2024 4th International Conference on Digital Futures and Transformative Technologies (ICoDT2) | 979-8-3503-9186-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICoDT262145.2024.10740205

Pre-trained Language Models for Decoding Protein Language: a Survey

Maheera Amjad*, Ayesha Munir †, Usman Zia ‡ and Rehan Zafar Paracha §
School of Interdisciplinary Engineering and Sciences (SINES), National University of Sciences and Technology (NUST), Islamabad, 44000, Pakistan

Email: *mamjad.phdbi23sines@student.nust.edu.pk, †amunir.mscse22srcms@student.nust.edu.pk, ‡usman.zia@sines.nust.edu.pk, §rehan@sines.nust.edu.pk

Abstract—Transformer-based language models, such as BERT, GPT-3/4, and T5, are revolutionizing natural language processing (NLP). These models are increasingly being explored in the biomedical domains such as gene or protein sequences, protein structures, medical records, drugs, and images (X-rays, MRIs, CT-scans). Protein structure prediction is one of the fast-growing research fields with applications to biotechnology. Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence. This survey paper investigates the application of pre-trained language models (PLMs) for protein structure prediction and analysis. PLMs can be trained on protein sequences and structures available in public repositories, enabling them to understand the ”protein language” composed of amino acids. This approach offers a powerful tool for various downstream analyses, including protein interaction prediction, protein function prediction, and drug discovery. This survey highlights the improved accuracy, enhanced interpretability, and reduced training data requirements offered by PLMs compared to traditional deep learning models. Moreover, it also focuses on comparison of PLMs in terms of their performance, datasets, and computational resources. Finally, future research directions focusing on addressing limitations, exploring new applications, and developing more efficient and interpretable models are discussed. This will pave the way for further advancements in protein science and related fields

Index Terms—Protein structure, protein sequence, pre-trained language models, transformers, natural language processing

1 Introduction

Transformer [1] based language models like Bidirectional Encoder Representations (BERT) [2], and T5 [3] have revolutionized this era by starting a new chapter. The power of transformers, transfer learning, and self-supervised learning are all combined in these models. Transformers are so popular due to their employment of a self-attention mechanism, which can easily simulate long-distance relationships and be run in parallel. The process of transfer learning [4] involves the model applying its knowledge from the source task to the target task. Computer vision models, for instance, are pre-trained using large labeled datasets and then applied to similar tasks with smaller labeled datasets [5], [6]. The two primary benefits of pre-trained models are: (1) they acquire universal language representations; and (2) there is no need to retrain the downstream models.

The self-attention mechanism identifies long-distance word relationships that recognize the representation of every token in input and their respective interactions. This mechanism is the reason of the outperformance of transformers as compared to Convolutional Neural Networks (CNN) and Recurrent Neural Networks(RNN) [1] [7] [8]. Transformers can further enhance their performance by the usage of multiple self-attention layers which are stacked over each other. This has made transformers the major and primary choice for pre-trained language models in various applications [9].

Pre-trained models for biological sequence data have reflected implicit knowledge by learning context-sensitive representation from unlabeled biological data [10]. The knowledge gained by these pre-trained models can be easily applied for further downstream analysis, including drug–target interaction (DTI) [11], enhancer-promoter interaction (EPI) [12], and protein classification [13].

This review provides an introduction to the transformers and their distinguishing self-attention feature. It also provides introduction to multi-modal models and how protein data fits into this category. It is followed by the introduction of proteins and how are they represented in these models. Finally, protein-specific tasks are presented with the available pre-trained models. The purpose of this review paper is to provide an insight to the pre-trained transformers used for various protein based predictions including their sequence, structure, families, mutations and others. In comparison to the existing survey papers, we have provided more targeted models available for protein-specific tasks. The goal is to provide the researchers with the knowledge of existing models which can be useful for their research. Furthermore, researchers can also identify the gap where new models can be trained to perform protein related or other biological sequence related tasks.

2 Transformers

Transformer [1] is a renowned deep learning model that has wide applications in different types of domains including Natural Language Processing (NLP), Computer Vision (CV), and speech processing. Transformer-based pre-trained models (PTMs) [7] are found to generate state-of-the-art results on several tasks. The Vanilla transformer [1] is composed of an encoder and decoder and is a sequence-to-sequence model. Most of the encoder block is composed of a position-wise Feed-Forward Network (FFN) and a multi-head self-attention module. A residual connection [14] and layer normalization [15] are used to generate a more complex model. A transformer can be usually used as an encoder-decoder, encoder-only, and decoder-only model. An encoder-decoder model is often used for sequence-to-sequence modeling [16]. In several NLP tasks, transformer-based Pre-trained language models (PLMs) such as BERT [2], RoBERTa [17], ALBERT [18], and T5 [3] demonstrated remarkable success. Training a downstream model from scratch is no longer required because of these models. The limitation of using a transformer is its effectiveness in processing long sequences mainly due to the complex computation of the self-attention modules and memory consumption [16].

2.1 Attention mechanism

Attention provides a fundamentally interpretable mechanism focusing on input regions of interest. Recent research shows that explainability and attention are not significantly associated [19]. In NLP state-of-the-art language models employ multi-head self-attention as shown in Fig. 1 which is a crucial component of Transformers including BERT (Bidirectional Encoder Representations from Transformers) [2] and XLNet [20]. BERT is an Encoder-only model as seen in Fig. 2. Based on each token's interaction with every other token in the input self-attention computes every token's representation. Consequently, the self-attention mechanism outperforms the CNN and RNN in handling long-distance word relationships [1] [7] [8]. Transformer employs the Query-Key-Value (QKV) architecture as its attention mechanism. Instead of applying a single attention mechanism, the transformer applies multi-head attention. A fully connected feed-forward module known as position-wise FFN functions independently at every position [16]. The primary component of a transformer is its multi-head self-attention layer, which connects each relevant token in the input sequence to accurately encode each word. [21]. The self-attention layer learns sequence-wide contextual data by requiring a sequence of tokens as input [22].

