Protein Language Models and Structure Prediction: Connection and Progression

A Systematic Survey

Bozhen Hu[†,‡], Jun Xia[‡], Jiangbin Zheng[‡], Cheng Tan[‡], Yufei Huang[‡], Yongjie Xu[‡], and Stan Z. Li[‡]

[†]Zhejiang University, Hangzhou, 310058, China
[‡]AI Division, School of Engineering, Westlake University, Hangzhou, 310030, China

hubozhen;xiajun;zhengjiangbin;
tancheng;huangyufei;xuyongjie@westlake.edu.cn
Correspondence: Stan.ZQ.Li@westlake.edu.cn

Abstract

The prediction of protein structures from sequences is an important task for function prediction, drug design, and related biological processes understanding. Recent advances have proved the power of language models (LMs) in processing the protein sequence databases, which inherit the advantages of attention networks and capture useful information in learning representations for proteins. The past two years have witnessed remarkable success in tertiary protein structure prediction (PSP), including evolution-based and single-sequence-based PSP. It seems that instead of using energy-based models and sampling procedures, protein language model (pLM)-based pipelines have emerged as mainstream paradigms in PSP. Despite the fruitful progress, the PSP community needs a systematic and up-to-date survey to help bridge the gap between LMs in the natural language processing (NLP) and PSP domains and introduce their methodologies, advancements and practical applications. To this end, in this paper, we first introduce the similarities between protein and human languages that allow LMs extended to pLMs, and applied to protein databases. Then, we systematically review recent advances in LMs and pLMs from the perspectives of network architectures, pre-training strategies, applications, and commonly-used protein databases. Next, different types of methods for PSP are discussed, particularly how the pLM-based architectures function in the process of protein folding. Finally, we identify challenges faced by the PSP community and foresee promising research directions along with the advances of pLMs. This survey aims to be a hands-on guide for researchers to understand PSP methods, develop pLMs and tackle challenging problems in this field for practical purposes.

Contents

List of Tables

1 Backgrounds

Proteins are the workhorses of life, playing an essential role in a broad range of applications ranging from therapeutics to materials. They are built from 20 different basic chemical building blocks (called amino acids), which fold into complex ensembles of 3-dimensional structures that determine their functions and orchestrate the biological processes of cells. However, predicting protein structure from amino acid sequence is challenging because small perturbations in the sequence of a protein can drastically change the protein's shape and even render it useless, while different amino acids can have similar chemical properties, so some mutations will hardly change the shape of the protein. What is worse, the polypeptide is flexible and can fold into a staggering number of different shapes (Kryshtafovych et al., 2019; Senior et al., 2020a). Thus, PSP has long been a central but incompletely tackled problem in the scientific community.

One way to find out the structure of a protein is to use an experimental approach, including X-ray crystallography, Nuclear Magnetic Resonance (NMR) Spectroscopy (Ikeya et al., 2019) and cryo-electron microscopy (cryo-EM) (Gauto et al., 2019). The experimental protein structures have been deposited in the Protein Data Bank (PDB) (wwPDB consortium, 2019). Unfortunately, laboratory approaches for structure determination are expensive and cannot be used on all proteins. This challenge makes the number of reported protein structures orders of magnitude lower than the size of datasets in other machine-learning application domains. For example, 190 thousand structures exist in PDB (Berman et al., 2000) versus 528 million protein sequences in UniParc (Consortium, 2013) and versus 10 million annotated images in ImageNet (Russakovsky et al., 2014).

In general, computational methods for predicting three-dimensional (3D) protein structures from protein sequences have traditionally taken two parallel paths, focusing on either physical interactions or evolutionary principles (Jumper et al., 2021). Since proteins generally fold into their lowest free energy states, the physics-based approach simulates the folding process of the amino acid chain using molecular dynamics based on the potential energy of the force field at a particular time or fragment assembly using the energy function, which concentrates on the physical interactions to form an energy-stable 3D structure. However, this approach has proved highly challenging for even moderately sized proteins due to the computational intractability of molecular simulation, the conditioned accuracy of fragment assembly, and the difficulty of producing sufficiently accurate models of protein physics (AlQuraishi, 2019; Susanty et al., 2021). On the other hand, thanks to the recent progress in protein sequencing (Ma, 2015; Ma and Johnson, 2012), a large number of protein sequences are now available. For example, the UniProt (Bateman, 2019) database contains over 200 million protein sequences with relevant information. These protein databases support to get multiple sequence alignment (MSA) of homologous proteins, which significantly benefits the development of evolutionary methods.

LMs have recently emerged as a powerful paradigm for learning "content-aware" data representations from large-scale sequence databases (Bepler and Berger, 2021), which are widely used for machine translation, question answering in NLP (Andreas et al., 2013) and are even extended to computer vision (Pham et al., 2013), molecules (Xia et al., 2022b), etc. Due to the similarities between protein and human languages, LMs are gradually modified into pLMs to deal with various protein data, specially matched for protein sequences to learn representations that can be used for PSP. Large-scale pLMs with self-supervised pre-training on tens of millions to billions of proteins (Bepler and Berger, 2022; Elnaggar et al., 2022; Rao et al., 2019; Rives et al., 2019) are the current state-of-the-art methods in predicting protein structure, function, and fitness from sequences. For example, the introductions of MSA input and pLMs have led to the vast success of AlphaFold2 (AF2) (Jumper et al., 2021) at the Critical Assessment of protein Structure Prediction (CASP) 14 competition (Kryshtafovych et al., 2019). Since the propose of AF2 and RosettaFold (Baek et al., 2021), discussions on PSP and related protein tasks have culminated, and more researchers have begun to develop pLMs and tackle challenging problems that have not yet been solved, including PSP without evolutionary information by the usage of pLMs, making

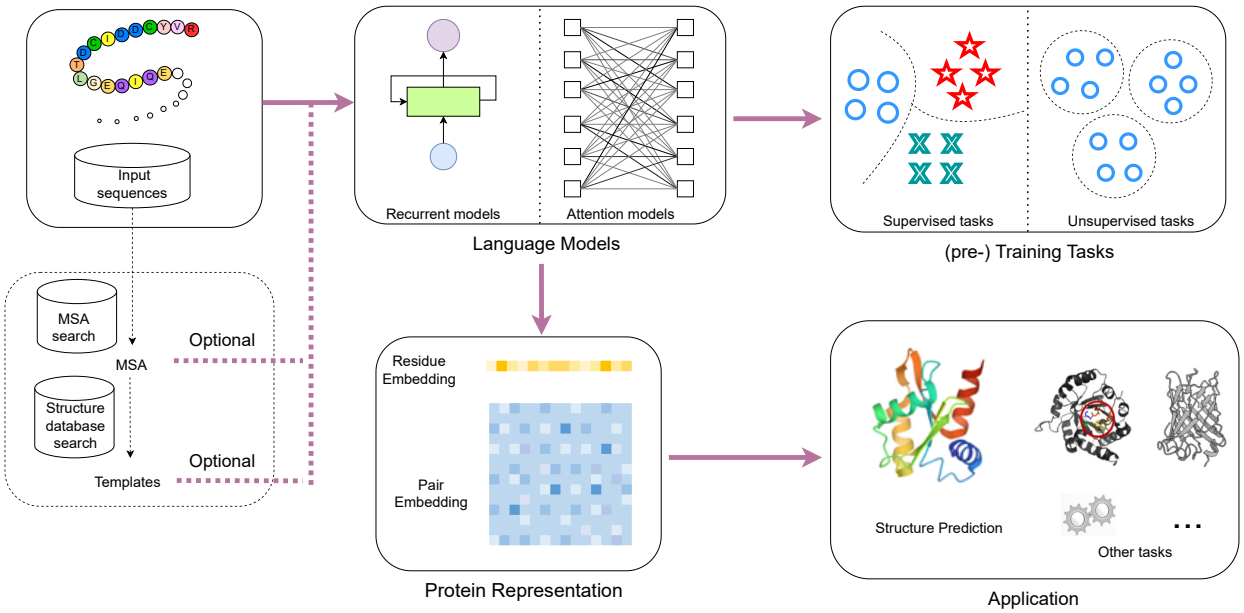


Figure 1: Schematic diagram of LMs for learning protein representations for PSP. Protein structures can be leveraged as labels in supervised tasks; dashed arrows represent optional.

accurate structure predictions for protein complexes (Evans et al., 2021), finding mechanisms behind protein folding, etc. Figure 1 shows the overall pipeline of pLMs for the prediction of protein 3D structure.

Protein representation learning, inspired by approaches in NLP, is an active area of research that learns representations used for various downstream tasks (Unsal et al., 2022). However, task-specific labelled proteins can be highly scarce because labelling proteins often requires time-consuming and resource-intensive lab experiments. To relieve and even tackle this problem, a pre-train and then fine-tune paradigm of LMs has been described in NLP. The knowledge is gained through pre-training a model on a source task and is used to improve the learning by fine-tuning the model on a new target task with fewer labels. Fortunately, self-supervised learning is applied to learn protein representation, like the masked language modelling tasks that reconstruct corrupted tokens given the surrounding sequence; well-known pre-trained sequence encoders include TAPE Transformer (Rao et al., 2019), ProteinBert (Ofer et al., 2021c), ProtTrans (Elnaggar et al., 2021) and ESM-1b (Rives et al., 2019) have been trained by predicting the masked residues in sequences. Besides, GearNet (Zhang et al., 2022c), STEPS (Chen et al., 2022b) are proposed for protein representations to exploit the topology information of structures, which are expected to contain more valuable features.

Although pLMs have been increasingly applied in protein representation learning and PSP, a systematic summary of this fast-growing field is still awaited. The following sections give an explanation of the common-used terms and notions, then summarize similarities between natural languages and proteins, next present commonly-used LMs and pLMs, show their applications, innovations, and differences, and finally introduce how different methods work for PSP, especially for the transformer-based pLMs that are used to reshape the protein representation learning area.

More importantly, we outline several works for single sequence PSP, structure prediction of antibodies, protein complexes, protein-ligands, protein-RNA, and protein conformational ensembles, etc., that are recently proposed and are future research directions. Besides, traditional physics- and machine-learning-based methods for protein structural feature prediction and PSP, are also presented. Moreover, we collect

abundant resources, including LMs and pLMs, methods for PSP, pre-training databases, and paper lists[1]. Finally, we provide possible future research directions by discussing the limitations and unsolved problems of existing methods.

To the best of our knowledge, this is the first survey including pLMs for structure prediction, presenting their connections and developments. We aim to help researchers develop more suitable algorithms and tackle essential, challenging, and urgent problems for proteins that can promote the development of biochemistry, biomedicine, and bioinformatics.

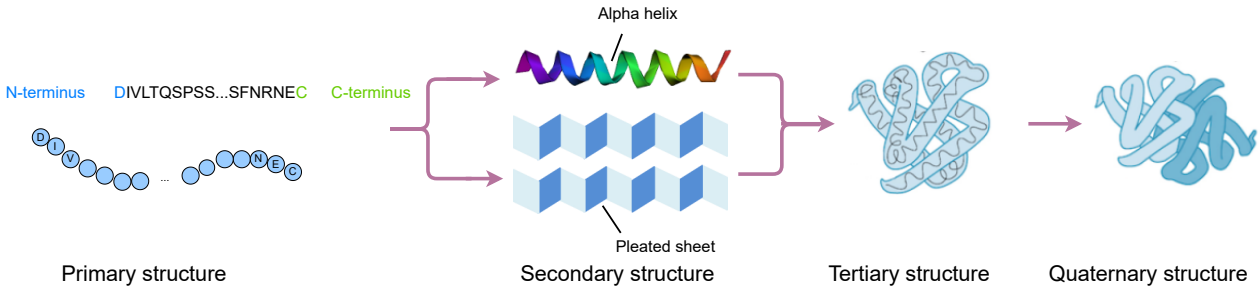


Figure 2: Four different levels of protein structures (Ihm, 2004; Patel and Shah, 2013).

2 Notions and Terms

- **Sequence/primary structure**: The linear sequence of amino acids in a peptide or protein (Sanger, 1952). Any sequence of polypeptides is reported starting from the single amine (N-terminus) end to carboxylic acid (C-terminus) (Hao et al., 2017) (Figure 2).
- **Secondary structure (SS)**: The 3D form of local segments of proteins. The two most common secondary structural elements are α-helix (H) and β-strand (E); 3-state SS includes H, E, C (coil region); 8 fine-grained states include three types for helix (G for 3_{10}-helix, H for α-helix, and I for π-helix), two types for strand (E for β-strand and B for β-bridge), and three types for coil (T for β-turn, S for high curvature loop, and L for irregular) (Wang et al., 2015).
- **Tertiary structure**: The 3D shape of a protein.
- **Quaternary structure**: The 3D arrangement of the subunits in a multisubunit protein (Chou and Cai, 2003).
- **Multiple sequence alignment (MSA)**: The result of the alignment of three or more biological sequences (protein or nucleic acid).
- **Sequence homology**: The biological homology between sequences (proteins or nucleic acids) (Koonin, 2005). MSA assumes all the sequences to be aligned may share recognizable evolutionary homology (Wang et al., 2018) and is used to indicate which regions of each sequence are homologous.
- **Coevolution**: The interdependence between the evolutionary changes of two entities (Ochoa and Pazos, 2014) plays an important role at all biological levels, which is evident between protein residues (Figure 3(a)).

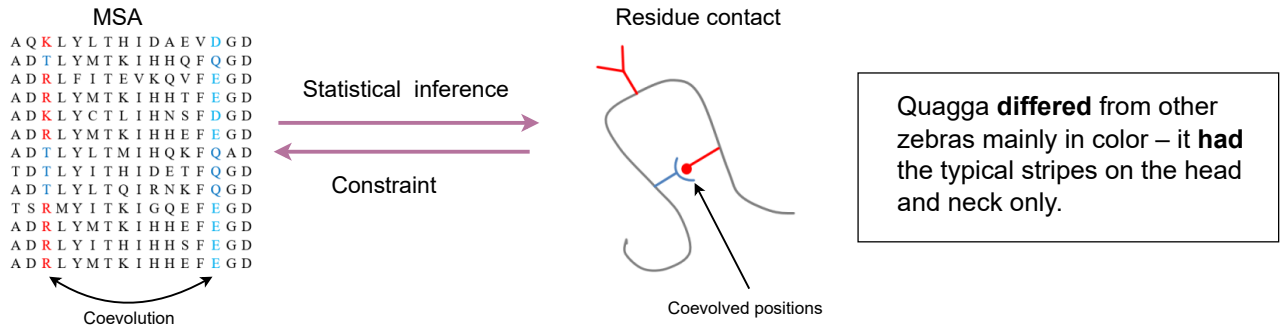
[1] https://github.com/bozhenhhu/A-Review-of-pLMs-and-Methods-for-Protein-Structure-Prediction

- **Templates**: The homologous 3D structures of proteins.
- **Contact map**: A two-dimensional binary matrix represents the residue-residue contacts of a protein within a distance threshold (Emerson and Amala, 2017), which produces a reduced representation of a protein structure.
- **Torsion angles**: Two essential torsion angles (dihedral angles) in the polypeptide chain, which describe the rotations of the chain around the bonds between $N - C_\alpha$ (φ) and $C_\alpha - C$ (ψ), respectively, each involving four atoms.
- **Protein structure prediction (PSP)**: The prediction of the 3D structure of a protein from its amino acid sequence.
- **Orphan proteins**: Proteins without any detectable homology (Basile et al., 2017) (MSAs of homologous proteins are not available).
- **Antibody**: A Y-shaped protein is produced by the immune system to detect and neutralize harmful substances, such as viruses and pathogenic bacteria.
- **RNA**: A polymeric molecule essential in various biological roles, most often single-stranded.
- **Protein complex**: A form of quaternary structure associated with two or more polypeptide chains.
- **Protein conformation**: The spatial arrangement of its constituent atoms that determines the overall shape (Blackstock, 1989).
- **Protein energy function**: Proteins fold into 3D structures in a way that leads to a low-energy state. Protein-energy functions are used to guide PSP by minimizing the energy value.
- **Monte Carlo methods**: A class of computational mathematical algorithms that use repeated random sampling to estimate the possible outcomes of an uncertain event.
- **Protein function prediction**: A Task that uses techniques to assign biological or biochemical roles to proteins. Gene Ontology (GO) annotations classify functions into three main categories of molecular function, biological process, and cellular component (Ashburner et al., 2000).
- **Protein stability prediction**: A task that uses methods to predict the impacts of amino acid mutations (substitutions).
- **Protein design**: A technique that design new proteins with novel purpose, behaviour from scratch, known structure or its sequence.
- **Protein structure refinement**: A task that aims to increase the accuracy of starting models (decoys), i.e., closer to their native states.
- **Supervised learning**: The use of labelled input-output pairs to learn a function that can classify data or predict outcomes accurately.
- **Unsupervised learning**: Models are trained without a labelled dataset and encouraged to discover hidden patterns and insights from the given data.
- **Natural language processing (NLP)**: The ability of computer programs to process, analyze, and understand the text and spoken words in much the same way humans can.
- **Language model (LM)**: A probability distribution of words or word sentences.
- **Embedding**: An embedding is a low-dimensional, learned continuous vector representation of discrete variables into which you can translate high-dimensional and real-valued vectors (words or sentences) (Li et al., 2016).

