GO prediction with Multiclass classification task

Kiana Seraj

November 2023

1 Introduction

Proteins perform majority of biological activities. It is, therefore, crucial to decipher the mechanism underlying their structural and functional properties. Like human language, protein sequences can be naturally represented as strings of letters. The protein alphabet consists of 20 common amino acids (AAs) (excluding unconventional and rare amino acids). While protein structure and function is dynamic and context-dependent (e.g. on cellular state, other molecules and PTMs), it is still defined by the underlying amino-acid sequence based on the central dogma of protein sciences, sequence $\rightarrow$ structure $\rightarrow$ function. Therefore, given these similarities, it seems natural to apply natural language processing (NLP) methods to protein sequences. Self-supervised learning enables AI systems to learn from vast amount of data which makes them to understand patterns of more subtle and less common representations. Additionally, *Fine-tuning* of *pre-trained* deep models have shown considerable promise for improving the model to solve the unseen problems, even on small data [13].

2 Language Models

Language models (LMs) have emerged as a useful *deep-learning* tool in analyzing protein sequences for further development of prediction methods of protein structures, features, and annotations from amino acid sequence information [8]. LMs use a *self-supervised* learning approach, where the training data is *unlabelled* and pseudo-labels are generated from the input unlabelled training data. This allows for predictions for a subset of the data, and is therefore useful when *pre-training* LMs or image representation such as contact maps.

Protein sequences can be treated, inherently, as linguistic and textual in nature. These sequences can be tokenized, much like a language can be described as a subset of letters or words; LLMs (Large Language Models), are able to tokenize amino-acids into individual amino-acids or grouped amino-acids, such as dimer or trimer, where more than one protein monomer is treated as a related structure. Tokenization is a fundamental step in Natural Language Processing (NLP) whereby a protein sequence is divided into "useful semantic units", and is followed by representing the protein sequence in either a local or distributed representation. Embedding of proteins are considered in the distributed representation [8].

In the context of NLP, the protein sequence can be encoded by deploying two different methods, namely: local representation and distributed representation [8] [12].

Local representation: Tokens may be one-hot encoded to vectors. For each token, as defined previously, an integer is assigned relating to its position in the sequence. The occurrence or presence of these defined tokens, is a binary representation. The information is thus treated as discrete and no overlap between tokens is considered.

Distributed representation: Tokens may also be represented in a continuous manner by removing the concept of one semantic unit. Therefore, each element in a vector will correspond to the whole representation and not a specific attribute or unit. This continuity introduces possibility to analyze related concepts between what was previously discrete and unrelated.

Word embedding techniques are some of the most popular distributed representations in NLP [8]. The resulting vectors, can be related to one another, with similarities in tokens resulting in similarities in their vector representations [16]. The continuous vector representation allows for analysis of the relationship between words, technically, vectors. The resulting vector representation conserves information of semantic similarity between vectors [8].

3 ESM-2

An evolutionary-scale model, a family of *transformer* protein language models, directly infers full atomic-level protein structure from primary sequence. The biological properties of a protein constrain the mutation to its sequences that are selected through evolution, recording biology into evolutionary patterns. Additionally, ESM has the potential to learn patterns in protein sequences across evolution and predict the *protein contact map* only through protein sequences. Therefore, motivates research on evolutionary-scale language models. ESM-2 is scaled up to more parameters, resulting in an acceleration of high-resolution structure prediction, this speed advantage makes the model to expand the structure prediction to metagenomic scale dataset [10].

3.1 ESM-2 vs ESM-1b

ESM-2 relative to previous generation model, ESM-1b, introduces improvements in architecture, training parameters, and increases computational resources and data. From 650 M parameters it is trained up to 15 billion parameters. The resulting ESM-2 outperforms previously ESM-1b. Also, on structure prediction benchmarks, it outperforms other protein language models [10]. The accuracy of the contact map prediction and perplexity are linked. Proteins undergoing large changes in the contact map accuracy, also undergo large changes in the *perplexity*, Fig 1.

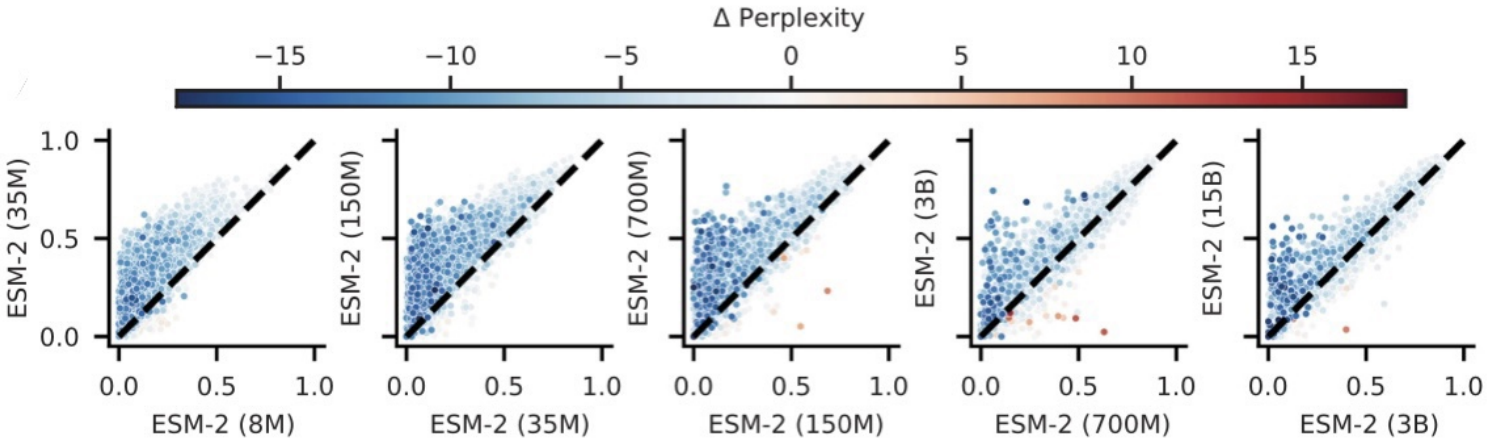

Figure 1: Comparing models from 8M to 15B parameters through unsupervised contact precision. Sequences with large changes in contact prediction performance also exhibit large changes in language model understanding measured by perplexity.