Every token embedding will be converted into appropriate query, key, and value vector, before calculating embeddings for tokens. It is performed by multiplying them by randomly initialized W^q, W^k, and W^v matrices. Then the attention head will calculate the query's dot product with all the keys, divide each by $\sqrt{dk}$, and then apply a softmax function as described in equation (1) to determine the weights assigned to these values [23]. The attention function is defined by mapping a set of key-value pairs and a query vector to an output vector containing the information for the complete sequence.

$$Attention(Q, K, V) = softmax\frac{QK^T}{\sqrt{dk}} \tag{1}$$

[22]

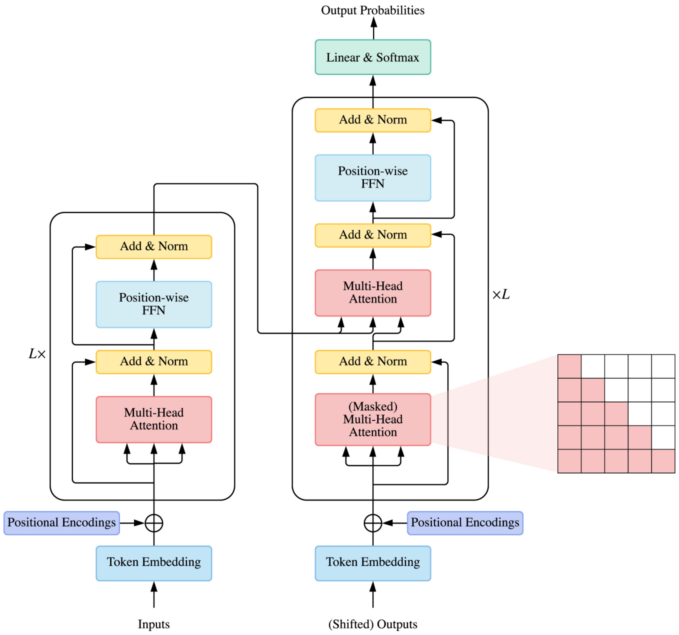

Fig. 1: An architecture of transformer [16]

2.2 Multi-Modal learning

Using multi-task and multi-modal deep learning models [24] [25] is an effective approach to handling complex problems. These models are developed to enhance prediction accuracy by utilizing multiple related tasks or information sources. By learning multiple related tasks at once, multi-task learning models seek to improve predictive performance and learning efficiency [24]. Multi-modal Deep-Learning models (DL) [25] attempt to combine data from multiple sources and modes. Multi-modal learning is performed by joining representations which uses complementarity and correlation between different modalities to enhance model performance [26].

The multi-stream models continue to learn a mapping between modal spaces while efficiently capturing the complexity of modality through uni-modal experts. By encouraging the sharing of various domain information, this method offers a flexible framework for downstream multi-modal activities [27].

PaLM-E [28], BLIP-2 [29], and LLaVA [30] serve as examples of how the PaLM-E-style model is a major development in the Vision Language (VL) multi-modal field. The PaLM-E-style model is modified for the biological domain, substituting biomolecule encoders in place of visual encoders. Since biomolecules naturally have 1D sequence representations, the uni-decoder may examine both virtual and sequence tokens, expanding our understanding of biomolecules in multiple ways. Consequently, the PaLM-E model can help Language Models (LMs) comprehend complicated 2D/3D biomolecular graphs and structures, in addition to efficiently utilizing the pre-trained biological models [31] [32].

3 Fundamentals of proteins

A biological sequence is a lengthy sequence that is represented by a series of distinct, fixed alphabets. For instance,

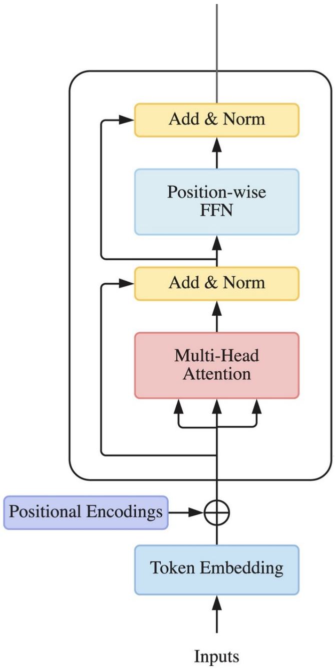

Fig. 2: Representation of BERT model [16]

the four letters ”A,” ”C,” ”T,” and ”G” in Deoxyribonucleic acid (DNA) represent various types of deoxynucleotides [33]. While the sequences of proteins and DNA are similar [34], the sequence of proteins is significantly more complex as proteins are polymers made up of 20 amino acids [35]. Proteins, like many other biological sequences, are also chains of fundamental units known as amino acids. These biological sequences are usually present in the form of folder-level, three-dimensional (3D) structures, where each structure has its dedicated function [36]. Thus, amino acid sequences serve as the foundation and primary source for many research goals, such as structure prediction [37], compound-protein interaction (CPI) [38], protein classification [13], protein function prediction [39], and sequence-based profiles [40]

3.1 Databases for protein data

Protein data is available in various formats but the most common and fundamental format of data is either in the sequence of amino acids or 3D structure. Various open-access databases provide these data for researchers. One of the most common databases is Protein Data Bank (PDB) [41], which is the largest collection of protein structures containing the three-dimensional structure of protein along with essential information. Although they also offer rich information about proteins categorized as families, superfamilies, common folds, and classes. Structural Classification of Proteins (SCOP) [42] and Structural Classification of Proteins—Extended (SCOPe) [43] also offer structural information of proteins. Protein family database(Pfam) [44] is a well-known database that uses multiple sequence comparisons and hidden Markov models to classify sequences into separate families.

