the *SeqVec* features. This naïve baseline also shows lower RMSE values in the same generalization experiment (figure 16).

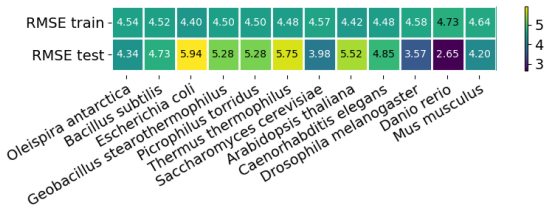

Figure 16: Root Mean Squared Error (RMSE) values of a naive linear regression baseline in the same leave-one-out experiment. The error values are better than those obtained by the Multilayer Perceptron (MLP) model with the *SeqVec* features.

6.3. Discussion

The publication of the *Meltome Atlas*, with a large amount of coherent and well-annotated records from diverse organisms, provided the requirements that the *ProTherm* database previously failed to meet. The *SeqVec* protein embeddings generated from this data set proved to be capable of modelling aspects of protein thermostability directly from sequence, but their use in the development of ML resulted in poor performances, only capable of differentiating between different organisms. This information is too general and can not be applied in a protein engineering approach.

The obtained *SeqVec* protein features could be used to accurately differentiate proteins originating from thermophile organisms from other organisms less resistant to high temperatures, suggesting that the *ELMo* model could identify the characteristics that make a protein resistant to high temperatures.

The ML models trained with the protein embeddings to predict the melting temperature of previously unseen proteins produced generally positive results with meaningful predictions when trained on the entire data set. However, the best performing MLP predictor was only slightly better than a simple baseline feature set that describes each protein by an amino acid frequency vector, and this model was also found to lack in generalizability. Since this experiment used a large enough data set, the reduced performance of the models can only be explained by the features used. It is noteworthy, however, that this experiment was based on the PCA-transformed protein embeddings, with only 100 dimensions. This was the only experiment performed in the thesis where enough data was available to discard this step. Additionally, since the algorithm chosen was the MLP, the feature reduction could have been performed by an initial layer of the perceptron. This was not studied, and could be somewhat responsible for the reduced performances obtained.

The issues with this model were not unexpected due to the difficulty of the task at hand, but with the surprisingly similar success of the baseline features we are forced to conclude that the *SeqVec* features do not provide a revolutionary approach for the development of new protein thermostability engineering methods that can predict the melting temperature directly from protein sequence. This is further evidenced by the high dispersion of the predictions around the average melting temperature of each organism, where simply predicting this value resulted in lower prediction errors.

This high dispersion was not studied in detail. By developing a model that uses the *SeqVec* embeddings together with the hand-crafted features that were here only used as baselines, more meaningful conclusions could have been drawn, perhaps showing that, with additional features, a useful ML prediction model could be developed with the *SeqVec* embeddings. But at the end of this experiment, with a model that predicts a general melting temperature value, randomly spread around the organism's average melting temperature, the delicate task that is protein design becomes impossible, and a more specific model that can accurately differentiate between very similar proteins, and not just separate the thermophile proteins, is still to be developed.

7. Conclusions

The objective of this thesis was to study the application of deep language models for protein sequence modelling. For this, the *SeqVec* model was used to produce protein features that were used to apply several ML models for the prediction of protein thermostability properties.

The first experiment with the *SeqVec* model produced successful protein secondary structure models, and in all experiments the protein embeddings showed a capacity to model aspects of protein thermostability, with an additional positive result in the capturing of three-dimensional conformation and amino acid interactions. The positive literature results on the application of deep language models for protein engineering, coupled with the advantages that these models have over conventional protein engineering methods, prove that this is an approach to biological sequence modelling with a lot of potential. It is, however, still behind state of the art performance, and would un-

doubtedly benefit from the preparation of larger and more well-curated databases of protein properties.

The results obtained in the development of ML models for protein thermostability prediction directly from wild-type sequences using the *SeqVec* protein embeddings revealed that such models are capable of identifying features related with thermostability, but are not yet adequate for the prediction of melting temperatures. The use of the *ProTherm* database confirmed its frequently mentioned issues, but not even with the larger *Meltome Atlas* could this prediction achieve performances useful for protein engineering. This is a difficult prediction task, with various factors influencing protein thermostability, that is not expected to be improved by introducing the use of transfer-learning with language models trained on protein sequences.