3.2 ESM-2 training

ESM-2 is trained *unsupervisedly* to predict across evolution, the identity of amino acids that have been randomly masked in protein sequences, including 15 % of positions in the sequence. The model is tasked with predicting the identity of the masked amino acids from the surrounding context. This method of masked language modelling causes the model to learn dependencies between amino acids. As biological structure is linked to the sequence patterns, it is expected that language model captures this biological structure through training over the masked amino acids in the protein sequence. Training on masked language modelling are known to develop attention patterns that corresponds to the residue-residue contact map of the protein.

ESM-2 is trained over sequences in UniRef, the model sees around 65 M unique sequences. Increased parameters, cause the model to demonstrate large improvements in the fidelity of its modeling of protein sequences. This fidelity can be measured by using perplexity , ranging from 1 for a perfect model to 20 for a model making prediction randomly. The 8 M parameter model has the perplexity of 10.45, and the 15 billion model reaches a perplexity of 6.37. ESM-2 is trained only on sequences, therefore any information about the structure develops must be the result of representing the pattern in sequence which eliminates the costly aspect of the structure prediction pipeline, using a *multiple sequence alignment* while greatly simplifying the neural architecture for inference. Resulting in the speed of up to 60 times on the inference forward pass and also removing the search for related proteins entirely [10].

3.3 Atomic-resolution structure prediction by EsmFold

EsmFold is developed by training a folding head for ESM-2, the sequence of proteins are inputted to ESM-2, then, the sequence is processed through the *feedforward* layers of the language model, then the models representations are passed to the folding head. The output of the folding blocks is passed to a transformer structure module, and three steps of recycling is performed before the representing of the final output Fig 2. ESMFold has a simplified architecture compared to AlphaFold, which deeply integrate the MSA into neural network architecture through an attention mechanism that operates across the rows and columns of the MSA.
The ESMFold approach results in a considerable improvement in prediction speed relative to AlphaFold. However, in structure prediction, AlphFold reaches an average *TM-score* more on CAMEO and CASP14.
Models based on the transformer and implementing self-attention have been shown to be efficient and successful in a wide range of tasks, AlphaFold2, in contrast, has a multi-block structure similar to the transformer but still addresses the problem differently (treating proteins as graphs and using multi-alignment sequences). In addition, the training of AlphaFold2 is much more complex than a transformer, meaning that with transformers we can scale structure prediction to much larger databases [10].

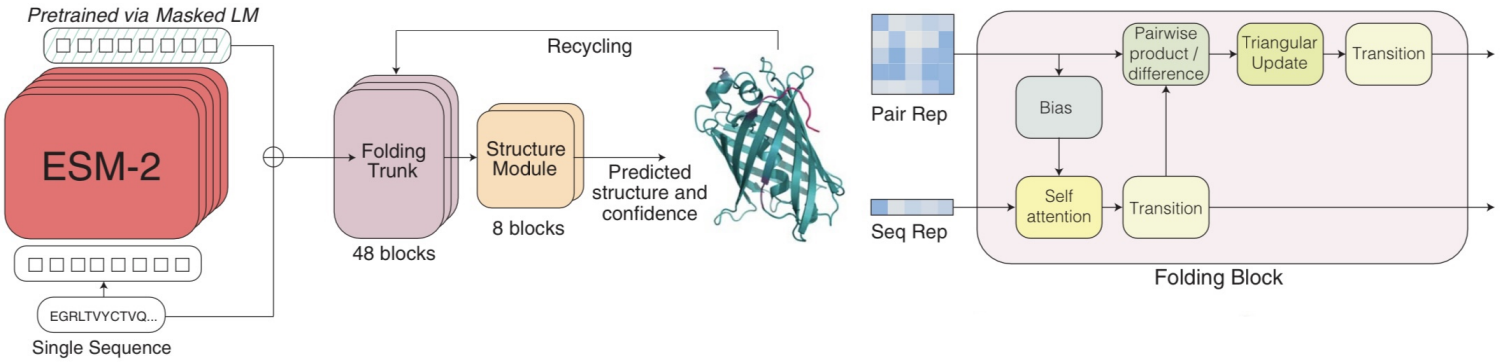

Figure 2: ESMFold model architecture

3.4 Structural characterization of metagenomics with ESM-2

The fast and high-resolution structure prediction by ESM-2 enables structure prediction for more than 617 M metagenomic protein sequences from MGnify90 database, including more than 225 M that are predicted with high confidence. This understanding increases by showing the right model enough protein sequences. As the representational capacity of the language model and the diversity of protein sequences seen in its training increase, we expect that deep information about the biological properties of the protein sequences could emerge, since those properties give rise to the patterns that are observed in the sequences. [10]

3.5 ESM Application

a.ADOPT:
Intrinsic disorder proteins possess no well-defined 3-D structure but rather adopt an ensemble of conformations in solution, yet they are functional [5]. ADOPT, a structural disorder predictor, benefits from the ESM transformer architecture and analysis of residue-level representations is used in the development of ADOPT. Hence, protein disorder is accurately predicted from sequence alone. ADOPT predicts whether a protein or a specific region is disordered with better performance than other predictors and faster than other methods. [15].
ADOPT Architecture
Composed of two blocks: first, The ESM *encoder*, which takes a protein input sequence and uses information from a large database of sequences to generate feature information for every residue in the sequence. Second, the *supervised* ML-based predictor, Lasso regression model(for binary case Logistic regression is used). The embedding vector of each protein sequence will be given to the predictor as input and it predicts the level of disorder of each residue, given in terms of Z-score. [15].

b.Zero-shot prediction of the effects of sequence variation on protein function:
Zero-shot learning or meta-learning in the context of language models means that the model develops a broad set of skills

and pattern recognition abilities at training time, and then uses those abilities at inference time to rapidly adapt to or recognize the desired task. The model does not need any fine tuning or gradient updates [2].