Universal Protein (UniProt) [45] is a combination of the three most popular databases: PIR, SIB, and EBI. It has annotations for around 120 million protein sequences. Another completely annotated protein sequence database with extensive annotation data and defined nomenclature is SWISS-PROT [46]. Uniref [47] was released in 2004 and is widely used for identifying similar sequences. Three distinct subsets (UniRef100, UniRef90, and UniRef50) make it unique as they contain protein sequences with 100%, 90%, and 50% similarity, respectively. Furthermore, details of other protein databases and their description are provided in Table 1

3.2 Protein sequence representation

Protein language models (PLMs) are basically natural language processing (NLP) based models which use protein sequences to build their understanding in order to perform various analyses [48].

Although words and protein sequences have some abstract parallelism, there are undoubtedly significant distinctions in their characteristics, grammar, and semantics. In terms of NLP and protein sequence, each amino acid could be easily replaced or interpreted as a word, these words combine to formulate a sentence which are the protein sequences. Thus, words are amino acids and protein sequence is a sentence. The most popular method is to use individual amino acids as a word to pass as input tokens to any model. Protein sequences have long-range connections, much like natural language, which makes them great match for analysis by modern NLP models like Transformers [49]. Figure 3A demonstrates the application of a Transformer language model for protein sequences, containing an Encoder and Decoder block. These blocks convert amino acid tokens (input) into the internal representation for model. To ensure clarity, we will refer to this internal representation as the protein sequence representation in a particular PLM [50]. Figure 3B explain all that process in more details. It represents how amino acid sequences are passed as input embeddings to the encoder in order to generate an internal representation, which is subsequently employed as a feature vector to the model [50].

Mapping amino acid sequences to protein 3D structure is challenging [51], but there are abundant sequenced proteins available for DL models. The largest open sources are UniProt [52], Pfam [53], and BFD [54]. PLMs are capable enough to learn extremely complicated correlations and patterns that appears between protein sequences over time and across the global biome by using self-supervised training techniques on massive protein databases [50]. PLM trained on large databases can be efficiently used for various downstream analyses of protein sequence and structure which may involve prediction of mutations in protein sequence or interactions with other structures. Such downstream analysis can be applied by using either feature-based or fine-tuning approach on the base mode [55]. The feature-based approach trains the model without labels, creating a feature vector for protein prediction tasks. The fine-tuned model is trained first without labels and then updated with relevant labels for downstream prediction tasks [55].

TABLE 1: List of most common protein sequence databases

Dataset	Year	Entities	Description	URL (Source)
PDB	1971	2.5-dimensional structure of biological macromolecules	Protein structure database, containing 3D structures obtained through experiments	http://www.rcsb.org
SWISS-PROT	1986	Protein sequence	Annotated protein sequence database	http://www.uniprot.org
SCOP	1994	Protein sequences and structures	Database of protein structure classification according to the spatial characteristics of protein domains	http://scop2.mrc-lmb.cam.ac.uk
Pfam	1995	Protein sequence	Protein family database, including annotations and sequences of 17 929 protein families	http://pfam.xfam.org/
UniProt	2002	Protein sequence	A database consisting of a large number of labeled and unlabeled primary protein sequences	http://www.uniprot.org
UniRef	2004	Protein sequence	Unlabeled big data protein sequence	http://www.uniprot.org
DisProt	2007	Protein sequence	The database of disordered proteins	https://www.disprot.org/
SCOPe	2012	Protein structural relationships	59 514 protein database (PDB) entries, including more than 65% of the protein structures in the PDB	http://scop.berkeley.edu/
BFD	2018	Protein sequences	Largest set of protein sequences	https://metaclust.mmseqs.org/
ProteinNet	2019	Protein sequences and structure	A standardized dataset for machine learning of protein structure	https://github.com/aqlaboratory/proteinnet

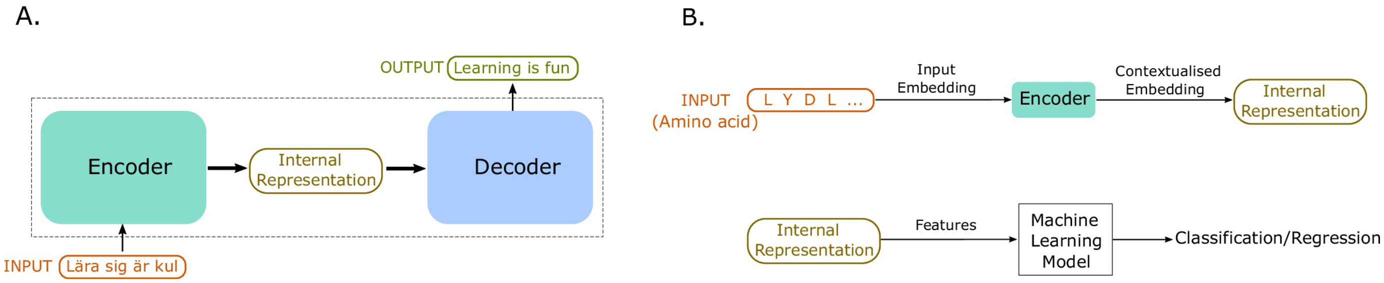

Fig. 3: (A) Sequence-to-sequence models conceptual foundation. It uses two blocks: an encoder and a decoder; to map an input sequence to an output sequence using the internal representations, which can easily convert the input sequence to the desired output language (B) An illustration of predicting properties using the Transformer language model. An encoder block is used to provide contextualized embeddings of the input sequence, which provides an internal representation of input embeddings. Subsequently, internal representation can be passed to the model and utilized to represent the features of the amino acids. Decoder block is typically not used after training; however, it is an essential part for natural language processing (NLP) applications [50]

4 Pre-trained transformer models for protein-specific tasks

Although biological sequence is considered a special and unique language, yet NLP has made it easy to comprehend the meaning of biological sequences along their relationships. Proteins are one of the most important component in biological processes. They are not irresponsible for controlling and managing biological responses in an individual but also have great heridetary role. Identification of the characteristics and their specific role can be performed using various bioinformatics tools, designed for each specific task. Yet, some of the tool are either too much time-consuming while some other might not provide accurate results. Various Machine learning (ML) and DL models have made it possible to execute several processes more effeciently. It has been further enhanced after the emergence of pre-trained transformers. Many downstream analysis of biological sequences has been made easier to process due to the availability of pre-trained models. Some of these analysis include prediction of protein sequence, protein structure, interactions of protein various other components, and many others as explained in following sections. There are many other fields and domains of protein analyses that still need to be explored for training new transformers or even using pre-trained transformers.

