Stockholm University

Unlocking protein sequences

Advances in protein structure and ligand-binding site prediction

Aditi Shenoy

Doctoral Thesis in Biochemistry towards Bioinformatics at Stockholm University, Sweden 2024

Unlocking protein sequences

Advances in protein structure and ligand-binding site prediction

Aditi Shenoy

Academic dissertation for the Degree of Doctor of Philosophy in Biochemistry towards Bioinformatics at Stockholm University to be publicly defended on Friday 26 January 2024 at 09.00 in Air & Fire, SciLifeLab, Tomtebodavägen 23A.

Abstract

The protein sequence determines how it will fold into its unique three-dimensional structure. Once folded, proteins perform their functions by interacting with other proteins or molecules called ligands within the cell. Experimental determination of protein structure and function is tedious. Computational approaches aim to accurately predict the properties of proteins to complement experimental efforts of understanding biochemical mechanisms within the cell. This thesis introduces computational techniques that predict the structure of protein complexes and identify protein residues involved in interactions with common biomolecules, such as metal ions and nucleic acids, based on sequence information.

AlphaFold, a method that predicted protein structure using sequence information with almost experimental accuracy, was a critical breakthrough that shaped the field of protein structure prediction. Subsequently, approaches such as FoldDock adapted the AlphaFold pipeline for dimer complexes. **Paper I** applies the FoldDock protocol to understand toxin-antitoxin systems. These protein complexes are highly evolutionary conserved, and high-confidence dimer predictions were generated. **Paper II** applies the FoldDock protocol to study protein-protein interactions in the human proteome. To verify the reliability of machine-learning-based computational methods, they must be tested on independent data different from the data used to train the method. **Paper III** involves generating and using a homology-reduced independent test set to benchmark the performance of protein complex structure predictors, including the recent AlphaFold release adapted for multi-chain proteins – AlphaFold-Multimer. A confidence score (pDockQ2) was proposed to estimate the quality of the interfaces within multimers. **Paper I, Paper II** and **Paper III** are associated with predicting and evaluating protein-protein interactions.

Representation learning involves finding effective representations of input data to maximise available information, making it easier to understand and process them for downstream prediction tasks. A recent advance in protein representation learning is Protein Language models (pLMs), where large language models are trained on a massive corpus of protein sequences. Highly contextualised and informative vector representations contained in the last hidden layer of the model have been used to predict numerous properties, such as ligand binding sites, subcellular localisation, and post-translational modifications, among others. **Paper IV** uses residue-level embeddings to predict whether a protein binds to one or more of the ten most common ions. It also predicts residue-level binding probabilities for multiple ions simultaneously. **Paper V** expands this approach beyond metals. It explores the impact of structure-informed features alongside sequence embeddings to predict whether a residue binds to nucleic acids, small molecules or metals. **Paper IV** and **Paper V** are associated with developing machine learning methods to predict and evaluate protein-ligand interactions.

In summary, the research conducted within this thesis offers valuable insights into three crucial levers to systematically harness the potential of machine learning for protein bioinformatics. These are (1) construction of homology-reduced non-redundant datasets, (2) finding optimal protein representations, and (3) rigorous evaluation and inference.

Stockholm 2024
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-224344

ISBN 978-91-8014-613-5
ISBN 978-91-8014-614-2

Stockholm University

Department of Biochemistry and Biophysics

Stockholm University, 106 91 Stockholm

UNLOCKING PROTEIN SEQUENCES

Aditi Shenoy

Stockholm University

Unlocking protein sequences

Advances in protein structure and ligand-binding site prediction

Aditi Shenoy

ISBN print 978-91-8014-613-5
ISBN PDF 978-91-8014-614-2

Cover illustration: Aditi Shenoy

Printed in Sweden by Universitetsservice US-AB, Stockholm 2023

ABSTRACT

The protein sequence determines how it will fold into its unique three-dimensional structure. Once folded, proteins perform their functions by interacting with other proteins or molecules called ligands within the cell. Experimental determination of protein structure and function is tedious. Computational approaches aim to accurately predict the properties of proteins to complement experimental efforts of understanding biochemical mechanisms within the cell. This thesis introduces computational techniques that predict the structure of protein complexes and identify protein residues involved in interactions with common biomolecules, such as metal ions and nucleic acids, based on sequence information.

AlphaFold, a method that predicted protein structure using sequence information with almost experimental accuracy, was a critical breakthrough that shaped the field of protein structure prediction. Subsequently, approaches such as FoldDock adapted the AlphaFold pipeline for dimer complexes. Paper I applies the FoldDock protocol to understand toxin-antitoxin systems. These protein complexes are highly evolutionary conserved, and high-confidence dimer predictions were generated. Paper II applies the FoldDock protocol to study protein-protein interactions in the human proteome. To verify the reliability of machine-learning-based computational methods, they must be tested on independent data different from the data used to train the method. Paper III involves generating and using a homology-reduced independent test set to benchmark the performance of protein complex structure predictors, including the recent AlphaFold release adapted for multi-chain proteins – AlphaFold-Multimer. A confidence score (pDockQ2) was proposed to estimate the quality of the interfaces within multimers. Paper I, Paper II and Paper III are associated with predicting and evaluating protein-protein interactions.

Representation learning involves finding effective representations of input data to maximise available information, making it easier to understand and process them for downstream prediction tasks. A recent advance in protein representation learning is Protein Language models (pLMs), where large language models are trained on a massive corpus of protein sequences. Highly contextualised and informative vector representations contained in the last hidden layer of the model have been used to predict

numerous properties, such as ligand binding sites, subcellular localisation, and post-translational modifications, among others.

Paper IV uses residue-level embeddings to predict whether a protein binds to one or more of the ten most common ions. It also predicts residue-level binding probabilities for multiple ions simultaneously.

Paper V expands this approach beyond metals. It explores the impact of structure-informed features alongside sequence embeddings to predict whether a residue binds to nucleic acids, small molecules or metals.

Paper IV and Paper V are associated with developing machine learning methods to predict and evaluate protein-ligand interactions.

In summary, the research conducted within this thesis offers valuable insights into three crucial levers to systematically harness the potential of machine learning for protein bioinformatics. These are (1) construction of homology-reduced non-redundant datasets, (2) finding optimal protein representations, and (3) rigorous evaluation and inference.

PUBLICATIONS

LIST OF INCLUDED PAPERS

PAPER I.

The structural basis of hyperpromiscuity in a core combinatorial network of type II toxin-antitoxin and related phage defense systems.
Ernits K, Saha CK, Brodiazhenko T, Chouhan B, **Shenoy A**, et al. *Proc Natl Acad Sci U S A*. 2023 Aug 15; 120(33):e2305393120.

PAPER II.

Towards a structurally resolved human protein interaction network.
Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, **Shenoy A**, et al. *Nat Struct Mol Biol.*. 2023 Feb; 30(2):216–25.

PAPER III.

Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes.
Zhu W*, **Shenoy A***, Kundrotas P, Elofsson A. *Bioinformatics*. 2023 Jul 1; 39(7):btad424.

PAPER IV.

M-Ionic: Prediction of metal ion binding sites from sequence using residue embeddings.
Shenoy A*, Kalakoti Y*, Sundar D, Elofsson A. bioRxiv; 2023. p. 2023.04.06.535847. doi: https://doi.org/10.1101/2023.04.06.535847 *(Submitted)*

PAPER V.

Impact of joint structure and sequence representations for ligand binding site prediction.
Shenoy, A. and Elofsson, A. *(Manuscript)* (2023).

* Contributed equally (Joint first authors)

LIST OF ADDITIONAL PAPERS

- Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, Chemes LB, .., **Shenoy A** et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic Acids Res. 2022 Jan 7;50(D1):D480–7.

- Bryant P, Pozzati G, Zhu W, **Shenoy A**, Kundrotas P, Elofsson A. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Nat Commun. 2022 Oct 12;13(1):6028.

- Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, .., **Shenoy A** et al. A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol. 2022 Nov; 29(11):1056–67.

AUTHOR CONTRIBUTIONS

PAPER I.

GA initiated the project. AS and AE generated AlphaFold models using the FoldDock pipeline for the toxin-antitoxin pairs. AS calculated pDockQ scores for these models and provided a list of potentially interacting complexes. These models were further experimentally validated by GA and VH's team. AS contributed to the methods section of the manuscript and generated Figure S4.

PAPER II.

AE conceptualised the project together with PBeltrao. AS contributed to minor data analysis of the protein complexes with GP and WZ. AE and PBeltrao wrote the initial draft of the manuscript. AS contributed to the final version by proofreading.

PAPER III.

AS and WZ created the benchmark datasets and generated and evaluated the models. AS and WZ together wrote and generated the figures for the manuscript. All authors analysed the results and contributed to the final manuscript.

PAPER IV.

AS conceptualised and initiated the project together with YK. AS prepared all the datasets, and YK implemented and optimised the model. AS conducted ablation studies and performed subsequent analyses. AE and DS provided inputs on experiments and analysis to perform. AS wrote the initial draft of the manuscript and conducted revisions. All authors contributed to the final manuscript.

PAPER V.

AS conceptualised and initiated the project. AS implemented and optimised the model and conducted all analyses with inputs from AE. AS wrote the initial draft of the manuscript.

ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
DL	Deep Learning
DCA	Direct Coupling Analysis
IPA	Invariant Point Attention
LLM	Large Language Model
LBS	Ligand Binding Site
ML	Machine Learning
MLM	Masked Language Modelling
MPNN	Message Parsing Neural Network
M	Million; (B - Billion)
MSA	Multiple Sequence Alignment
MI	Mutual Information
NLP	Natural Language Processing
PSSM	Position-Specific Scoring Matrix
PAE	Predicted Aligned Error
PDB	Protein Data Bank
pLMs	Protein Language Models
PLI	Protein-Ligand Interactions
PPI	Protein-Protein Interactions
RL	Representation Learning
SASA	Solvent Accessible Surface Area
3D	Three-Dimensional
TA	Toxin-Antitoxin
2D	Two-Dimensional

Metrics:

AUPR	Area under the PR curve
AUROC	Area under the ROC curve
FPR	False Positive Rate
MCC	Matthew's Correlation Coefficient
PR	Precision-Recall
ROC	Receiver Operating Characteristics
TPR	True Positive Rate

CONTENTS

PREFACE

If one zooms into the basic unit of life, a cell, what would you expect to see?

Let us use the city of Stockholm for this thought experiment.

About ~1 million people live in Stockholm. About ~70% of the population is employed. So, for the most part, each resident gets up in the morning, has a cup of coffee and goes to work. Every individual has a distinct job to do. They work in grocery stores, banks, construction, restaurants, or some sort of enterprise. Stockholm has numerous islands, each well connected through trains, buses and trams. Many people depend on a group of other individuals who work in the public transport system to get from their homes to places of work. All the day-to-day proceedings of a city result from each individual doing their job every single day. In case certain individuals fall sick or supporting machinery ceases to function as expected, chaos erupts, and there are disruptions in the way the city functions. But for the most part, everybody and everything works smoothly.

Just as humans are the workhorses of any city, proteins are complex biomolecules that play an essential role in most biological processes. Millions of proteins within the cell are constantly in flux, interacting with numerous other biomolecules to perform their respective functions. Proteins move between different territories (or organelles) of the cell. Specific groups of proteins (e.g. chaperone, transport, enzymatic or signalling proteins) enable other proteins to function. If a protein fails to work correctly, its associated functions are disrupted, leading to various diseases. It is marvellous to think about how, by and large, all these complex biomolecules constantly work together with clockwork precision, allowing for perfect harmony within the cell. This thesis is written in awe of these remarkable molecules and attempts to serve as a modest contribution to the scientific endeavour of understanding proteins.

1

INTRODUCTION

Proteins are indispensable entities of a biological cell. These complex biomolecules are involved in various processes ranging from signalling, transport, metabolism and catalysis to even rudimentary functions like cell division. Proteins consist of a string of amino acids that are bound together by peptide bonds. To be rendered functional, the unfolded sequence of residues folds into a three-dimensional (3D) structure. The folded structure then goes on to interact with other proteins or ligands in the cellular environment, though exceptions exist [1].

Protein folding is the mechanism proteins employ to self-assemble into 3D structural units from their primary amino acid sequence. Dill et al. [2] succinctly described the protein folding problem with three questions: (i) What physical forces drive the folding process? (ii) How do proteins instantaneously fold in the 'right way'? (iii) Can computational approaches predict protein structures using sequence information?

1.1 THREE-DIMENSIONAL FUNCTIONAL UNITS

The driving forces of protein folding include the hydrophobic effect and atomic interactions such as hydrogen bonding and Coulombic and van der Waals interactions [3]. Additionally, weak (and more frequent) interactions in the protein backbone contribute to secondary structure formation, and strong (and less frequent) interactions in the protein side chains contribute to ligand binding or post-translational modifications [3]. However, despite consensus on physical forces driving protein folding, its mechanism has been a point of contention for decades [4, 5].

Based on pioneering experiments conducted to refold a denatured ribonuclease protein under normal physiological conditions, Anfinsen postulated the 'thermodynamic hypothesis' [6]. It suggested that the native conformation of the protein had the lowest Gibbs free energy and that the interatomic interactions between the amino acids in the protein sequence were sufficient to determine the 3D folded structure [6]. This funnel-shaped landscape was the classical view of the conformational energy landscape of proteins. Think of it as a roulette wheel. If one were

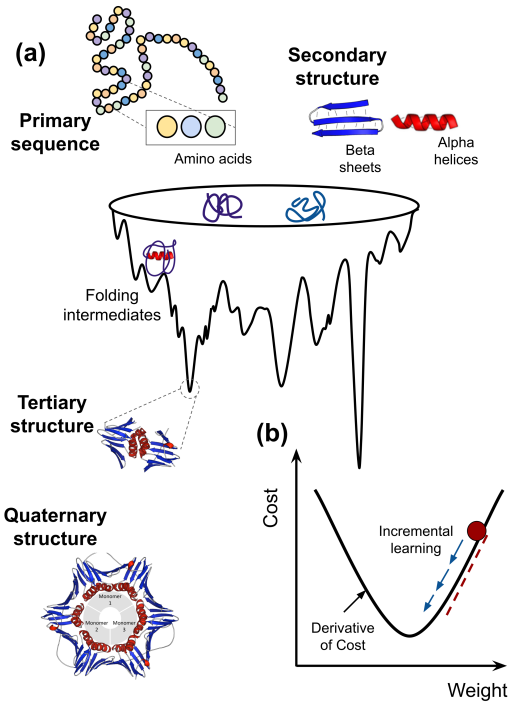

Figure 1.1: (a) Levels of the protein structure[1] and visualisation of the protein folding landscape (folding funnel hypothesis) (b) Visualisation of gradient descent.

to roll the ball in the spinning roulette wheel, it would eventually arrive at the lowest state and come to a halt.