Evolution encodes information about protein functions into patterns in protein sequences, and forms a landscape that reveals how function constrains sequences. ESM, captures this pattern through masked language training and becomes able to score sequence variations with zero-shot learning. The AlphaFold's transformer module, named Evoformer, does not work at all in the zero-shot mutation effect prediction task [11].

4 Evolution-aware and Evolution-free Protein Language Models

The key question in this part is that whether protein language models trained on MSAs can be as good as sequence-based models in function prediction tasks or not.
AlphaFold, trained on experimental 3D protein structures from the Protein Data Bank (PDB) can approach the resolution of experimental structures for most protein sequences. Its multiple sequence alignment representation module, Evoformer, like MSA-Transformer, takes a family of evolutionary-related and aligned protein sequences as input [7].
In contrast, ESM takes individual protein sequences. Thus, in here former model is referred to as evolution-aware PLM and the latter as evolution-free PLM [6].
Evolution-aware PLMs are superior to evolution-free ESM-1b model only in the structure prediction tasks, but in general, are worse than ESM-1b in most function prediction tasks. A better structure protein language model does not have a better representation for predicting function.
Only in stability prediction, a fitness prediction task, Evoformer is better than ESm-1b, as it has a closer relationship to protein structure [6].

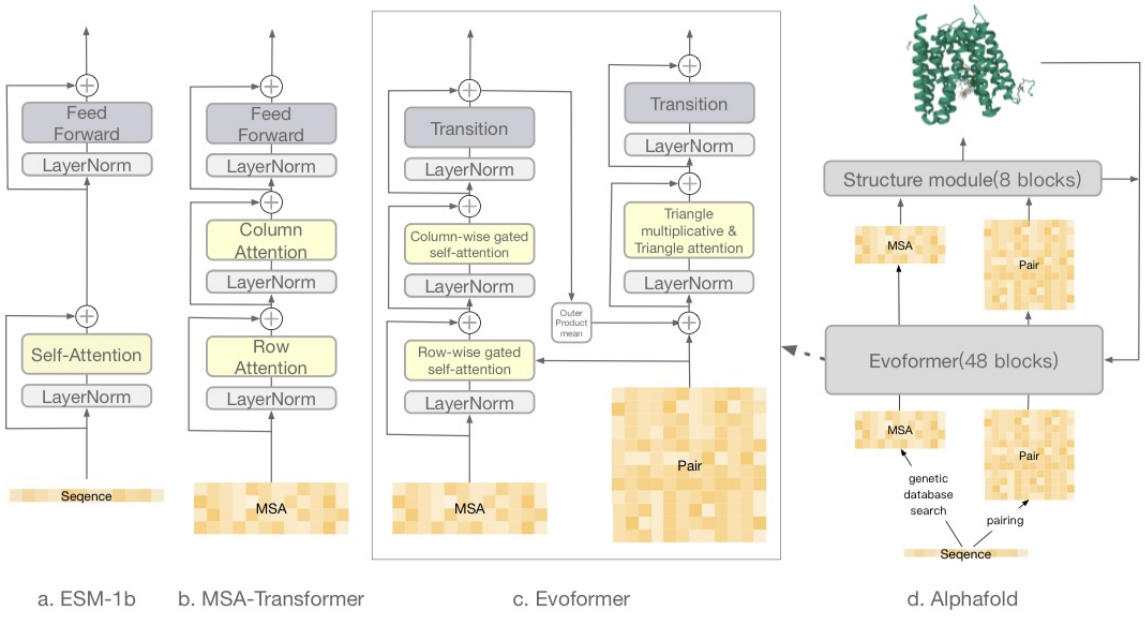

Figure 3: Core modules of three PLMs

5 DeepFRI

Protein structure and function predictions are a key area of research, which has and is progressing rapidly. With the massive influx of protein sequence databases and datasets, ML (Machine Learning) techniques serve to leverage the available data to predict protein functions. With the goal of predicting protein functions, DeepFRI utilizes a Graph Convolution Network. The protein structures as well as a language model are used to 'leverage sequence features' [8].

A *homology model* predicts 3-D representation of the structure of a protein molecule with the use of a known a protein molecule(s). This allows for a deeper and more diverse training dataset, as the experimental structural data of these newly introduced structures is unavailable and generated computationally.

The inclusion of homology models in the training dataset serves to increase the number of training inputs per function, and subsequently increase the number and learning ability of available predicted functions. The increase in training samples serves to increase the performance and accuracy as per the author's: the F_{max} shows an increase from 0.455 to 0.545 with the inclusion of the homology models.

Class activation mapping allows function predictions at an unprecedented resolution, allowing site-specific annotations at the residue-level in an automated manner

Gradient-weighted Class Activation Maps (grad-CAMs) provide further insight and precision within the scale of a protein residue. The relevant protein residues are learned and used in predicting the 'Molecular Function branch of the GO hierarchy' [4] (Discussed further in this report). Effectively, the CAMs aid in increasing the resolution of the model in predicting functionality from protein-level [4] to residue-level, allowing for identification of functional sites.

The folding of a protein is responsible for a 'wide variety of functions within the cell' [4]. Both disordered and ordered regions may be functional, however the ordered 3D conformations form majority of the domain [4] [17]. The importance of protein's structural features range from 'binding specificity and conferring mechanical stability, to catalysis of biochemical reactions, transport, and signal transduction' [4].

5.1 Gene Ontology (GO) and Enzyme Commission (EC) Numbers

Gene Ontology is a comprehensive functional classification and is hierarchically composed of three ontologies [8]:

- MF (Molecular Function)
- BP (Biological Processes)
- CC (Cellular Component)

As discussed in the context of NPL, classifying the structure and functionality of the influx of protein sequences proves to be a challenging task and very useful in the context of understanding proteins and their potential uses and applications.