- **Convolution Neural Networks (CNNs)**: A class of neural networks that consist of convolutional operations to capture the local information found in the data.
- **Recurrent Neural Networks (RNNs)**: A class of neural networks where connections between nodes form a directed or undirected graph along a temporal sequence.
- **Attention models**: A class of neural networks used to focus on specific components of a complex input and categorize the whole dataset sequentially (Lin et al., 2017).
- **Tansfer learning**: A machine learning method where a model developed for one task is reused for a model to solve a different but related task (Pan and Yang, 2010; Weiss et al., 2016), which has two major activities, i.e., pre-training and fine-tuning.
- **Pre-training**: A strategy in AI refers to training a model with one task to help it form parameters that can be used in other tasks.
- **Fine-tuning**: A method that takes the weights of a pre-trained neural network, which are used to initialize a new model being trained on the same domain.
- **Autoregressive language model**: A feed-forward model predicts the future word from a set of words given a context (Bond-Taylor et al., 2021).
- **Masked language model**: A language model masks some of the words in a sentence and predicts which words should replace those masks.
- **Bidirectional language model**: A language model learns to predict the probability of the next token in the past and future directions (Jahan et al., 2021).
- **Multi-task learning**: A machine learning paradigm in which multiple tasks are solved simultaneously while exploiting commonalities and differences across tasks (Bepler and Berger, 2021).
- **Sequence-to-Sequence (Seq2Seq)**: A family of machine learning approaches train models to convert sequences from one domain to sequences in another domain.
- **Knowledge graph**: A semantic network uses a graph-structured data model or topology to integrate data (McCusker et al., 2018).
- **Knowledge distillation**: The process of transferring the knowledge from a large model or set of models to a single smaller model (Gou et al., 2020).
- **Multi-modal learning**: Training models by combining information obtained from more than one modality (Skocaj et al., 2012; Wang et al., 2022c).
- **Residual neural network**: An artificial neural network (ANN) in which skip connections or shortcuts are used to jump over some layers, e.g., the deep residual network, ResNet (He et al., 2016).

3 Protein and Language

LMs are increasingly applied to large-scale protein sequence databases to learn embeddings for protein structure or function prediction recently (Asgari and Mofrad, 2015; Yang et al., 2018; Young et al., 2017). One important reason is that human languages and proteins share common characteristics. Such as the hierarchical organization (Ferruz and Höcker, 2022; Ofer et al., 2021a), which means that the four different levels of protein structures (see Figure 2) analogy to letters, words, sentences, and texts of human languages to a certain degree. It illustrates that proteins and languages typically comprise modular elements that can be reused and rearranged. Moreover, the rules of protein folding, e.g., the hydrophilicity and hydrophobicity of amino



(a) Protein (b) Language

Figure 3: Comparisons of protein and language. (a) Relationship between a MSA and the residue contact of one protein in the alignment. The positions that coevolved are highlighted in red and light blue. Residues within these positions where changes occurred are shown in blue. Given such a MSA, one can infer correlations statistically found between two residues that these sequence positions are spatially adjacent, i.e., they are contacts (Ochoa and Pazos, 2014; Vorberg, 2017; Zerihun and Schug, 2018). (b) One grammatically complex sentence contains long-distance dependencies (shown in bold).

acids, the principle of minimal frustration (Bryngelson et al., 1995) and the "folding funnel" landscapes of proteins (Leopold et al., 1992), etc., are similar to language grammars of linguistics.

Figure 3a shows a MSA and its statistical inference for residue contact of one protein. This illustrates that long-range dependencies exist between two residues; they may be far apart in the sequence but close in space resulting in coevolution. Similarly, long-distance dependencies, which pose a problem for machine translation and RNNs, also appeared in languages. Figure 3b shows an example of language grammar rules that require agreement in the category of words that might be far from each other (Choshen and Abend, 2019). All these similarities illustrate that researchers can deal with protein data using successful methods in NLP.

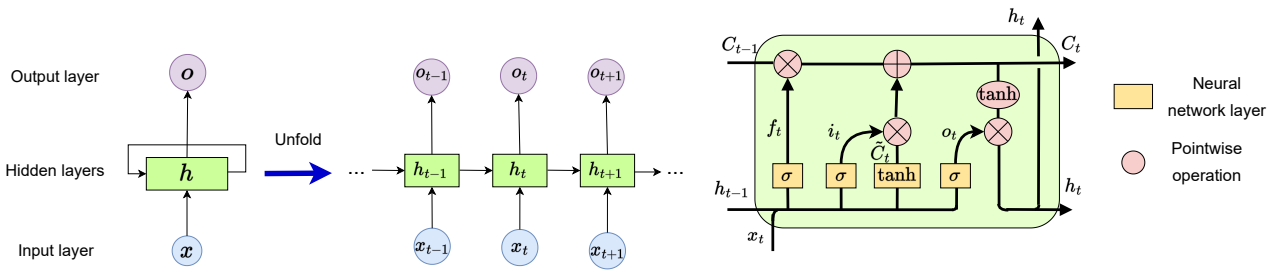
However, proteins are not human languages, despite these and other similarities. For example, the training of LMs often requires a massively large corpus, which needs tokenization, i.e., splitting the text into individual tokens or directly using words as tokens, which serves computational goals and, ideally, can also fulfil linguistic goals in NLP (Alley et al., 2019; Asgari and Mofrad, 2015; Madani et al., 2021; Ofer et al., 2021b; Yang et al., 2018). Compared with algorithms in NLP, protein tokenization methods still remain at a low level without a well-defined and biologically meaningful protein token algorithm (Vu et al., 2022). This may be a direction for unlocking the secrets of proteins.

4 Language Models

This section firstly introduces encoder architectures of LMs broadly falling into two categories: recurrent neural networks (RNNs) and attention mechanisms, especially for long short-term memory (LSTM) (Lample et al., 2016) and transformers (Vaswani et al., 2017). Then we present the commonly-used pre-trained LMs and their developments.

4.1 Recurrent Neural Networks (RNNs) and Long Short-term Memory (LSTM)

The first example of LM was studied by Andrey Markov, who proposed the Markov chain in 1913 (Hayes et al., 2013; Li, 2022). After that, some machine learning methods, hidden Markov models (HMMs) and their



(a) RNNs (b) LSTM cell

Figure 4: Graphical explanation of RNNs and LSTM.

variants particularly, are described and applied as fundamental tools in many fields, including biological sequences (Bishop and Thompson, 1986) where the goal is to recover a data sequence that is not immediately observable (Chiu and Rush, 2020; Domingos, 2015; Gales et al., 2008; Nicolai and Sachs, 2013; Stigler et al., 2011; Wong et al., 2013).

Neural networks started to produce superior results in various NLP tasks since 2010s (Ferruz and Höcker, 2022). RNNs allow previous outputs to be used as inputs while having hidden states to exhibit temporal dynamic behaviours. Therefore, RNNs can use their internal states to process variable-length sequences of inputs that are useful and applicable in NLP tasks (AGMLS et al., 2009).

RNNs are typically shown in Figure 4a. In each timestep t, the input $x_t \in \mathbb{R}^l$, hidden $h_t \in \mathbb{R}^d$ and output state vectors $o_t \in \mathbb{R}^d$, where the superscripts l and d refer to the number of input features and the number of hidden units, respectively, are formulated as follows:

$$h_t = g\left(W_x x_t + W_h h_{t-1} + b_h\right)$$
$$o_t = g\left(W_y h_t + b_y\right)$$

Where $W_x \in \mathbb{R}^{d\times l}$, $W_h \in \mathbb{R}^{d\times d}$ and $W_y \in \mathbb{R}^{d\times d}$ are the weights associated with the input, hidden and output vectors in the recurrent layer, and $b_h \in \mathbb{R}^d$, $b_y \in \mathbb{R}^d$ are the bias, which are shared temporally, g is the activation function.

In order to deal with the vanishing gradient problem (Hochreiter, 1991) that can be encountered when training traditional RNNs, LSTM networks were developed to process sequences of data. They presented superior capabilities in learning long-term dependencies (Lample et al., 2016) with various applications such as time series prediction (Schmidhuber et al., 2005), protein homology detection (Hochreiter et al., 2007), drug design (Gupta et al., 2018), etc. Unlike standard LSTM, bidirectional LSTM (BiLSTM) adds one more LSTM layer, reversing the information flow direction. This means it is capable of utilizing information from both sides and is also a powerful tool for modelling the sequential dependencies between words and phrases in a sequence (Ma et al., 2021).

The LSTM architecture aims to provide a short-term memory that can last more timesteps, shown in Figure 4b, σ and $\tanh$ represent the sigmoid and tanh layer. Forget gate layer in the LSTM is to decide what information is going to be thrown away from the cell state at timestep t, $x_t \in \mathbb{R}^l$, $h_t \in (-1, 1)^d$ and $f_t \in (0, 1)^d$ are the input, hidden state vectors and forget gate's activation vector.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

Then, the input gate layer decides which values should be updated, and a tanh layer creates a vector of new candidate values, $\tilde{C}_t \in (-1, 1)^d$ that could be added to the state, $i_t \in (0, 1)^d$ is the input gate's activation

vector.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

Next, we combine old state $C_{t-1} \in \mathbb{R}^d$ and new candidate values $\tilde{C}_t \in (-1, 1)^d$ to create an update to the new state $C_t \in \mathbb{R}^d$.

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

Finally, the output gate layer decides what parts of the cell state to be outputted, $o_t \in (0, 1)^d$.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$h_t = o_t \odot \tanh(C_t)$$

where $\{W_f, W_i, W_c, W_o\} \in \mathbb{R}^{d \times l}$, $\{U_f, U_i, U_c, U_o\} \in \mathbb{R}^{d \times d}$ and $\{b_f, b_i, b_c, b_o\} \in \mathbb{R}^d$ are weight matrices and bias vector parameters in the LSTM cell, $\odot$ means the pointwise multiplication.

4.2 Attention Mechanism and Transformer

Traditional Sequence-to-Sequence (Seq2Seq) models use RNNs or LSTMs as encoders and decoders (Radford et al., 2017) to process sequence and extract features for tasks. The final state of RNNs or LSTMs (better to say, encoders) must hold information for the entire input sequence, which may cause information loss. Therefore, traditional RNNs and LSTMs were soon superseded by attention mechanisms (Bahdanau et al., 2014a; Han et al., 2021), which help look at all hidden states from the encoder sequence for making predictions and were first applied for sequence modelling in machine translation (Bahdanau et al., 2014b).

The attention layer can access all previous states and learn their importance to weight them. Based solely on attention mechanisms, Google Brain released Transformer (Vaswani et al., 2017) in 2017, a new network architecture dispensing with recurrence and convolutions entirely. This led to the development of pre-trained models, e.g., Bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) and Generative pre-trained transformer (GPT) (Brown et al., 2020; Radford et al., 2018a,b), which were trained with large language datasets. Unlike RNNs, Transformer processes the entire input all at once, using stacked self-attention layers for both the encoder and decoder. Each layer consists of a multi-head attention module followed by a feed-forward module with a residual connection and normalization. The vanilla single-head attention is called "Scaled Dot-Product Attention" and operates as follows:

$$\text{Att}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V$$

where $Q, K, V \in \mathbb{R}^{l \times d}$ are d-dimensional vector representations of l words in sequences of queries, keys and values, respectively. Multi-head attention allows the model to attend to information from different representation subspaces in parallel jointly.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h}) W^O$$
$$\text{where head}_\text{i} = \text{Att}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d \times d_i}, W_i^K \in \mathbb{R}^{d \times d_i}, W_i^V \in \mathbb{R}^{d \times d_i}$ and $W^O \in \mathbb{R}^{hd_i \times d}, d_i = d/h$, there are h parallel attention layers or heads. Moreover, in Transformer, the position-wise feed-forward networks consist of two linear transformations with a ReLU activation in between, and positional encoding is added to the input embedding at the bottoms of the encoder and decoder stacks to make use of the order of the sequence.

Table 1: Representative pre-trained LMs in general domains

Model	Network	Objective	#Params.	Comments
ELMo (Peters et al., 2018)	LSTM	Bidirectional LM	93.6M	the first deep contextualized word representation
GPT (Radford et al., 2018a)	Transformer	Autoregressive LM	110M	a pre-trained LM for predicting the next word
BERT (Devlin et al., 2018)	Transformer	Masked LM	340M	the most commonly-used LM for predicting masked tokens
Transformer-XL (Dai et al., 2019)	Transformer	Autoregressive LM	257M	enabling learn dependencies beyond a fixed length
XLM (Lample and Conneau, 2019)	Transformer	Multi-task	665M	cross-lingual pre-training
Udify (Kondratyuk and Straka, 2019)	BERT	Multi-task	∼340M	leveraging a multilingual BERT self-attention model
GPT-2 (Radford et al., 2022)	Transformer	Autoregressive LM	1.5B	larger model, more training data
Grover (Zellers et al., 2019)	GPT2	Autoregressive LM	1.5B	defending against the general neural fake news
XLNet (Yang et al., 2019b)	BERT	Autoregressive LM	∼340M	more training data, integrates ideas from Transformer-XL
RoBERTa (Liu et al., 2019d)	BERT	Masked LM	355M	more training data, dynamic masking
CTRL (Keskar et al., 2019)	Transformer	Autoregressive LM	1.63B	trained to control particular aspects of the generated text
Megatron-LM (Shoeybi et al., 2019)	Transformer	Autoregressive LM	8.3B	a large transformer model
ALBERT (Lan et al., 2019)	BERT	Masked LM	223M	a lite BERT
DistillBERT (Sanh et al., 2019)	BERT	Masked LM	65M	a distilled version of BERT
SpanBERT (Joshi et al., 2019)	BERT	Masked LM	∼340M	presenting and predicting the masked span
MASS (Song et al., 2019)	Transformer	Seq2Seq LM	∼307M	masked Seq2Seq pre-training
MT-DNN (Liu et al., 2019c, 2020)	BERT	Multi-task	∼340M	for multiple natural language understanding tasks
MT − DNN_{KD} (Liu et al., 2019b)	MT-DNN	Multi-task	∼340M	incorporating knowledge distillation
ERNIE (Zhang et al., 2019)	BERT	Masked LM	∼114M	incorporating knowledge graphs
KnowBERT (Peters et al., 2019)	BERT	Masked LM	∼110M	incorporating knowledge bases into BERT
KEPLER (Wang et al., 2019)	RoBERTa	Masked LM	∼125M	incorporating knowledge embedding
VideoBERT (Sun et al., 2019a)	BERT	Multimodal model	∼340M	modelling between the visual and linguistic domain
VisualBERT (Li et al., 2019)	BERT	Multimodal model	∼110M	modelling a broad range of vision and language tasks
ERNIE (Sun et al., 2019b)	BERT	Masked LM	∼110M	a new masking strategy
PEGASUS (Zhang et al., 2020c)	Transformer	Masked & Seq2Seq LM	568M	for abstractive text summarization
Unicoder-VL (Li et al., 2020b)	BERT	Multimodal model	∼110M	cross-modal learning
UNILM (Bao et al., 2020)	BERT	Bidirectional & Seq2Seq LM	∼110M	for natural language understanding and generation tasks
Turing-NLG (Rasley et al., 2020)	Transformer	Autoregressive LM	17B	a hugely large LM
ELECTRA (Clark et al., 2020)	BERT	Generator & Discriminator	335M	token detection
GPT-3 (Brown et al., 2020)	GPT-2	Autoregressive LM	175B	extending the model size
T5 (Raffel et al., 2020)	Transformer	Seq2Seq LM	11B	producing new text as output
Switch Transformer (Fedus et al., 2021)	Transformer	Masked LM	1.6T	increasing the pre-training speed
BEIT (Bao et al., 2021)	Transformer	Masked image model	307M	a vision Transformer
MT-NLG (Smith et al., 2022)	Transformer	Autoregressive LM	530B	the largest publicly monolithic transformer

All examples report the largest model of their public series. Network displays high-level backbone models preferentially if they are used to initialize parameters. #Param. means the number of parameters; M, millions; B, billions; T, trillions; & , and; ∼ means estimated data. Related terminologies are listed in Section 2.