Similarly, a vast conformational energy landscape must be sampled for proteins to arrive at their native conformation having minimum Gibbs free energy (ΔG). Levinthal noted that exhaustively sampling the entire protein landscape to arrive at the correct native conformation would extend into billions of years. Assuming an average protein length of 150 residues, n = 3 degrees of freedom (two rotations along the backbone and one along the sidechain) and a picosecond to fold into each conformation, sampling the entire landscape would take 3^{150} x 10^{-12} seconds ($\sim 10^{48}$ years)[7]. However, in reality, the folding occurs in a matter of seconds or less (Levinthal's paradox) [8]. The mutual exclusivity between thermodynamic control (finding global energy minimum) and kinetic control (instantaneous folding) led to the theorisation of 'new views' on folding pathways [4]. The rugged folding funnel hypothesis [4, 9] (Figure 1.1 (a)[1]) allows for multi-states

[1]Components of protein structure adapted from Wikipedia figure created by Thomas Shafee and licensed under CC BY 4.0 license

(multiple local minima) as well as many alternative routes to arrive at the lowest energy state (global minima).

1.2 COMPUTATIONAL PROTEIN STRUCTURE PREDICTION

For over 50 years, various computational solutions have been proposed to derive the tertiary structure of a protein from its primary amino acid sequences (often referred to as the 'Protein Structure Prediction' problem). Experimental techniques such as X-ray Crystallography, Nuclear Magnetic Resonance (NMR), and Cryo-Electron Microscopy (Cryo-EM) have been used to resolve protein structures. These experimentally determined structures are stored in the Protein Data Bank (PDB) [10]. Early solutions to generate reliable models of protein structures were based on homology. In this case, experimentally resolved protein structures having sequence similarity to the query protein were identified from PDB and used as 'templates' (e.g. fold recognition [11] or homology modelling [12]). This approach was referred to as template-based modelling.

Template-free modelling (*Ab initio* approaches), on the other hand, attempts to predict structures based on only the sequence information without using global similarity to proteins in PDB. These approaches aim to learn the 'principles' of protein folding and are especially useful when homologous structures are unknown (or yet to be experimentally determined). The primary limitations of these *de novo* approaches lie in the vast computational resources needed for modelling and their ability to learn an appropriate energy function in a relatively short timescale. Moreover, their performance drops as the protein size increases, making them unreliable for proteins with more than 150 residues [13].

A crucial factor contributing to the development of successful predictors was the adoption of Machine Learning (ML). This class of algorithms iteratively learns the input data and arrives at a solution using an optimisation algorithm (e.g. Gradient Descent [14, 15]) that finds the lowest error between the prediction and the ground truth. Chapter 2 delves into ML concepts, workflow and principles used in their application to structural biology problems.

Analogous to the protein energy landscape, there are numerous paths the algorithm can explore to arrive at its solution. The minimum energy state is determined using a step-wise approach (Figure 1.1 (b)). These algorithms are data-hungry. They perform better with large and diverse training datasets. The exponential growth in the number of structures deposited in PDB [16] and the advancements in computation (Moore's law) [17] have been opportune for the considerable progress seen in *ab initio*

protein structure prediction. However, it is possible that these ML models do not learn the patterns or principles of the input data but instead memorise it. When the models are unable to generalise, they are said to be 'overfitting', and these ML models become unreliable. Therefore, the performance of these methods is assessed on how well they perform on 'unseen' data. Bearing this concept in mind, the need for an independent assessment of protein structure predictors was realised. As a result, a double-blind assessment of the performance of the proposed predictors (called Critical Assessment of Protein Structure Prediction - CASP) was established in 1994 [18]. This competition has been conducted biennially ever since to drive the development and assessment of protein structure predictors.

1.3 ALPHAFOLD REVOLUTION

In 2020, during CASP14, a giant breakthrough was made in *ab initio* protein structure prediction. AlphaFold [19] brought a paradigm shift in structural biology and was considered a revolutionary advancement in predicting protein structures with close to experimental accuracy [20]. The abundance of structures in PDB and the availability of massive computing in parallel with the development of ML algorithms has led to the success of AlphaFold. Based on structures deposited in PDB, AlphaFold learned from an exhaustive set of the most represented native conformations and broadly understood biophysical constraints that govern a stable conformation. By using optimisation algorithms like gradient descent, the error between the prediction and ground truth is propagated back into the network, and AlphaFold is able to find the proteins' minimum energy configuration.

In an effort to make the models openly available to the scientific community, 200 million (M) structures, including 48 complete proteomes, were made available through the AlphaFold Protein Structure Database (AlphaFoldDB) [21]. Details on the AlphaFold method and its applications are discussed in Chapter 4.

1.4 THE ERA OF LANGUAGE MODELS

Alongside the exponential increase in computing hardware, biological sequence databases are growing at a rate faster than Moore's law [17]. There is approximately 1000 times more protein sequence information in UniProt [22] than experimentally resolved structure information in PDB (as of November 2023). With the advent of AlphaFold, this difference is

less pronounced. However, the untapped potential of sequences in biological databases is yet to be fully exploited. Gaining popularity over the last ten years, a subarea of research in ML called Natural Language Processing (NLP) has focussed on understanding language. A relatively recent and widely adopted general-purpose application of NLP has been Large Language Models (LLMs). Based on knowledge learned by 'pre-training' the algorithm on a massive amount of unstructured and unlabeled data, LLMs have shown promise in fields like text generation, translation and summarisation, to name a few [23]. Since a protein can be thought of as a sentence (composed of a string of letters or amino acids), it was only a matter of time before NLP was applied to protein sequences to understand the semantics of the 'proteins language'. The key concepts and algorithms behind LLMs are described in Chapter 2. Their application to protein prediction problems has been expanded in Chapters 4 and 5.

1.5 AIM AND SCOPE OF THIS THESIS

To comprehend how proteins carry out their destined functions, it is crucial to analyse how proteins (i) interact with each other and (ii) interact with other molecules within the cell. The underlying relationship between sequence, structure and function has been explored using ML algorithms. Individual protein sequences or a group of protein sequences aligned together (referred to as a multiple sequence alignment) have been commonly used as input to train the prediction tasks. Various other protein representations used to train these tasks have been described in Chapter 3. The papers included in this thesis are divided into two categories based on the type of input data used and the objective of training the ML model. These categories use protein sequences:

- to predict multi-chain protein structures (Paper I and Paper II) and evaluate the performance of these methods on an independent benchmark dataset (Paper III).
- to predict ligand binding sites using protein language models (Paper IV and Paper V).

This thesis contributes to understanding the language of proteins for structure and function prediction.

2

MACHINE LEARNING FOR STRUCTURAL BIOLOGY

Machine learning (ML) broadly describes a group of algorithms that can learn to detect patterns in datasets. These algorithms aim to generate a mathematical model that can make inferences based on input features. The step where the model wrangles with the data and tries to make sense of it is called 'training' the model. The parameters that the model learns during the training process are assumptions made by the model to fit the data it has seen. The model uses these learned parameters to generate a prediction for new data. Models could be trained to predict discrete variables (classification) or continuous variables (regression).

2.1 DATA AND ALGORITHMS

Algorithms (also called models) have varying flexibility based on the number of degrees of freedom it has. Restrictive models have fewer degrees of freedom and make more prior assumptions. If the assumptions suit the data used to train the model, a small number of data samples is sufficient to fit the model. Examples of such models are linear regression, logistic regression or k-means clustering. If there are more degrees of freedom, the model has fewer assumptions. In this case, more data is required to learn the mathematical relationships between the input samples. Examples of flexible models are neural networks or decision trees. The No Free Lunch Theorem (NFL) [24] states that all algorithms perform equally well if their performance, over all possible problems, is averaged. So, if one used infinite data and no prior assumptions to constrain the search space, the model's performance would be no better than random. Therefore, the performance improvement of models hinges on how well the prior assumptions of the model match the underlying principles of the data and the objective being trained on.

Based on the input data, the type of algorithm and the learning objective, there are numerous approaches one could pursue to apply machine learning to their problem of interest. These approaches could be broadly classified into three categories: (i) supervised learning, (ii) unsupervised learning, and (iii) reinforcement learning. Additional subcategories exist

under or at the intersection of these approaches (for example, semi-supervised, self-supervised, and representation learning). The following section will dive into the details of a few learning approaches, how they differ from the other approaches and their subcategories relevant to and within the scope of this thesis.

2.2 LEARNING APPROACHES

In supervised learning, the model learns the objective function that maps the input data to the corresponding output data (target). For example, let us take a dataset of protein sequences; each sequence is directed to a location in the cell. Based on a signature within the sequence, the protein gets localised to the nucleus, cytoplasm or any of the other cellular organelles. So, the protein sequence is the input data, and the organelle associated with each protein sequence is the target label. The objective of the supervised machine learning model is, thus, to predict the class (in this case, organelle) by training a classifier based on the features in the input protein sequence.

In contrast, in unsupervised learning, the model explores the characteristics of input data without the presence of any target labels. This approach aims to understand the underlying patterns within the dataset. Common ways of using unsupervised learning are to cluster the datasets into groups having similar features (i.e. clustering) or reduce the number of input features required to separate the data into groups by retaining only the essential information (i.e. dimensionality reduction). For example, given a set of protein features (e.g. protein length, surface accessibility, number of hydrophobic or hydrophilic residues, etc., in a tabular form), the unsupervised machine learning model is trained to identify groups within the input dataset (e.g. soluble/membrane proteins or secondary structure elements).

A new field of semi-supervised learning emerged at the intersection of supervised and unsupervised learning. In this approach, the model learns the patterns within an unlabelled dataset in an unsupervised manner. Next, the insights learnt by the model in the first step are used to train a supervised task using a much smaller corpus of labelled data. This approach has shown promise when the amount of labelled training data is limited. A notable example of this type of training is in large language models, which are explained later in this chapter.

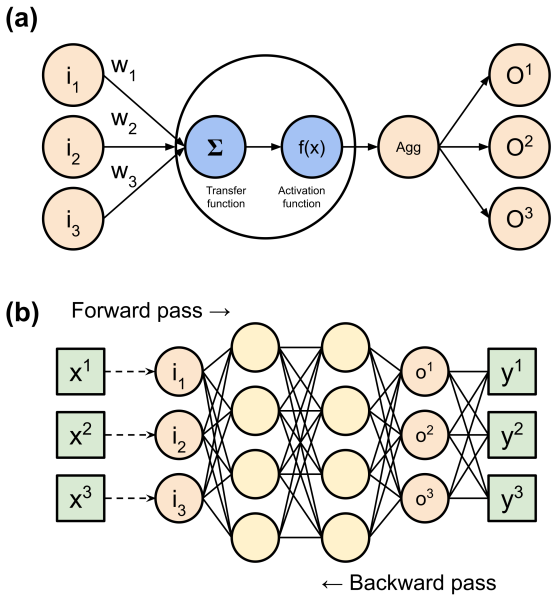

Figure 2.1: (a) Representation of a single artificial neuron (ANN) where i_1, i_2, i_3 are the input neurons; w_1, w_2, w_3 are the weights associated with each input and o^1, o^2 and o^3 are the output neurons of the ANN. (b) Representation of a Multi-layer perceptron (MLP) where x^1, x^2, x^3 are the input features into i_1, i_2 and i_3 input neurons respectively. The output neurons o^1, o^2 and o^3 produce 3 outputs in this network i.e. y^1, y^2, y^3

2.3 DEEP LEARNING

While constructing ML models, an important design choice is whether to employ manual or automatic feature extraction. In the first case, the input data is pre-processed, and domain-informed features are manually extracted prior to training the classifier model. Alternatively, we can allow the model to identify features that are important for the prediction problem automatically. This kind of automatic feature extraction is seen in models belonging to a subset of ML called Deep Learning (DL). These complex models comprising multiple non-linear computational blocks can arrive at representations with many abstraction levels. DL models are made up of an interconnected network of artificial neurons consisting of input, hidden and output neurons (also called Artificial Neural Network (ANN) [25]). Each neuron calculates the weighted (w_i) sum of one or more inputs (x_i), adds a bias b, and produces an output y. An activation

function f is applied to this output to allow the modelling of non-linear relationships. The most popular non-linear function is a rectified linear unit (ReLU) given by $g(z) = \max\{0, z\}$ [26]. ReLU produces efficient sparse representations and reduces the likelihood of vanishing gradients. The expression for each neuron can be written as Equation 2.1

$$y = f(\sum_{i=1}^{n}(x_i w_i + b) \tag{2.1}$$

In the forward pass of the neural network, the output of the input neuron is fed into the neurons in all the hidden layers. Subsequently, the hidden layers' outputs become input to the final neuron. The training process of the neural network occurs in the backward pass (also termed back-propagation). The final result y from the output neuron (predicted output) is compared with the target (ground truth) value, and an error (also termed loss) is obtained by applying a loss function (e.g. mean squared error or cross-entropy loss) to these two values. The error at the end of each forward pass is used to adjust the weight associated with each neuron in the backward pass.