4.1 Protein sequence prediction

The connection between protein sequences and the spatial structure is one the basic requirement for performing any analyses of proteins. It can be easily understood by analyzing protein sequences, which allows theoretical foundation for additional study on the structure and function of proteins [56] [57]. trRosettaX-Single [58] is a transformer based PLM. It performs single-sequence protein structure prediction. This approach is known for incorporating sequence embedding from a supervised transformer PLM. Benchmark experiments demonstrated that it performs well on modelled protein structures with an average template modeling score (TM-score of 0.79). It has also been found to beat AlphaFold2 and RoseTTAFold on orphan proteins (proteins which lack any sort of annotation) [58].

While trRosettaX-Single outperforms other models and performs well on single-sequence of protein, but it lacks the ability to incorporate complex knowledge of protein sequence including the phylogeny and evolutionary properties. These properties are of great importance for predicting the protein fold in regards to orthologs and paraloge structures.

4.2 Protein structure prediction

Protein structure prediction can be divided into two primary stages: (a) secondary structure prediction (α -helix, β sheet, or coil) and (b) tertiary structure prediction (3D).

It is possible to split these primary prediction jobs into smaller ones. For example, since two residues in a sequence can be spatially close to one another in the 3D configuration, predictions for 2D interactions may be generated and then used progressively for 3D structure prediction. [59]. 3D protein structure with correct prediction is the most important task as it controls all the downstream analysis including all forms of protein interactions with other target structures. Several distinct DL models have been used to undertake the secondary structure prediction such as Prottrans [60] and ProteinBERT [61] Multiple sequence alignment (MSA) transformer [62], Protein Profiler [63], DeepProSite [64], SPOT-1D-Single [65], xTrimoPGLM [66], trRosettaX-Single [58].

AlphaFold, a state-of-the-art model from DeepMind, is very effective at predicting the 3D structures of proteins based just on their amino acid sequences [67]. It can be accessed using European Bioinformatics Institute's portal, (http://alphafold.ebi.ac.uk/) to predict protein structure. Furthermore, AlphaFold2 (AF2) has successfully solved the fold recognition problem by generating structural alignments of predicted human proteins to PDB library (41,000 proteins). Correlation coefficient between confidence and TM-score for AF2 is 0.67 [68].

Prottrans [60] is a combination of pre-training multiple models for examining their enhanced performance. It was trained on two auto-regressive and four auto-encoder models. Transformer-XL and XLNet were chosen out of the auto-regressive models while auto-encoder models chosen includes BERT, Albert, Electra, and T5. Model was was trained on 393 billion amino acids, collected from UniRef and BFD. They revealed the high and fast information absorbing capability of larger models with less training and thus low computation. The correlation between performance and samples and the need for sufficient model size emphasizes the importance of substantial computing resources.

ProteinBERT [61] is observed to provide close to the state-of-art performance despite provided with limited information. ProteinBert exhibited 0.70% accuracy for 3D structure prediction and 0.06% accuracy for remote homology prediction without pre-training but it was increased substantially after pre-training the model. 0.74% accuracy for 3D structure prediction and 0.22% accuracy for remote homology prediction was observed after pre-training.

xTrimoPGLM is a pre-trained transformer and pre-training was performed using 100B parameters and 1 trillion training tokens. It was trained specifically to predict 3D structures of proteins. xTrimoPGLM outperforms other protein understanding benchmarks and provides an advanced 3D structural prediction model. It generates de novo protein sequences and performs programmable generation after supervised fine-tuning, contributing to the evolving landscape of foundation models in protein science [66].

trRosettaX-Single [58] utilizes its own two-dimensional predicted geometry to reconstruct three-dimensional structures by energy reduction. A combination of sequence profiler and ESM-1b Transformer model representation was used to train and predict protein secondary structure. It is found to perform twice as fast as AF2, utilizing a significantly smaller amount of computational power (<10%). Using 2,000 designed proteins and network hallucination, trRosettaX-Single produces very confident structural models [69]. Using the feature combination of representations from ESM-1b [69], one-hot encoding, and SPOT-1D-Single [65] to train a neural network classifier, there is some further research on contact predictions. This demonstrated benefits over techniques based on evolutionary profiles and above employing just the ESM-1b representation [70].

Oligopeptides, known as bitter peptides, are usually produced during the fermentation of food and the hydrolysis of proteins [71], which are beneficial for the development of drugs since they can make patients more receptive to taking the medications by reducing the bitterness. Without using any structural knowledge, BERT4Bitter was developed to predict bitter peptides entirely from the original amino acid sequence [72].

Most of the models are trained on single sequence but MSA transformer [62] takes multiple sequences as input. In conventional biology protein sequences are aligned together to identify evolutionary related sequences based on their family and superfamily information. MSA transformer makes use of this technique and thus integrates MSA of proteins and masked language modelling to train and finally predict the sequences. As, MSA is 2D input (aligned protein sequences), there has to be two types of positional embeddings. 1D sequence embedding was provided to model for distinguishing various aligned positions while another positional embedding was provided to each MSA column to distinguish between various sequences.