4.3 Pre-trained Language Models

In order to train effective deep neural models focusing on storing knowledge for specific tasks with limited human-annotated data, transfer learning with a pre-training phrase and a fine-tuning stage has been adopted (Han et al., 2021; Pan and Yang, 2010; Thrun and Pratt, 1998). In recent years, we have witnessed a rapid development of pre-trained LMs that have been widely used in NLP and computer vision, etc. Due to its prominent nature, the transformer gradually becomes a standard neural architecture for natural language understanding and generation. It also serves as the most commonly-used backbone neural architecture for pre-trained models as they have achieved state-of-the-art results on almost all NLP tasks. This indeed subverted our current perception of the performance of deep learning models, thus drawing more attention. An overview of some typical pre-trained LMs in general domains is shown in Table 1.

4.3.1 GPT and BERT

With transformers as architectures and LM learning as objectives, BERT and GPT are the two landmarks that completely open the door towards the era of large-scale, deeply pre-trained LMs.

GPT optimizes standard autoregressive language modelling during pre-training, which uses a transformer to model the conditional probability of each word and therefore, is powerful at predicting the next token in a sequence. Formally, given an unsupervised corpus of tokens $\mathcal{X} = \{x_0, x_1, \ldots, x_n, x_n + 1\}$, GPT applies a standard language modelling objective to maximize the following likelihood:

$$\mathcal{L}(\mathcal{X}) = \sum_{i=1}^{n+1} \log P\left(x_i \mid x_{i-k}, \ldots, x_{i-1}; \Theta\right)$$

Where k is the size of the context window, and the conditional probability P is modelled using a network decoder with parameters Θ.

BERT uses a multi-layer bidirectional transformer encoder as its architecture. In the pre-training phase, BERT adopts the strategies of next-sentence prediction to understand sentence relationships with the help of a binary classifier and masked language modelling (MLM), which is powerful and applied in most self-supervised pre-training tasks. Formally, given a corpus consisting of tokens $\mathcal{X} = \{x_0, x_1, \ldots, x_n, x_n + 1\}$, BERT maximizes the following likelihood:

$$\mathcal{L}(\mathcal{X}) = \textstyle\sum_{x \in mask(\mathcal{X})} \log P(x \mid \tilde{\mathcal{X}}; \Theta)$$

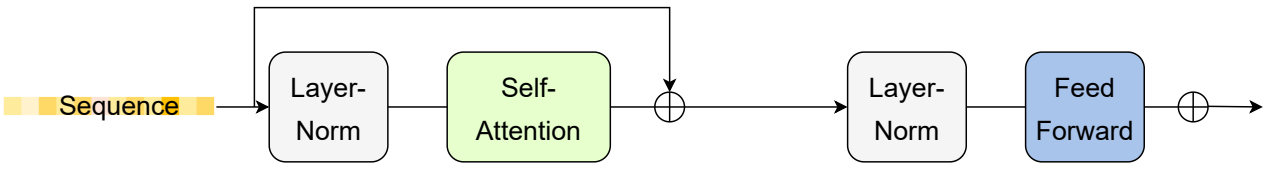
where $mask(\mathcal{X})$ are the masked tokens, $\tilde{\mathcal{X}}$ is the result after masking some tokens in $\mathcal{X}$, and the probability P is modeled by the transformer encoder with parameters Θ.

4.3.2 Post GPT and BERT Era

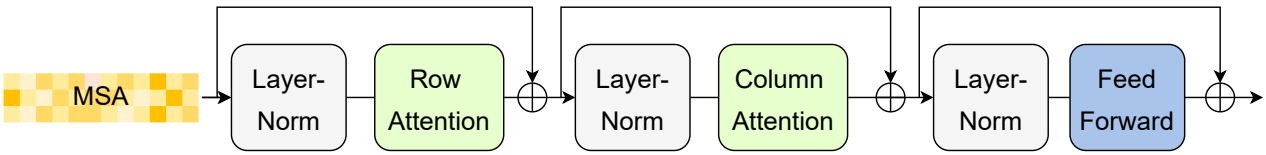
After GPT and BERT, various of their improvements and variants have been proposed, as shown in Table 1. For example, researchers increased the size of models and datasets (Liu et al., 2019d; Yang et al., 2019b), as large transformer models became the de facto standard in NLP on the basis of scaling laws, which can govern the dependence of overfitting on the model and dataset size for a given compute budget (Hoffmann et al., 2022; Kaplan et al., 2020). The new masking strategies were proposed like entity-level masking, phrase-level masking (Sun et al., 2019b), and span masking that masks the tokens consecutively according to the span length (Joshi et al., 2019), entity-level masking and phrase-level masking. Incorporating different data sources has also been developing as an important direction, such as utilizing multilingual corpora, and knowledge graphs (Lample and Conneau, 2019; Peters et al., 2019). Furthermore, because Pre-trained LMs are not data-hungry to labelled data and present better performance, they gradually stepped into different domains, including financial, computer vision and biomedical applications (Araci, 2019; Bengio et al., 2022; Lee et al., 2020; Mikolov et al., 2013; Wang et al., 2022a), etc.

5 Protein Language Models

As stated before, protein sequences corresponding to strings of amino-acid letters are a natural fit to most LMs, which are able to capture complex dependencies among these amino acids (Ofer et al., 2021a). pLMs have been developed and emerged as promising approaches for learning protein sequences. In this section, we first introduce LSTM pLMs, then present transformer pLMs with different implementation strategies and applications elaborately, especially the pLMs aimed at PSP. We listed a group of representative pLMs in Table 2. The (pre-) training databases that appeared in this table are listed in Table 4, where the CullPDB (Wang and Dunbrack, 2005) is a secondary structure prediction dataset.



(a) ESM-1b



(b) ESM-MSA

Figure 5: Core modules of ESM-1b and ESM-MSA.

5.1 LSTM Protein Language Models

Klausen et al. (2018) trained a combination of convolutional and LSTM neural networks to predict protein structural features, solvent accessibility (SA), secondary structure (SS), structural disorder, and torsion angles (φ, ψ) for each residue of the input sequences. SPIDER3-Single (Heffernan et al., 2018) modelled on the single sequence instead of relying on evolutionary information from MSAs. Having similar training objectives and backbone architectures, such kind of examples are DeepPrime2Sec (Asgari et al., 2019), SPOT-1D-Single (Singh et al., 2021a). Furthermore, DeepBLAST (Morton et al., 2020), SPOT-1D-LM (Singh et al., 2021c) and SPOT-Contact-Single (Singh et al., 2021b) are the usages of pre-trained pLMs to get embeddings for downstream tasks like contact map prediction, function prediction, etc.

However, Rao et al. (2019) (TAPE) benchmarked a group of protein models across a panel of tasks, concluding that there exist opportunities for specific innovative design of protein models and training methods for vanilla LSTMs and Transformers. Methods specified for protein data processing have been researched and tested.

Without structural or evolutionary data, UniRep (Alley et al., 2019) summarized arbitrary protein sequences into fixed-length vectors by multiplicative long-/short-term-memory (mLSTM) (Krause et al., 2016). Analogous to UniRep, UDSMProt (Strodthoff et al., 2020) and SeqVec (Heinzinger et al., 2019) used LSTM or its variants to remember long-range dependencies for protein sequences to get rich representations that can be transferred and retrieved afterwards. ProSE (Bepler and Berger, 2021) extra added structural supervision by residue-residue contact loss and structural similarity prediction loss to better capture the semantic organization of proteins and improved the ability to predict protein functions instead of using the masking loss only. Besides, Bepler and Berger (2019) proposed a soft symmetric alignment mechanism to measure similarities between sequence embeddings. CPCProt (Lu et al., 2020) learned protein embeddings by formalizing the InfoNCE loss for the principle of mutual information maximization. All these models demonstrate the LSTM-like pLMs' ability to capture some biological properties of proteins.

5.2 Transformer Protein Language Models

Elnaggar et al. (2021) have trained six successful LMs (T5, ELECTRA, ALBERT, XLNet, BERT, and Transformer-XL, listed in Tabel 1) on protein sequences containing 393 billion amino acids using many resources (5616 GPUs and one TPU Pod). ESM-1b (Rives et al., 2019) employed a deep Transformer (shown in Figure 5a) and a masking strategy to build up complex representations that incorporate context from across the sequence. The results of ProtTrans (Elnaggar et al., 2021) and ESM-1b implied that large-scale pLMs have the advantage of learning

the grammar of proteins even without using evolutionary information. Like approaches in NLP that change masking strategy, McDermott et al. (2021) updated the random masking scheme with a fully differentiable adversarial masking model. PMLM (He et al., 2022) considered the dependency among masked tokens to capture the correlations (coevolution) of inter-residues, which was demonstrated to improve the performance on the TAPE contact benchmark.

5.2.1 Multi-source Knowledge Enhancement of Protein Representations

Extra information, including MSAs, functions, structures and biological priors, may enrich protein embeddings. In detail, firstly, MSA Transformer (Rao et al., 2021a) extended the transformer LMs to deal with sets of sequences as input by alternating attention over rows and columns, shown in Figure 5b. Its internal representations enable high-quality unsupervised structure learning with an order of magnitude fewer parameters than contemporaneous pLMs. Secondly, ProteinBERT (Ofer et al., 2021c) was pre-trained on protein sequences and Gene Ontology (GO) annotations which can be encoded as a binary vector. The learned embeddings contain information from both sequence and GO annotation to predict diverse protein functions (Ashburner et al., 2000); likewise, OntoProtein (Zhang et al., 2022b) considered GO as a factual knowledge graph, which was used to enhance protein representations. Finally, Mansoor et al. (2021) encoded two types of protein information (sequence and structure) through joint training in a semi-supervised manner. Bepler and Berger (2021) have carried out multi-task with structural supervision, leading to an even better-organized embedding space. STEPS (Chen et al., 2022b) tried to correlate the embeddings learned from sequence and structure by pseudo bi-level optimization.

However, Liu et al. (2019a) indicated that not all external knowledge is beneficial for downstream tasks. Thanks to the construction of benchmarks which can give relatively fair results of different methods. PEER benchmarks (Xu et al., 2022) were built for protein sequence understanding, including protein function and structure prediction, protein-protein interaction prediction, protein-ligand interaction prediction tasks, etc. Its results showed that selecting suitable auxiliary tasks can boost different models' performance. Therefore, it is necessary to inject external features and design algorithms carefully and adequately.

5.2.2 Transformer Models Designed for Protein Structure Prediction

Early pLMs tend to predict structural features (Asgari et al., 2019; Heffernan et al., 2018; Heinzinger et al., 2022; Høie et al., 2022; Kandathil et al., 2020; Klausen et al., 2018; Luo et al., 2020; Rao et al., 2021b; Singh et al., 2021b), like SS, SA, torsion angles, remote homology, and contact map, etc., which are useful for constructing protein 3D structures. Current pLMs tend to predict PSP end-to-end, which are introduced as follows.

Evoformer is the core module (encoder) of the famous network, AF2 (Jumper et al., 2021), which is repeated 48 times with no shared weights (shown in Figure 6). It uses a variant of axial attention (Ho et al., 2019), including row-wise gated self-attention with pair bias and column-wise gated self-attention, to process the MSA representation, which is transitioned by a 2-layer neural network, then used to update the pair representation by an outer product mean block, containing linear transforms, outer product, and mean, etc. In order to make the pair representation in the embedding space satisfy the demand of consistency, like the triangle inequality on distances, a triangle multiplicative update block and a triangle self-attention block were designed. The former updates the pair representation by combining information within each triangle of graph edges, while the latter operates self-attention around graph nodes. These ingenious designs let the output of Evoformer produce more insightful patterns for accurate PSP. Besides, Hu et al. (2022) have shown that pLMs, especially those trained for PSP, like Evoformer, are valuable and general-purpose for various structure and function tasks.

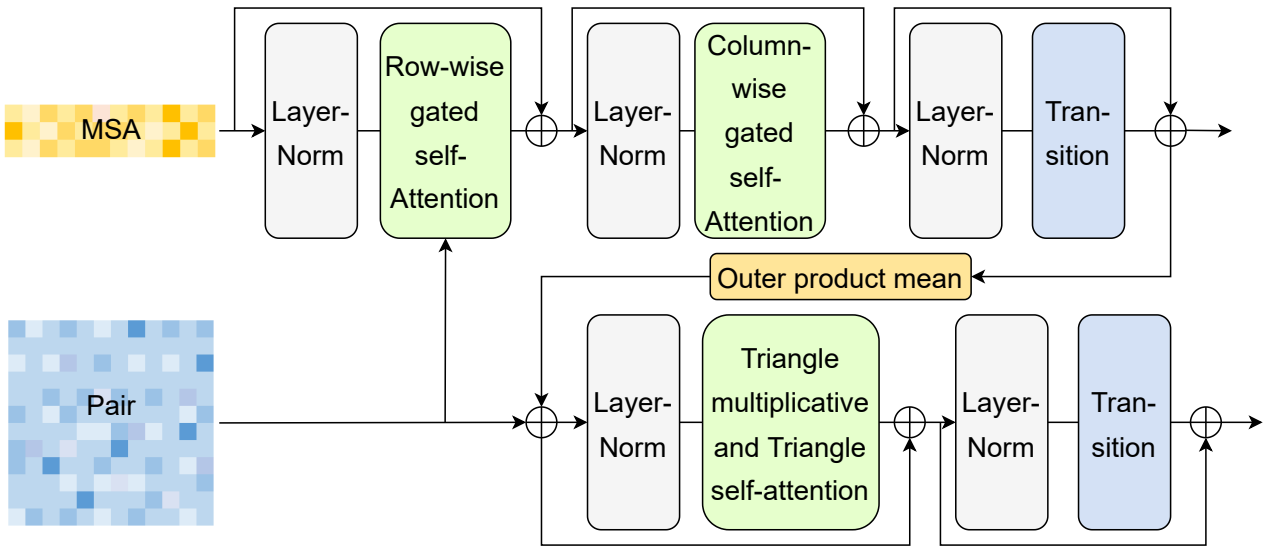


Figure 6: Evoformer block.

AminoBERT (Chowdhury et al., 2022) adopted a transformer to learn latent information from a single sequence, whose representation is inputted in a recurrent geometric network to generate the backbone structure of a protein. ESM-2 (Lin et al., 2022) is the largest pLM to date, extending ESM-1b and having parameters up to 15 billion, which has achieved the lowest validation perplexity and highest TM-score (Zhang and Skolnick, 2005) on CASP14 compared with other smaller ESM models. The results indicated that improving the model sizes of pLMs is able to improve the PSP performance with little or no evolutionary information. OmegaPLM (Wu et al., 2022b) was trained with a stack of efficient GAU layers (Hua et al., 2022) to get single- and pairwise-residue embeddings, which are expected to contain structural and functional information by different masking strategies, including random masking, sequential masking, and span masking (Joshi et al., 2019). In GAU, gate operation is applied after the attention aggregation and uses $\text{relu}^2(\cdot)$ to replace the conventional $\text{softmax}(\cdot)$, which performs better in terms of both computation speed and convergence rate in original experiments. AntiBERTy (Ruffolo and Gray, 2022) was pre-trained on antibody sequences to produce contextual embeddings for subsequent antibody structure prediction.

5.2.3 Other Applications

In terms of sequence generation, ProGen (Madani et al., 2020) was trained on sequences conditioned on a set of protein properties like function or affiliation with a particular organism, and the training database contains about 280 million protein sequences and associated properties from different datasets. Compared with MSA Transformer (Rao et al., 2021a), MSA2Prot (Bepler and Ram, 2022) is a MSA-to-protein transformer, developed axial attention and cross attention for transformer encoder and decoder to model sequence probabilities autoregressively. Other methods include ProtGPT2 (Ferruz et al., 2022), RITA (Hesslow et al., 2022), and ProGen2 Nijkamp et al. (2022) (up to 6.4 billion parameters), etc.