The objective of the optimisation algorithm - Gradient Descent [14, 15] is to minimise the loss (also called cost) by updating the model parameters (i.e. weights and biases). The delta Δw (indicating the change in weights) is the derivative of the loss function with respect to each of the weights in the network and using the learning rate η to moderate the rate at which the Δw moves in the direction of the gradient towards the local minimum of the loss function (Figure 1.1 (b)).

NOVEL DL ARCHITECTURES

The development of four highly specialised and revolutionary architectures has led to the success of DL across a broad spectrum of applications. These are Convolutional Neural Networks (CNN) [27–29] (for Computer Vision tasks such as object detection, segmentation and recognition in images), Recurrent Neural Networks (RNN) [30] and Long short-term memory Networks (LSTM) [31] (for sequence translation and speech recognition) and Transformers [32] (computer vision, NLP and more). When Nvidia adapted their graphic processing units (GPU) to perform matrix computations in deep learning in 2009, the speed for algorithms increased significantly ($\sim$100 fold). This hardware advancement allowed the training of deeper and more efficient neural networks.

The DL revolution took off after the success of CNNs in classifying 1.2M images into 1000 classes in the ImageNet contest [33]. CNNs successfully create an internal representation by detecting the local features in a multidimensional vector (most commonly 2D images). RNNs are most valuable for tasks involving sequential inputs. These networks process one element at a time, and their hidden units retain information about the previously processed elements in the sequence. Though powerful, RNNs do not work well with very long sequences and are subject to exploding or vanishing gradients [34]. The LSTM networks outperformed RNNs since they could process long sentences better. However, when the Transformer was introduced in 2017, its performance superseded all the above methods. This architecture transformed the fields of computer vision, NLP, speech or audio processing, as well as multimodal processing. The following section will focus on the Transformer architecture and describe the mechanism that makes it so powerful.

2.4 TRANSFORMERS

The Transformer model was built for language modelling and machine translation. There are four components responsible for the effectiveness of transformers: (1) Input embedding, (2) Encoder-Decoder stack, (3) Self-attention mechanism, and (4) Positional encoding.

2.4.1 *Input representation:*

The input sequence must be converted to a machine-readable format. First, the sentence is pre-processed by splitting it into words or tokens. Each token is then encoded as a numerical vector and is converted to continuous vector representations (also called 'embeddings') using an embedding algorithm. The most straightforward way (but often ineffective) to create embeddings is by replacing each word with a number according to a predefined word-number mapping. This one-hot representation, however, scales with the number of unique words in the dataset, is too sparse and does not consider neighbouring words.

Early methods in NLP relied on counting the frequency of occurrence of certain words in a sequence (Bag-of-Words or BoW) or by counting combinations of groups of words (k-mers or n-grams). These representations lose the order of the words in the sequence. Methods based on global token co-occurrences (GloVe [35]) and local co-occurrences (Word2vec [36]) retain contextual information about the sequence. Two critical approaches

in Word2vec were introduced in [36] where the current word was predicted based on the neighbourhood (Continuous BoW), or the surrounding context was predicted based on a respective word (Skip-gram).

Though these methods take context information into account, they cannot discern between the same words occurring in different contexts. Peters et al. developed contextualised embeddings by training a bidirectional LSTM. The ELMo (Embeddings from Language Models) vectors encode for semantic meaning of words in context [37].

2.4.2 *Encoder-Decoder structure:*

The sequence embeddings enter a stack of encoders to produce an intermediate representation that is subsequently fed into the decoder, which generates the output sequence one element at a time. This model is autoregressive, which means the model considers all previous elements while predicting the next element in the sequence. Each encoder layer consists of a multi-head attention followed by a fully connected feed-forward network. Six blocks of such encoder layers form the encoder block. In addition to the layers in the encoder layer, the decoder layer contains an additional component which runs multi-head attention on the output of the encoder block. The decoder block also contains six decoder layers. The underlying mechanism that models relationships between all the words in the sentence and computes a contextualised representation is called 'self-attention' and is described in the next section [32].

2.4.3 *Attention*

In the translation process, the model learns to focus on (or pay 'attention' to) certain words in the sentence. The objective of the attention function is to map a set of vectors (i.e. Query q, key k and value v) to an output. These three vectors are obtained by multiplying the input embedding for each word with the weight matrices associated with the query (W_Q), key (W_K) and values (W_V), respectively. These weight matrices are updated in the backpropagation step after each epoch.

Once the q, k and v vectors are generated, it is time to calculate a score that weighs certain words more than others in the context of the sentence. The score is obtained by computing a dot product of the query with all the keys in the sequence and normalising the result using a softmax function. Attention mechanisms have been previously introduced in Bahdanau attention [38] (additive) and Luong attention [39] (multiplicative). If the weighted sum of the softmax result is multiplied by the vector v, we get

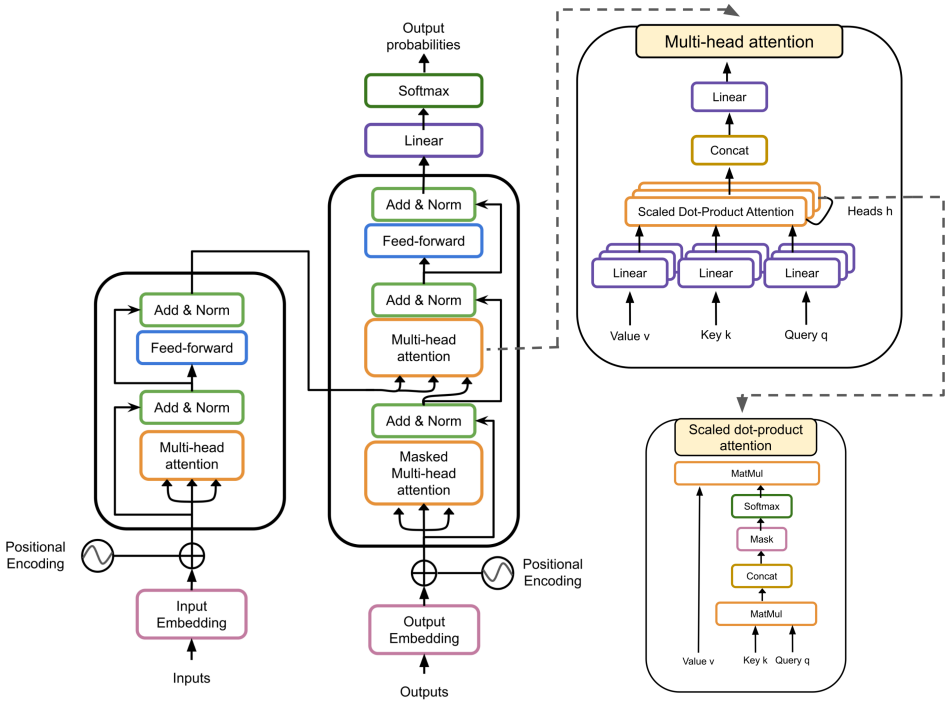

Figure 2.2: Graphical illustration of the transformer architecture [32]. The inputs are fed into the encoder stack, and the outputs are obtained from the decoder stack. Details of the multi-head attention and the scaled dot product attention are illustrated alongside the transformer.

the Bahdanau attention; otherwise, if the product of the softmax result is multiplied by the vector v, we obtain the Luong attention. The main difference between these attention mechanisms and the one proposed in the Transformer architecture is that the dot-product of the query and keys is scaled by a factor (d_k) before applying the softmax function ('Scaled Dot-Product Attention'). This modification was found to improve speed and memory requirements [32].

Multi-head attention is used to aggregate information from various subspaces jointly. It is obtained by calculating multiple attention scores on multiple linear projections of the q, k and v vectors in parallel. Each combination of linear projections used to calculate Scaled Dot-Product Attention is called an 'attention head'. The concatenated outputs from each head result in the multi-head attention output.

2.4.4 *Positional encoding*

To preserve sequence order information in addition to the weighted score for each word in the sequence (attention), positional encodings are used. These are the absolute or relative positions of the word in the sequence, and since it has the same dimensions as the input embedding, they are directly added to each other.

2.5 APPLICATION OF TRANSFORMERS

Transformers have shown much success and promise in NLP, especially in tasks involving machine translation, language modelling and generation. These models are capable of capturing long-range dependencies, thereby understanding the rules that govern languages when trained on a corpus of text data. These so-called Language Models (LM) are probabilistic. That means they provide a probability distribution of the likely next words to occur conditioned on all the previous words in the sequence.

Large Language Models (LLMs) are LMs trained on incredibly massive amounts of sequence information (e.g. all of Wikipedia or UniProt). Since LLMs can generate sentences based on a given prompt, they are also called generative transformer language models [40]. Models learn semantics in a language by leveraging the distributional hypothesis, which states that words occurring in similar contexts have similar meanings [41]. Instead of static embeddings generated by the methods described in section 2.4.1, new methods, where the transformer architecture is modified based on the training objective, have become prevalent.

The decoder-only architecture (Generative Pre-trained Transformer or GPT) showed that generative pre-training of a language model (unsupervised learning) was followed by discriminative fine-tuning improved prediction on the downstream tasks [42]. These LMs were trained in an autoregressive fashion (unidirectional and based on only preceding tokens).

An encoder-only architecture trained with a bidirectional masked language modelling (MLM) objective is called Bidirectional Encoder Representations from Transformers (BERT) [43]. The MLM objective involves randomly masking 10-15 % of the tokens, and the model is trained to guess the missing token based on the context. Further optimisation of the BERT model can be seen in ALBERT [44], RoBERTa [45], DistilBERT [46] and ELECTRA [47].

The objective of BERT models is to create meaningful word representations while that of GPT to fine-tune the pre-trained generative model

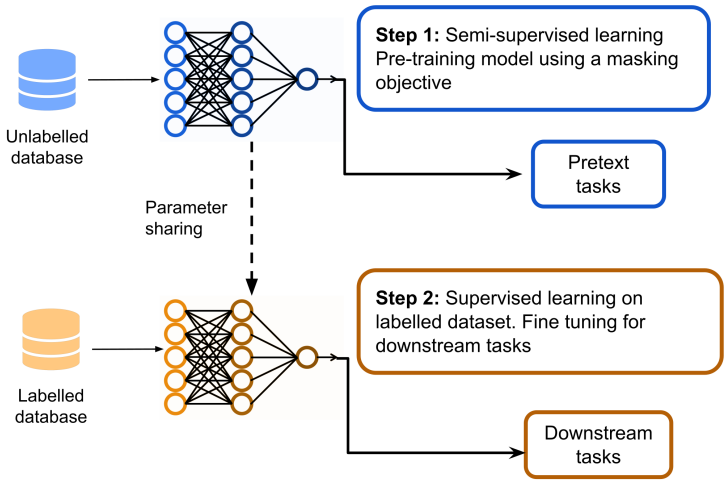

Figure 2.3: Illustration of the steps involved in self-supervised learning inspired by [49]. Step 1 involves learning from an unlabelled database in an unsupervised way. In Step 2, the model uses what was learned in Step 1 to predict specific downstream prediction tasks.

for task-agnostic predictions. An encoder-decoder architecture (called T5) also uses an MLM objective and is pre-trained on multiple text-to-text tasks in a supervised and semi-supervised manner [48].

The two-step approach of pre-training and fine-tuning is graphically represented in Figure 2.3. Most protein sequences in UniProt are unannotated, and they are ideal for semi-supervised learning. The unsupervised pre-training on all known protein sequences learns from the entire protein landscape and captures the semantic organisation of proteins [40]. Since BERT and T5 can create highly contextualised embeddings and are able to capture long-range dependencies, they are ideal for proteins since contacts in the 3D protein structure could, in fact, be between residues that are not adjacent to each other in the protein sequence.

2.6 MACHINE LEARNING WORKFLOW AND PRACTICALITIES

The broad factors affecting the performance of ML models are (1) the number of features and training samples, (2) effective representations, (3) choice of algorithm, and (4) rigorous evaluation. The number of features (attributes) of a dataset represents its dimension. If the number of attributes is high (in order of 100 or more), it is referred to as a high-dimensional dataset. A common problem that arises when dealing with

high-dimensional data is that when the number of features in the data increases, the number of training samples needed (in order to make the model generalisable) increases exponentially. This problem is called the curse of dimensionality. A rule of thumb is that the number of dimensions d should be much lesser than the number of samples n in the dataset ($d \ll n$, usually by a factor 50 [50]). This is why domain-informed feature extraction is crucial before training ML models.

The next factor is selecting an effective representation of the input dataset to feed into the model network. The area of Representation Learning (RL) deals with arriving at a representation containing informative priors that makes the downstream learning task easier. RL provides a way to explore and interpret data based on information captured in the high-dimensional space (unsupervised learning). A common use case of RL is to use the representations for subsequent prediction tasks (transfer learning) [51]. Chapter 3 discusses the various protein representations used for protein analysis.

The choice of ML algorithm depends on the training objective, type of data (input and output), number of features and the acceptable tradeoff between interpretability and accuracy. ML algorithms are generally simpler and shallow, making them more explainable (but less powerful) than DL algorithms (often called black box models).

The final (and often most important) aspect of the ML workflow is conducting a rigorous evaluation of the performance of the model. A crucial prerequisite is to construct independent training, validation and testing datasets. A model is fit based on the training dataset. A validation dataset is used to compare and test the iterative development in the model architecture or the choice of hyperparameters. Once a satisfactory model with the optimal parameters is found, the results are reported using an independent test set. Reporting results based on the training or validation dataset does not indicate how well the model is able to generalise to new (and previously unseen) data. Furthermore, there is a risk of overfitting if there is an overlap between the training and validation or testing datasets.