Protein Profiler [63] is another pre-trained model which uses multiple sequence data as input to process and predict not only secondary structure, but also evolutionary information, homology and stability of structure. It exhibited 0.74% accuracy for secondary structure prediction, 0.23% for evolutionary or homology of structure, and 0.55% accuracy for predicting stability of structure as compare to Transformer which exhibited 0.73%, 0.21% and 0.73% accuracy for respective tasks. It clearly outperformed in secondary structure and homology prediction task, while, not giving good results for predicting of structure stability. Most state-of-the-art models for protein-related prediction tasks have been derived using features of profile-to-profile comparisons obtained from protein MSAs utilizing tools such as PSI-BLAST [73] and HMMER [74]. By modeling protein families and domains, a protein profile emerges by converting MSAs into a scoring system of amino acid sites that depends upon how frequently particular positions occur. Nevertheless, these techniques are restricted by insertions and deletions gaps in the protein sequences [75]. Additionally, they exhibit average performance in generating profiles and MSA for sequences with few or no homologs [76].

Although these models are found to outperform traditional structure modelling and prediction models yet, they have some limitations such as handling disordered regions of a protein sequence that don't fold into a stable 3D structure. BERT-based models have a specific limit to input size, hence for structure prediction of longer protein sequences, input sequence will be truncated thus losing contextual information which is necessary for structure prediction. MSA transformer was trained on multiple sequence alignments, thus incorporating evolutionary relationship between sequences, but the input does not contain any structural data to incorporate structural relationship between sequences. Furthermore, these models might be biased for predicting certain sequence region if that was over-expressed in training datasets.

4.3 Predicting mutations in protein sequence

Mutations in protein sequences make them diverse and change their role in biological processes. A single mutation can either make a protein structure unstable, can be responsible to cause deadly disease, or can make a protein stable enough to combat another disease. Hence, identification and prediction of mutations is a significant task that aids not only evolutionary studies but also various other downstream analyses. After pre-training a Transformer and fine-tuning on paired protein sequences, pre-trained and fine-tuned transformer models such as MutFormer [77] can predict pathogenic missense mutations, outperforming a range of existing tools [78]. Having BERT as the base classic architecture, MutFormer contains three major components which includes embedding, convolutions, and the transformer body. It was fine-tuned using a dataset of 84K pathogenic missense SNVs (Single nucleotide variations)and SNPs (Single nucleotide polymorphism) with over 0.1% allele frequency. A training and independent validation set was generated using ANNOVAR. Total 6 models were pre-trained as listed in Table 2

TABLE 2: MutFormer pre-trained model sizes [77]

Model Name	Hidden Layers	No. of parameters
MutBERT8L	8	58M
MutBERT10L	10	72M
MutFormer8L	8	62M
MutFormer10L	10	76M
MutFormer12L	12	86M
MutFormer8L (with integrated convs)	8	64M

Despite, the model being trained for different sizes of parameters and hidden layers, the most important requirement is the input data of mutations used to train the model. MutFormer is not considering the constant update of these mutations and variations. Further, predictions of the mutations by the model still require clinical validation as these predictions are based on statistical calculations and not any experimental evidence.

4.4 Prediction of homolog proteins

Identification of homologs having evolutionary-related functions is important in microbiology and medicine because it can be used to find new emerging antibiotic-resistant genes [79]. Computational tools like MMseqs2 [80], Pfam profile [81], and PSI-BLAST [82] align evolutionary-related protein positions using the information of conserved sequence patterns from evolutionary constraints that are responsible to preserve the structure and sequence of protein. These are the conventional methods for homology prediction, but they are unable to identify remotely similar sequences [83].

ProtTrans [60], ProteinBERT [61], and Protein Profiler [63] can predict homology in addition to the structure prediction as explained earlier. Additionally, ProtTrans [60], was used to anticipate the profile of proteins, leading to a novel technique [84]. Masked token of the protein sequence were supplied to the model in order to predict the characteristics of the given protein. The model then estimated the highly probable amino acids in those masked regions of a provided protein sequence. Sequence profiles from Homology-derived Secondary Structure of Proteins (HSSP) dataset were projected and compared. This approach was found useful and can help researchers obtain prediction profiles for new sequences, as evidenced by the strong analogy between the two profiles (predicted and HSSP database). Contrastive learning was performed on different samples which was locating embedding space among different samples and pushed apart and comparable samples were brought together [85]. Embeddings from the pre-trained Transformer model ProtT5 [86] were transferred to a new embedding space using FNN. The homologous sequences and more distant relations were identified by utilizing the similarity between pairings, which was measured using Euclidean distance in the embedding space. Additionally, structural hierarchies were also revealed by the contrastive learning approach that gives proteins their structural similarities [85].

While, these models can accurately and efficiently predict homology, yet they lack interpretability of these predictions as the output is mainly generated based on probabilities and statistical elucidations.

4.5 Protein binding site prediction

Protein binding sites are the fundamental regions for the activity of protein in any biological process, such as binding to peptides, drugs, membrane proteins and other biological components. Most of the existing computational techniques are limited due to the large amount of data, high-throughput prediction, and availability of 3D structure of protein. However, DeepProSite [64] is one of the state-of-art model in predicting protein binding sites only using protein sequence, while achieving exceptional predictions. This model is introduced as topology-aware Graph Transformer model which is able to use pre-trained language models and ESMFold for binding site predictions. Graph transformers can easily handle various node and edge types to capture interactions between nodes [87]. This ability allows DeepProSite to capture the interactions between residues of protein sequence and predict the binding sites in the 3D structure [64].

As, DeepProSite highly depends on the ProSite database for the data, thus incorporating any biased present in the database including large number of particular motifs or functional sites. While it takes only protein sequence and generates state-of-art results, but to identify the protein binding sites 3D structure of protein is also of great importance and can add value in predicting interacting residues while binding to ligand or any other entity.