Other than the models mentioned above, a group of pLMs are developed and trained for different purposes. For example, ESM-1v (Meier et al., 2021), Tranception (Notin et al., 2022) enabled the prediction of the effects of mutations. Structured Transformer (Ingraham et al., 2019), ESM-IF1 (Hsu et al., 2022), Fold2Seq (Cao et al., 2021), PeTriBERT (Dumortier et al., 2022), ProteinMPNN (Dauparas et al., 2022), and (Gao et al., 2022a,b; Tan et al., 2022), etc., were proposed for protein design, which is also referred to as the inverse protein

Table 2: List of representative pLMs

Model and Repository	Approach	Input	Network	#Embedding	#Param.	Pre-training Database
NetSurfP-2.0 (Klausen et al., 2018)	Supervised	MSA, Structure	CNN, BiLSTM	2048	N/A	PDB, UniRef30
SPIDER3-Single (Heffernan et al., 2018)	Supervised	Sequence, Structure	LSTM-BRNN (Heffernan et al., 2017)	1024, 512	N/A	12442 proteins
SeqVec (Heinzinger et al., 2019)	Unsupervised	Sequence	ELMo	1024	∼93.6M	UniRef50
UniRep (Alley et al., 2019)	Unsupervised	Sequence	mLSTM (Krause et al., 2016)	1900	∼18.2M	UniRef50
SSA (Bepler and Berger, 2019)	Supervised	Sequence, Structure	BiLSTM	100, 512	N/A	Pfam, SCOP
DeepPrime2Sec (Asgari et al., 2019)	Supervised	MSA, Sturcture	ELMo, CNN, BiLSTM	N/A	N/A	UniRef50, Swiss-Prot, CullPDB
Ingraham et al. (2019)	Unsupervised	Structure	Transformer	128	N/A	CATH4.2
TAPE (Rao et al., 2019)	Unsupervised	Sequence	LSTM Transformer	2048 768	N/A 38M	Pfam
ESM-1b (Rives et al., 2019)	Unsupervised	Sequence	Transformer	1280	650M	UniParc
UDSMProt (Strodthoff et al., 2020)	Unsupervised	Sequence	LSTM	400	∼24M	Swiss-Prot
$\text{CPCProt}_{\text{GRU\_large}}$ (Lu et al., 2020)	Unsupervised	Sequence	GRU (Cho et al., 2014)	1024	8.4M	Pfam
$\text{CPCProt}_{\text{LSTM}}$ (Lu et al., 2020)	Unsupervised	Sequence	LSTM	2048	71M	Pfam
Sturmfels et al. (2020)	Supervised	MSA	Transformer	N/A	N/A	Pfam
ProGen (Madani et al., 2020)	Unsupervised	Sequence, Property	Transformer	1028	1.2B	∼280M proteins
ProtTXL (Elnaggar et al., 2021)	Unsupervised	Sequence	Transformer-XL	1024	562M	BFD100, UniRef100
ProtBert (Elnaggar et al., 2021)	Unsupervised	Sequence	BERT	1024	420M	BFD100, UniRef100
ProtXLNet (Elnaggar et al., 2021)	Unsupervised	Sequence	XLNet	1024	409M	UniRef100
ProtAlbert (Elnaggar et al., 2021)	Unsupervised	Sequence	ALBERT	4096	224M	UniRef100
ProtElectra (Elnaggar et al., 2021)	Unsupervised	Sequence	ELECTRA	1024	420M	UniRef100
ProtT5 (Elnaggar et al., 2021)	Unsupervised	Sequence	T5	1024	11B	UniRef50, BFD100
SPOT-1D-Single (Singh et al., 2021a)	Supervised	Sequence, Structure	BiLSTM, ResNet (He et al., 2016)	256	N/A	39120 proteins
ProSE (Bepler and Berger, 2021)	Supervised	Sequence, Structure	BiLSTM	1024	1M	UniRef90, SCOPe
MSA Transformer (Rao et al., 2021a)	Unsupervised	MSA	Transformer	768	100M	UniRef50, UniClust30
ESM-1v (Meier et al., 2021)	Unsupervised	Sequence	ESM-1b	1280	650M	UniRef90
ESM-IF1 (Hsu et al., 2022)	Supervised	Sequence, Structure	GVP (Jing et al., 2020), Transformer	512	142M	UniRef50, CATH
ProteinBERT (Ofer et al., 2021c)	Supervised	Sequence GO annotation	Transformer	128, 512	∼16M	UniRef90
Fold2Seq (Cao et al., 2021)	Supervised	Sequence, Structure	Transformer	256	N/A	CATH4.2
AminoBERT (Chowdhury et al., 2022)	Unsupervised	Sequence	Transformer	3072	N/A	UniParc
Evoformer (Jumper et al., 2021)	Supervised	MSA, Structure	Attention network	384, 128	93M	PDB, BFD, UniClust30, etc.
OmegaPLM (Wu et al., 2022b)	Unsupervised	Sequence	GAU (Hua et al., 2022)	1280	670M	UniRef50
OntoProtein (Zhang et al., 2022b)	Supervised	Sequence, GO	ProtBert, BERT	1024	N/A	ProteinKG25
PeTriBERT (Dumortier et al., 2022)	Unsupervised	Sequence, Structure	BERT	3072	<40M	AlphaFoldDB
MSA2Prot (Bepler and Ram, 2022)	Unsupervised	MSA	Transformer	768	N/A	Pfam
PMLM (He et al., 2022)	Unsupervised	Sequence	Transformer	1280	715M	UniRef50
ProGen2 (Nijkamp et al., 2022)	Unsupervised	Sequence	Transformer	4096	6.4B	UniRef90, BFD30
Tranception (Notin et al., 2022)	Unsupervised	Sequence	Transformer	1280	700M	UniRef100
ProtGPT2 (Ferruz et al., 2022)	Unsupervised	Sequence	GPT-2	1280	738M	UniRef50
RITA (Hesslow et al., 2022)	Unsupervised	Sequence	GPT-3	2048	1.2B	UniRef100
ESM-2 (Lin et al., 2022)	Unsupervised	Sequence	Transformer	5120	15B	UniRef50

All examples report the largest model of their public series, the model name with colour is linked with GitHub or server page. Approach and database are listed for the pre-training stage, and the latter is elaborated on in Section 8. Input is classified into protein sequence, MSA, structure (structural features or coordinates), and function. Network displays high-level backbone models preferentially if they are used. #Embedding means the dimension of embeddings; #Param., the number of parameters of network; M, millions; B, billions; T, trillions; N/A, null; &, and; <, less than; ∼ estimated data.

folding problem, meaning recovering the native sequence of a protein from its tertiary structure (coordinates). ProtTucker (Heinzinger et al., 2022) utilized pLM and contrastive learning Chen et al. (2020b); Xia et al. (2022a) strategy to improve the ability to recognize distant homologous, and other tasks, including profile prediction (Sturmfels et al., 2020), evolutionary velocity prediction (Hie et al., 2021), enzymatic active sites identification (Dassi et al., 2021), etc.

6 Methods of Protein Structure Prediction (PSP)

This section mainly introduces different methods that work for different levels of prediction of protein structures, including traditional methods and deep learning methods. Among them, pLM-based models function importantly and significantly influence protein tasks.

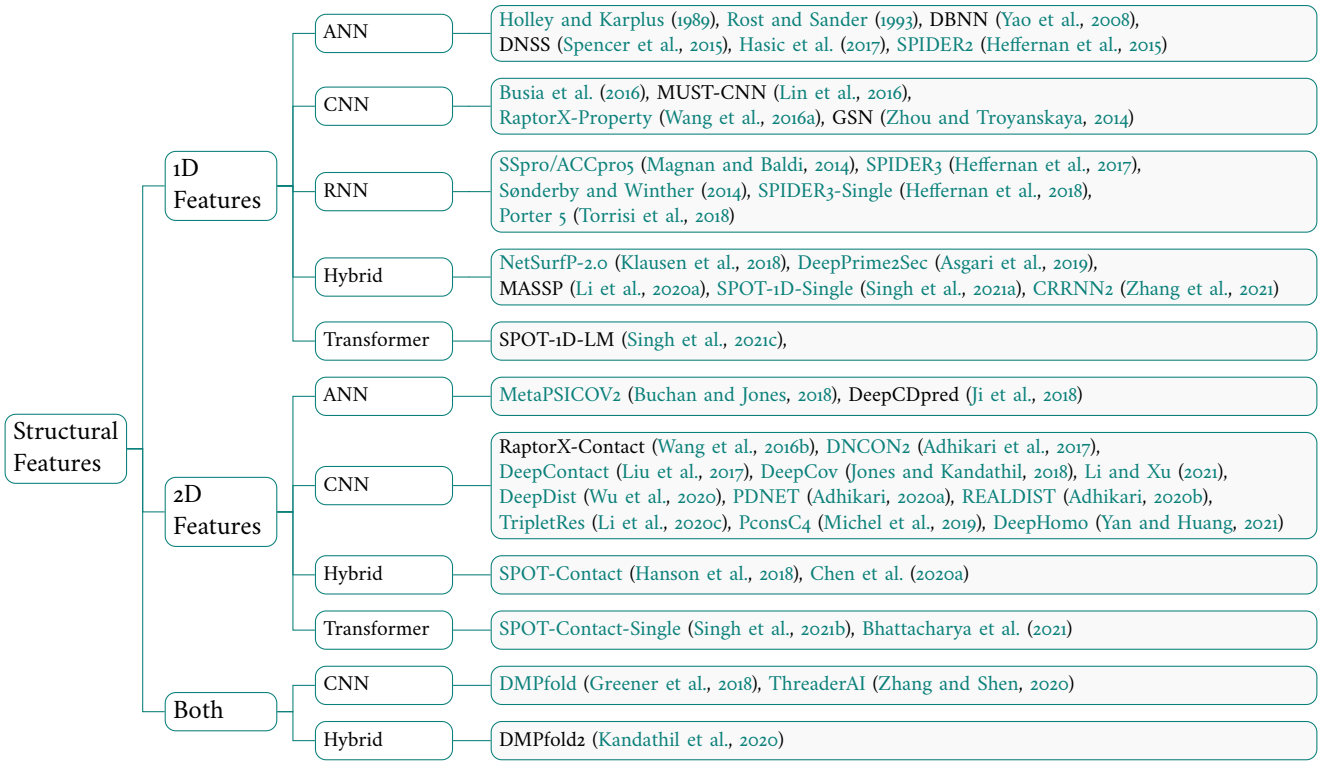


Figure 7: Taxonomy of structural feature prediction methods. ANN, Artificial Neural Network; Hybrid means models use CNN- and RNN-based methods simultaneously. The model name with colour is linked with the official GitHub or server page.

6.1 Structural Features Prediction

Structural features include 1D features (SS, SA, torsion angles, contact density, etc.) and 2D features (contact map and distance map) (Torrisi et al., 2020), which are useful for predicting protein structures. In the early stages of PSP, instead of predicting atom coordinates directly, researchers tried to predict structural features to evaluate their methods and aid the process of PSP. Methods of structural feature prediction are classified in Figure 7.

In the late 1980s, Holley and Karplus (1989) have already worked for protein secondary structure prediction (SSP) based on a neural network. DBNN (Yao et al., 2008) combined dynamic Bayesian networks and neural networks to achieve improvements in SSP. DeepCNF (Wang et al., 2015) integrated conditional random fields (CRF) and shallow convolutional neural networks to predict SS of 3-state (H, E, C) and 8-state (G, H, I, E, B, T, S, L). Heffernan et al. (2017) employed LSTM and bidirectional recurrent neural networks (BRNNs), which are capable of capturing long-range interactions. MASSP (Li et al., 2020a) used multi-tiered neural networks, which are composed of a CNN and a LSTM neural network, to get 1D structural attributes. As for 2D structural features prediction, RaptorX-Contact (Wang et al., 2016b) integrated sequence and evolutionary coupling information by two deep residual neural networks to predict contact maps. Chen et al. (2020a) presented an attention-based convolutional neural network for protein contacts, including a sequence attention module and a regional attention network. Li and Xu (2021) studied the inter-atom distance and inter-residue orientation by a ResNet network and then built 3D structure models constrained by the predicted mean and deviation through PyRoseatta (Chaudhury et al., 2010). Similarly, DeepDist (Wu et al., 2020), PDNET (Adhikari, 2020a), REALDIST (Adhikari, 2020b), etc., were also proposed to learn the real-value inter-residue distance by designing regression models, which is more difficult than multi-class classification problem (Xu, 2018) but

stepping forward for PSP.

However, the structural feature prediction tasks do not generally have significant practical values since there is already highly accurate 3D structure data from AF2. Is it a thing of the past? Considering it from another perspective, it can still provide a reference for the results of various proposed methods and work for finding the relationships between protein sequence, structure, and function by mining more protein grammars. For example, Bhattacharya et al. (2021) have introduced a simplified attention layer, factored attention, to find the role of attention, which achieved nearly identical performance to the Potts model with far fewer parameters. Particularly, DeepHomo (Yan and Huang, 2021) aimed to predict inter-protein residue-residue contacts across homo-oligomeric protein interfaces by integrating multi-source information to remove the potential intra-protein contact noises that exist in the MSA. Geoformer (Wu et al., 2022b) refined the details of contact prediction to illustrate its effectiveness in resolving the problem of triangular inconsistency.

6.2 Traditional Methods for PSP

Simons et al. (1997) generated native-like structures from fragments of unrelated protein structures using Bayesian scoring functions and a simulated annealing procedure in the 1990s. TOUCHSTONE II (Zhang et al., 2003) proposed a parallel hyperbolic sampling algorithm used in the Monte Carlo simulation processes to accelerate the conformational search faster. Rohl et al. (2004) utilized the Rosetta algorithm for *de novo* PSP, which can assemble fragments by a Monte Carlo strategy. Such a strategy was also adopted in QUARK (Xu and Zhang, 2012). Cavalli et al. (2007) used chemical shifts as structural restraints to help determine the conformations of proteins. Based on backbone chemical shifts, Schmitz et al. (2012) stepped further on the quest for reliable PSP using pseudocontact shifts. EdaFold (Simoncini et al., 2012) is a fragment-based PSP method via estimation of distribution algorithm that is learned from previously generated decoys. Stepping on this, a cluster-based model and an energy-based variation were provided by Simoncini et al. (2017). UniCon3D (Bhattacharya et al., 2016) is a generative, probabilistic model using united-residue conformational search, sampling lower energy conformations with higher accuracy than traditional random sampling.

Structural features like SS and contact maps appear to help the PSP. DCAFold (Sulkowska et al., 2012) integrated contacts estimated from direct coupling analysis with an accurate knowledge of local information to fold proteins. FragFold (Kosciolek and Jones, 2014) used fragment assembly with both statistical potentials and predicted contacts. RASREC (Braun et al., 2015) integrated evolutionary information in the form of intra-protein residue-residue contacts. CONFOLD (Adhikari et al., 2015) developed a distance geometry algorithm using structural features as restraints. Ovchinnikov et al. (2016) used structural information during Rosetta conformational sampling and refinement to improve the model's accuracy. Besides, RBO Aleph (Mabrouk et al., 2015) leveraged evolutionary and physicochemical information to predict contacts, used in conformational space search, afterwards, similar instances are SCDE (Zhang et al., 2020a), TDFO (Zhang et al., 2020b).

6.3 Deep Learning Methods for PSP: Past, Present, and Future

6.3.1 Structural Features-Based Methods

DESTINI (Gao et al., 2019) has two main components: a fully convolutional residual neural network for contact prediction with a template-based structural modelling procedure. Other than the distance map between C_β atoms, Yang et al. (2019a) and Li and Xu (2021) predicted orientations between residues via residual networks, which theoretically fully define the relative positions of the backbone atoms of two residues. The protein 3D structure is obtained from these inter-residue geometrics by energy minimization, then refined by full-atom relaxation; the pipeline is shown in Figure 8. AlphaFold (Senior et al., 2020b) optimized the potential

and constructed predicted distances between pairs of residues through a simple gradient descent algorithm, which can generate structures directly, ignoring exhausting sampling procedures. Structural features-based PSP models commonly have two steps: contact predicting and structure modelling, which have so far been restricted to the accuracy of individual components, even though scientists seek to utilize different information (MSAs, templates, biochemical properties, etc.). In contrast, end-to-end differentiable models have obtained remarkable achievements (LeCun et al., 2015).