According to the authors: "Understanding the functional roles and studying the mechanisms of newly discovered proteins is one of the most important biological problems in the post-genomic era" [4].

The DeepFRI model proves to be a method applicable to hundreds of thousands of protein sequences of unknown origin [4].

They have enabled task-specific feature extraction directly from protein sequence (or the corresponding 3D structure), overcoming the limitations of standard feature-based machine learning (ML) methods [4].

5.2 Network Schematic

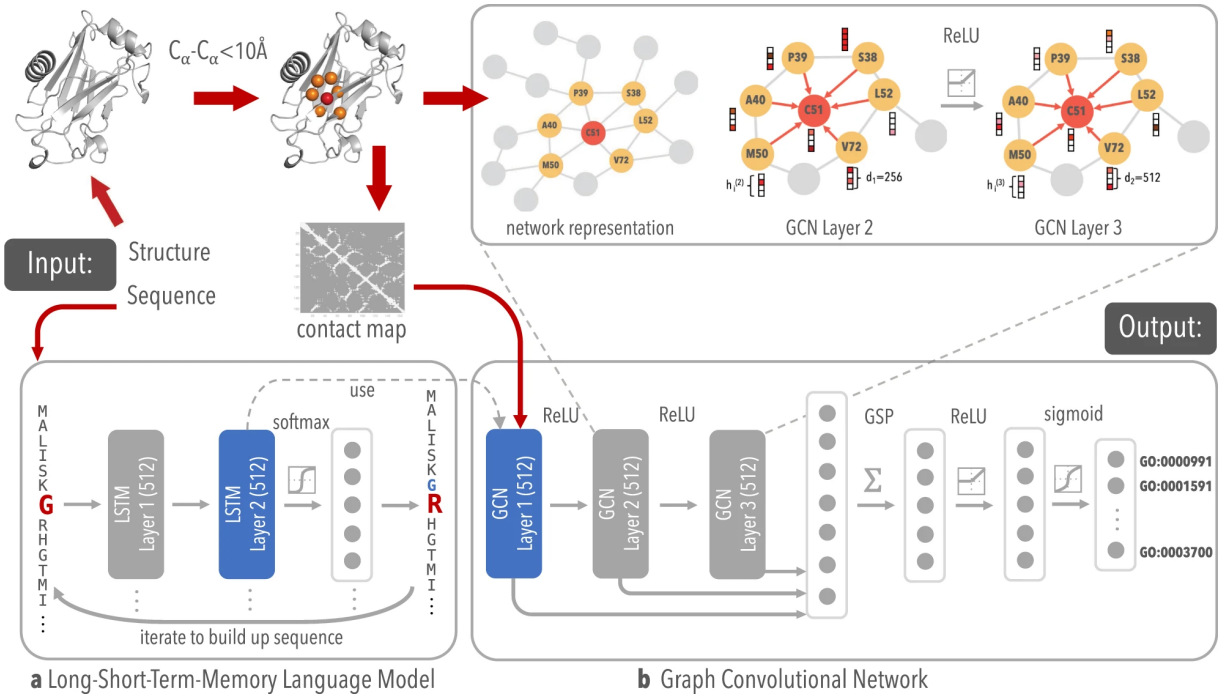

Figure 4: Schematic overview of the DeepFRI method

"DeepFRI's model uses a two-stage architecture, inputting a protein structure and a sequence representation from a pre-trained, task-agnostic language model, represented as graphs derived from amino acid interactions in the 3D structure" [4].

The first stage consists of a language model (self-supervised) with *Recurrent Neural Network (RNN)* architecture with *long short-term memory (LSTM-LM)* [8]. The LSTM is useful in this context to model long term dependencies, as discussed in 6.6. The choice of language model allows for extraction of residue-level features directly from only protein sequences.

The second stage requires the input residue-level features from the LSTM, as well as, the contact map for the protein. This stage consists of a GCN, that combines a deep architecture with a network representation of the protein. The contact map input allows for this network representation of the protein.

Different models are trained for predicting different GO terms. We choose to focus on the Molecular Function of the GO terms.

The network outputs the probability of each function (MF-GO term) being activated.

The authors make an argument for the inclusion of contact maps alongside LSTM-LM features, by comparing the Precision-Recall curves, as seen in Figure 5. DeepFRI outperforms a baseline model in which contact maps are used without an LSTM. Furthermore, DeepFRI outperforms a baseline model that only makes use of LM features without a contact map. This supports the authors' claims that the inclusion of both the LSTM and contact maps improve the predictive functional performance of such a model [4].

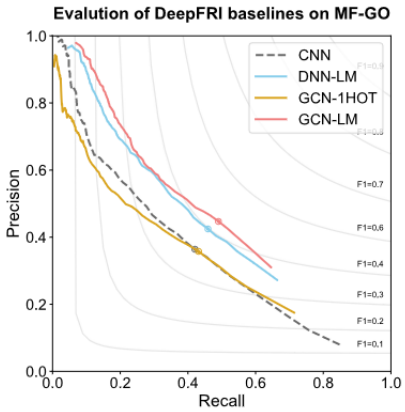

Figure 5: Precision-Recall curves for different methods; CNN: sequence only. DNN-LM: Language model only. GCN-LM: DeepFRI architecture where sequence and contact maps are used

5.3 DeepFRI Method

Method followed in constructing the network for predicting the GO terms, specifically:

5.3.1 Preparation of data

Atomic coordinates (3-D) are collected from the PDB. Contact maps are constructed, and protein residues are considered to be in contact based on a governing rule, such as the distance between corresponding C_α atoms [4].

GO terms are used as functional labels for training the models. Protein sequences collected are split into training. validation, and testing in an 80%, 10%, 10% ratio, respectively.