4.6 Protein interactions prediction

Biological functions, including drug signaling, information transfer, and transcriptional regulation, are the predominant protein-protein interactions (PPIs). The conventional wet-lab experiment approach is expensive and time-consuming for gathering PPI data [88]. Attention mechanisms and transformer models have been utilized lately in PPI prediction, for robust protein sequence representation and complex models using biological information [26]. Several models have been trained for PPI prediction tasks including ADH-PPI [89], SDNN-PPI [88], HANPPIS [90], TransformerGO [91], GraphsformerCPI [92], CFAGO [93].

ADH-PPI [89] is a hybrid model using long short-term memory, convolutional, and self-attention layers. This model was pre-trained on benchmark datasets which includes *Saccharomyces cerevisiae (S.cervisiae)* and *Helicobacter pylori (H.pylori)* [94]. SDNN-PPI [88] also makes use of self-attention mechanism and tested on four independent datasets including *Caenorhabditis elegans (C.elegans)*, *Escherichia coli (E.coli)*, *Homo sapiens (H.sapiens)* and *Mus musculus (M.musculus)*. Furthermore, the task-agnostic Transformer is pre-trained to perform a combination of protein prediction tasks including protein family classification and protein interaction prediction [95]. It was observed to outperform when compared with other deep learning models such as DeepPPI (a deep neural network for protein-protein prediction) [96] provided 94.43% accurate predictions for *S.cervisiae* while SDNN outperformed with 95.48% accuracy [88].

HANPPIS [90] is a hierarchical attention network structure that uses six predominant features for predicting PPI sites. The addition of these features of protein sequence, amino acid position, secondary structure, position-specific scoring matrix (PSSM), pre-training vector, and hydrophilic properties helped in improving the model and resulted in effective results as compared to existing methods [90].

TransformerGO [91] also uses attention mechanism and produce dense graph embeddings for GO keywords to learn complex semantic relationships between positive (one of the proteins is activated) and negative interactions (one of the proteins is deactivated). STRING-DB datasets [97] of *S.cervisiae* and H.sapiens were used to test the performance of model and it was observed that it outperformed models with an accuracy of 0.961% for *S.cervisiae* and 0.974% for *H.sapiens*. GraphsformerCPI [92] is used for predicting interactions between compound and protein thus having a major role in drug designing. The architecture of this model includes multiple stacks of encoder-decoder, where each layer contains two encoders and one decoder. Semantic features from encoder layers are used by decoders to identify affinity probabilities protein residues and atoms of compounds. It was also observed to outperform traditional deep learning networks.

PPI networks can also be easily predicted using the Structural Gated Attention Deep (SGAD) [98] model which was evaluated using 11 independent datasets and one combined dataset. Multigranularity semantic fusion-based transformer [99] can also support the prediction of PPI. A global semantic representation is generated by Transformer by using a phrase contains indicators about two biomedical entities which are under consideration for identifying PPI. Meanwhile, a multichannel model is used to extract semantic information close to the target entity, for local feature extraction. Tests conducted on five PPI datasets demonstrate that the approach provides a notable enhancement over the most recent approaches [99].

CFAGO is a protein function prediction method with the integration of predicting PPI networks for single-species using a multi-head attention mechanism [93]. A multiple kernel ensemble attention method [100] and Lexically aware Transformer-based Bidirectional Encoder Represention model (LBERT) [101] were tested on the PPI dataset alongside other datasets and were found to outperform previous methods.

Most of these models outperforms the previous one, yet they are trained on similar datasets including mainly: *S.cervisiae*, *H.sapiens*, *C.elegans*, *E.coli*, and *M.musculus*. Use of similar datasets may lead to a biased generation of protein interactions. Furthermore, these models predict interactions based on the already known interactions so they might not be able to predict any novel interactions that might exist in complex structures or other organisms. Testing of the interactions were also performed on the existing protein structures, hence novel structures were not considered during their predictions.

4.7 DNA-protein binding site prediction

The identification of protein-DNA binding sites is the initial step in developing novel drugs and gaining a mechanistic knowledge of biological processes like transcription and repair. Protein-DNA interactions are important in biological systems. Using the predicted structural models by AF2 [102], GraphSite [103] has achieved accuracy in predicting DNA-binding residues, building on the recent breakthrough in protein structure prediction made by the program. BindTransNet is a transferable Transformer-based approach for DNA-protein binding prediction across cell types. It uses a Transformer Encoder and transfer learning to extract some shared long-range dependencies between different motifs available in cross-cell types [104].

4.8 Drug-target predictions

The therapeutic effects of active compounds are determined by drug-target interactions (DTIs), which regulate target biological processes. Identifying molecules with relevant binding activity is essential in early stages of drug research [105]. However, low or high-throughput bioassays are time-consuming, labor-intensive, and unfeasible due to the vast compound and protein space [106]. By reducing the size of the search area and eliminating couples that are unlikely to bond, DTI prediction can help. The discipline has broadened to include the discovery of new pharmaceuticals, the repurpose of existing drugs, and the identification of novel proteins that may function as interaction partners for established drugs [107]. DTITR [108], MolTrans [109], GraphormerDTI [110], DeepMGT-DTI [111], TAG-DTI [112], GEFormerDTA [113] utilized Transformer based attention mechanism to predict DTIs.

DeepTTA [114] uses a multilayer neural network to predict anti-cancer medication responses based on transcriptome data and a transformer for drug representation learning. To be more precise, DeepTTA predicts drug responses using chemical substructures of pharmaceuticals and transcriptomic gene expression data [114].