Constructing data sets with protein sequences presents a special issue. It is not enough to remove overlapping sequences in the training and validation datasets. Protein sequences having more than 30 % sequence identity are homologous (have a common ancestor) [52] and can often have similar structures (and possibly similar functions). So, it is crucial to remove homologous sequences within and across training and testing samples. Homology partitioning makes sure that all similar sequences (with sequence identity up to a certain threshold (20-30 %)) are either in the

		Actual: Positive	Actual: Negative
Predicted	Positive	True Positive (**TP**)	False positive (**FP**)
Predicted	Negative	False Negative (**FN**)	True Negative (**TN**)

Figure 2.4: Confusion matrix summarising four possible outcomes of predicted values against expected values.

training set or testing set and do not span across both. There is an overrepresentation of certain protein families, so in order to not propagate this bias of imbalance to the model, only the representative sequence after clustering all the proteins is taken into account (homology reduction). When both these approaches are performed, we obtain a homology-reduced, non-redundant dataset that is ready for model training.

Cross-validation (CV) is a statistical technique of sampling different subsections of the data to estimate the performance of the model across the entire dataset. The procedure (often called k-fold CV) involves splitting the dataset into k folds (also called splits or partitions). One of the folds is held out for testing the model that has been trained on the remaining k-1 folds.

2.7 EVALUATION METRICS

Based on whether the ML model is trained on a regression or classification task, there are different metrics used to evaluate the performance of the models. In the case of regression tasks, scores such as mean squared error (MSE) or root mean squared error (RMSE) are used to estimate how far the prediction is from the ground truth. The performance of classification tasks is assessed based on the components of a confusion matrix (Figure 2.4).

The per-class performance is assessed using Precision (proportion of positive predictions that were actually positive), recall (proportion of actual positives that were correctly identified), F1-score, and Matthew's Correlation Coefficient (MCC) given by the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.4}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (FP + TN) * (FN + TN)}} \tag{2.5}$$

MCC is a robust measurement, especially for imbalanced classes, since it takes the number of examples within each class into consideration [53]. Varying the threshold between positive and negative classes determines the tradeoff between True Positive Rate (TPR) and False Positive Rate (FPR). These tradeoffs can be visualised using receiver operating characteristics (ROC) and the area under the ROC curve (AUROC) score [54]. Similarly, the tradeoff between precision and recall at various thresholds is seen in precision-recall (PR) curves and area under the PR curves (AUPR) score.

3

REPRESENTATIONS OF PROTEINS

There are numerous ways to convert proteins into a form suitable for training machine-learning models. A protein representation is a numerical vector representing a protein in a 'machine-readable' form. Algorithms can be trained using one or more of such representations for various prediction tasks. Protein representations can broadly be classified into three categories: (i) sequence-based, (ii) structure-based representations and (iii) a combination of sequence and structure representations. The sections below describe various residue-level representations, and the subsequent chapters summarise applications of these representations on protein prediction tasks.

3.1 SEQUENCE-BASED REPRESENTATIONS

One-hot encoding is a conventional method for encoding a protein sequence. It is a numerical vector consisting of zeros and ones, where the amino acid in question is assigned '1', and the other amino acids have '0' assigned to them. As a result, each protein sequence (length L) is converted into an Lx20 dimensional vector (i.e. one dimension for the 20 residue types). The one-hot encoded sequence, however, contains no direct information about the sequence's physicochemical properties or amino acid frequencies. Global sequence descriptors (e.g. amino acid composition, pseudo amino acid composition, hydrophobicity, and polarity, among others [55]) that capture physicochemical properties of protein sequences have been shown to predict secondary structure content [56], subcellular localisation [57] and protein folds [58].

However, the limited information in single sequences translated to fairly low prediction performance. Moreover, to predict structure and function with residue-level granularity, it is vital to know the properties of the individual amino acids and their neighbourhood information. Therefore, subsequent methods focussed on generating amino-acid encodings that also contained physical, chemical and biological properties. Since homologous proteins (with >30% sequence similarity) are known to have similar

three-dimensional folds [59], the information contained in groups of evolutionarily related sequences can be capitalised to generate context-rich amino acid encodings.

3.1.1 *Evolution-informed representations*

As the number of protein sequences increases in public protein databases (e.g., UniProt [22] and Protein Data Bank (PDB) [10]), the probability of finding evolutionarily related sequences for most query proteins increases. If we take two or more such sequences and align them, a Multiple Sequence Alignment (MSA) can be generated. The MSA has been pivotal in predicting protein structures over the last 50 years. One of the early applications of using the information in the MSA is for secondary structure prediction [60, 61]. To feed the information from MSAs into machine learning models, they must first be converted into numerical matrices. These matrices could be sequence profiles (containing the frequency of each amino acid occurring in the alignment) or, more specifically, a Position-Specific Scoring Matrix (PSSM) (containing the log odds score of each amino acid in the alignment).

The co-evolutionary signal contained in the MSAs captures the conversation patterns in the alignment [62] and has been a critical component of 3D protein structure prediction. Extracting co-evolutionary signals involves looking for patterns within the MSA where columns of residues tend to co-occur (or 'co-evolve') with certain other columns. The main limitation of PSSMs is that calculating the fraction of occurrence leads to loss of information that exists in the MSA. Additionally, MSAs made up of a few sequences have low statistical power, and downstream neural networks are not able to differentiate between informative and non-informative PSSMs. Non-obvious patterns that simultaneously affect multiple columns can be learned more effectively directly from the raw MSA using neural networks.

The authors of rawMSA [63] used the techniques from NLP to create more informative representations of MSA. They converted the input MSA into high-dimensional vectors using word2vec [36] using a two-layer neural network (embedding layer). The MSA encoder from CopulaNet [64], an end-to-end framework to predict inter-residue distances directly from MSA, uses its intermediate MSA embeddings for downstream prediction tasks.

Deep and diverse alignments are an information-rich repository and have been shown to be a crucial component for improved performance in protein structure prediction [19]. The quality of the prediction is highly dependent on the quality of the MSA (high signal-to-noise ratio), and various

factors need to be optimised (alignment tools, hyperparameters and databases [65]) for high-quality MSAs. Moreover, with the exponential growth of sequences in protein databases, generating MSAs is computationally expensive and time-consuming. Versatile representations for large-scale protein prediction tasks are an active area of research. As described in Chapter 2, explicit feature extraction can only take prediction performance so far. Neural networks have the potential to not only represent data more efficiently but also make more informed decisions by interpreting patterns in high-dimensional data.

3.1.2 *Protein Language Models (pLMs)*

Evolutionary pressures have directed sampling of the vast protein space. Based on the large number of protein sequences known today, evolutionary constraints of the amino acids that retain the proteins' structural and functional properties can be studied. Self-supervised learning approaches like pLMs have emerged as valuable tools in gauging these constraints by learning from a large corpus of unrelated protein sequences. As described in [40], language models are more data-driven than classical methods motivated by biological priors. To understand the underlying principles of the protein language, the amino acids are first tokenised (each amino acid is a 'token'), converted into a numerical representation (embedding) and then trained using a variety of NLP architectures (described in Chapter 2). Once these rules of the unlabeled and unrelated protein sequences are learned, the intermediate numerical representations obtained from the last hidden layer of the pre-trained LLMs are fine-tuned for downstream prediction tasks.

UniRep [66] representation was trained on a multiplicative LSTM averaged over its 1900 hidden units using 24 million (M) UniRef50 protein sequences, and the model contained 18M parameters.

SeqVec [67] adapted the ELMo model [34] to model protein sequences. In addition to the fact MSAs are expensive to generate, the authors of SeqVec [67] argued that certain proteins (e.g. intrinsically disordered proteins or less characterised proteins from the Dark Proteome) that generate shallow alignments cannot be studied by methods which depend on MSA information. However, this argument also extends to pLMs since the representation of proteins in the training database while pre-training the language models is still scarce. So, the biases that exist in the databases will also permeate into the large language models. The SeqVec model was trained on UniRef50 (~33M sequences); the model contained 93M parameters and was analysed on residue-level (word predictions) and

protein-level (sentence predictions). The results from these studies produced context-dependent embeddings (e.g. could be differentiated based on biophysical properties of amino acids) as compared to the embeddings generated from word2Vec architectures (e.g. ProtVec [68]).

TAPE [69] used a 12-layer transformer network (38M parameters) and was trained on 31M sequences from the Pfam database [70]. Further models were trained on larger protein sequence databases and deeper transformer models with even more training parameters.

The Evolutionary Scale Modelling (ESM) [71] representation was trained on ten times the number of protein sequences as previous models (~250 M UniParc [22] sequences) and had 652.4M model parameters. This deep contextualised pLM was trained with an MLM objective to learn the dependencies between amino acids. ESM embeddings were shown to encode biological properties, remote homology, protein family alignment, mutational effects and long-range contact information that inform secondary and tertiary structure prediction. Despite these large models of 650-700M parameters, the ESM authors believe they underfit the protein sequence database, suggesting the need to incorporate more sequences through metagenomic data and the potential for further increased model scale [71]. A subsequent release (ESM2) was trained with metagenomic sequences and the number of model parameters scaled to 15 billion (B).

ProtTrans [72] tested the limits of scaling up the number of proteins and parameters that can be used for training very large language models. The largest protein sequence database, Big Fantastic Database (BFD) [73], consisting of UniProt [22] and proteins from multiple metagenomic sequencing projects, was used to train models in ProtTrans. The training data comprised eight times the amount of data used for training previous pLMs. ProtTrans is suite containing numerous trained language models, including ProtTXL (based on Transformer-XL [74]), ProtBert (based on BERT model [43]), ProtXLNet (based on XLNet model [75]), ProtAlbert (based on ALBERT model [44]), ProtElectra (based on ELECTRA [47]), ProtT5 (based on T5 transformer model [48]). The parameters for these models range from 402M to 11B. Effects of auto-regressive and auto-encoding pre-training were compared. The success of combining pLM with explicit evolutionary information was reported, especially for predicting secondary structure, subcellular localisation and transmembrane regions. Protein-BERT [76], PRoBERTa [77] and ProGen [78] are additional pLMs that vary in training datasets and model parameters.

A large number of parameters does not necessarily mean better predictive models. Inverse-scaling of language models [79] state that flaws in training data or training objectives could lead to decreased performance

in spite of increased model scale. Therefore, very large pLMs such as ProtT5 (having 3B or 11B parameters) do not guarantee better predictive performance.

MSA Transformer: Statistical information from MSAs provide constraints on residue contacts that are used to infer the 3D structure of proteins (elaborated in Chapter 4). The advantage of these methods is that the model is fit on individual families of protein sequences, and covariation signals can be directly extracted from each case. In case of pLMs, which are trained on single sequences, many more parameters are needed to learn the co-evolutionary information. The MSA Transformer [80] combines the advantages of MSA-based statistical measures with pLMs. This pLM is trained on 26M MSAs generated for each UniRef50 sequence, and the model with 100M parameters is able to generalise across protein families. The MSA Transformer has proven successful in unsupervised contact prediction [80] and identifying phylogenetic relationships [81].

Evoformer: The input to AlphaFold [19] is the amino acid sequence, the MSA and template information. The core model comprises two parts: the Evoformer module and the Structure module. The Evoformer is an attention-based network responsible for creating a refined representation of the MSA embeddings and pairwise distances. The structure module is responsible for predicting the proteins' 3D structure using the refined pair and MSA representations. Without the use of templates, the Evoformer can be thought of as a sophisticated pLM where the pairwise distances within the protein and MSA representations are iteratively refined during each of its 48 blocks. Similar to the MSA Transformer, the Evoformer uses random masking on the input MSAs. Through rounds of row-wise and column-wise self-attention, these MSA-based pLMs consider relationships within the MSA and generate an 'evolution-aware' representation capable of predicting contact maps and secondary structure [82].

3.2 STRUCTURE-BASED REPRESENTATIONS

Residues that are widely separated within a protein sequence could, in fact, be in contact (distance less than 8Å) in 3D space as a folded structure. Though sequence representations described above contain prior knowledge about structure, explicitly taking structural features into account while representing proteins could allow for insights into the relative position of atoms, highlight possible ligand binding pockets or even represent residues that are exposed or buried within the structure.

Early structure representations included assigning a secondary structure via an objective algorithm called DSSP [83]. Based on H-bonding patterns and H-bond bridges between two residues that are not adjacent to each other, three classes of secondary structure were determined (alpha helices, beta sheets and coils). These 3-state secondary structure predictors were extended to 8-states having more granularity. Another feature unique to folded structures and that is also obtained from the DSSP software is solvent-accessible surface area (SASA) [84]. The relative solvent accessible area (RSA) (given for each residue by SASA/max(SASA) for the residue) gives the extent to which a residue is exposed or buried within the structure [85]. RSA has been used to derive hydrophobicity scales and is used for the prediction of protein-interface and protein-domain interactions [86, 87]. Other structural features used for protein property prediction include net charge and distribution of cationic patches on the protein surface [88].

3.2.1 *Graph representation*

Graphs allow for encoding diverse relationships between heterogeneous objects [89]. Therefore, either sequence or structural information can be represented in a graph format where nodes are used to represent atoms or residues in a protein and edges can be used to represent the relationships between the nodes.

Geometric learning is a sub-field of DL which is capable of generalising convolutional operations on efficient graph representations (graph convolutional networks (GCNs) [90]. The message parsing neural network (MPNN) applies an aggregation function to k-nearest neighbours (kNN), followed by a transformation function (e.g. sum, self-attention), where nodes are residues and edges encode for the relationship between the residues [91].