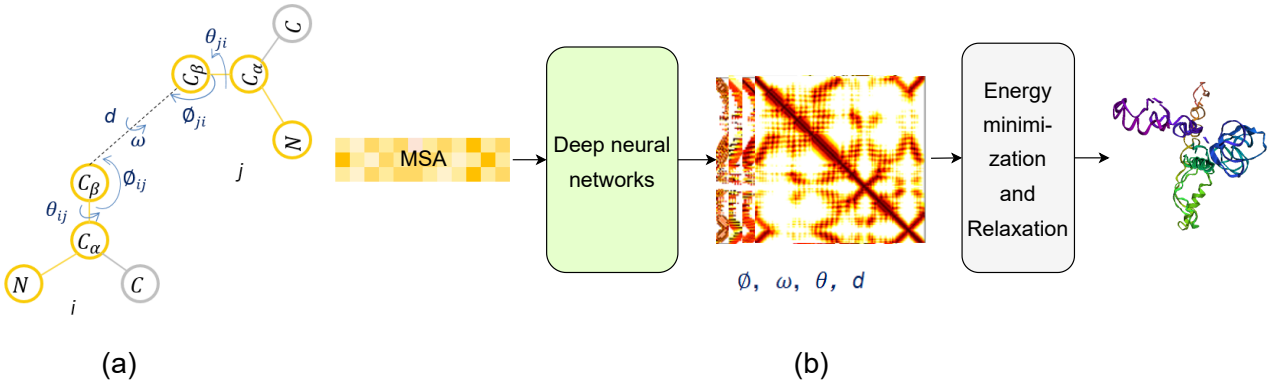


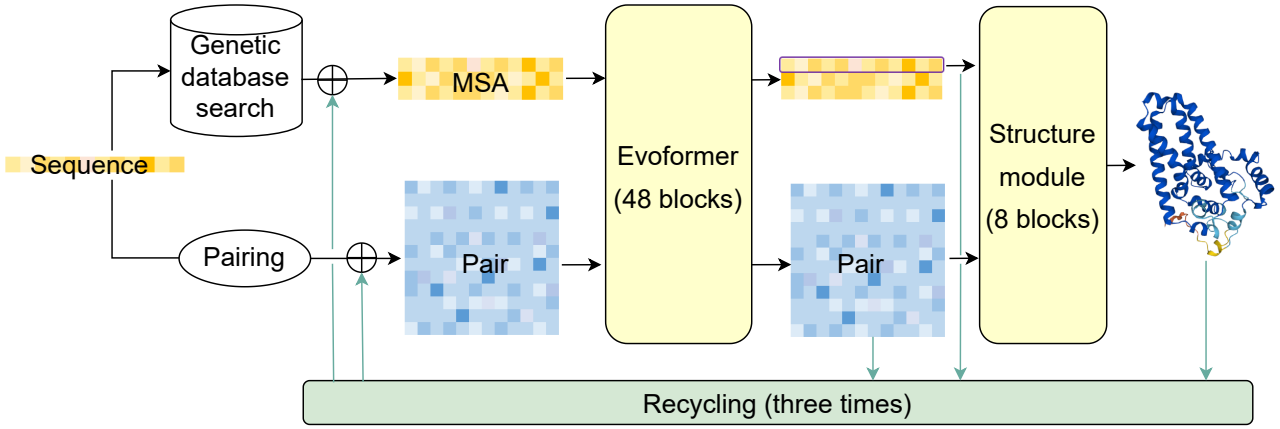
Figure 8: A pipeline of predicting structural features and protein 3D structure. (a) Inter-residue geometrics, including distances (d) and orientations, three dihedral ($\omega, \theta_{ij}, \theta_{ij}$) and two planar ($\phi_{ij}, \phi_{ji}$) angles. (b) Outline of the PSP based on structural features from MSA via energy minimization and full-atom relaxation.

6.3.2 End-to-end PSP

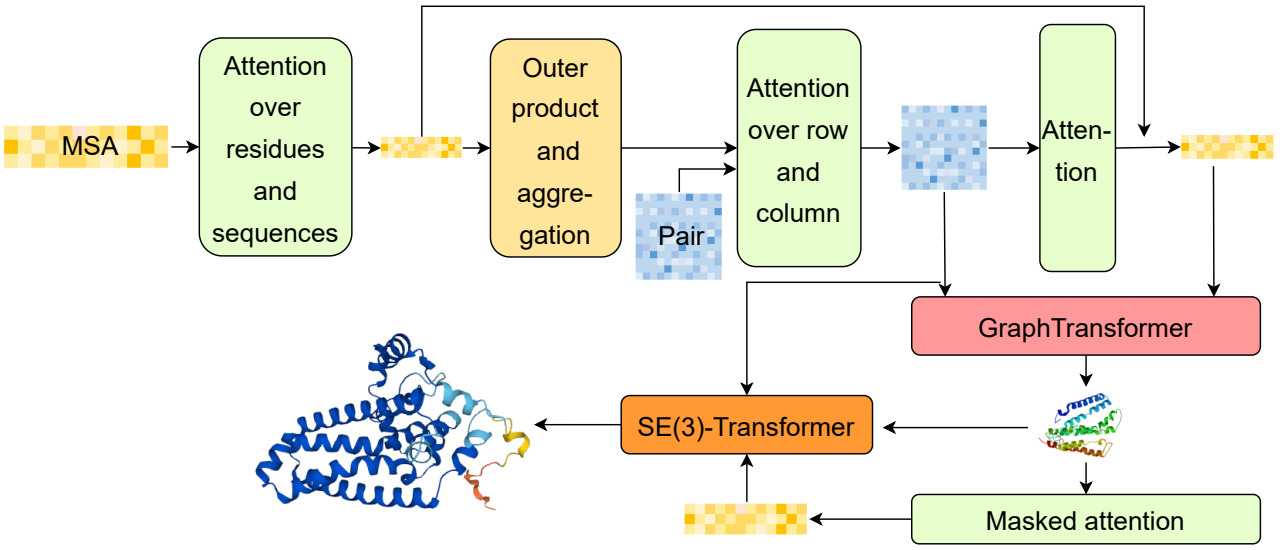
RGN (AlQuraishi, 2018) predicted torsional angles from protein sequences fed into computational units based on LSTM, which are sequentially translated into Cartesian coordinates to generate the predicted structure. Costa et al. (2021) harvested individual and contextual residual embeddings produced by MSA Transformer, assigned to nodes and edges to Graph Transformer (Shi et al., 2020). The birth of the highly-accurate model, AlphaFold2 (AF2) (Jumper et al., 2021) has stimulated the development of end-to-end models for PSP.

AF2 can produce even near-experimental results in most cases, whose accuracy was vastly higher than that of other competing methods in CASP14. The overall model architecture of AF2 is exhibited in Figure 9a. In addition to the above-mentioned Evoformer as mentioned earlier (listed in Table 2, shown in Figure 6), AF2 also has a decoder, the structure module that has eight layers with shared weights with the pair and first row MSA representations from the Evoformer as input. Each layer updates the single representation and the backbone frames, parameterized as Euclidean transforms. The structure module mainly includes the Invariant Point Attention (IPA) module, which is a form of attention that acts on a set of frames and is invariant under global Euclidean transformations because of the invariant operation, L2-norm of the global transformed vector. Moreover, AF2 executes the network three times with minor extra training time, embedding the previous outputs as additional inputs, making the network more profound and relatively important. There are various models with incremental improvements on different aspects concerning AF2, reducing training time for FastFold (Cheng et al., 2022) and Uni-Fold (Li et al., 2022), faster MSA generation for ColabFold (Mirdita et al., 2022), denoising the searched MSA or generating virtual MSA for EvoGen (Zhang et al., 2022a) that is useful for proteins lacking sequence homology, and re-implementation of AF2(HelixFold (Wang et al., 2022b), MEGA-Fold (Liu et al., 2022b), OpenFold (Ahdritz et al., 2022)).

As for another famous PSP work, RoseTTAFold (Baek et al., 2021) (as shown in Figure 9b), it is a three-track



(a) AF2



(b) RosettaFold

Figure 9: A simplified schematic of AF2 and RosettaFold architectures.

model with attention layers in which information flows back and forth at the 1D, 2D, and 3D levels between sequences, distances, and coordinates, which is mainly consisted of seven modules. Firstly, the MSA features are processed by attention over rows and columns, and then the processed features are aggregated by the outer product that can obtain the correlation (coevolution) between two residues in each sequence to update pair features, which are refined via axial attention. Next, the MSA features are also updated based on attention maps derived from pair features, which had a good agreement with the true contact map. Using these learned MSA and pair features as the node and edge embeddings and building a fully-connected graph, Graph Transformer is employed to estimate the initial 3D structure, and new attention maps can be derived from the current structure to update MSA features. Finally, 3D coordinates are refined by SE(3)-Transformer (Fuchs et al., 2020) based on updated MSA and pair features.

Because of the remarkable success of AF2 and RosettaFold, a set of research has emerged, e.g., the protein-peptide binders identification (Chang and Perez, 2022), applying to small molecules (Hekkelman et al., 2021), antibody structure prediction (Ruffolo and Gray, 2022), protein complex structure prediction (Bryant et al., 2021; Evans et al., 2021), RNA 3D structure prediction (Shen et al., 2022), and generating multiple protein conformations (Stein and Mchaourab, 2022), so on and so forth, which are introduced in the following two

subsections. A set of representative methods for PSP and related tasks are shown in Table 3.

6.3.3 Single Sequence Structure Prediction via pLMs

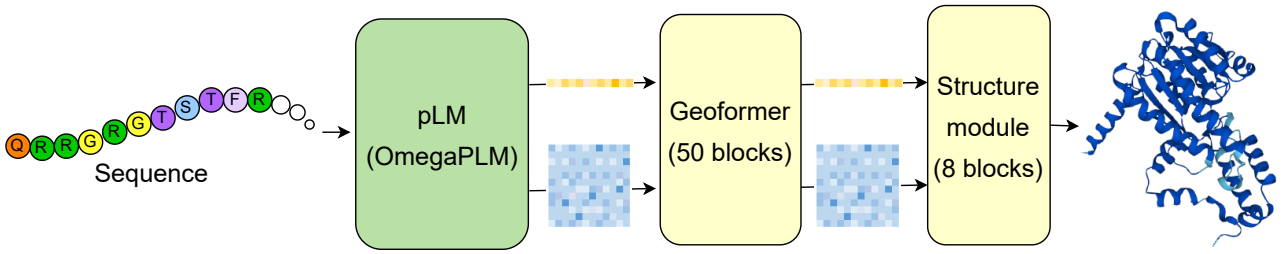


Figure 10: Model architecture of OmegaFold.

These mainstream PSP pipelines heavily rely on MSAs, which becomes a kind of bottleneck for the reason that it is time-consuming and resource-intensive to search for MSAs and templates. Besides, a protein sequence can theoretically determine its structure; MSA-aware models may memorize the determined structures of similar proteins for PSP, making it difficult for us to understand the mechanism of folding in reality. In order to predict structure from only sequences without MSAs, a set of recent works deemed pLMs can capture grammatically structural information from protein sequence databases. OmegaFold (Wu et al., 2022b) utilized OmegaPLM to obtain the residue and pairwise embeddings from a single sequence, which are fed into Geoformer. Different from the triangle multiplicative and triangle attention modules applied in Evoformer to enforce the edge representation to satisfy the triangle inequality of distances, Geoformer considers maintaining geometric consistency in the high-dimensional space as well as in the Euclidean space with the help of the structure module implemented by AF2. The model architecture is demonstrated in Figure 10. The results of contact predictions illustrated that accuracy improves while inconsistency drops with stacked Geoformer layers. Thus, OmegaFold enables accurate predictions on orphan proteins, similar to AF2 structures. Other than OmegaFold, a contemporaneous work, ESMFold (Lin et al., 2022), has trained large-scale pLMs (ESM2) to learn more structural and functional properties to replace the role of MSAs; similar examples are HelixFold-Single (Fang et al., 2022), RGN2 (Chowdhury et al., 2022). While trRosettaX-Single (Wang et al., 2022d) used pre-trained pLMs (ESM-1b) and retrained it based on supervised learning (s-ESM-1b), distilling knowledge from a pre-trained MSA-based network (Res2Net_MSA) to the student network (Res2Net_Single) to predict distance and orientations.

6.3.4 Diversified Progress in the Protein Field

Roney and Ovchinnikov (2022) found that AlphaFold has the ability to learn a highly accurate biophysical energy function and find a low-energy conformation using co-evolutionary information, which is important for single-sequence PSP using physical principles (Roney, 2022). However, AF2 structures may contain protein segments that are placed with uncertainty (Akdel et al., 2022) that need manual inspection instead of purely relying on machines. Most native proteins folded into stable conformations with the lowest free energy (Cb and Scheraga, 1975). Therefore, both Atom Transformer (Du et al., 2020) and GraphEBM (Liu et al., 2022a) are the energy-based models for protein side-chain conformation with different networks, a transformer model and a GNN (Busbridge et al., 2019; Sanyal et al., 2020) enhanced model, which predicted scores of side-chain conformations for given structures.

Based on AlphaFold-Multimer, ColAttn (Chen et al., 2022a) made use of pLMs to identify interologs of a complex by estimating the column attention weight matrix. Elofsson et al. (2022) completed a large protein complex assembly using a Monte Carlo tree search to address the problem of AlphaFold-multimer, whose performance declined rapidly for proteins with three or more chains. Colossal-AI (Bian et al., 2021) team has developed pipelines (xTrimoMultimer, xTrimoABFold and xTrimoDock) based on pLMs for the structure prediction of protein complexes and antibodies. To predict the structures of protein-nucleic acid complexes, Baek et al. (2022) proposed RoseTTAFoldNA based on RoseTTAFold, which is a unified framework for protein-DNA, protein-RNA complexes, and RNA tertiary structures.

When having a coarse-grained representation of protein structures, EquiFold (Lee et al., 2022) predicted protein structures via iterative refinement using an SE(3)-equivariant (Liao and Smidt, 2022) neural network. Structure refinement also takes coarse structures as input and output refined coordinates. This process has been integrated into some PSP models, like AF2 and RoseTTaFold. Different from DeepACCNet (Hiranuma et al., 2020) and GNNRefine (Jing and Xu, 2021), which used predicted distances to guide PyRosetta (Chaudhury et al., 2010) protein structure refinement, a SE(3)-equivariant graph transformer network that is equivariant to the rotation and translation of coordinates was developed in ATOMRefine (Wu and Chen, 2022) for all-atom structural refinement end-to-end, where initial structures were generated by AF2. In contrast, GNNRefine performs relatively poorly on AF2 structures. Researchers have become adept at working out more challenging problems. For example, DCGAN (Anand and Huang, 2018) has applied generative adversarial networks (GANs) to generate pairwise distance maps from corrupted protein structures and thus recover robust 3D structures. Wu et al. (2022a) trained a denoising diffusion-based generative model with only a vanilla transformer, which generated high-quality, and diverse protein structures inspired by the huge success of diffusion models in a wide range of data modalities (Rombach et al., 2022; Rouard and Hadjeres, 2021), such kind of generative models can benefit structure refinement, protein conformation and structure prediction by generating biologically plausible and robust structures, it is worth to expect more successes achieved by diffusion model in the biomedical field.

7 Discussion: Limitations and Future Trends

With the million-level protein sequence databases, pLMs have been becoming larger and larger (billion-level parameters for ESM-2), which are dominated by big companies in reality. For example, DeepMind used 128 TPU v3 cores to train and fine-tune AF2 over one week. Training such deep learning models might only be accessible to large companies such as Google; it seems hard for academic research groups to learn protein embeddings from the start, which is also a burden for the theme of a green environment. Considering this problem, researchers can put more attention to methods' innovations.

Firstly, why not better utilize these large-scale pre-trained pLMs? How to best leverage them is still under-explored compared to pre-trained language models in NLP; it means developing suitable and special transfer learning methods to adapt knowledge to various downstream tasks. Knowledge distillation is inspiring; MSA, structure and function information can be distilled and transferred (Costa et al., 2021), which is used in NLP (Yang et al., 2020). Secondly, a large-scale pLM cannot tackle all problems, e.g., some reports have indicated that AF2 does not appear to be well suited to predict the impact of mutations on proteins (Buel and Walters, 2022; Pak et al., 2021). Structure-triggered tasks cannot be directly transferred to other prediction tasks (Hu et al., 2022). Furthermore, Nijkamp et al. (2022) found bigger models may not translate into better zero-shot fitness performance. Therefore, there still exists space for different models to tackle various problems. Thirdly, ESM-2 concluded that the improvement of low-scale models is easily saturated with high evolutionary depth, while with the low evolutionary depth, it continues when models' sizes increase, which illustrates that combining extra information suitably, including injecting biological and physical priors, using evolutionary

Table 3: Representative methods for PSP and related tasks.