DTITR [108] is also a Transformer based architecture which uses protein sequence and structure data to predict binding affinities between drug and its target. Using self-attention mechanism it not only captures the chemical and biological nature of drug-target but also their pharmacological properties for DTIs. This model was evaluated on Davis dataset [115] 31,824 interactions between 72 kinase inhibitors (compounds/drug) and 442 kinases (proteins/target). PDB structures of proteins were obtained from UniProt [52] and SMILES of compounds were obtained from PubChem [116]. As the model was observed to outperform exhibiting 0.192 mean squared error (MSE), 0.438 root mean squared error (RMSE), and 0.907 concordance index (CI) as compared to DeepDTA (compound-target predictor based on LSTM and CNN) [117] which resulted 0.215 MSE, 0.464 RMSE, and 0.891 CI. MolTrans [109] also takes SMILES of drug and protein sequence as input tokens to predict drug-target predictions. For Davis dataset [115], it outperforms with AUC (area under the curve) of 0.895, while, for DeepDTA it was 0.876. GraphormerDTI [110] also uses similar input data. It was observed to outperform MolTrans which had 0.876 value for AUC while GraphormerDTI exhibited 0.893 AUC. DeepMGT-DTI [111] also performed in a similiar architecture but for drug SMILES they used KEGG database [118] in addition to PubChem [116] TAG-DTI [112] was found to outperform DTITR [108]. TAG-DTI exhibited 0.185 MSE, 0.430 RMSE, and 0.917 CI while DTITR exhibited 0.192 MSE, 0.438 RMSE, and 0.907 CI. GEFormerDTA [113] was compared with various deep learning models and outperformed all of them. For DeepDTA [117] 0.196 MSE, 0.442 RMSE and 0.866 CI was observed while GEFormerDTA exhibited 0.212 MSE, 0.461 RMSE and 0.895 CI.

Similar to models for predicting protein interactions, these models also outperform each other but still they are trained on data obtained from similar databases and datasets but do not encounter any novel drug-target interactions. This may lead to a biased predictions by models.

4.9 Post-translational modifications prediction

Post-translational modification is the biological mechanism of altering proteins [119], where chemical groups are added to amino acid side chains [120]. These modifications are responsible for affecting protein functions, properties, conformation, stability, and interactions [121]. Some of the most important post-translational modifications include methylation, phosphorylation, glycosylation, ubiquitination, and acetylation [122]. Protein acetylation, is a result of covalent bonds, involving acetyl group to a lysine residue [123], thus playing distinct roles in cellular functions.

BERT-Kcr was trained, and lysine crotonylation was predicted [124] using BERT embeddings [2] of the amino acids. TransPTM [125] predicts a non-histone acetylation site. The acetylation of non-histone proteins involves a wider variety of enzymes and target proteins [126]. Att-Lys lysine acylation predictor [127] is a multi-head self-attention-based model that takes substrate composition features of the rice protein's amino acid sequence and the word vector features of the protein's twenty amino acids, which were trained using the Word2vec method. PTransIPs [128] identify phosphorylation sites that use ProtTrans [60] and EMBER2 to extract sequence and structure embeddings, enhancing model performance. It uses Transformer architecture, convolutional neural networks, and TIM loss function for model design and training, proving universal for other peptide bioactivity tasks.

Conclusion

5 Conclusion and Future directions

This paper focuses on the review of pre-trained models used for protein-related data. In general, this review introduced the transformers and gave a brief introduction of their distinguished features. It is followed by the introduction of proteins, popular databases, and the most common protein-related tasks where these methods are used.

In particular, this review illustrated the contribution of pre-trained models for specifically proteins data, which can be either plain protein sequences or a complex 3D structure. It also includes information on protein data including protein family, protein homologs, mutations in protein sequences, interactions of protein with other proteins, drugs, DNA or others, and post-translational modifications. This review provides the most popular models trained on these and other protein-specific tasks which are listed in Table 3 along their basic information. We have compared the models with the other deep learning models and also highlighted which models outperformed when compared with each other. It was observed in drug-target prediction models that every new model outperformed the previous one.

For future researchers have a wide opportunity to either enhance the performance of these existing models or even introduce their new pre-trained models in a similar way. This is not the only way but there are various protein specific tasks which are still not explored, yet being the most important tasks. To the maximum knowledge gathered so far researchers can pre-train models for predicting biological pathways and the relationship of protein in these pathways. Various biological pathway databases including KEGG are available and contains all the relevant information for the biological pathways. Further there are databases which are specific to particular organism, that must also be incorporated to avoid any bias in the model. Models can be pre-trained to identify the modifications needed in the protein structure to make them more effective for drugs or vice-versa. In a similar way antibody-drug conjugates can be efficiently generated with high affinity for their target. These are some of the many other protein-related tasks which need to explored for transformers and particularly pre-trained transformers.

In particular future models need to encounter data from multiple databases to avoid any biased predictions of structures, mutations, interactions or any other protein related task. Further bias can also be improved by adding data of the available organisms instead of the most common organisms. Future models also should incorporate novel predictions with their validation results which can be helpful for their experimental testing.