There are two prerequisites for protein graph representation. First, they need to have rotational and translational invariance. To satisfy the condition of invariance, when a function (e.g. GCN) is applied to a graph, the resulting value should not change when the graph has been rotated or translated in 3D space. Second, the graphs should be 'locally informative', i.e. the edge features from k neighbours should be enough to reconstruct node information for the residue of interest [91]. ProteinMPNN [92] enhanced the performance of the base MPNN by including distances between N, Cα, C, O, and virtual Cβ atoms instead of only using distances between Cα atoms. Another enhancement was replacing the fixed left-to-right

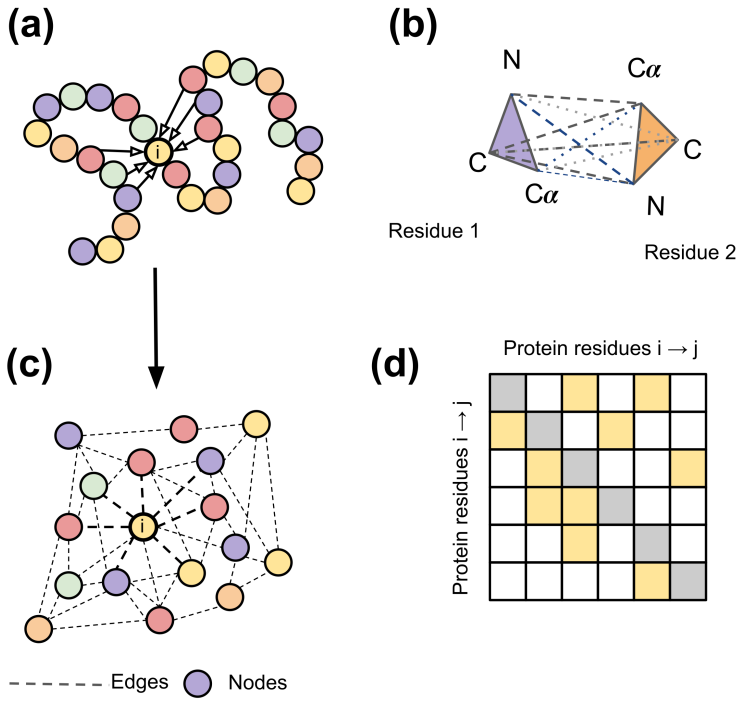

Figure 3.1: (a) Illustration about how neighbouring residues in a folded 3D protein structure can influence properties of a residue of interest *i* (b) Interactions between residues are essentially, interactions between the atoms in the residues. The interatomic distance between the residues provides information on which neighbours are the closest in 3D space (c) A folded 3D structure can be represented as a graph where the nodes can be residues and the edges can be the distance between the respective residues. A graph is defined by an adjacency matrix represented in (d). This is an all-vs-all matrix and the elements with the shortest distances (e.g. yellow boxes) comprise the *k*-nearest neighbours. Figure inspired by [91, 92]

autoregressive decoding to an order agnostic random decoding. The per-residue representations from ProteinMPNN, therefore, contain interatomic distances and information about its local neighbourhood (Figure 3.1).

3.2.2 *Structure-based pLMs*

Though it is possible to detect homology from protein sequences, structural similarity between 3D structures is more sensitive. With the increased availability of structural data (214M structures in AlphaFoldDB [21] and 617M metagenomic structures in ESM Atlas [93]), there is a need for

faster structural alignment tools. Catering to this need, Foldseek [94] was developed to increase speed by looking at the global similarity between protein structures. Foldseek uses a 20-state structural alphabet (3Di) to describe tertiary interactions. Each state represents the geometric conformation of residue i with respect to its spatially closest residue j [94]. Therefore, this method essentially converts structure into an intermediate sequence and performs sequence similarity between 3Di sequences. With the intention to combine 1D sequence information with 3D structural information, ProstT5 [95] trained a pLM model on Foldseeks' 3Di alphabet.

Other structure-based representations that have been used in protein structure prediction (especially in the latest CASP14) have been Voronoi tessellations [96], 2D manifolds, and point clouds (atoms as points) [89].

3.3 SEQUENCE- AND STRUCTURE-BASED REPRESENTATIONS

One of the early works which produced sequence representations based on structure information was [97]. In this study, using weak supervision based on global structural similarity, bidirectional LSTMs were used to learn protein sequence embeddings. Subsequently, numerous methods employing creative ways to combine structure and sequence information were published. DeepFRI [98] independently trained a LSTM language model using the sequence information and concatenated it in the first layer of a GCN using the graph representation of the structure. On the other hand, in spite of leveraging information from both sequence and structure, LM-GVP [99] co-models the language model and graph representation together.

AlphaFold produces a sequence representation (based on sequence and MSA) and can use structural templates as constraints in its pairwise representation. This method also uses combined representations for improved structure prediction, though the effect of templates is not large. An approach that has been explored in Paper v is to extract structure-based descriptors and study the performance of joint sequence and structure representations.

The representations described in this chapter have been used in one way or the other for various protein prediction tasks. The two main prediction tasks that are relevant to this thesis are protein structure prediction (Chapter 4) and protein binding site prediction (Chapter 5). The following chapters give an overview of the field and summarise how input representations can affect prediction performance.

4

PROTEIN STRUCTURE PREDICTION

As introduced in Chapter 1, the computational protein structure prediction problem involves finding tertiary structures from the primary amino acid sequence. This can be done with (template-based) or without (template-free) the use of templates. The correlations between amino acids in multiple columns in an MSA are essential for inferring evolutionary conserved contacts in 3D structures (coevolution) [100]. Briefly touched upon in the earlier chapter, classical statistical measures such as Mutual Information (MI) [101] have been used to find these coevolutionary signals in MSAs. A distance map (or contact map) is a two-dimensional (2D) matrix giving the distance between all pairs of residues in a protein tertiary structure. The presence of contact is given if two residues are less than 8 Å apart.

A significant leap was made in contact prediction when measures like Direct Coupling Analysis (DCA) [102–104], Sparse Inverse Covariance [105] and Pseudo Likelihood Maximisation [106] were able to distinguish between direct and indirect correlations. These methods, however, need very deep MSA with a large number of homologous sequences to be effective for contact-assisted tertiary structure prediction. This paved the way for contact-map prediction using deep learning approaches based on DCA (e.g. RaptorX, which used CNNs on PSSMs [107] or PconsC4 which used CNNs on a combination of sequential features [108]. The trRosetta [109] architecture pushed the performance of earlier methods further by training a multi-objective deep neural network to predict Cβ–Cβ distances and inter-residue orientations. Results from accurate contact map prediction provide constraints for accurate tertiary structure prediction.

4.1 ALPHAFOLD

Contrary to previous approaches, which divided the protein prediction task into first predicting contacts and then building a 3D structure from it, AlphaFold used an end-to-end training approach (Figure 4.1). Three MSAs

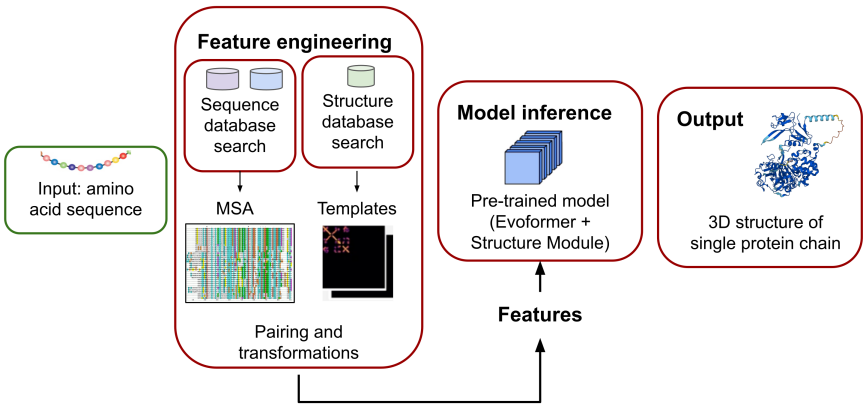

Figure 4.1: Illustration of the AlphaFold inference workflow to obtain structures for monomers. (Figure inspired by [19] and Google Vertex AI blog on AlphaFold)

serve as input to AlphaFold. These are generated from (a) JackHMMER [110] run on MGnify [111] (b) JackHMMER run on UniProt and (c) HHBlits [112] run on [Big Fantastic Database (BFD) + UniClust30]. The core trunk of the network (Evoformer with 48 blocks) visualises protein structures as graphs in 3D space.

The pair representation (L x L) encodes relations between all residues, where L is the length of the protein. The MSA representation (N x L) encodes the coevolutionary signal, where N is the number of sequences in the alignment. These representations are constantly updated by specific functions to each stack. Axial self-attention is applied to the MSA representations and triangular multiplicative updates, and triangular self-attention is applied to the pair representations. Spatial and evolutionary relationships are learned as a result of the outer product mean and attention applied to both representations, thereby allowing for continuous information exchange between the pair and MSA representations.

The output of the Evoformer (single and pair representation) is converted into a 3D representation (one N-Cα-C backbone frame per residue) in the structure module. The geometry-aware attention called 'Invariant Point Attention' (IPA), which acts on sets of frames, is responsible for making the representation equivariant to rotations and translations. The entire network focuses on local and iterative refinement of the structure. By producing structures with almost experimental accuracy, AlphaFold has more or less solved the problem of single-chain (monomer) protein structure prediction.

4.2 PLM-BASED STRUCTURE PREDICTORS

AlphaFold, however, takes considerable time and computational resources to generate models. A tremendous amount of resources is needed to generate a large number of models (e.g. human proteome (~65,000 models) took ~100,000 core hours on Sweden's National Supercomputer Centres' supercomputing resource Berzelius). Moreover, MSAs need to be generated for each protein that needs to be predicted.

In pLMs, since the model has been trained once, the saved parameters can be used directly on the protein sequence, leading to 10-60 times faster model generation. These are especially useful for large-scale predictions of proteomes or all metagenomic proteins (ESM Metagenomic Atlas with over 617M proteins [93]). The most popular pLM-based methods are ESMFold [93] and OmegaFold [113]. So far, the pLM-based protein structure predictors are not at par with AlphaFolds' performance for protein complexes (PAPER III).

4.3 MULTI-CHAIN PROTEIN STRUCTURE PREDICTORS

Proteins seldom exist alone in cells. They often interact with other proteins, forming a multi-chain (also called multimeric) complex. These multimers could either consist of all the same chains (homomers) or different chains (heteromers). Depending on the number of chains in the complex, multimers are classified as dimers (two chains), trimers (three chains), tetramers (four chains), pentamers (five chains) and hexamers (six chains), among others.

After AlphaFolds' release, researchers modified the inputs to the single-chain prediction pipeline. By providing two protein chains as input separated by a glycine linker, it was possible to model multimers by leveraging the potential coevolutionary signal between the individual chains. The three most popular and widely used methods include FoldDock [114], ColabFold [115], and AF2Complex [116]. FoldDock and ColabFold turned off templates and only relied on joint MSA representation. The combined MSA was obtained by pairing sequences belonging to the same organism (paired alignment) and arranging the MSA as disconnected blocks along a diagonal (block alignments). AF2Complex, on the other hand, showed that using only templates without any alignments was enough to predict accurate structures.

4.4 METRICS TO EVALUATE PROTEINS STRUCTURES

The reliability of prediction methods can be estimated by assessing the accuracy of their predicted models. The 'correctness' of the predicted models can be estimated by comparing them against the native ground truth (experimental structure). The metrics commonly used to estimate the quality of the models, such as RMSD, TM - score, and DockQ, measure how far the predicted structure is from the native structure [117].

- Root Mean Squared Error (RMSD) quantitatively measures (in Å) the similarity between two superimposed atomic coordinates.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {d_i}^2} \tag{4.1}$$

  where the distance between each pair of atoms (d_i) is averaged over n atom pairs.

  The limitation of RMSD is that it is sensitive to local variations between the two structures. For example, a single flexible loop differing between the structures does not allow accurate superimposition, resulting in a sizable global backbone RMSD [117]. RMSD is strongly affected by outliers. It also depends on the size of proteins and inflates distance for proteins of different sizes.

- Template-modelling score (TM-score) ignores outliers and eliminates the dependence on the size of the protein by normalising over the protein length.

$$TM - score = max[\frac{1}{L_{target}} \sum_{i=1}^{L_{aligned}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{target})}\right)^2}] \tag{4.2}$$

  where L_{target} is the length of the native structure, and $L_{aligned}$ is the length of the aligned region of the model. d_0 is a scaling factor given by $1.24(\sqrt[3]{L_{target} - 15}) - 1.8$

- DockQ is a continuous protein-protein docking model quality score [118] that combines three quality measures - F_{nat}, iRMS, LRMS. F_{nat} is the fraction of native contacts found in the interface of the predicted model. iRMS is the interface RMSD, where an interface is defined at a separation of 10 Å. LRMS is the Ligand RMSD calculated between

the backbone of the ligand chain superimposed on the receptor chain [118].

$$DockQ = \frac{(F_{nat} + RMS_{scaled}(LRMS, d_1) + RMS_{scaled}(iRMS, d_2)}{3} \tag{4.3}$$

where d_1 and d_2 are scaling parameters. RMS_{scaled} is given by:

$$RMS_{scaled}(RMS, d_i) = \frac{1}{1 + \left(\frac{RMS}{d_i}\right)^2} \tag{4.4}$$

where d_i is the scaling factor (d_1 optimised to 8.5Å for LRMS and d_1 optimised to 1.5Å for iRMS). The DockQ score ranges between 0 and 1. According to the CAPRI criteria, a DockQ score of above 0.23 is considered an acceptable model.

In cases where experimental structure is unavailable, we still need a measure that can estimate the quality of the predicted model. AlphaFold provides a predicted TM-score (pTM) as output to estimate the quality of the predicted model. The authors of FoldDock developed the predicted DockQ (pDockQ) score to estimate the quality of protein-protein interactions (PPI). This score was based on features from the AlphaFold model (e.g. residue-level confidence score plDDT and number of contacts). Therefore, the pDockQ score could be used to identify high-confidence dimeric complexes where predictions with scores above 0.5 were considered very confident.

4.5 IDENTIFYING PPI USING FOLDDOCK

The FoldDock pipeline and pDockQ score showed success in discriminating between interacting and non-interacting heterodimeric protein pairs with an AUC between 0.76 and 0.82 on E. coli and mammalian test sets [114]. The high performance of this method was applied to Toxin-Antitoxin (TA) systems in PAPER I and the human proteome in PAPER II. The following section is a primer on TA systems.