5.3.2 LSTM Language Model (LSTM-LM)

The LSTM-LM is trained to predict an amino-acid residue in the context of its position is a sequence [4]. The language model is pretrained on $\sim$ 10M sequences from Pfam, and serves as a fully unsupervised and allows for the 2nd LSTM layer to be used as feature extractor.

The LSTM layer serves to map protein sequences, to sequences of vector representations, allowing similar residue features to be related in the vector embedding space. For each residue, a 512-dimensional embedding feature vector, this is used to pass to the GCN.

5.3.3 GCN

The GCN receives as input the 512-dimensional feature vector, as well as the constructed contact map. The GCN has 3 layers, and captures convolutions over 3rd order neighbourhoods of residues, corresponding to features between neighbouring residues. Moreover, the graph-based structure from the contact maps, serves to be important. The use of a GCN is powerful where the data is represented as one or more graphs [4].

The GCN layers output a residue-level feature vector, and then a fixed length feature vector is constructed by means of Generalized Sum Pooling (GSP), as well as activation functions to determine which GO term should be activated.

5.3.4 Training & Validation

A temporal holdout validation is used on the trained model. GO annotations from two different time points are compared (2019 and 2020). There exist examples where DeepFRI outperformed other methods and predicted GO terms significantly more accurately.

6 Transformer + DeepNN Model: TransformerFRI

6.1 Network Schematic

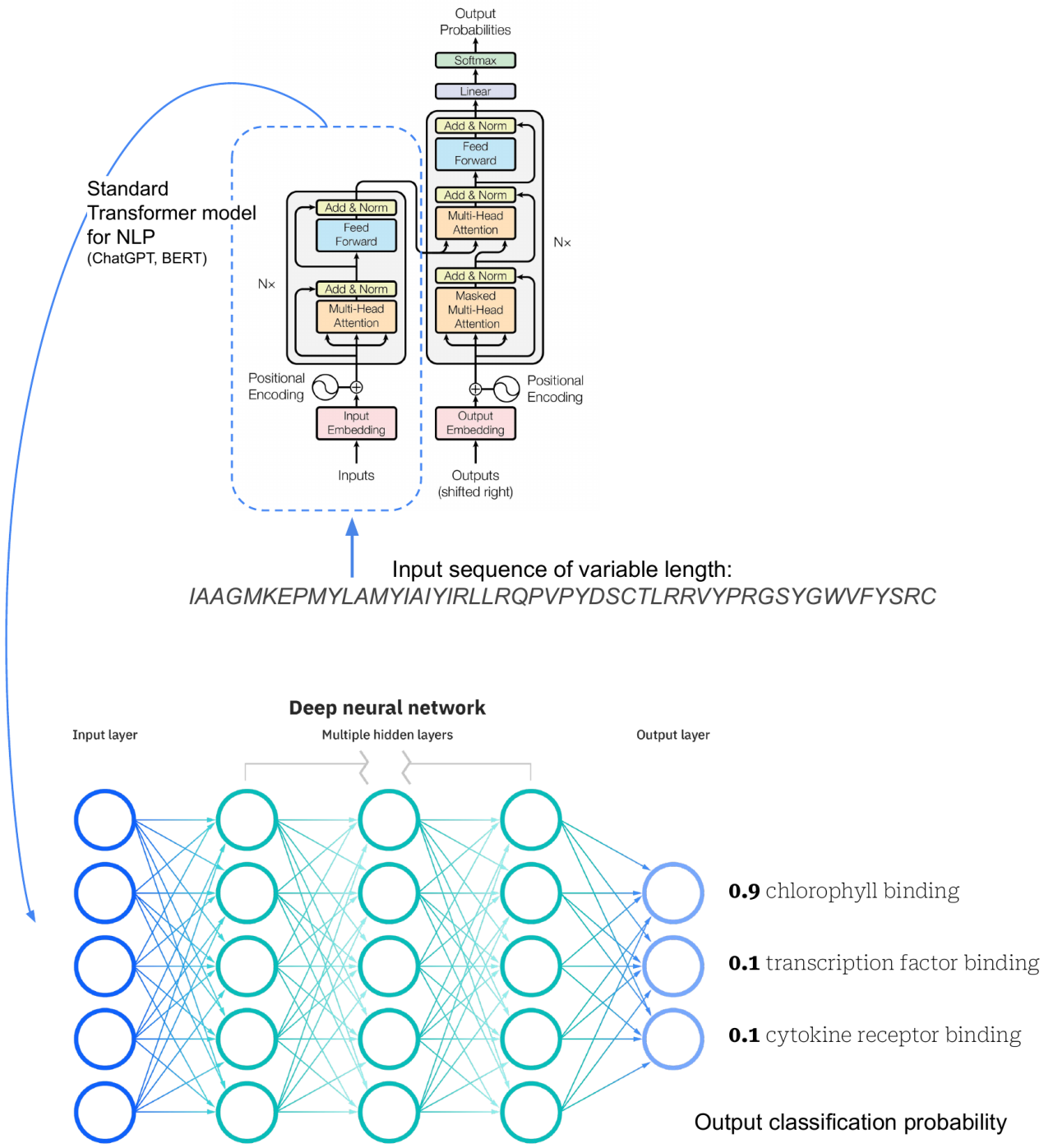

Figure 6: Schematic overview of our method: TransformerFRI

Our model consists of a Transformer, of the same architecture used in the likes of BERT and ChatGPT. In order to take a protein sequence of varying length, the encoding by the Transformer allows for a protein sequence to be encoded into a dense vector representation of fixed length. This approach to our model allows for the data (protein sequences), to be preprocessed and prepared for use in a Deep Neural Network (DNN). Subsequently, the DNN outputs a classification probability of a protein's affinity to bind to the chosen binding sites. Based on the data available (subset of that used in DeepFRI), we have chosen the following 4 activations (binding sites): oxidoreductase activity acting on diphenols,

chlorphyll binding, transcription factor binding, and cytokine receptor binding. The choice was based on the completeness in the data available, most prevalent binding sites were chosen as training data, where only one of the 4 bindings occur.