Model	Evolution	Energy	Main Network	Task
RGN (AlQuraishi, 2018)	✓	✗	LSTM	PSP
Costa et al. (2021)	✓	✗	MSA Transformer Graph Transformer (Shi et al., 2020)	PSP
AF2 (Jumper et al., 2021)	✓	✗	Evoformer, IPA	PSP
RoseTTAFold (Baek et al., 2021)	✓	✗	Graph Transformer SE(3)-Transformer (Fuchs et al., 2020)	PSP
FastFold (Cheng et al., 2022)	✓	✗	AF2	PSP
ColabFold (Mirdita et al., 2022)	✓	✗	AF2	PSP
HelixFold (Wang et al., 2022b)	✓	✗	AF2	PSP
MEGA-Fold (Liu et al., 2022b)	✓	✗	AF2	PSP
Uni-Fold (Li et al., 2022)	✓	✗	AF2	PSP
EvoGen (Zhang et al., 2022a)	✓	✗	Attention	MSA Generation
ESMFold (Lin et al., 2022)	✗	✗	ESM-2	Single Sequence PSP
OmegaFold (Wu et al., 2022b)	✗	✗	OmegaPLM, Geoformer	Single Sequence PSP
HelixFold-Single (Fang et al., 2022)	✗	✗	AF2	Single Sequence PSP
RGN2 (Chowdhury et al., 2022)	✗	✗	AminoBERT	Single Sequence PSP
trRosettaX-Single (Wang et al., 2022d)	✗	✓	s-ESM-1b, Res2Net_Single	Single Sequence PSP
EquiFold (Lee et al., 2022)	✗	✗	SE(3)-Transformer	Antibody Structure Prediciton
IgFold (Ruffolo and Gray, 2022)	✗	✗	Graph Transformer, AntiBERTy, IPA	Antibody Structure Prediciton
AlphaFold-Multimer (Evans et al., 2021)	✓	✗	AF2	Complex Structure Prediciton
Bryant et al. (2021)	✓	✗	AF2	Complex Structure Prediciton
ColAttn (Chen et al., 2022a)	✓	✗	MSA Transformer, AlphaFold-Multimer	Complex Structure Prediciton
Elofsson et al. (2022)	✓	✗	AlphaFold-Multimer	Complex Structure Prediciton
RoseTTAFoldNA (Baek et al., 2022)	✓	✗	RoseTTAFold	Structure Prediction of RNA, Protein -DNA & Protein-RNA Complexes
E2Efold-3D (Shen et al., 2022)	✓	✗	Transformer, IPA	RNA Structure Prediction
DCGAN (Anand and Huang, 2018)	✗	✓	GAN	Structure Generation
Wu et al. (2022a)	✗	✗	Transformer	Structure Generation
Atom Transformer (Du et al., 2020)	✗	✓	Transformer	Protein Conformation
SPEACH_AF (Stein and Mchaourab, 2022)	✓	✗	AF2	Protein Conformation
GraphEBM (Liu et al., 2022a)	✗	✓	DimeNet (Klicpera et al., 2020)	Protein Conformation
GNNRefine (Jing and Xu, 2021)	✗	✗	GNN	Structure Refinement
ATOMRefine (Wu and Chen, 2022)	✗	✗	SE(3) Graph Transformer	Structure Refinement

✓ and ✗ appeared in the column of Evolution, and Energy represents this model whether used this information or not. The model name with colour is linked with codes or its server page. Most of these present methods are pLM-aware models. &: and.

information, different levels of structures, GO, etc., can reduce models' size and boost their performance. These different types of information (heterogeneous data) are involved with data mining and task design. Thus, multi-task or multi-modal learning is a direction that deserves to have more exploration (Bepler and Berger, 2021). Finally, since sequence mutations can cause genetic diseases, their effects on function form a landscape that reveals how function constrains sequence. Predicting robust protein structures and understanding the functional effects of sequence mutations are critical.

The architectures of most current pLMs are the same as LMs' in NLP, i.e., people use these LMs in the protein communities without much adaptation and modification, which hinders the development of pLMs and the understanding of protein grammars. Techniques present in LMs are also used in pLMs, like masking strategies. Proteins and Languages have similarities, but they are not the same as we have mentioned in Section 3. Therefore, designing proper pLMs and developing suitable methods for protein data considering its properties is an important and urgent problem.

Despite the proliferation of LMs, these pLM-based models still lack interpretability, hindering people from

understanding the mechanism behind protein folding. AF2-like models cannot provide the detailed understanding of molecular and chemical interactions that is important for studies of molecular mechanisms and structure-based drug design. Thus, model interpretability helps researchers find protein grammars useful in biomedical applications. Visualization is a tool that can record the process of protein folding or how the protein functions during various activities. Therefore, developing a high-quality protein tokenization method or combining it with other machine learning methods is appealing, like Bayesian modelling, Monte Carlo strategy, energy functioning, Markov random field, direct coupling analysis (Weigt et al., 2009), or diffusion models.

Due to the different large databases of proteins, people need to be concerned about the wrong information in the databases. Some research groups choose to build new datasets to satisfy their unique needs; data leakage, bias, and ethics should be considered in this situation. For instance, ProteinKG25 was created for utilizing protein sequence and functions (Zhang et al., 2022b) better to get meaningful embeddings. On the other hand, building complete, reliable, and just benchmarks is essential to evaluating various models and promoting the appearance of solid methods.

Scientists have been developing a unified model for DNA, RNA, and protein engineering in the biomedical domain (Wang et al., 2022a), BERT-RBP (Yamada and Hamada, 2021) adopted BERT architecture to predict RNA-protein interactions, still has a distance to go to interpret their relationships well. Single protein sequences, protein-ligand complexes, and RNA complexes structure prediction and conformation modelling have been appealing to a group of researchers. These problems are more complicated than those in PSP, and they are not yet solved. Even for PSP, there is still space for models predicting structures with ultra-high accuracy (resolution less than 0.5Å).

8 Databases

We follow the mapping for LMs from NLP to proteins in ProtTrans (Elnaggar et al., 2022), which interpret proteins as "sentences". Therefore, the number of proteins that needed disk space for datasets and official web pages are listed in Table 4.

In detail, UniProt (Consortium, 2013) provides a comprehensive, high-quality, and freely accessible resource of protein information, where UniProt databases include UniProtKB, UniRef (Suzek et al., 2015), UniPac databases. The UniProt Knowledgebase (UniProtKB) is comprised of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former provides the reviewed relevant information about a particular protein, including protein and gene names, function, protein-protein interactions, etc. Proteins in the latter largely have no experimental data available, while these unreviewed records are enriched with automatic annotation and classification. The UniRef databases (UniProt Reference Clusters) (Suzek et al., 2015) provide clustered sets of sequences from the UniProt Knowledgebase (UniProtKB) and provide complete coverage of sequence space at several resolutions (100%, 90%, 50% identity) in a hierarchical fashion. UniRef50 clusters are generated using UniRef90 clusters, which are based on UniRef100 clusters. UniRef databases have a broad range of usages, like functional annotation, family classification, systems biology, structural genomics, phylogenetic analysis, and mass spectrometry. The UniProt Archive (UniParc) is the most comprehensive publicly accessible non-redundant protein sequence database without record annotation.

The Protein Data Bank (PDB) (wwPDB consortium, 2019) is a database for the experimentally determined three-dimensional structural data of biological macromolecules, including the sequence and 3D structures of proteins. The number of structures in the PDB has grown at an approximately exponential rate based on historical experience, and there are 170k multi-chain protein structure files as of October 2021. However, a million-level protein structure prediction named as PSP (Liu et al., 2022b) was presented based on PDB after

the appearance of AF2 with 570k true structure sequences and 745k complementary distillation sequences.

The CATH (Orengo et al., 1997) is a novel public hierarchical classification database of protein domain structures for public download, where these domains are obtained from protein three-dimensional structures deposited in the PDB. The domains are classified into four main levels: protein class (C), which describes the secondary structure composition of each domain, i.e., all alpha, all beta, mixed alpha-beta, or few secondary structures; at the architecture (A) level, the arrangement of secondary structures is summarized; at the topology/fold (T) level, the sequential connectivity is taken into account; at the homologous superfamily (H) level, proteins have a demonstrable evolutionary relationship. Structured Transformer (Ingraham et al., 2019) uses CATH4.2, processed and cluster-split into training, validation, and test sets, which contain 18025 chains in the training set, 1637 chains in the validation set, and 1911 chains in the test set. The latest release of CATH-Gene3D (v4.3) has 151 million protein domains classified into 5,841 superfamilies and is to predict the locations of structural domains on millions of publicly available protein sequences.

The Big Fantastic Database (BFD) is a sequence profile database and is one of the largest metagenomic databases, as it keeps one copy of duplicates from UniProt and other big protein databases. These sequences were clustered by a sequence identity cut-off of 30% and a coverage threshold of 90% using MMseqs2/Linclust (Hou et al., 2022; Steinegger et al., 2019b; Steinegger and Söding, 2018), e.g., the BFD30 dataset was clustered to 65 million clusters at 30% sequence identity. Searching for a protein against BFD is slow but more sensitive. Sometimes it is convenient to use MSA, computed from each of the BFD clusters.

Pfam (El-Gebali et al., 2019) is a database of protein families used extensively in bioinformatics. It aims to provide a complete and accurate classification of protein families and domains, which includes annotations and alignments generated by hidden Markov models (HMMs). *pfamseq* is a profile HMM that is queried against a sequence database based on UniProtKB and used to find homologues for a Pfam entry. The recent version of Pfam contains 19,632 families; each family includes descriptions, alignments, architectures, etc.

The AlphaFold Protein Structure Database (AlphaFoldDB) is a collection of protein structure predictions created by DeepMind in partnership with EMBL-EBI. These 3D structures are predicted by AlphaFold, which achieves accuracy competitive with experiments. The latest database release contains over 200 million entries for UniProt.

ProteinKG25 is a large-scale Knowledge Graph (KG) dataset with aligned descriptions and protein sequences, respectively, to Gene Ontology (GO) terms and protein entities (Zhang et al., 2022b). Go terms are seen as graph nodes, and the relationships between the terms are edges. ProteinKG25 combines these two different structures, GO and Gene Annotation into a unified KG for training LMS that incorporates GO information. It includes most of the triplets (Protein-GO triplets and GO-GO triplets) and a few entities (proteins and gene terms) and relations.

Uniclust databases (Mirdita et al., 2017) cluster UniProtKB sequences at the level of 90%, 50% and 30% pairwise sequence identity, formed Uniclust90, Uniclust50, Uniclust30. Compared with UniRef, which relied on the CD-HIT software for clustering, Uniclust databases used the developed software suite MMseqs2, which had higher consistency in the functional annotation. Furthermore, Uniclust sequences are annotated with matches to other commonly used datasets (Pfam, SCOP (Hubbard et al., 1997) and PDB), which may not be annotated in UniProt with HHblits (Steinegger et al., 2019a).

The Structural Classification of Proteins (SCOP) database (Hubbard et al., 1997) was created by manual classification of protein structural domains based on similarities of their structures and evolutionary relationships. Similar to CATH and Pfam, the original hierarchical organizations of SCOP are class, fold, superfamily, and family. The new version, SCOP2, was released in 2020 to provide a better database for protein structure annotation and classification. Structural Classification of Proteins—extended (SCOPe) extends the SCOP database of protein structural relationships. The ASTRAL compendium provides tools to help research protein structure

Table 4: Information of Protein Databases

Dataset	#Proteins	Disk Space	Description	Link
UniProtKB/Swiss-Prot	500K	0.59GB	knowledgebase	https://www.uniprot.org/uniprotkb?query=*
UniProtKB/TrEMBL	229M	146GB	knowledgebase	https://www.uniprot.org/uniprotkb?query=*
UniRef100	314M	76.9GB	clustered sets of sequences	https://www.uniprot.org/uniref?query=*
UniRef90	150M	34GB	90% identity	https://www.uniprot.org/uniref?query=*
UniRef50	53M	10.3GB	50% identity	https://www.uniprot.org/uniref?query=*
UniParc	528M	106GB	Sequence	https://www.uniprot.org/uniparc?query=*
PDB	190K	50GB	3D structure	https://www.wwpdb.org/ftp/pdb-ftp-sites
CATH4.3	N/A	1073MB	hierarchical classification	https://www.cathdb.info/
BFD	2500M	272GB	sequence profile	https://bfd.mmseqs.com/
Pfam	47M	14.1GB	protein families	https://www.ebi.ac.uk/interpro/entry/pfam/
AlphaFoldDB	214M	23 TB	predicted 3D structures	https://alphafold.ebi.ac.uk/
ProteinKG25	5.6M	147MB	a KG dataset with GO	https://drive.google.com/file/d/1iTC2-zbvYZCDhWM_wxRufCvV6vvPk8HR
Uniclust30	N/A	6.6GB	clustered protein sequences	https://uniclust.mmseqs.com/
SCOP	N/A	N/A	structural classification	http://scop.mrc-lmb.cam.ac.uk/
SCOPe	N/A	86MB	extented version of SCOP	http://scop.berkeley.edu

K, thousand; M, million, disk space is in GB or TB (compressed storage as text), which is estimated data influenced by the compressed format.

and evolution. The SCOPe has corrected errors in SCOP and incorporated the Astral database (Chandonia et al., 2017). The latest release of SCOPe ASTRAL 2.08 has more than 50K protein sequences with known structures and SCOP classifications.

Critical Assessment of Protein Structure Prediction (CASP) is a worldwide experiment for PSP that will have been conducted 15 times by the end of 2022. Research groups from all over the world participate in CASP to objectively test their structure prediction methods. By categorizing different themes, like quality assessment, model refinement, domain boundary prediction, protein complex structure prediction, etc., and selecting target proteins, CASP researchers can identify what progress has been made and highlight the future efforts that may be most productively focused on.

9 Conclusion

This paper systematically summarizes recent advances in pLMs, PSP, and related tasks, including background, why and how pLMs are used in protein representation learning, existing pLMs and PSP methods, the relationships between pLMs and PSP, and how pLMs function in the development of protein folding. A set of databases are introduced. Furthermore, we also discuss some limitations, possible tackling directions, and future trends. Finally, we expect that LMs and pLMs can aid more in the specific biochemical, biomedical, and bioinformatic domains.

References

Adhikari, B. (2020a). A fully open-source framework for deep learning protein real-valued distances. *Scientific Reports*.

Adhikari, B. (2020b). Realdist: Real-valued protein distance prediction. *bioRxiv*.

Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). Confold: Residue-residue contact-guided ab initio protein folding. *Proteins*.

Adhikari, B., Hou, J., and Cheng, J. (2017). Dncon2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*.

AGMLS, F., Bunke, R., and Schmiduber, J. (2009). A novel connectionist system for improved unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 31(5).

Ahdritz, G., Bouatta, N., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M. M., Zhang, S., Ojewole, A., Guney, M. E., Biderman, S., Watkins, A. M., Ra, S., Lorenzo, P. R., Nivon, L., Weitzner, B., Ban, Y.-E. A., Sorger, P. K., Mostaque, E., Zhang, Z., Bonneau, R., and AlQuraishi, M. (2022). Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*.

Akdel, M., Pires, D. E., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., Bryant, P., Good, L. L., Laskowski, R. A., Pozzati, G., et al. (2022). A structural biology community assessment of alphafold2 applications. *Nature Structural & Molecular Biology*, pages 1–12.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*.

AlQuraishi, M. (2018). End-to-end differentiable learning of protein structure. *Cell systems*.

AlQuraishi, M. (2019). End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301.

Anand, N. and Huang, P.-S. (2018). Generative modeling for protein structures. *Learning*.

Andreas, J., Vlachos, A., and Clark, S. (2013). Semantic parsing as machine translation. *meeting of the association for computational linguistics*.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Asgari, E. and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE*.

Asgari, E., Poerner, N., McHardy, A. C., and Mofrad, M. R. K. (2019). Deepprime2sec: Deep learning for protein secondary structure prediction from the primary sequences. *bioRxiv*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S. E., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*.

Baek, M., Mchugh, R., Anishchenko, I., Baker, D., and Dimaio, F. (2022). Accurate prediction of nucleic acid and protein-nucleic acid complexes using rosettafoldna.

Bahdanau, D., Cho, K., and Bengio, Y. (2014a). Neural machine translation by jointly learning to align and translate. *arXiv: Computation and Language*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014b). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bao, H., Dong, L., and Wei, F. (2021). Beit: Bert pre-training of image transformers. *Learning*.

Bao, H., Dong, L., Wei, F., Wang, W., Yang, N., Liu, X., Wang, Y., Piao, S., Gao, J., Zhou, M., and Hon, H.-W. (2020). Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Preprint*.