4.5.1 *Toxin-Antitoxin (TA) systems*

TA systems consist of two proteins, a toxin and an antitoxin, that are encoded by adjacent or overlapping genes. The expression of toxin leads to cell death, and the antitoxin binds to the toxin so as to neutralise it. Bacterial TAs play a role in cell cycle arrest, tolerating environmental

stresses such as antibiotic treatments, lack of nutrients or even phage defence mechanisms [119]. Based on the mechanism used by the antitoxin to neutralise the toxin and whether the TA are protein or RNA-based, there are different types of TA systems. Type I systems involve a non-coding RNA antitoxin binding to the mRNA that encodes for the toxin, thus inhibiting its translation of the toxin. Type II systems are the most common group and involve toxin proteins that are inhibited by directly being bound by an antitoxin protein. Type III systems involve an mRNA-based antitoxin binding to a toxin protein. Under stressful conditions, the antitoxin gets degraded, and the toxin is no longer tightly regulated, leading to catastrophic responses. Therefore, in normal conditions, it is essential for TA to be a very stable complex with a strong coevolutionary signal.

SUMMARY OF PAPER I:

Type II TA systems are modular in nature, and there are promiscuous antitoxin domains that can swap partners and neutralise structurally and evolutionarily unrelated toxins. In PAPER I, a gene neighbourhood conservation algorithm (FlaGs) [120] was iteratively (NetFlax) used to identify a network of potential TA partner swapping. The structures of the TA pairs in this network (3,597 protein pairs) were predicted using the FoldDock pipeline. Despite possible higher-order interactions occurring in TA systems, since at least one interface exists between each toxin and antitoxin, only binary TA pairs were predicted. Based on the plDDT residue-level confidence score from AlphaFold and the number of contacts at the interface, the overall interaction confidence score pDockQ was calculated for the dimer complexes. The distribution of the pDockQ of the true TA pairs was much better than the random distribution with non-interacting TA pairs, with about 60% of the models predicted accurately. Through experimental validation of the predicted structures, the mechanism of interaction (direct as well as indirect neutralisation) for the hyperpromiscuous Panacea-containing antitoxins (PanA) was proposed using mutagenesis and functional assays.

4.5.2 *Human PPI interactions*

The complete set of proteins expressed in humans is referred to as the human proteome. Large-scale efforts such as [121][122] have aimed to identify these proteins experimentally. Creating a network to map how

these proteins interact can provide insights about their function. Experimental techniques such as yeast two-hybrid (Y2H) and mass spectrometry have been used to identify PPIs. Human Reference Interactome (HuRI) [123] and the Human Protein Complex Map (hu.MAP v.2.0) [124] have used these experimental methods to create a large-scale reference map of PPIs in human proteomes. Identifying the structures of these PPI allows for studying protein mechanisms, the impact of disease-causing mutations and post-translational modifications [125]. However, only 5% of these human PPIs interactions have been structurally characterised.

SUMMARY OF PAPER II:

In PAPER II, the FoldDock protocol was used for large-scale structure prediction of the human interactome. The PPIs from the HuRI (with 55,586 binary interactions) and hu.MAP (with 10,207 PPIs) were selected for this study. Predictions were made for 65,484 protein pairs, and the pDockQ score (with a 0.5 threshold) was used to identify the high-confidence interactions. Protein pairs from the hu.MAP, on average, generated higher confidence models than protein pairs from the HuRI set. They both reported better pDockQ than complexes created using randomly shuffled protein pairs. A PPI network using these high-confidence models was generated, serving as a resource for further experimental analysis of stable complexes.

4.6 ALPHAFOLD-MULTIMER

AlphaFold was extended to predict multi-chain complexes (AlphaFold-Multimer) [126] by adapting their model with support for pairing MSA alignments between different protein chains and taking stoichiometry into account. The MSAs of chain A and chain B are paired using the UniProt species annotation. The permutation of identical chains that align best with the ground truth is chosen since optimising over all possible permutations is not feasible. A positional encoding describing which chain a residue belongs to is added to the initial pair representation to ensure breaking symmetry between identical chains. Similar to the quality score of pTM, AlphaFold-Multimer provides an additional score called interface pTM (ipTM) to score the quality of the predicted interface. A weighted combination of pTM and ipTM gives the overall ranking confidence score of the model. Additionally, AlphaFold-Multimer also provides the relative error between the chains through a Predicted Aligned Error (PAE) matrix. Though the authors of AlphaFold-Multimer tested their results on recent

PDB proteins, it lacked a rigorous benchmark on a homology-reduced test dataset. In case the Recent PDB proteins they tested on were very similar to multimers they used for training, the performance of this method would be overestimated. Hence, there was a need for an independent validation of protein complex predictors.

SUMMARY OF PAPER III:

A homology-reduced and structurally non-redundant benchmark dataset using structures after the training date of AlphaFold-Multimer was created to benchmark the performance of AlphaFold-Multimer (PAPER III). The dataset was classified as per oligomeric state, structural redundancy within each group was removed, and homology-reduction with respect to the AlphaFold-Multimer training set was performed. Global (TM-score) and local (DockQ) similarity scores were used to evaluate the models. However, since each interface in higher order complexes generates one DockQ score, an analysis on reporting the DockQ score for multimers was conducted ($DockQ_i$ vs $DockQ_{ij}$). Therefore, in PAPER III, the difference between pairwise and multi-chain evaluation of chains was highlighted. AlphaFold-Multimer performed better than other multimer predictors, but the success rates for larger complexes were lower, albeit not by much. The pDockQ score overestimated the performance of the predicted models for multimers. Therefore, an updated confidence score (pDockQ2) was proposed that more accurately estimated the quality for each interface in a multimer.

5

LIGAND BINDING SITE (LBS) PREDICTION

Proteins bind to biomolecules (e.g. nucleic acids and small molecules) or ions (e.g. metal ions) at specific sites to carry out their functions. These distinct sites, called ligand binding sites (LBS), are a feature of the properties of the surrounding residues and their intermolecular forces. Ligands are usually found in 3D pockets of the protein structure and have been a prime research focus for molecular docking, compound design, affinity predictions and molecular dynamics simulations [127]. Identifying ligand binding sites with high binding affinity are 'druggable' targets and have applications in drug design [128]. Since experimental methods to identify such sites are time-consuming and expensive, computational prediction of protein-ligand interactions (PLI) can accelerate the screening by reducing the search space. The challenges of predicting PLI are (a) the need for manual curation and verification of data, (b) incomplete or unreliable training data and (c) the lack of reliable scoring metrics evaluating protein and ligand poses simultaneously [129].

5.1 PLI DATABASES

BioLip [130] is a commonly used database to obtain annotations for creating training, validation and testing datasets to train models for binding-site prediction. It is a semi-manually curated database of interactions based on PDB entries and manual review to filter only biologically relevant interactions. The residues are annotated as a ligand binding site if the distance between any atoms of the ligand molecules and at least one of the atoms of the amino acid residues is not more than the sum of the atom radii + 0.5 Å. The annotations from BioLip have been used to train the models in Paper IV and Paper V. Binding MOAD [131] and LIGand attachment site (LigASite) [132] are other databases with binding-site information.

5.2 COMPUTATIONAL LBS PREDICTION

Computational approaches to predict binding sites can broadly be classified into three groups depending on what type of feature they use: (a) structure (or template), (b) sequence, and (c) both sequence and structure-based.

5.2.1 *Traditional approaches*

Template (or ligand) based approaches are similarity-based approaches. These methods look through a list of candidate structures or molecules and predict interactions based on similarity in structural pockets (or biological compounds). Molecular docking [133] is a key tool for finding binding conformations by scouring the protein surface. Based on the binding strength, it scores and ranks docking poses of ligands in proteins' 3D structures. Approaches which aim to find the energetically favourable binding surfaces (energy-based) or find cavities based on the geometry of the molecular surface (geometry-based) have also been used to identify LBS [127].

Sequence-based methods rely on the assumption that ligand-binding residues are evolutionarily conserved to maintain their function. Many methods have exploited evolutionary information such as PSSM, secondary structure and propensity of residues binding to specific ligands, in addition to sequence properties like amino acid composition and physicochemical properties for sequence-based binding site prediction [134]. Since the properties of a residue are often influenced by its neighbouring residues, many methods (e.g. TargetS [135], MIonSite [136] and IonSeq [137]) use a sliding window approach to construct input features using the above amino acid properties from adjacent residues to train their ML algorithms.

Structure-based methods are generally more accurate than sequence-based methods. Until recently, it was not possible to use structure-based methods for proteins that lacked similar structures in PDB. The release of AlphaFold has mitigated this problem. Large-scale screening for multiple ligands across millions of protein structures requires substantial time and computational resources. So accurate and fast sequence-based predictors are of research interest. Sequence-based methods can identify the residues involved in binding but do not describe the orientation of the ligand in the binding pocket. Hybrid approaches utilise evolutionary conserved sequence features and protein surface properties to identify binding pockets. For example, COACH [138] combines the predictions from TM-SITE (uses

binding substructures) and S-SITE (uses sequence profiles) to identify LBS. IonCon [137] is yet another hybrid method.

5.2.2 *Recent advances in LBS prediction*

With the increase in PDB structures, the structure and sequence approaches to identify LBS have seen a leap in performance. Increased data and robust DL networks have enabled the prediction of classes having too few samples. The superimposition of query proteins on similar templates helps identify potential binding pockets. Methods like MIB2 [139] rely on fragment transformation methods [140] to identify local structural motifs. AlphaFill [141] annotates AlphaFold models with small molecules or ions by using structural and sequence similarity to known experimental structures. Metals comprise more than 30% of the BioLip database, making it the most prevalent category for binding site prediction. mebipred [142] is a sequence-based MLP classifier using amino acid properties and counts of metal-binding 5-mers.

5.2.3 *pLM-based LBS predictors*

The pLMs representations contain contextual information, are fast to generate and have shown good performance in a variety of protein prediction tasks (e.g. subcellular localisation [143], post-translational modification [144] and signal peptide prediction [145]). pLM embeddings have also been used as input to deep neural networks to predict binding sites. The authors of bindEmbed21DL [146] noted that ProtT5 embeddings outperformed MSA-based prediction of residues binding to metals, nucleic acids or small molecules. LMetalSite [147] used ProtTrans embeddings as input to a transformer model followed by one MLP network for each ion to predict binding residues for the four most frequent ions.

SUMMARY OF PAPER IV:

M-Ionic (PAPER IV) provides the probability of a protein binding to a metal ion. It also gives the residue-level binding probabilities to multiple ions simultaneously. The study involved creating a homology-reduced training, validation and independent test set. Instead of passing the ESM2 embeddings of the entire protein into the network, M-Ionic was trained by batching residues together, irrespective of whether they belonged to the same protein or not. M-Ionic predicts the binding probabilities for six additional ions compared to LMetalSite [147]. Extensive validation was

performed by studying the impact of explicit evolutionary and structural information. The impact of mutating binding residues on prediction performance was also investigated. A key insight obtained from this study is that residue-level embeddings contain neighbourhood information.

Methods such as [148–150] combine pLM representations with structure-based representations for LBS prediction. Proteins are represented as graphs where the node features are sequence-based pLM representations or are structure-based descriptors obtained from AlphaFold models. A graph-based neural network is then trained to identify whether the nodes bind to a certain ligand (node classification problem).

SUMMARY OF PAPER V:

Combining sequence and structure representations has the potential to complement each other and overcome their individual limitations. In PAPER V, structural descriptors were appended to the residue-level ProtT5 embeddings for proteins from the bindEmbed21DL [146] datasets. The joint representations were trained to differentiate between metal, small molecule and nucleic acid binding residues. The impact of structure-informed residue-level sequence representations on LBS prediction was examined.

5.2.4 *Diffusion-based and co-modelling approaches*

Despite the success of pLM-based approaches, current methods that have shown promising results use diffusion-based generative modelling or end-to-end co-modelling approaches. Diffusion models jointly learn the distribution of protein and ligand conformations, thus sampling through the distribution of docked ligand poses. DiffDock [151] learns the reverse diffusion process by progressively refining over random poses with updates in translation, rotation and torsion angles. Another approach called RoseTTAFold ALL-Atom (RFAA) [152] uses an atomic graph representation for co-modelling ligands and proteins in complexes. RFAA uses a uniform molecular atoms format with information about element types, bond types and chirality and a label indicating whether the atom belongs to a ligand or protein. The molecular input is passed through a modified 3D Track SE3 transformer [153] as in RoseTTAFold [154]. Rigorous and independent benchmarking on these methods is yet to be done.

6

FUTURE PERSPECTIVES

Over the last few years, AlphaFold caused a tectonic shift in the structural biology field. By essentially solving the protein folding problem for monomers, it paved the way to solve more complex challenges. The categories introduced in CASP15 are indicative of the next challenges being tackled by the community. These include the prediction of protein assemblies, interactions with disordered regions, protein-ligand interactions and RNA structure prediction. Proteins do not have a fixed state (as captured by structures seen in PDB); they are dynamic. So, understanding how proteins move within the cell (protein dynamics) is a key area of interest. Proteins also exist in various conformations. Studies [155–158] have shown that sampling the MSA can generate potential protein conformations. Most recently, we have seen a convergence of experimental tools (e.g. Cryo-EM) and computational tools for understanding protein dynamics through multiple conformations [159]. A prediction of open and closed conformations, for example, could provide insights to unravel the functions that govern the cellular mechanisms.

As the size of databases and computational power increases, the reliance on machine learning methods to make sense of the datasets also increases. So far, ML advances have always superseded the significant advances in structural biology. Though there is quick adoption of the latest ML algorithms to structural biology problems, creating domain-informed, standardised datasets like ProteinNet [160] would only increase the pace of development. Bridging together experts from various disciplines to work on the same problem using a multi-disciplinary approach proved critical for the success of AlphaFold.