6.2 TransformerFRI

This outline explains the theory behind our architecture choice, as well as explaining the key differences between our approach, TransformerFRI, compared to the reference paper, DeepFRI.

6.3 DeepFRI

Subsequently, the architecture of the DeepFRI model is considered.

In the original paper, the authors define an architecture that requires two separate sets of input data: the sequence of the protein, and the contact map. These two inputs are then fed into a Graph Convolutional Neural (GCN) network, from which the function is predicted.

6.4 Protein sequence

Protein sequences have proven to be inherently difficult to model, due to having characteristic varying lengths. Protein sequences can have lengths of varying amino-acids, generally 50-2000 amino-acids long [1].

These sequences may be treated like human language, with the equivalent of a sentence consisting of words, so too a sequence containing amino-acids; analogous to words in a sentence, determined by our tokenization choice.

6.5 Contact Maps

When attempting to recreate the experiments in the original paper, acquiring the contact maps proves to be a computationally heavy operation. For reference, using a paid, high-CPU instance comprising 12 threads, resulted in only around 10000 contact maps being acquired (orders of magnitude less than DeepFRI). Furthermore, these contact maps are varying in dimensionality, and therefore need niche and targeted modelling to be able to preprocess the data efficiently and thoroughly, and prepare this data for a network with fixed input length.

6.6 Language Models, a history

In classical machine learning, Recurrent Neural Networks (RNNs) were used to model language (or any data) of varying lengths. The models are similar to a traditional neural network, but instead of a single input, the model learns sequences of inputs that are fed into the model sequentially. At the end of the sequence, a special 'reset' token is sent to the model, such that the model knows that the sequence is complete.

This approach was a big revelation at the time, and was the birth of language models as we know them today. The big takeaway from these models is their ability to take in binary-encoded sequences of varying lengths, and to normalise this input into a dense vector of defined and fixed dimensionality.

The downside of this approach in the context of analyzing protein sequences in a meaningful manner, is an RNNs propensity for vanishing gradients. When the input data consists of long sequences, gradients can become infinitesimally small as they are backpropagated through time, resulting in an inability for the network to 'learn'. The network therefore, fails to retain or learn long-term dependencies. In the above context the following terms are of importance:

- **vanishing gradient and backpropagation**: An iterative approach of adjusting parameters using the derivative of the loss function. Considering a long sequence input (such as a protein sequence), gradients are propagated backwards through many time-steps (for each discrete token in a sequence), and therefore, by the chain-rule, the gradient may approach zero, rendering the network incapable of learning.
- **long-term dependency**: The relationship between distant input data elements, in our case: distant amino-acids, monomers (or n-mers, depending on our tokenization).

RNNs do have a number of downsides, namely their tendency to lose information in very long sequences (vanishing gradients during back propagation of the neural networks, as discussed above) and, their general bias towards more

recent words of the sequence compared to older parts; and, the difficulty in training them, both from a hyper parameter perspective but also from a hardware perspective: the fact that a single input (i.e. a sentence) requires multiple forward and (as well as backward passes during training) during inference necessitates more hardware resources, and generally slower inference time.

Long Short-Term Memory models (LSTMs) address the concern of vanishing gradients that the RNN poses, by including a memory component that is better able to learn long sequences, as its name suggests. This also better models long-term dependencies, but structurally still suffers from the architecture issues that RNNs have, where one sequence requires multiple passes through the model.

Language Models (LMs) have become popular in recent times due to the adaptation of the Transformer architecture, which addresses many of the issues presented by classical approaches. Namely, it gets the benefits of the LSTMs' ability to model long term dependencies, but also addresses the issue of needing multiple passes through the model to predict or train on a complete sequence.

The Transformer does this by having a fixed, very large input dimensionality, where sequences smaller than this defined dimensionality are padded with 0s. The Transformer has two mechanisms that are fundamental in its operation, Positional Encoding, which allows the model to learn where the beginning and end of a sequence are; and, Multi-Head Attention, which is similar to other developments in attention (for example, attention layers have also been applied to recurrent models), but differs slightly in that there are parallel attention matrices across your input that are concatenated and linearly transformed into the final dense vector representation.

6.7 Why TransformerFRI

In other literature, there has been work done to generate a contact map directly from a protein sequence. While the authors acknowledge this work, they still independently process the protein sequence using an LSTM, and then still incorporate the contact map as an additional input.

Since the protein sequence does have signal in it (enough to generate a contact map in other literature), an interesting approach would be using the protein sequence as the only input to such a model.

In order to further push this development, we try a Transformer architecture (only the first half, the encoder), paired with a standard neural network acting as a classifier on top of the encoder. Our model output is a classification probability, as opposed to most LMs which would predict another sequence. As such, we do not need the decoder portion of the Transformer (the Transformer encodes a sequences into a dense space, and decodes this dense space back into a new sequence; we only need the former portion).

Due to non-access to supercomputer-level hardware, we test our hypothesis on a subset of the available data. We limit the data by instances where we have single-value classifications across functions (i.e. a sequence results in only a single activation being the output, not multiple). We also only isolate activations where we have enough data, at least 100 samples.

6.8 Results of TransformerFRI

From our results, we can see that the model does indeed work.
The confusion matrix suggests the following accuracy per predicted label:

Binding site	Accuracy	Number of sequences used in training
oxidoreductase activity acting on diphenols	72.41%	683
chlorophyll binding	42.86%	260
transcription factor binding	92.86%	273
cytokine receptor binding	79.31%	318

Our results suggest further investigation into our model could yield interesting and useful results, in determining probable binding sites of various proteins, solely inputting their sequence.