More frequently used in protein thermostability engineering, is the prediction of the effect that a mutation will have on the protein's free Gibbs energy of unfolding. Using the *SeqVec* features, we developed ML models that achieved higher Matthews correlation coefficients than some well established models. Our models also achieved useful precision metrics, proving that transfer-learning methods can compete with current literature. The application of these models to guide mutagenesis studies is expected to see an increase in use and a substantial increase in performance if more data is made available for their training.

8. Acknowledgements

An acknowledgement is made to the study supervisors, Prof. Dr. Marcel J. T. Reinders and Prof. Dr. Ana Luísa Nobre Fred, for the support and input, and also to Stavros Makrodimitris and the people at the Delft Bioinformatics Lab, without whom this work would not have been possible.

References

[1] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[2] S. Mazurenko, Z. Prokop, and J. Damborsky, "Machine learning in enzyme engineering," *ACS Catalysis*, vol. 10, no. 2, pp. 1210–1223, 2019.

[3] K. K. Yang, Z. Wu, and F. H. Arnold, "Machine-learning-guided directed evolution for protein engineering," *Nature methods*, vol. 16, no. 8, pp. 687–694, 2019.

[4] M. Musil, H. Konegger, J. Hon, D. Bednar, and J. Damborsky, "Computational design of stable and soluble biocatalysts," *ACS Catalysis*, vol. 9, no. 2, pp. 1033–1054, 2018.

[5] H. P. Modarres, M. Mofrad, and A. Sanati-Nezhad, "Protein thermostability engineering," *RSC advances*, vol. 6, no. 116, pp. 115252–115270, 2016.

[6] C.-W. Chen, M.-H. Lin, C.-C. Liao, H.-P. Chang, and Y.-W. Chu, "istable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules," *Computational and structural biotechnology journal*, 2020.

[7] U. Consortium, "Uniprot: a worldwide hub of protein knowledge," *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[8] A. S. Schwartz, G. J. Hannum, Z. R. Dwiel, M. E. Smoot, A. R. Grant, J. M. Knight, S. A. Becker, J. R. Eads, M. C. LaFave, H. Eavani, *et al.*, "Deep semantic protein representation for annotation, discovery, and engineering," *BioRxiv*, p. 365965, 2018.

[9] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC bioinformatics*, vol. 20, no. 1, p. 723, 2019.

[10] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," *Nature methods*, vol. 16, no. 12, pp. 1315–1322, 2019.

[11] A. Rives, S. Goyal, J. Meier, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *bioRxiv*, 2019.

[12] D. B. Duong, L. Gai, A. Uppunda, D. Le, E. Eskin, J. J. Li, and K.-W. Chang, "Annotating gene ontology terms for protein sequences with the transformer model," *bioRxiv*, 2020.

[13] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song, "Evaluating protein transfer learning with tape," in *Advances in Neural Information Processing Systems*, pp. 9689–9701, 2019.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 01 2000.

[16] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "Cd-hit suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[17] J. M. Dana, A. Gutmanas, N. Tyagi, G. Qi, C. O'Donovan, M. Martin, and S. Velankar, "SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins," *Nucleic Acids Research*, vol. 47, pp. D482–D489, 11 2018.

[18] M. S. Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, and A. Sarai, "Protherm and pronit: thermodynamic databases for proteins and protein–nucleic acid interactions," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D204–D206, 2006.

[19] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, and M. Rooman, "Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0," *Bioinformatics*, vol. 25, no. 19, pp. 2537–2543, 2009.

[20] Y. Chen, H. Lu, N. Zhang, Z. Zhu, S. Wang, and M. Li, "Premps: Predicting the effects of single mutations on protein stability," *bioRxiv*, 2020.

[21] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992.

[22] L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in *Proceedings of the 2003 SIAM international conference on data mining*, pp. 47–58, SIAM, 2003.

[23] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein structure prediction using rosetta," in *Methods in enzymology*, vol. 383, pp. 66–93, Elsevier, 2004.

[24] A. Jarzab, N. Kurzawa, T. Hopf, M. Moerch, J. Zecha, N. Leijten, Y. Bian, E. Musiol, M. Maschberger, G. Stoehr, *et al.*, "Meltome atlas—thermal proteome stability across the tree of life," *Nature methods*, vol. 17, no. 5, pp. 495–503, 2020.