Learning effective representations of data remains an active area of research for generative artificial intelligence (AI). Representation learning helps unlock patterns in a large corpus of data by representing words in an abstract, high-dimensional space. The principles of RL and NLP applied to the protein language have shown promising results but are not yet comparable to the performance of structure-based methods. Sequence

pLMs contain evolutionary information [146]. However, pLMs do not consistently outperform evolution-based methods and are not generalisable in all contexts. Larger pLMs with more parameters also do not seem to improve performance substantially (e.g. ProtT5 [72]). Therefore, it appears that sequence-based pLMs have reached their limit, and other modalities with biological priors are needed to boost prediction performance further. As noted in [161], no single modality is enough to predict all protein prediction tasks. Effectively combining evolutionary information, pLM embeddings, and structure information could help understand how proteins fold and function within the cell.

Biologists often have a framework to understand how systems work [162]. The system is broken apart, the outcome of the intervention is observed, and a hypothesis for further testing is generated. The approach to understanding how AlphaFold and pLMs work has been similar - to study edge cases where these systems fail. However, by designing modular and transparent systems from the onset can go a long way in understanding the choices made by the algorithm. As a result, experimentalists can be better informed while using computational methods to solve biological problems. Therefore, explainable AI in structural biology would be an exciting area to look out for.

Though numerous innovative adaptations of AlphaFold have allowed researchers to tackle the upcoming challenges in structural biology, the ultimate goal would be to model all interactions within the biological cell. One could envision this as creating a digital cell with an all-vs-all dynamic approach, taking into account all proteins, ligands and other molecules simultaneously. Further, as generative AI gains momentum, generative adversarial networks and variational autoencoders are showing promise for designing proteins with specific properties [163]. Unconditional language models have learned the protein landscape and can generate sequences in low-probability regions. These models can be conditioned with desired properties to computationally limit the landscape (machine learning-assisted directed evolution). Imagine prompting a general-purpose algorithm to simulate proteins with a particular property, stability and interaction in a digital environment akin to the biological cell. Although this can seem improbable at the moment, as the scientific goalpost keeps moving forward, such a scenario may be closer than expected.

7

SAMMANFATTNING

För att förstå vad som händer inuti en biologisk cell, så kan vi jämföra med Stockholms stad. Omkring 1 miljon människor bor i Stockholm. Cirka 70 % av befolkningen är sysselsatta. Vanligen går varje invånare upp på morgonen, tar en kopp kaffe och går till jobbet. Varje individ har ett särskilt jobb att utföra. De arbetar i livsmedelsbutiker, banker, byggverksamhet, restauranger eller något slags företag. Stockholm har många öar, samanbundna med bror där det går tåg, bussar och spårvagnar. Många människor är beroende av andra individer, som t.ex. arbetar i kollektivtrafiken, för att ta sig från sina hem till arbetsplatser. Dagliga handlingar i en stad är resultatet av att varje individ utför sitt jobb varje dag. Om några individer blir sjuka eller maskiner upphör att fungera som förväntat, utbryter kaos och det uppstår störningar i hur staden fungerar. Men för det mesta fungerar allt smidigt.

Precis som människor är arbetshästar i vilken stad som helst, är proteiner arbetshästarna i en cell. De är komplexa biomolekyler som spelar en viktig roll i de flesta biologiska processer. Proteiner består av en kedja av aminosyror (även kallade rester) som är sammanbundna av peptidbindningar. För att fungera så måste proteinkedjan veckas ihop för att få en specifik tredimensionell (3D) struktur. Proteiner är involverade i olika processer, allt från signalering, transport, metabolism och katalys till rudimentära funktioner som celldelning. Om ett protein inte fungerar korrekt störs dess funktioner, vilket kan leda till sjukdomar.

Proteinernas sekvens avgör hur det kommer att veckas till sin unika tredimensionella struktur. När de väl är veckande är miljontals proteiner i cellen konstant i rörelse och interagerar med många andra biomolekyler för att utföra sina respektive funktioner. Proteiner rör sig mellan olika delar av cellen. Experimentell bestämning av proteinstruktur och funktion är tidskrävande. Beräkningsmetoder syftar till att förutsäga egenskaperna hos proteiner för att komplettera experimentella metoder för att förstå biokemiska mekanismer inom cellen. Denna avhandling introducerar

beräkningstekniker som förutsäger strukturen av proteinkomplex och identifierar proteinrester involverade i interaktioner med andra biomolekyler, såsom metalljoner och nukleinsyror, direkt från sekvensinformationen. Forskningen som utförs inom denna avhandling ger värdefulla insikter om hur man använder maskininlärning för proteinbioinformatik.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Arne. Thank you for this incredible opportunity to do a PhD in your lab. I am grateful for your perceptive supervision, candour and compassion. Thank you for the freedom to explore and for the countless opportunities to learn and grow into the person I am today.

I sincerely thank Petras for his meticulous insights and valuable suggestions that contributed to advancing my research. I am grateful to Belén, Lukas, Pia, and Samuel (Flores) for their advice and helpful discussions. I also thank my opponent, Geoff Barton, and the committee members for taking the time to read my work.

To David and Marco - I deeply appreciate you helping me get started in the lab. Thank you for being patient, supportive and very encouraging mentors. I am grateful to Sudha, Claudio, John, Erik, and Saman for creating a warm and welcoming environment. Gabriele, thank you for being the most cheerful person in the lab. A very special thanks goes to my work sister, Wensi. Thank you for always being there. I am grateful to Sarah and Samuel (Fromm) for making my last year the most enjoyable. I also sincerely thank my collaborators, especially Yogesh, Gemma, and Bhanu. I would be remiss not to mention colleagues and friends at DBB and SciLifeLab who have contributed to a wholesome PhD.

I am incredibly grateful to have dear friends who have kept me motivated, grounded and sane over the years. A big thank you to Hélène, Pablo and Ana - your company brings me so much joy. I am grateful to Imran and Kätlin for all the heartfelt conversations. I cherish the moments we spend together. Thank you to my mentor and friend Renuka for supporting and inspiring me throughout. I am thankful to Tanja, Lukas, Shubhaa, Savreen, Fabio, and Mariana for all the delightful times together. Around the world, yet extremely close to my heart - Aditya, Sudharshinee, Aathira, Arun, Akshat, Niharika, and Prakhar - thank you for always being there.

I am immensely grateful to my entire family for their endless love. A special mention to Mummy, Papa, and Apeksha for their constant support. Thank you for taking me in wholeheartedly and for graciously providing me with another place to call home.

To the extraordinary duo who made everything possible with their abundant love, patience, understanding, and encouragement - Amma and Papa - thank you for being my safe haven no matter how far apart we are. The qualities of discipline, grit, balance and faith you instilled in me were key in getting me this far.

Lastly, and most importantly, to my dear husband and best friend, Mukund - you continually inspire me with your drive, discipline, and childlike curiosity. Thank you for being my numero uno champion and for picking me up (often literally!) amidst the highs and lows of my academic journey. You make every day and everything better.

1. Dunker, A. K., Silman, I., Uversky, V. N. & Sussman, J. L. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology. Catalysis and regulation / Proteins* **18,** 756–764. ISSN: 0959-440X (1st Dec. 2008).
2. Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **338,** 1042–1046 (23rd Nov. 2012).
3. Newberry, R. W. & Raines, R. T. Secondary Forces in Protein Folding. *ACS chemical biology* **14,** 1677–1686. ISSN: 1554-8929 (16th Aug. 2019).
4. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Structural Biology* **4,** 10–19. ISSN: 1545-9985 (Jan. 1997).
5. Sorokina, I., Mushegian, A. R. & Koonin, E. V. Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process? *International Journal of Molecular Sciences* **23,** 521. ISSN: 1422-0067 (4th Jan. 2022).
6. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (New York, N.Y.)* **181,** 223–230. ISSN: 0036-8075 (20th July 1973).
7. Brändén, C.-I. & Tooze, J. *Introduction to protein structure* 2nd ed. 410 pp. ISBN: 978-0-8153-2304-4 978-0-8153-2305-1 (Garland Pub, New York, 1999).
8. Levinthal, C. How to fold graciously. *Mossbauer Spectroscopy in Biological Systems Proceedings, vol. 67, no. 41, pp. 22–26* (1969).
9. Wolynes, P. G., Luthey-Schulten, Z. & Onuchic, J. N. Fast-folding eriments and the topography of protein folding energy landscapes. *Chemistry & Biology* **3,** 425–432. ISSN: 1074-5521 (1st June 1996).
10. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28,** 235–242. ISSN: 0305-1048 (1st Jan. 2000).
11. Jones, D. T., Taylort, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **358,** 86–89. ISSN: 1476-4687 (July 1992).
12. Martí-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure* **29,** 291–325. ISSN: 1056-8700 (2000).
13. Pearce, R., Li, Y., Omenn, G. S. & Zhang, Y. Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. *PLOS Computational Biology* **18,** e1010539. ISSN: 1553-7358 (16th Sept. 2022).
14. Cauchy, M. A. Me´thode ge´ne´rale pour la re´solution des syste‘mes d’e´quations simultane´es.
15. Curry, H. B. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics* **2,** 258–261. ISSN: 0033-569X, 1552-4485 (1944).

16. Burley, S. K. *et al.* Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students. *Biomolecules* **12,** 1425. ISSN: 2218-273X (4th Oct. 2022).

17. Moore, G. Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE* **86,** 82–85. ISSN: 0018-9219, 1558-2256 (Jan. 1998).

18. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics* **23,** ii–iv. ISSN: 1097-0134 (1995).

19. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 1–7. ISSN: 1476-4687 (1st Aug. 2021).

20. Callaway, E. After AlphaFold: protein-folding contest seeks next big breakthrough. *Nature* **613,** 13–14 (13th Dec. 2022).

21. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50,** D439–D444. ISSN: 0305-1048 (D1 7th Jan. 2022).

22. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** D158–D169. ISSN: 0305-1048 (D1 4th Jan. 2017).

23. Clusmann, J. *et al.* The future landscape of large language models in medicine. *Communications Medicine* **3,** 1–8. ISSN: 2730-664X (10th Oct. 2023).

24. Wolpert, D. & Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1,** 67–82. ISSN: 1089778X (Apr. 1997).

25. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61,** 85–117. ISSN: 0893-6080 (1st Jan. 2015).

26. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444. ISSN: 1476-4687 (May 2015).

27. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology* **195,** 215–243. ISSN: 0022-3751 (Mar. 1968).

28. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36,** 193–202. ISSN: 1432-0770 (1st Apr. 1980).

29. Atlas, L., Homma, T. & Marks, R. *An Artificial Neural Network for Spatio-Temporal Bipolar Patterns: Application to Phoneme Classification Neural Information Processing Systems* **0** (American Institute of Physics, 1987).

30. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536. ISSN: 1476-4687 (Oct. 1986).

31. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computation* **9,** 1735–1780. ISSN: 0899-7667 (15th Nov. 1997).

32. Vaswani, A. *et al. Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 5998–6008 (Curran Associates, Inc., 2017).

33. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60,** 84–90. ISSN: 0001-0782, 1557-7317 (24th May 2017).

34. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* **5,** 157–166. ISSN: 1045-9227, 1941-0093 (Mar. 1994).

35. Pennington, J., Socher, R. & Manning, C. *GloVe: Global Vectors for Word Representation Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* EMNLP 2014 (eds Moschitti, A., Pang, B. & Daelemans, W.) (Association for Computational Linguistics, Doha, Qatar, Oct. 2014), 1532–1543.

36. Mikolov, T., Chen, K., Corrado, G. & Dean, J. *Efficient Estimation of Word Representations in Vector Space* 6th Sept. 2013.

37. Peters, M. E. *et al. Deep contextualized word representations* 22nd Mar. 2018.

38. Bahdanau, D., Cho, K. & Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate* 19th May 2016.

39. Luong, M.-T., Pham, H. & Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv:1508.04025 [cs]* (20th Sept. 2015).

40. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Systems* **12,** 654–669.e3. ISSN: 2405-4712 (16th June 2021).

41. Harris, Z. S. Distributional Structure. *WORD* **10,** 146–162. ISSN: 0043-7956, 2373-5112 (Aug. 1954).

42. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training.

43. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (24th May 2019).

44. Lan, Z. *et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* 8th Feb. 2020.

45. Liu, Y. *et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach* 26th July 2019.

46. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter* 29th Feb. 2020.

47. Clark, K., Luong, M.-T. & Le, Q. V. ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS (2020).

48. Raffel, C. *et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* 19th Sept. 2023.

49. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4,** 279–287. ISSN: 2522-5839 (Mar. 2022).

50. Alwosheel, A., van Cranenburgh, S. & Chorus, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* **28,** 167–182. ISSN: 1755-5345 (1st Sept. 2018).

51. Detlefsen, N. S., Hauberg, S. & Boomsma, W. Learning meaningful representations of protein sequences. *Nature Communications* **13,** 1914. ISSN: 2041-1723 (8th Apr. 2022).

52. Rost, B. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection* **12,** 85–94. ISSN: 1741-0134, 1741-0126 (Feb. 1999).

53. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21,** 6. ISSN: 1471-2164 (2nd Jan. 2020).

54. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27,** 861–874. ISSN: 01678655 (June 2006).

55. Dubchak, I., Muchnik, I. & Kim, S. H. Protein folding class predictor for SCOP: approach based on global descriptors. *Proceedings. International Conference on Intelligent Systems for Molecular Biology* **5,** 104–107. ISSN: 1553-0833 (1997).

56. Eisenhaber, F., Frömmel, C. & Argos, P. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins: Structure, Function, and Bioinformatics* **25,** 169–179. ISSN: 1097-0134 (1996).

57. Fujiwara, Y. & Asogawa, M. Prediction of subcellular localizations using amino acid composition and order. *Genome Informatics. International Conference on Genome Informatics* **12,** 103–112. ISSN: 0919-9454 (2001).