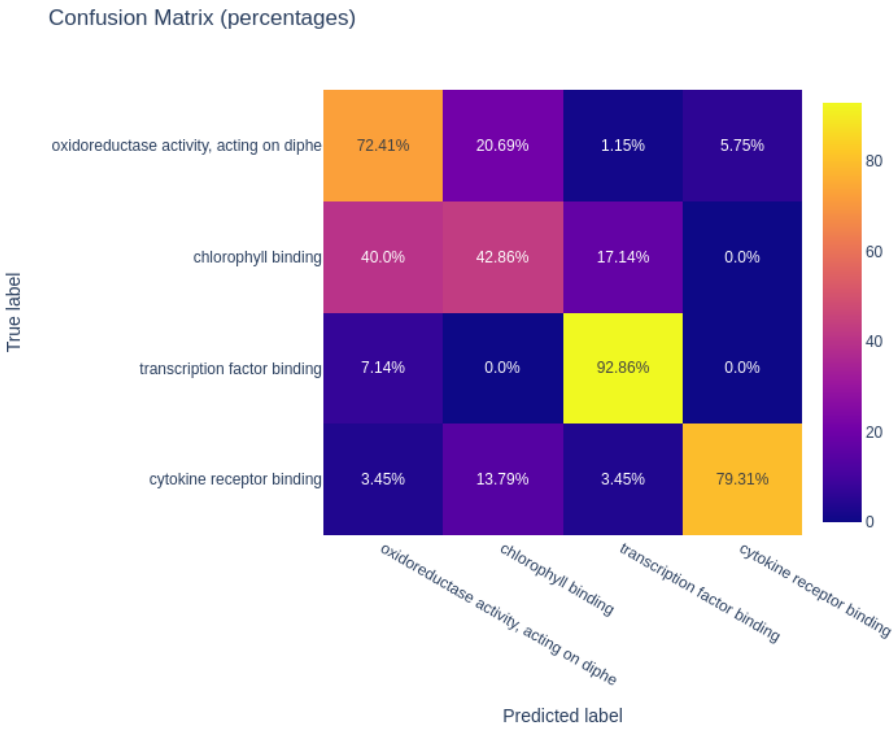

Figure 7: Confusion Matrix for TransformerFRI

7 Conclusions

Deep learning and NLP methods are making inroads into protein research and the trend in NLP is towards deeper and larger language models as GPT3 and beyond. The emergence of atmoic-level structure in protein language models shows a high-resolution picture of protein structure encoded by evolution into protein sequences that can be captured with unsupervised learning.

DeepFRI [4] proves to be a useful tool in identifying protein functions, with the use of a protein sequence along with its associated contact map.

Our method, TransformerFRI, makes use of a computationally less taxing approach of predicting the Molecular Function (MF) GO term, with solely an input of the protein sequence.

Despite the limited scale to which we were able to compute, our results suggest a lightweight method to determine the Molecular Function of a protein is worth further interest and research. Such a development can prove useful in fields such as pharmacology, toxicology, biological signalling, immunity, etc.

8 Glossary

Attention: A model's ability to determine how relevant tokens are relative to other tokens in the same input when creating an output. This captures contextual relationships in an input. An attention weight can be assigned to each token, allowing the model to determine which tokens require more or less "attention" than others.

DNN (Deep Neural Network): A Neural Network consists of an input later, initial data, and subsequent hidden layers. These layers consists multiple "neurons" (a neuron is simply a fundamental computational unit and is equivalent to a node). The network utilizes labelled data to train the relationship between nodes in adjacent layers - known as weights and biases. Moreover, activation functions add non-linearity to the network, allowing for detailed and problem-specific activation conditions. Without activation functions, the network may as well have one hidden layer, and the linearity would simply suggest a Linear Regression model. We consider a network to be a DNN when there are many hidden layers.

Encoder: The encoder's first sublayer carries out multi-head self-attention. The self-attention is applied multiple times (multi-head), using different weight matrices. Multi-head attention occurs in parallel, and can capture a wider range of relationships between tokens of the same input sequence compared with single-head, where only one weight matrix is used. The encoder then outputs a dense vector representation of fixed length.
Discussed in 6.6.

Feed-forward neural network: An artificial neural network in which the connections between nodes does not form a cycle. It is the simplest form of neural network as information is only processed in one direction. While the data may pass through multiple hidden nodes, it always moves in one direction and never backwards [14].

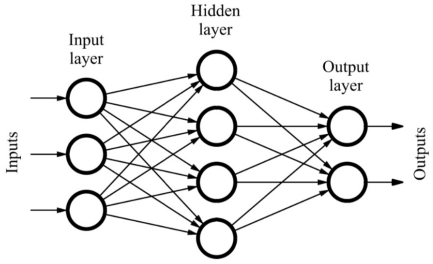

Figure 8: Sample of a feed-forward neural network

Fine-tuning: In deep learning, fine-tuning is an approach to transfer learning in which the weights of a pre-trained model are trained on new data. Fine-tuning can be done on the entire neural network, or on only a subset of its layers, in which case the layers that are not being fine-tuned are "frozen" (not updated during the backpropagation step) [18].

Homology model: Also known as a comparative model or homology modeling, is a computational method used in structural biology and bioinformatics to predict the three-dimensional structure of a protein or biomolecule based on its similarity to the structure of a related, experimentally determined reference molecule. This approach relies on the principle that evolutionarily related proteins or molecules share similar structural characteristics and functions.

Labeled data:A dataset in which each data point (e.g., text, image, or any other input) is accompanied by specific, assigned annotations. These labels indicate the desired or correct output or category associated with each data point. Labeled data is essential for supervised machine learning tasks, as it serves as the training and evaluation dataset for models. It allows algorithms to learn and make predictions based on the provided labels, ensuring that they can generalize and make accurate predictions on new, unseen data.

Long Short-Term Memory:A recurrent neural network (RNN), overcomes traditional RNNs shortage in capturing long-term dependencies, ideal for sequence prediction tasks.

