# Protein Function Prediction

Chan Cheuk Ka, Cheung Ho Lun

Proteins are the building blocks of all organisms. They play a vital role in all biological processes. The structural features and 3D shape of proteins closely determine their functions. Knowing the functions of these proteins is useful for drug development because it allows the design of specific drugs to treat patients. As different people will respond differently to drugs due to genetic differences, exploring the function of different proteins can help realise personalised medicine by giving specific treatment.

There are several challenges in predicting the functions of proteins. Firstly, sequences with minor differences can have drastically different functions, and vice versa. Secondly, some proteins do not have easily identified features for some types of functions. Lastly, many proteins can have more than one function, causing the prediction to be more complicated. These factors limit the accuracy of the prediction; only less than 1% of known protein sequences have determined functions [*]. It is also challenging to identify the function of rarer proteins because they are under-represented in research [†].

Bioinformatics is an excellent tool for solving the aforementioned problems. It uses computer science to handle large amounts of bio-related information. For instance, motif-based methods can predict protein binding sites and deep learning algorithms can train on existing data and predict unseen protein features. Gene ontology (GO) provides a structured vocabulary for describing gene and protein functions.

Deep learning can lead to faster, more scalable, and more accurate predictions, and there is a clear trend in deep learning methods outperforming sequence-based methods. In this report, we will explore four novel methods of protein function prediction.

## 1 DEEPred [‡]

### 1.1 Architecture

The architecture of DEEPred is a stack of 1000 simple feed-forward multi-task deep neural networks with a preprocessed feature vector as input and the confidence for five GO terms as output. DEEPred consists of a stack of 1000 multi-task DNNs, each responsible for five GO terms. They opted for the DNNs only to predict five terms because they found that predicting for more gave a worse performance while predicting for fewer terms diminished the utility of using a multi-task network. To determine which five GO terms each DNN should predict, the authors divided them into ten different levels based on their broadness, where each DNN can only predict terms within the same broadness level. They used this approach to mitigate output bias towards more common terms.

To extract the most important features from the sequence to use as the actual input for the model, they incorporated a preprocessing step, where feature vectors are created. They compared the efficacy of three different preprocessing methods and ultimately chose sub-sequence profile mapping (SPMap), where the sequence is split into sub-sequences before hierarchical clustering. SPMap was used to generate the feature vectors for all inputs.

[*] S. Fujita *et al.*, *Computational and Structural Biotechnology Journal*, 2024, doi: 10.1016/j.csbj.2024.11.028

[†] C. J. Jeffery, *Frontiers in Bioinformatics*, 2023, doi: 10.3389/fbinf.2023.1222182.

[‡] A. Sureyya Rifaioglu *et al.*, *Scientific Reports*, 2019, doi: 10.1038/s41598-019-43708-3

### 1.2 Discussion

The main advantages of DEEPred are that *a*) the authors did 100,000 automated hyper-parameter optimisations to find the best ones, *b*) the training data include both experimental low-noise data as well as electronic high-noise data, which makes the model noise-tolerant, and *c*) it has higher scalability since training many small models to predict only five terms in parallel is faster and more accurate than training one very large model to predict every term, with the bonus of being able to retrain each part independently to optimise performance.

### 1.3 Results

DEEPred achieved the highest F1 score in molecular function (MF) of 83%, but only 65% in cellular component (CC) and 55% in biological process (BP). Compared to other models, DEEPred is among the top three in every category and has the highest overall performance in the benchmark when combined.

## 2 DeepGraphGO [§]

### 2.1 Architecture

The bulk of the DeepGraphGO model architecture consists of a graph neural network (GNN), which inputs feature vectors and outputs GO terms along with their corresponding confidence values. Similarly to DEEPred, DeepGraphGO also generates feature vectors with a preprocessing step. It uses the InterProScan database to generate feature vectors with information about protein domains, protein families, functional sites, and motifs.

The feature vectors are condensed into a low-dimensional representation before being used to map them onto a graph using protein-protein interaction (PPI) networks and performing graph convolutions using two GCLs. Seven types of PPI networks (neighbourhood, fusion, co-occurrence, co-expression, experiment, database, text mining) comprising seventeen species, fetched from a database called STRING, were used to generate the master PPI foundation graph.

### 2.2 Discussion

The major advantages of DeepGraphGO are that *a*) it can predict protein functions in many species since it uses multi-species information in the model, compared to other GNNs where every species requires training a separate model, *b*) it is easily expandable by transferring learning onto a larger PPI network, and *c*) it has a lot more context using the PPI network than only the amino acid sequence, so it can infer unseen proteins from similar ones to achieve better performance. Despite this, GNNS also have an inherent limitation, that being they are very slow to train.

### 2.3 Results

DeepGraphGO achieved the best scores among other models. It can comprehensively provide all GO terms without any missing terms, whereas other models often miss the more challenging terms. Even for more difficult proteins and multi-species prediction, DeepGraphGO scores the highest due to its ability to infer function from similar data points and its multi-species network.

[§] R. You *et al.*, *Bioinformatics*, 2021, doi: 10.1093/bioinformatics/btab270

# 3 DeepFRI [¶]

## 3.1 Architecture

DeepFRI is a two-stage graph convolution network (GCN) that predicts protein functions by extracting features from sequences and protein structures. The first stage is a self-supervised language model with long short-term memory (LSTM). The language model is first pre-trained on a set of protein domain sequences from protein databases. The model acts as a sequence feature extractor for the second stage of DeepFRI. The second stage is a GCN that propagates the residue-level features and constructs the final protein-level feature representations.

## 3.2 Results

The authors compared the performance of DeepFRI with other methods, including sequence-based and deep-learning methods. They found that DeepFRI has a higher protein-centric F-score, outperforming other methods in MF and BP. It robustly predicted MF-GO terms of proteins with low sequence identity to the training set, meaning that DeepFRI learns structure-function relationships more robustly than other methods.

They also explored DeepFRI's performance in training different models. In terms of the F-score, DeepFRI performed better for native structures than DMPFold models and Rosetta models. Additionally, DeepFRi demonstrated a significant denoising capability due to the high correlation between GCN features extracted from contact maps. By comparing the MF-GO terms with different methods, we can observe that DeepFRI outperformed other methods for more MF-GO terms with fewer training samples.

## 3.3 Discussion

Although DeepFRI can identify some site-specific predictions, it still has some limitations. It has lower performance for unseen protein models, which limits the accuracy of rare protein functions. Another limitation is that DeepFRI cannot fully capture long-distance structural correlations within proteins, which restricts the depth of the networks.

# 4 Dense Homologue Retriever (DHR) [‖]

Professor Yu Li of CUHK proposed a fast and sensitive method for detecting protein homologues using deep dense retrieval. The first step involves converting protein sequences into a specialised "vector" using a protein language model. This vector summarises essential information about the protein, like its structure and function. To find related proteins, the method compares the vectors of the protein being searched for with the vectors of proteins in a database. If two vectors are similar, it means the proteins are likely homologues. This method skips alignment and compares the vector representations, which makes it much faster and able to find distant homologues. The model is trained under contrastive learning to assign similar vectors to related proteins and very different vectors to unrelated ones, which helps improve the accuracy.

[¶] V. Gligorijević *et al.*, *Nature Communications*, 2021, doi: 10.1038/s41467-021-23303-9
[‖] L. Hong *et al.*, *Nature Biotechnology*, 2024, doi: 10.1038/s41587-024-02353-6