58. Fischer, D. & Eisenberg, D. Protein fold recognition using sequence-derived predictions. *Protein Science : A Publication of the Protein Society* **5,** 947–955. ISSN: 0961-8368 (May 1996).

59. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5,** 823–826. ISSN: 0261-4189 (Apr. 1986).

60. Rost, B. & Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences* **90,** 7558–7562 (15th Aug. 1993).

61. Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics* **40,** 502–511. ISSN: 1097-0134 (2000).

62. Pereira, J. & Alva, V. How do I get the most out of my protein sequence using bioinformatics tools? *Acta Crystallographica Section D Structural Biology* **77,** 1116–1126. ISSN: 2059-7983 (1st Sept. 2021).

63. Mirabello, C. & Wallner, B. rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. *PLOS ONE* **14,** e0220182. ISSN: 1932-6203 (15th Aug. 2019).

64. Ju, F. *et al.* CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nature Communications* **12,** 2535. ISSN: 2041-1723 (5th May 2021).

65. Zhang, C. *et al.* DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics.*

66. Alley, E. C. *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16,** 1315–1322. ISSN: 1548-7105 (Dec. 2019).

67. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20,** 723. ISSN: 1471-2105 (17th Dec. 2019).

68. Asgari, E. & Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE* **10,** e0141287. ISSN: 1932-6203 (10th Nov. 2015).

69. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. *arXiv:1906.08230 [cs, q-bio, stat]* (19th June 2019).

70. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Research* **36,** D281–D288. ISSN: 0305-1048 (Database issue Jan. 2008).

71. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118.** ISSN: 0027-8424, 1091-6490 (13th Apr. 2021).

72. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44,** 7112–7127. ISSN: 1939-3539 (Oct. 2022).

73. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* **16,** 603–606. ISSN: 1548-7105 (July 2019).

74. Dai, Z. *et al. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context* 2nd June 2019.

75. Yang, Z. *et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding* 2nd Jan. 2020.

76. Brandes, N. *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38,** 2102–2110. ISSN: 1367-4803 (12th Apr. 2022).

77. Nambiar, A. *et al. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks* preprint (Bioinformatics, 16th June 2020).

78. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* **41,** 1099–1106. ISSN: 1546-1696 (Aug. 2023).

79. McKenzie, I. R. *et al. Inverse Scaling: When Bigger Isn't Better* 15th June 2023.

80. Rao, R. *et al.* MSA Transformer. *bioRxiv,* 2021.02.12.430858 (13th July 2021).

81. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nature Communications* **13,** 6298. ISSN: 2041-1723 (22nd Oct. 2022).

82. Hu, M. *et al. Exploring evolution-based & -free protein language models as protein function predictors* 13th June 2022.

83. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637. ISSN: 0006-3525, 1097-0282 (Dec. 1983).

84. Lee, B. & Richards, F. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology* **55,** 379–IN4. ISSN: 00222836 (Feb. 1971).

85. Tien, M. Z. *et al.* Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE* **8,** e80635. ISSN: 1932-6203 (21st Nov. 2013).

86. Chothia, C. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology* **105,** 1–12. ISSN: 0022-2836 (25th July 1976).

87. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *Journal of Molecular Biology* **196,** 641–656. ISSN: 0022-2836 (5th Aug. 1987).

88. Bhardwaj, N. *et al.* Structural Bioinformatics Prediction of Membrane-Binding Proteins. *Journal of molecular biology* **359,** 486–495. ISSN: 0022-2836 (2nd June 2006).

89. Laine, E., Eismann, S., Elofsson, A. & Grudinin, S. Protein sequence-to-structure learning: Is this the end(-to-end revolution)? *Proteins: Structure, Function, and Bioinformatics* **89,** 1770–1786. ISSN: 1097-0134 (2021).

90. Kipf, T. N. & Welling, M. *Semi-Supervised Classification with Graph Convolutional Networks* 22nd Feb. 2017.

91. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design.

92. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** 49–56 (7th Oct. 2022).

93. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (17th Mar. 2023).

94. Van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology,* 1–4. ISSN: 1546-1696 (8th May 2023).

95. Heinzinger, M. *et al. ProstT5: Bilingual Language Model for Protein Sequence and Structure* 25th July 2023.

96. Olechnovič, K. & Venclovas, Č. VoroIF-GNN: Voronoi tessellation-derived protein–protein interface assessment using a graph neural network. *Proteins: Structure, Function, and Bioinformatics* **n/a.** ISSN: 1097-0134.

97. Bepler, T. & Berger, B. *Learning protein sequence embeddings using information from structure* 16th Oct. 2019.

98. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nature Communications* **12,** 3168. ISSN: 2041-1723 (26th May 2021).

99. Wang, Z. *et al.* LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports* **12,** 6832. ISSN: 2045-2322 (27th Apr. 2022).

100. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18,** 309–317. ISSN: 0887-3585 (Apr. 1994).

101. Gomes, M., Hamer, R., Reinert, G. & Deane, C. M. Mutual information and variants for protein domain-domain contact prediction. *BMC Research Notes* **5,** 472. ISSN: 1756-0500 (31st Aug. 2012).

102. Weigt, M. *et al.* Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106,** 67–72 (6th Jan. 2009).

103. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108,** E1293–E1301 (6th Dec. 2011).

104. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* **6,** e28766. ISSN: 1932-6203 (7th Dec. 2011).

105. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190. ISSN: 1367-4803 (15th Jan. 2012).

106. Ekeberg, M. *et al.* Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* **87,** 012707 (11th Jan. 2013).

107. Wang, S. *et al.* Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology* **13,** e1005324. ISSN: 1553-7358 (5th Jan. 2017).

108. Michel, M., Menéndez Hurtado, D. & Elofsson, A. PconsC4: fast, accurate and hassle-free contact predictions. *Bioinformatics* **35,** 2677–2679. ISSN: 1367-4803 (1st Aug. 2018).

109. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117,** 1496–1503 (21st Jan. 2020).

110. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11,** 431. ISSN: 1471-2105 (18th Aug. 2010).

111. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* **48,** D570–D578. ISSN: 0305-1048 (D1 8th Jan. 2020).

112. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9,** 173–175. ISSN: 1548-7105 (Feb. 2012).

113. Wu, R. *et al. High-resolution de novo structure prediction from primary sequence* 22nd July 2022.

114. Bryant, P., Pozzati, G. & Elofsson, A. *Improved prediction of protein-protein interactions using AlphaFold2* 1 (10th Mar. 2022), 1265.

115. Mirdita, M. *et al. ColabFold - Making protein folding accessible to all* (bioRxiv, 8th Feb. 2022), 2021.08.15.456425.

116. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nature Communications* **13,** 1744. ISSN: 2041-1723 (1st Apr. 2022).

117. Kufareva, I. & Abagyan, R. Methods of protein structure comparison. *Methods in molecular biology (Clifton, N.J.)* **857,** 231–257. ISSN: 1064-3745 (2012).

118. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE* **11** (ed Levy, Y. K.) e0161879. ISSN: 1932-6203 (25th Aug. 2016).

119. Lee, K.-Y. & Lee, B.-J. Structure, Biology, and Therapeutic Application of Toxin–Antitoxin Systems in Pathogenic Bacteria. *Toxins* **8,** 305. ISSN: 2072-6651 (Oct. 2016).

120. Saha, C. K. *et al.* FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation. *Bioinformatics* **37,** 1312–1314. ISSN: 1367-4803 (9th June 2021).

121. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581. ISSN: 1476-4687 (May 2014).

122. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419 (23rd Jan. 2015).

123. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580,** 402–408. ISSN: 1476-4687 (Apr. 2020).

124. Drew, K., Wallingford, J. B. & Marcotte, E. M. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Molecular Systems Biology* **17,** e10016. ISSN: 1744-4292 (1st May 2021).

125. Beltrao, P. *et al.* Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell* **150,** 413–425. ISSN: 0092-8674, 1097-4172 (20th July 2012).

126. Evans, R. *et al. Protein complex prediction with AlphaFold-Multimer* 10th Mar. 2022.

127. Zhao, J., Cao, Y. & Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal* **18,** 417–426. ISSN: 2001-0370 (17th Feb. 2020).

128. Dhakal, A., McKay, C., Tanner, J. J. & Cheng, J. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Briefings in Bioinformatics* **23,** bbab476. ISSN: 1477-4054 (1st Jan. 2022).

129. Leemann, M. *et al.* Automated benchmarking of combined protein structure and ligand conformation prediction. *Proteins: Structure, Function, and Bioinformatics,* prot.26605. ISSN: 0887-3585, 1097-0134 (26th Oct. 2023).

130. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **41,** D1096–D1103. ISSN: 0305-1048 (D1 1st Jan. 2013).

131. Hu, L. *et al.* Binding MOAD (Mother Of All Databases). *Proteins: Structure, Function, and Bioinformatics* **60,** 333–340. ISSN: 1097-0134 (2005).

132. Dessailly, B. H., Lensink, M. F., Orengo, C. A. & Wodak, S. J. LigASite—a database of biologically relevant binding sites in proteins with known apo -structures. *Nucleic Acids Research* **36,** D667–D673. ISSN: 0305-1048 (suppl_1 1st Jan. 2008).

133. Morris, G. M. & Lim-Wilby, M. *Molecular Modeling of Proteins* (ed Kukol, A.) 365–382 (Humana Press, Totowa, NJ, 2008). ISBN: 978-1-59745-177-2.

134. Ye, N. *et al.* A Comprehensive Review of Computation-Based Metal-Binding Prediction Approaches at the Residue Level. *BioMed Research International* **2022,** 8965712. ISSN: 2314-6133 (31st Mar. 2022).

135. Yu, D.-J. *et al.* Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **10,** 994–1008. ISSN: 1545-5963 (July 2013).

136. Qiao, L. & Xie, D. MIonSite: Ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information. *Analytical Biochemistry* **566,** 75–88. ISSN: 0003-2697 (1st Feb. 2019).

137. Hu, X., Dong, Q., Yang, J. & Zhang, Y. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics* **32,** 3260–3269. ISSN: 1367-4803 (1st Nov. 2016).

138. Yang, J., Roy, A. & Zhang, Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **29,** 2588–2595. ISSN: 1367-4803 (15th Oct. 2013).

139. Lu, C.-H. *et al.* MIB2: metal ion-binding site prediction and modeling server. *Bioinformatics* **38** (ed Martelli, P. L.) 4428–4429. ISSN: 1367-4803, 1367-4811 (15th Sept. 2022).

140. Lu, C.-H. *et al.* The fragment transformation method to detect the protein structural motifs. *Proteins: Structure, Function, and Bioinformatics* **63,** 636–643. ISSN: 1097-0134 (2006).

141. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods* **20,** 205–213. ISSN: 1548-7105 (Feb. 2023).

142. Aptekmann, A. A. *et al.* mebipred: identifying metal-binding potential in protein sequence. *Bioinformatics* **38,** 3532–3540. ISSN: 1367-4803 (15th July 2022).

143. Stärk, H., Dallago, C., Heinzinger, M. & Rost, B. *Light Attention Predicts Protein Location from the Language of Life* preprint (Bioinformatics, 26th Apr. 2021).

144. Pokharel, S. *et al.* Improving protein succinylation sites prediction using embeddings from protein language model. *Scientific Reports* **12,** 16933. ISSN: 2045-2322 (8th Oct. 2022).

145. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology* **40,** 1023–1025. ISSN: 1546-1696 (July 2022).

146. Littmann, M. *et al.* Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports* **11,** 23916. ISSN: 2045-2322 (13th Dec. 2021).

147. Yuan, Q. *et al.* Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Briefings in Bioinformatics* **23,** bbac444. ISSN: 1477-4054 (1st Nov. 2022).

148. Zeng, Y. *et al.* Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics* **39,** btad187. ISSN: 1367-4811 (1st Apr. 2023).

149. Roche, R. *et al.* EquiPNAS: improved protein-nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *bioRxiv,* 2023.09.14.557719 (16th Sept. 2023).

150. Wu, F. *et al.* Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology* **6,** 1–8. ISSN: 2399-3642 (25th Aug. 2023).

151. Corso, G. *et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking* 4th Oct. 2022.

152. Krishna, R. *et al. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom* 9th Oct. 2023.

153. Fuchs, F. B., Worrall, D. E., Fischer, V. & Welling, M. *SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks* 24th Nov. 2020.

154. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science (New York, N.Y.)* **373,** 871–876. ISSN: 1095-9203 (20th Aug. 2021).

155. Stein, R. A. & Mchaourab, H. S. SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. *PLoS Computational Biology* **18,** e1010483. ISSN: 1553-734X (22nd Aug. 2022).

156. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11** (eds Robertson, J. L., Swartz, K. J. & Robertson, J. L.) e75751. ISSN: 2050-084X (3rd Mar. 2022).

157. Sala, D., Engelberger, F., Mchaourab, H. S. & Meiler, J. Modeling conformational states of proteins with AlphaFold. *Current Opinion in Structural Biology* **81,** 102645. ISSN: 0959-440X (1st Aug. 2023).

158. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature,* 1–3. ISSN: 1476-4687 (13th Nov. 2023).

159. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods* **18,** 176–185. ISSN: 1548-7105 (Feb. 2021).

160. AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics* **20,** 311. ISSN: 1471-2105 (11th June 2019).

161. Bordin, N. *et al.* Novel machine learning approaches revolutionize protein knowledge. *Trends in Biochemical Sciences* **48,** 345–359. ISSN: 0968-0004 (1st Apr. 2023).

162. Lazebnik, Y. Can a biologist fix a radio?–Or, what I learned while studying apoptosis. *Cancer Cell* **2,** 179–182. ISSN: 1535-6108 (Sept. 2002).

163. Khakzad, H. *et al.* A new age in protein design empowered by deep learning. *Cell Systems* **14,** 925–939. ISSN: 2405-4712 (15th Nov. 2023).