Basile, W., Sachenkova, O., Light, S., and Elofsson, A. (2017). High gc content causes orphan proteins to be intrinsically disordered. *PLoS computational biology*, 13(3):e1005375.

Bateman, A. (2019). Uniprot: A worldwide hub of protein knowledge. *Nucleic Acids Research*.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., Kandola, J., Hofmann, T., Poggio, T., and Shawe-Taylor, J. (2022). A neural probabilistic language model.

Bepler, T. and Berger, B. (2019). Learning protein sequence embeddings using information from structure. *Learning*.

Bepler, T. and Berger, B. (2021). Learning the protein language: Evolution, structure, and function. *Cell systems*.

Bepler, T. and Berger, B. (2022). Learning the protein language: Evolution, structure, and function.

Bepler, T. and Ram, S. (2022). Few shot protein generation.

Berman, H. M., Westbrook, J. D., Feng, Z., Gilliland, G. L., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*.

Bhattacharya, D., Cao, R., and Cheng, J. (2016). Unicon3d: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics*.

Bhattacharya, N., Thomas, N., Rao, R., Daupras, J., Koo, P. K., Baker, D., Song, Y. S., and Ovchinnikov, S. (2021). Single layers of attention suffice to predict protein contacts. *bioRxiv*.

Bian, Z., Liu, H., Wang, B., Huang, H., Li, Y., Wang, C., Cui, F., and You, Y. (2021). Colossal-ai: A unified deep learning system for large-scale parallel training. *arXiv preprint arXiv:2110.14883*.

Bishop, M. and Thompson, E. A. (1986). Maximum likelihood alignment of dna sequences. *Journal of molecular biology*, 190(2):159–165.

Blackstock, J. C. (1989). Guide to biochemistry.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. (2021). Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Braun, T., Leman, J. K., and Lange, O. F. (2015). Combining evolutionary information and an iterative sampling strategy for accurate protein structure prediction. *PLOS Computational Biology*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *neural information processing systems*.

Bryant, P., Pozzati, G., and Elofsson, A. (2021). Improved prediction of protein-protein interactions using alphafold2 and extended multiple-sequence alignments. *bioRxiv*.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195.

Buchan, D. W. A. and Jones, D. T. (2018). Improved protein contact predictions with the metapsicov2 server in casp12. *Proteins*.

Buel, G. R. and Walters, K. J. (2022). Can alphafold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology*.

Busbridge, D., Sherburn, D., Cavallo, P., and Hammerla, N. Y. (2019). Relational graph attention networks. *arXiv preprint arXiv:1904.05811*.

Busia, A., Collins, J., and Jaitly, N. (2016). Protein secondary structure prediction using deep multi-scale convolutional neural networks and next-step conditioning. *arXiv: Learning*.

Cao, Y., Das, P., Chen, P.-Y., Chenthamarakshan, V., Melnyk, I., and Shen, Y. (2021). Fold2seq: A joint sequence(1d)-fold(3d) embedding-based generative model for protein design. *international conference on machine learning*.

Cavalli, A., Salvatella, X., Dobson, C. M., and Vendruscolo, M. (2007). Protein structure determination from nmr chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America*.

Cb, A. and Scheraga, H. A. (1975). Experimental and theoretical aspects of protein folding. *Advances in Protein Chemistry*.

Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2017). Scope: Manual curation and artifact removal in the structural classification of proteins - extended database. *Journal of Molecular Biology*.

Chang, L. and Perez, A. (2022). Alphafold encodes the principles to identify high affinity peptide binders.

Chaudhury, S., Lyskov, S., and Gray, J. J. (2010). Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*.

Chen, B., Xie, Z., Xu, J., Qiu, J., Ye, Z., and Tang, J. (2022a). Improve the protein complex prediction with protein language models.

Chen, C., Wu, T., Guo, Z., and Cheng, J. (2020a). Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. *Proteins*.

Chen, C., Zhou, J., Wang, F., Liu, X., and Dou, D. (2022b). Structure-aware protein self-supervised learning.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Cheng, S., Wu, R., Yu, Z., Li, B., Zhang, X., Peng, J., and You, Y. (2022). Fastfold: Reducing alphafold training time from 11 days to 67 hours.

Chiu, J. T. and Rush, A. M. (2020). Scaling hidden markov language models. *arXiv preprint arXiv:2011.04640*.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Choshen, L. and Abend, O. (2019). Automatically extracting challenge sets for non local phenomena in neural machine translation. *ArXiv*, abs/1909.06814.

Chou, K.-C. and Cai, Y.-D. (2003). Predicting protein quaternary structure by pseudo amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 53(2):282–289.

Chowdhury, R., Bouatta, N., Biswas, S., Rochereau, C., Church, G. M., Sorger, P. K., and AlQuraishi, M. (2022). Single-sequence protein structure prediction using language models from deep learning. *Nature Biotechnology*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *Learning*.

Consortium, U. (2013). Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic Acids Research*.

Costa, A., Ponnapati, M., Jacobson, J. M., and Chatterjee, P. (2021). Distillation of msa embeddings to folded protein structures with graph transformers. *bioRxiv*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *meeting of the association for computational linguistics*.

Dassi, L. K., Manica, M., Probst, D., Schwaller, P., Teukam, Y. G. N., and Laino, T. (2021). Identification of enzymatic active sites with unsupervised language modeling. *ChemRxiv*.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., Haas, R. J. D., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. (2022). Robust deep learning based protein sequence design using proteinmpnn.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.

Du, Y., Meier, J., Ma, J., Fergus, R., and Rives, A. (2020). Energy-based models for atomic-resolution protein conformations. *Learning*.

Dumortier, B., Liutkus, A., Mas, A., and Krouk, G. (2022). Petribert : Augmenting bert with tridimensional encoding for inverse protein folding and design.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The pfam protein families database in 2019. *Nucleic Acids Research*.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2021). Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2022). Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing.

Elofsson, A., Bryant, P., Pozzati, G., Zhu, W., Shenoy, A., and Kundrotas, P. (2022). Predicting the structure of large protein complexes using alphafold and monte carlo tree search.

Emerson, I. A. and Amala, A. (2017). Protein contact maps: A binary depiction of protein 3d structures. *Physica A-statistical Mechanics and Its Applications*.

Evans, R., O'Neill, M. J., Pritzel, A., Antropova, N., Senior, A. W., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R. D., Clancy, E., Kohli, P., Jumper, J. M., and Hassabis, D. (2021). Protein complex prediction with alphafold-multimer. *bioRxiv*.

Fang, X., Wang, F., Liu, L., He, J., Lin, D., Xiang, Y., Zhang, X., Wu, H., Li, H., and Song, L. (2022). Helixfold-single: Msa-free protein structure prediction by using protein language model as an alternative.

Fedus, W., Zoph, B., and Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv: Learning*.

Ferruz, N. and Höcker, B. (2022). Controllable protein design with language models. *Nature Machine Intelligence*, pages 1–12.

Ferruz, N., Schmidt, S., and Höcker, B. (2022). A deep unsupervised language model for protein design.

Fuchs, F. B., Worrall, D. E., Fischer, V., and Welling, M. (2020). Se(3)-transformers: 3d roto-translation equivariant attention networks. *neural information processing systems*.

Gales, M., Young, S., et al. (2008). The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304.

Gao, M., Zhou, H., and Skolnick, J. (2019). Destini: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports*.

Gao, Z., Tan, C., and Li, S. Z. (2022a). Alphadesign: A graph protein design method and benchmark on alphafolddb.

Gao, Z., Tan, C., and Li, S. Z. (2022b). Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*.

Gauto, D. F., Estrozi, L. F., Schwieters, C. D., Effantin, G., Macek, P., Sounier, R., Sivertsen, A. C., Schmidt, E., Kerfah, R., Mas, G., et al. (2019). Integrated nmr and cryo-em atomic-resolution structure determination of a half-megadalton enzyme complex. *Nature communications*, 10(1):1–12.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2020). Knowledge distillation: A survey. *International Journal of Computer Vision*.

Greener, J. G., Kandathil, S. M., and Jones, D. T. (2018). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications*.

Gupta, A., Müller, A. T., Huisman, B. J., Fuchs, J. A., Schneider, P., and Schneider, G. (2018). Generative recurrent networks for de novo drug design. *Molecular informatics*, 37(1-2):1700111.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., Yuan, J., Zhao, W. X., and Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*.

Hanson, J., Paliwal, K. K., Litfin, T., Yang, Y., and Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*.

Hao, X.-H., Zhang, G.-J., and Zhou, X.-G. (2017). Conformational space sampling method using multi-subpopulation differential evolution for de novo protein structure prediction. *IEEE Transactions on NanoBioscience*, 16(7):618–633.

Hasic, H., Buza, E., and Akagic, A. (2017). A hybrid method for prediction of protein secondary structure based on multiple artificial neural networks. *international convention on information and communication technology electronics and microelectronics*.

Hayes, B. et al. (2013). First links in the markov chain. *American Scientist*, 101(2):252.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. *european conference on computer vision*.

He, L., Zhang, S., Wu, L., Xia, H., Ju, F., Zhang, H., Liu, S., Xia, Y., Zhu, J., Deng, P., Shao, B., Qin, T., and Liu, T.-Y. (2022). Pre-training co-evolutionary protein representation via a pairwise masked language model.

Heffernan, R., Paliwal, K. K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*.

Heffernan, R., Paliwal, K. K., Lyons, J., Singh, J., Yang, Y., and Zhou, Y. (2018). Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *Journal of Computational Chemistry*.

Heffernan, R., Yang, Y., Paliwal, K. K., and Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019). Modeling the language of life – deep learning protein sequences. *bioRxiv*.

Heinzinger, M., Littmann, M., Sillitoe, I., Bordin, N., Orengo, C., and Rost, B. (2022). Contrastive learning on protein embeddings enlightens midnight zone.

Hekkelman, M. L., d. de Vries, I., Joosten, R. P., and Perrakis, A. (2021). Alphafill: enriching the alphafold models with ligands and co-factors. *bioRxiv*.

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. (2022). Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*.

Hie, B., Yang, K. K., and Kim, P. S. (2021). Evolutionary velocity with protein language models. *bioRxiv*.

Hiranuma, N., Park, H., Baek, M., Anishchanka, I., Dauparas, J., and Baker, D. (2020). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications*.

Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. (2019). Axial attention in multidimensional transformers. *arXiv: Computer Vision and Pattern Recognition*.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1).

Hochreiter, S., Heusel, M., and Obermayer, K. (2007). Fast model-based protein homology detection without alignment. *Bioinformatics*, 23(14):1728–1736.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., De, D., Casas, L., Hendricks, L., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. V. D., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J., Vinyals, O., and Sifre, L. (2022). Training compute-optimal large language models.

Høie, M. H., Kiehl, E. N., Petersen, B., Nielsen, M., Winther, O., Nielsen, H., Hallgren, J., and Marcatili, P. (2022). Netsurfp-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning.

Holley, L. H. and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the United States of America*.

Hou, Q., Pucci, F., Pan, F., Xue, F., Rooman, M., and Feng, Q. (2022). Using metagenomic data to boost protein structure prediction and discovery. *Computational and Structural Biotechnology Journal*, 20:434–442.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. (2022). Learning inverse folding from millions of predicted structures. *bioRxiv*.

Hu, M., Yuan, F., Yang, K. K., Ju, F., Su, J., Wang, H., Yang, F., and Ding, Q. (2022). Exploring evolution-based & -free protein language models as protein function predictors.

Hua, W., Dai, Z., Liu, H., and Le, Q. V. (2022). Transformer quality in linear time.

Hubbard, T., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. (1997). Scop: a structural classification of proteins database. *Nucleic Acids Research*.

Ihm, Y. (2004). A threading approach to protein structure prediction: studies on tnf-like molecules, rev proteins, and protein kinases.

Ikeya, T., Güntert, P., and Ito, Y. (2019). Protein structure determination in living cells. *International Journal of Molecular Sciences*, 20(10):2442.

Ingraham, J., Garg, V. K., Barzilay, R., and Jaakkola, T. S. (2019). Generative models for graph-based protein design. *Learning*.

Jahan, M. S., Khan, H. U., Akbar, S., Farooq, M. U., Gul, S., and Amjad, A. (2021). Bidirectional language modeling: A systematic literature review. *Scientific Programming*.

Ji, S., Oruç, T., Mead, L., ur Rehman, M. F., Thomas, C. M., Butterworth, S., and Winn, P. J. (2018). Deepcdpred: Inter-residue distance and contact prediction for improved prediction of protein structure. *PLOS ONE*.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., and Dror, R. O. (2020). Learning from protein structure with geometric vector perceptrons. *Learning*.

Jing, X. and Xu, J. (2021). Fast and effective protein model refinement using deep graph neural networks. *Nature Computational Science*.

Jones, D. T. and Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2019). SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Jumper, J. M., Evans, R. O., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R. D., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D. L., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*.

Kandathil, S. M., Greener, J. G., Lau, A. M., and Jones, D. T. (2020). Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments. *bioRxiv*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv: Learning*.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv: Computation and Language*.

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., Sommer, M. O. A., Winther, O., Nielsen, M., Petersen, B. O., and Marcatili, P. (2018). Netsurfp-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins*.

Klicpera, J., Groß, J., and Günnemann, S. (2020). Directional message passing for molecular graphs. *Learning*.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. *empirical methods in natural language processing*.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, 39(1):309–338.

Kosciolek, T. and Jones, D. T. (2014). De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLOS ONE*.

Krause, B., Lu, L., Murray, I., and Renals, S. (2016). Multiplicative lstm for sequence modelling. *Learning*.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *north american chapter of the association for computational linguistics*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *neural information processing systems*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *Learning*.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lee, J. H., Yadollahpour, P., Watkins, A., Frey, N. C., Leaver-Fay, A., Ra, S., Cho, K., Gligorijevic, V., Regev, A., Bonneau, R., Design, P., and Genentech (2022). Equifold: Protein structure prediction with a novel coarse-grained structure representation.

Leopold, P. E., Montal, M., and Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725.

Li, B., Mendenhall, J. L., Capra, J. A., and Meiler, J. (2020a). A multi-task deep-learning system for predicting membrane associations and secondary structures of proteins. *bioRxiv*.

Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. (2020b). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *national conference on artificial intelligence*.

Li, H. (2022). Language models: past, present, and future. *Communications of the ACM*, 65(7):56–63.

Li, J. and Xu, J. (2021). Study of real-valued distance prediction for protein structure prediction with deep learning. *Bioinformatics*.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv: Computer Vision and Pattern Recognition*.

Li, S., Chua, T.-S., Zhu, J., and Miao, C. (2016). Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 666–675.

Li, Y., Zhang, C., Bell, E. W., Zheng, W., Zhou, X.-G., Yu, D.-J., and Zhang, Y. (2020c). Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLOS Computational Biology*.

Li, Z., Liu, X., Chen, W., Shen, F., Bi, H., Ke, G., and Zhang, L. (2022). Uni-fold: An open-source platform for developing protein folding models beyond alphafold.

Liao, Y.-L. and Smidt, T. (2022). Equiformer: Equivariant graph attention transformer for 3d atomistic graphs.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Costa, A. D. S., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Lin, Z., Lanchantin, J., and Qi, Y. (2016). Must-cnn: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. *national conference on artificial intelligence*.

Liu, D., Chen, S., Zheng, S., Zhang, S., and Yang, Y. (2022a). Se(3) equivalent graph attention network as an energy-based model for protein side chain conformation.

Liu, S., Zhang, J., Chu, H., Wang, M., Xue, B., Ni, N., Yu, J., Xie, Y., Chen, Z., Chen, M., Liu, Y., Patra, P., Xu, F., Chen, J., Wang, Z., Yang, L., Yu, F., Chen, L., and Gao, Y. Q. (2022b). Psp: Million-level protein sequence dataset for protein structure prediction.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., and Wang, P. (2019a). K-bert: Enabling language representation with knowledge graph. *national conference on artificial intelligence*.

Liu, X., He, P., Chen, W., and Gao, J. (2019b). Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Liu, X., He, P., Chen, W., and Gao, J. (2019c). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496. Association for Computational Linguistics.

Liu, X., Wang, Y., Ji, J., Cheng, H., Zhu, X., Awa, E., He, P., Chen, W., Poon, H., Cao, G., and Gao, J. (2020). The Microsoft toolkit of multi-task deep neural networks for natural language understanding. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019d). Roberta: A robustly optimized bert pretraining approach. *arXiv: Computation and Language*.