Multiple sequence alignment:A bioinformatics technique used to compare and analyze multiple biological sequences, such as DNA, RNA, or protein sequences. The primary objective of MSA is to identify regions of similarity and difference among these sequences, enabling researchers to infer evolutionary relationships, identify conserved motifs, and

study structural and functional properties.

Perplexity: One of the metrics for evaluating language models, measuring a language model's uncertainty of a sequence and is defined as the exponential of the negative log-likelihood of the sequence. The metric applies specifically to classical language models (sometimes called autoregressive or causal language models) and is not well defined for masked language models like BERT. Therefore, in ESM model evaluation, exponential of negative pseudo-log-likelihood of a sequence is used. For brevity, this estimate is referred to as the "perplexity," as it can be interpreted in a similar manner [9].

$$\text{PseudoPerplexity}(x) = \exp\left\{-\frac{1}{L}\sum_{i=1}^{L}\log P(x_i|x_{(j\neq i)})\right\} \tag{1}$$

Pre-training:In deep learning, during pretraining, a model is trained on a large and general dataset, typically without any task-specific objectives. This phase allows the model to learn foundational features, representations, and patterns from the data, which can be leveraged in subsequent fine-tuning for specific tasks. Pretraining is commonly used in transfer learning and is a crucial step in developing models that can adapt to various tasks and domains.

Protein Contact Map: Protein contact map is a two-dimensional representation of the three-dimensional layout of protein structure. The construction of the contact map is as follows:
1. C_α atom of each amino acid is considered as vertices of the protein contact network.
2. The distances between every pair of residues are determined using Euclidean distance.
3. To determine whether any two residues are connected, the distance between the residues should be less than or equal to a cut-off value (5 Å, 7 Å, 8.5 Å).
4.If any two residues are connected, then the matrix cell values are set to 1 or else 0 if they are not connected [3].

Recurrent Neural Network: The opposite of a feed forward neural network is a recurrent neural network, in which certain pathways are cycled [14]. Simple RNNs are commonly trained through backpropagation, in which they may experience either a 'vanishing' or 'exploding' gradient problem, limiting effectiveness in applications that require the network to learn long-term relationships.

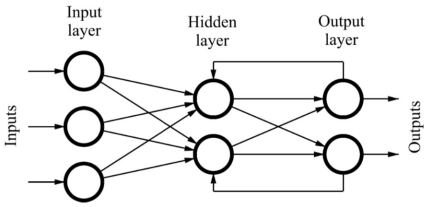

Figure 9: Sample of a recurrent neural network

Self-attention: A specific type of attention where the model weighs the importance of each element within the same input. The output is therefore the same as the input, the model simply learns the importance of different elements within the same sequence input.

Self-supervised learning: A machine learning paradigm where an algorithm learns from the data itself without requiring external labels or annotations. It gains the ability to capture rich, context-aware representations by performing tasks such as predicting missing parts of a sequence or the next word in a sentence.

Supervised learning: A type of machine learning in which an algorithm is trained on a labeled dataset. The goal of supervised learning is to generalize from the training data to make accurate predictions or classifications on new, unseen data. Common applications of supervised learning include image classification, speech recognition, and regression tasks.

TM-score: The template modeling score or TM-score is a scoring function for the automated evaluation of topological similarity of protein structures [19]. TM-score has the value in (0,1], where 1 indicates a perfect match between two structures. Following strict statistics of structures in the PDB, scores below 0.17 correspond to randomly chosen unrelated proteins whereas structures with a score higher than 0.5 assume generally the same fold in SCOP/CATH.

Transformer (decoder is not discussed): An architecture of a deep learning model, that has proved very useful in NPL. A transformer can be thought of as encoder-decoder model. The decoder is not used in our model.

Unlabeled data:In contrast to labeled data, consists of raw input data without associated labels. Unlabeled data can be valuable for unsupervised learning, which discovers hidden patterns, structures, or relationships within the data without relying on labeled examples.

Unsupervised learning:A machine learning approach where an algorithm is trained on data without explicit supervision or labeled outputs. Instead, the algorithm seeks to discover underlying patterns, structures, or relationships within the data. Common techniques in unsupervised learning include clustering, dimensionality reduction, and generative modeling.

References

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. The shape and structure of proteins. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Isaac Arnold Emerson and Arumugam Amala. Protein contact maps: a binary depiction of protein 3d structures. *Physica A: Statistical Mechanics and its Applications*, 465:782–791, 2017.

[4] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

[5] Johnny Habchi, Peter Tompa, Sonia Longhi, and Vladimir N Uversky. Introducing protein intrinsic disorder. *Chemical reviews*, 114(13):6561–6588, 2014.

[6] Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 2022.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[8] Lukasz Kurgan. *Machine Learning in Bioinformatics of Protein Sequences: Algorithms, Databases and Resources for Modern Protein Bioinformatics*. World Scientific, 2022.

[9] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[11] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[13] Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021.

[14] Ramon Quiza and J Paulo Davim. Computational methods and optimization. *Machining of hard materials*, pages 177–208, 2011.

[15] Istvan Redl, Carlo Fisicaro, Oliver Dutton, Falk Hoffmann, Louie Henderson, Benjamin MJ Owens, Matthew Heberling, Emanuele Paci, and Kamil Tamiola. Adopt: intrinsic protein disorder prediction through deep bidirectional transformers. *NAR Genomics and Bioinformatics*, 5(2):lqad041, 2023.

[16] Nguyen-Quoc Trinh-Trung-Duong Nguyen, Khanh Le, Quang-Thai Ho, Dinh-Van Phan, and Yu-Yen Ou. Using word embedding technique to efficiently represent protein sequences for identifying substrate specificities of transporters. *Analytical Biochemistry*, 577:73–81, 2019.

[17] Christine Vogel, Carlo Berzuini, Matthew Bashton, Julian Gough, and Sarah A Teichmann. Supra-domains: evolutionary units larger than single protein domains. *Journal of molecular biology*, 336(3):809–823, 2004.

[18] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

[19] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.