Liu, Y., Palmedo, P., Ye, Q., Berger, B., and Peng, J. (2017). Enhancing evolutionary couplings with deep convolutional neural networks. *Cell systems*.

Lu, A. X., Zhang, H., Ghassemi, M., and Moses, A. M. (2020). Self-supervised contrastive learning of protein representations by mutual information maximization. *bioRxiv*.

Luo, J., Cai, Y., Wu, J., Cai, H., Yang, X., and Lin, Z. (2020). Self-supervised representation learning of protein tertiary structures (ptsrep): Protein engineering as a case study. *bioRxiv*.

Ma, B. (2015). Novor: real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*.

Ma, B. and Johnson, R. (2012). De novo sequencing and homology searching. *Molecular & Cellular Proteomics*.

Ma, C., Dai, G., and Zhou, J. (2021). Short-term traffic flow prediction for urban road sections based on time series analysis and lstm_bilstm method. *IEEE Transactions on Intelligent Transportation Systems*.

Mabrouk, M., Putz, I., Werner, T., Schneider, M., Neeb, M., Bartels, P., and Brock, O. (2015). Rbo aleph: leveraging novel information sources for protein structure prediction. *Nucleic Acids Research*.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. (2021). Deep neural language modeling enables functional protein generation across families. *bioRxiv*.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. (2020). Progen: Language modeling for protein generation.

Magnan, C. and Baldi, P. (2014). Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*.

Mansoor, S., Baek, M., Madan, U., and Horvitz, E. (2021). Toward more general embeddings for protein design: Harnessing joint representations of sequence and structure. *bioRxiv*.

McCusker, J. P., Erickson, J., Chastain, K., Rashid, S., Weerawarana, R., and McGuinness, D. (2018). What is a knowledge graph. *Semantic Web Journal*.

McDermott, M. B. A., Yap, B., Hsu, T.-M. H., Jin, D., and Szolovits, P. (2021). Adversarial contrastive pre-training for protein sequences. *arXiv: Computation and Language*.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*.

Michel, M., Hurtado, D. M., and Elofsson, A. (2019). Pconsc4: fast, accurate and hassle-free contact predictions. *Bioinformatics*.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. *international conference on learning representations*.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). Colabfold: making protein folding accessible to all. *Nature Methods*, pages 1–4.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*.

Morton, J. T., Cem, S., R, B., Berenberg, D., Gligorijevic, and Bonneau, R. (2020). Protein structural alignments from sequence. *bioRxiv*.

Nicolai, C. and Sachs, F. (2013). Solving ion channel kinetics with the qub software. *Biophysical Reviews and Letters*, 8(03n04):191–211.

Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N., Madani, A., and Research, S. (2022). Progen2: Exploring the boundaries of protein language models.

Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A., Marks, D. S., and Gal, Y. (2022). Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval.

Ochoa, D. and Pazos, F. (2014). Practical aspects of protein co-evolution. *Frontiers in cell and developmental biology*, 2:14.

Ofer, D., Brandes, N., and Linial, M. (2021a). The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758.

Ofer, D., Brandes, N., and Linial, M. (2021b). The language of proteins: Nlp, machine learning & protein sequences. *Computational and structural biotechnology journal*.

Ofer, D., Brandes, N., Linial, M., Rappoport, N., and Peleg, Y. (2021c). Proteinbert: A universal deep-learning model of protein sequence and function. *F1000Research*.

Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). Cath – a hierarchic classification of protein domain structures. *Structure*.

Ovchinnikov, S., Kim, D. E., Wang, R. Y.-R., Liu, Y., DiMaio, F., and Baker, D. (2016). Improved de novo structure prediction in casp11 by incorporating coevolution information into rosetta. *Proteins*.

Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., Kondrashov, F. A., and Ivankov, D. N. (2021). Using alphafold to predict the impact of single mutations on protein stability and function. *bioRxiv*.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.

Patel, M. and Shah, H. B. (2013). Protein secondary structure prediction using support vector machines (svms). *2013 International Conference on Machine Intelligence and Research Advancement*, pages 594–598.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *north american chapter of the association for computational linguistics*.

Peters, M. E., Neumann, M., Logan, R. L., Schwartz, R., Joshi, V., Singh, S., and Smith, N. A. (2019). Knowledge enhanced contextual word representations. *empirical methods in natural language processing*.

Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. (2013). Dropout improves recurrent neural networks for handwriting recognition. *international conference on frontiers in handwriting recognition*.

Radford, A., Jozefowicz, R., and Sutskever, I. (2017). Learning to generate reviews and discovering sentiment. *arXiv: Learning*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018a). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018b). Language models are unsupervised multitask learners.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2022). Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. (2019). Evaluating protein transfer learning with tape. *bioRxiv*.

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. (2021a). Msa transformer. *bioRxiv*.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. (2021b). Transformer protein language models are unsupervised structure learners. *international conference on learning representations*.

Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. (2020). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *knowledge discovery and data mining*.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*.

Rohl, C. A., Strauss, C. E. M., Misura, K. M., and Baker, D. (2004). Protein structure prediction using rosetta. *Methods in Enzymology*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Roney, J. (2022). *Evidence for and Applications of Physics-Based Reasoning in AlphaFold*. PhD thesis.

Roney, J. P. and Ovchinnikov, S. (2022). State-of-the-art estimation of protein model accuracy using alphafold.

Rost, B. and Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences of the United States of America*.

Rouard, S. and Hadjeres, G. (2021). Crash: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis. *international symposium/conference on music information retrieval*.

Ruffolo, J. A. and Gray, J. J. (2022). Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophysical Journal*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*.

Sanger, F. (1952). The arrangement of amino acids in proteins. In *Advances in protein chemistry*, volume 7, pages 1–67. Elsevier.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv: Computation and Language*.

Sanyal, S., Anishchenko, I., A, D., Baker, D., and Talukdar, P. P. (2020). Proteingcn: Protein model quality assessment using graph convolutional networks. *bioRxiv*.

Schmidhuber, J., Wierstra, D., and Gomez, F. J. (2005). Evolino: Hybrid neuroevolution/optimal linear search for sequence prediction. In *Proceedings of the 19th International Joint Conferenceon Artificial Intelligence (IJCAI)*.

Schmitz, C., Vernon, R. B., Otting, G., Baker, D., and Huber, T. (2012). Protein structure determination from pseudocontact shifts using rosetta. *Journal of Molecular Biology*.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020a). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

Senior, A. W., Evans, R., Jumper, J. M., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020b). Improved protein structure prediction using potentials from deep learning. *Nature*.

Shen, T., Hu, Z., Peng, Z., Chen, J., Xiong, P., Hong, L., Zheng, L., Wang, Y., King, I., Wang, S., Sun, S., and Li, Y. (2022). E2efold-3d: End-to-end deep learning method for accurate de novo rna 3d structure prediction.

Shi, Y., Huang, Z., Wang, W., Zhong, H., Feng, S., and Sun, Y. (2020). Masked label prediction: Unified message passing model for semi-supervised classification. *international joint conference on artificial intelligence*.

Shoeybi, M., Patwary, M. M. A., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv: Computation and Language*.

Simoncini, D., Berenger, F., Shrestha, R., and Zhang, K. Y. J. (2012). A probabilistic fragment-based protein structure prediction algorithm. *PLOS ONE*.

Simoncini, D., Schiex, T., and Zhang, K. Y. J. (2017). Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *Proteins*.

Simons, K. T., Kooperberg, C., Huang, E. S., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*.

Singh, J., Litfin, T., Paliwal, K. K., Singh, J., Hanumanthappa, A. K., and Zhou, Y. (2021a). Spot-1d-single: Improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. *Bioinformatics*.

Singh, J., Litfin, T., Singh, J., Paliwal, K. K., and Zhou, Y. (2021b). Spot-contact-single: Improving single-sequence-based prediction of protein contact map using a transformer language model. *bioRxiv*.

Singh, J., Paliwal, K. K., Singh, J., and Zhou, Y. (2021c). Spot-1d-lm: Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *bioRxiv*.

Skocaj, D., Leonardis, A., and Kruijff, G.-J. M. (2012). *Cross-Modal Learning*, pages 861–864. Springer US, Boston, MA.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R. Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., and Catanzaro, B. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.

Sønderby, S. K. and Winther, O. (2014). Protein secondary structure prediction with long short term memory networks. *arXiv: Quantitative Methods*.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. *arXiv: Computation and Language*.

Spencer, M., Eickholt, J., and Cheng, J. (2015). A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Stein, R. A. and Mchaourab, H. S. (2022). Speach_af: Sampling protein ensembles and conformational heterogeneity with alphafold2. *PLOS Computational Biology*.

Steinegger, M., Meier, M., Mirdita, M., Voehringer, H., Haunsberger, S. J., and Soeding, J. (2019a). Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*.

Steinegger, M., Mirdita, M., and Söding, J. (2019b). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*.

Steinegger, M. and Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*.

Stigler, J., Ziegler, F., Gieseke, A., Gebhardt, J. C. M., and Rief, M. (2011). The complex folding network of single calmodulin molecules. *Science*, 334(6055):512–516.

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2020). UDSMProt: universal deep sequence models for protein classification. *Bioinformatics*, 36(8):2401–2409.

Sturmfels, P., Vig, J., Madani, A., and Rajani, N. F. (2020). Profile prediction: An alignment-based pre-training task for protein sequence models. *arXiv: Learning*.

Sulkowska, J. I., Morcos, F., Weigt, M., Hwa, T., and Onuchic, J. N. (2012). Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences of the United States of America*.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. (2019a). Videobert: A joint model for video and language representation learning.

Sun, Y., Shuohuan, W., Yukun, L., Feng, S., Chen, X., Zhang, H., Tian, X., Danxiang, Z., Tian, H., and Wu, H. (2019b). Ernie: Enhanced representation through knowledge integration. *arXiv: Computation and Language*.

Susanty, M., Rajab, T. E., and Hertadi, R. (2021). A review of protein structure prediction using deep learning. In *BIO Web of Conferences*, volume 41. EDP Sciences.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*.

Tan, C., Gao, Z., Xia, J., and Li, S. Z. (2022). Generative de novo protein design with global context. *arXiv preprint arXiv:2204.10673*.

Thrun, S. and Pratt, L. (1998). Learning to learn: introduction and overview. *Learning to learn*.

Torrisi, M., Kaleel, M., and Pollastri, G. (2018). Porter 5: state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*.

Torrisi, M., Pollastri, G., and Le, Q. (2020). Deep learning methods in protein structure prediction. *Computational and structural biotechnology journal*.

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A. C., and Doğan, T. (2022). Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vorberg, S. (2017). *Bayesian statistical approach for protein residue-residue contact prediction*. PhD thesis, lmu.

Vu, M. H., Akbar, R., Robert, P. A., Swiatczak, B., Sandve, G. K., Greiff, V., Trygve, D., and Haug, T. (2022). Advancing protein language models with linguistics: a roadmap for improved interpretability.

Wang, B., Xie, Q., Pei, J., Tiwari, P., Li, Z., and fu, J. (2022a). Pre-trained language models in biomedical domain: A systematic survey.

Wang, G. and Dunbrack, R. L. (2005). Pisces: recent improvements to a pdb sequence culling server. *Nucleic Acids Research*.

Wang, G., Fang, X., Wu, Z., Liu, Y., Xue, Y., Xiang, Y., Yu, D., Wang, F., and Ma, Y. (2022b). Helixfold: An efficient implementation of alphafold2 using paddlepaddle.

Wang, H., Zhang, J., Chen, Y., Ma, C., Avery, J., Hull, L., and Carneiro, G. (2022c). Uncertainty-aware multi-modal learning via cross-modal random network prediction. *arXiv preprint arXiv:2207.10851*.

Wang, S., Li, W., Liu, S., and Xu, J. (2016a). Raptorx-property: a web server for protein structure property prediction. *Nucleic Acids Research*.

Wang, S., Peng, J., Ma, J., and Xu, J. (2015). Protein secondary structure prediction using deep convolutional neural fields. *arXiv: Biomolecules*.

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2016b). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*.

Wang, W., Peng, Z., and Yang, J. (2022d). Single-sequence protein structure prediction using supervised transformer protein language models.

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., and Tang, J. (2019). Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*.

Wang, Y., Wu, H., and Cai, Y. (2018). A benchmark study of sequence alignment methods for protein clustering. *BMC bioinformatics*, 19(19):95–104.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*.

Weiss, K. R., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*.

Wong, K.-C., Chan, T.-M., Peng, C., Li, Y., and Zhang, Z. (2013). Dna motif elucidation using belief propagation. *Nucleic acids research*, 41(16):e153–e153.

Wu, K. E., Yang, K. K., van den Berg, R., Zou, J. Y., Lu, A. X., and Amini, A. P. (2022a). Protein structure generation via folding diffusion.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J. (2022b). High-resolution de novo structure prediction from primary sequence.

Wu, T. and Chen, C. (2022). Atomic protein structure refinement using all-atom graph representations and se(3)-equivariant graph neural networks.

Wu, T., Guo, Z., Hou, J., and Cheng, J. (2020). Deepdist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics*.

wwPDB consortium (2019). Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528.

Xia, J., Wu, L., Chen, J., Hu, B., and Li, S. Z. (2022a). Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM Web Conference 2022*, pages 1070–1079.

Xia, J., Zhu, Y., Du, Y., and Li, S. Z. (2022b). Pre-training graph neural networks for molecular representations: Retrospect and prospect. In *ICML 2022 2nd AI for Science Workshop*.

Xu, D. and Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins*.

Xu, J. (2018). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*.

Xu, M., Zhang, Z., Lu, J., Zhu, Z., Zhang, Y., Ma, C., Liu, R., and Tang, J. (2022). Peer: A comprehensive and multi-task benchmark for protein sequence understanding.

Yamada, K. and Hamada, M. (2021). Prediction of rna-protein interactions using a nucleotide language model. *bioRxiv*.

Yan, Y. and Huang, S.-Y. (2021). Accurate prediction of inter-protein residue-residue contacts for homo-oligomeric protein complexes. *Briefings in Bioinformatics*.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2019a). Improved protein structure prediction using predicted inter-residue orientations. *Proceedings of the National Academy of Sciences of the United States of America*.

Yang, K. K., Wu, Z., Bedbrook, C. N., and Arnold, F. H. (2018). Learned protein embeddings for machine learning. *Bioinformatics*.

Yang, Z., Cui, Y., Chen, Z., Che, W., Liu, T., Wang, S., and Hu, G. (2020). Textbrewer: An open-source knowledge distillation toolkit for natural language processing. *arXiv: Computation and Language*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019b). Xlnet: Generalized autoregressive pretraining for language understanding. *neural information processing systems*.

Yao, X.-Q., Zhu, H., and She, Z.-S. (2008). A dynamic bayesian network approach to protein secondary structure prediction. *BMC Bioinformatics*.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2017). Recent trends in deep learning based natural language processing. *arXiv: Computation and Language*.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Zerihun, M. B. and Schug, A. (2018). Biomolecular structure prediction via coevolutionary analysis: A guide to the statistical framework.

Zhang, B., Li, J., Quan, L., and Lyu, Q. (2021). Multi-task deep learning for concurrent prediction of protein structural properties. *bioRxiv*.

Zhang, G.-J., Laifa, M., Xiaoqi, W., and Zhou, X.-G. (2020a). Secondary structure and contact guided differential evolution for protein structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Zhang, G.-J., Xiaoqi, W., Laifa, M., Liujing, W., Hu, J., and Zhou, X.-G. (2020b). Two-stage distance feature-based optimization algorithm for de novo protein structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Zhang, H. and Shen, Y. (2020). Template-based prediction of protein structure with deep learning. *bioRxiv*.

Zhang, J., Liu, S., Chen, M., Chu, H., Wang, M., Wang, Z., Yu, J., Ni, N., Yu, F., Chen, D., Yang, Y. I., Xue, B., Yang, L., Liu, Y., and Gao, Y. Q. (2022a). Few-shot learning of accurate folding landscape for protein structure prediction.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020c). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. (2022b). Ontoprotein: Protein pretraining with gene ontology embedding.

Zhang, Y., Kolinski, A., and Skolnick, J. (2003). Touchstone ii: a new approach to ab initio protein structure prediction. *Biophysical Journal*.

Zhang, Y. and Skolnick, J. (2005). Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). Ernie: Enhanced language representation with informative entities. *meeting of the association for computational linguistics*.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. (2022c). Protein representation learning by geometric structure pretraining.

Zhou, J. and Troyanskaya, O. G. (2014). Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *international conference on machine learning*.