TABLE 3: List of models

Model	Base Model	Input	Param	Pre-training dataset	Description	Year
trRosettaX-Single [58]	Transformer	protein sequence	100B	55 human-designed proteins of size between 50 and 300 amino acids	single sequence protein structure prediction	2022
ProteinBERT [61]	BERT	protein sequence	16M	∼ 106M UniRef9	predicts 3D secondary structure and remote homology	2022
MSA transformer [62]	masked language model	aligned protein sequences	100M	4.3 TB of 26 million MSAs, with an average of 1192 sequences per MSA	protein structure and function	2021
Protein Profiler [63]	ProtAlbert Transformer	protein sequence	-		predicts secondary structure, fold, superfamily, family and protein contacts	2022
DeepProSite [64]	ProtT5	protein sequence	3B	UniRef50	predicts protein binding sites	2023
SPOT-1D-Single [65]	hybrid LSTM	protein sequence	-	39 120 protein sequences	predicts protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures	2021
xTrimoPGLM [66]	GLM-130B	protein sequence	100B	940 million unique protein sequences with almost 200 billion residues,	advanced 3D structure prediction	2024
BERT4Bitter [72]	BERT	protein seqence	-	BTP640	identify bitter peptides	2021
MutFormer [77]	BERT	protein sequence	MutFormer 8L-62M param MutFormer 10L-76M param MutFormer 12L-86M param	human reference protein sequences	predicts missense mutations.	2023
ADH-PPI [89]	hybrid LSTM, convolutional, and attention layers	protein sequence	-	Saccharomyces cerevisiae (S. cerevisiae) and Helicobacter pylori (H.pylori) datasets	predicts protein-protein interactions	
TransformerGO [91]	Transformer	Gene Ontology (GO) terms	-	GO graph and GO annotations PPI for Homo Sapiens (120 386) and S. cerevisiae (252 984)	captures semantic similarity between GO sets	2022
HANPPIS [90]	SeqVec-ELMO	protein sequence	-	Dset_186, Dset_72, and Dset_164 containing: 1,923, 5,517 and 6,096 active sites and 16,217, 30,702 and 27,585 non-interactive sites	predicts interaction sites of proteins	2021
CFAGO [93]	Encoder-Decoder model	protein sequence	-	GO, UniProt, pfam	identifies universal protein representation and protein functions	2023
DTITR [108]	Transformer	protein sequence and drug SMILES	-	interactions of proteins with inhibitors and kinases SMILES from PubChem	identifies drug-target interaction	2022
MolTrans [109]	Transformer	-	MINER DTI dataset BIOSNAP dataset	identifies drug-target interaction	2021	
GraphormerDTI [110]	12 stacked Graph Transformer layers and 3 stacked 1D-CNN layers	protein sequence and drug SMILES	-	DrugBank, Davis and KIBA	identifies drug-target interaction	2024
DeepMGT-DT [111]	Transformers	protein sequence and drug SMILES	-	KEGG and PubChem database	identifies drug-target interaction	2022
TAG-DTI [112]	Transformers	protein sequence and drug SMILES	-	442 kinases and 72 kinase inhibitors protein sequences of Davis dataset from UniProt	identifies drug-target interaction	2024
GEFormerDTA [113]	Transformers	protein sequence and drug SMILES	-	kinase Davis dataset and the KIBA dataset	identifies drug-target interaction	2024
DeepTTA citejiang2022deeptta	Transformer and multilayer neural network	protein sequence and drug SMILES	-	GDSC dataset	identifies drug-target interaction	2022
TransPTM [125]	Transformer	protein sequence	-	NHAC, which includes 11 subsets with sequences of 11 to 61 amino acids	predicts non-histone acetylation site	2024
PTransIPs [128]	Transformer	protein sequence and structure	-	-	identifies phosphorylation sites	2024
Att-Lys lysine acylation predictor [127]	Multi-head self attention	protein sequence and word vector od amino acids	-	rice proteins	predicts lysine acylation	2022

References

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

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[7] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872–1897, 2020.

[8] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus: A survey of transformer-based pretrained models in natural language processing," *arXiv preprint arXiv:2108.05542*, 2021.

[9] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammu: a survey of transformer-based biomedical pretrained language models," *Journal of biomedical informatics*, vol. 126, p. 103982, 2022.

[10] B. Song, Z. Li, X. Lin, J. Wang, T. Wang, and X. Fu, "Pretraining model for biological sequence data," *Briefings in functional genomics*, vol. 20, no. 3, pp. 181–195, 2021.

[11] B. Playe and V. Stoven, "Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity," *Journal of cheminformatics*, vol. 12, no. 1, p. 11, 2020.

[12] Z. Hong, X. Zeng, L. Wei, and X. Liu, "Identifying enhancer–promoter interactions with neural network based on pre-trained dna vectors and attention mechanism," *Bioinformatics*, vol. 36, no. 4, pp. 1037–1043, 2020.

[13] N. Strodthoff, P. Wagner, M. Wenzel, and W. Samek, "Udsmprot: universal deep sequence models for protein classification," *Bioinformatics*, vol. 36, no. 8, pp. 2401–2409, 2020.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[15] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[16] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI open*, vol. 3, pp. 111–132, 2022.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[19] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

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

[22] S. Zhang, R. Fan, Y. Liu, S. Chen, Q. Liu, and W. Zeng, "Applications of transformer-based language models in bioinformatics: a survey," *Bioinformatics Advances*, vol. 3, no. 1, p. vbad001, 2023.

[23] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. I, 2017.

[24] N. Vithayathil Varghese and Q. H. Mahmoud, "A survey of multi-task deep reinforcement learning," *Electronics*, vol. 9, no. 9, p. 1363, 2020.

[25] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[26] M. Lee, "Recent advances in deep learning for protein-protein interaction analysis: A comprehensive review," *Molecules*, vol. 28, no. 13, p. 5169, 2023.

[27] Q. Pei, L. Wu, K. Gao, J. Zhu, Y. Wang, Z. Wang, T. Qin, and R. Yan, "Leveraging biomolecule and natural language through multi-modal learning: A survey," *arXiv preprint arXiv:2403.01528*, 2024.

[28] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[29] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, pp. 19730–19742, PMLR, 2023.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[31] Y. Luo, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao, and Z. Nie, "Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine," *arXiv preprint arXiv:2308.09442*, 2023.

[32] H. Cao, Z. Liu, X. Lu, Y. Yao, and Y. Li, "Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery," *arXiv preprint arXiv:2311.16208*, 2023.

[33] S. G. Gregory, K. Barlow, K. McLay, R. Kaul, D. Swarbreck, A. Dunham, C. Scott, K. Howe, K. Woodfine, C. Spencer, *et al.*, "The dna sequence and biological annotation of human chromosome 1," *Nature*, vol. 441, no. 7091, pp. 315–321, 2006.

[34] P. Christen and E. Hofmann, *EJB Reviews 1991*. Springer Science & Business Media, 2013.

[35] S. Karlin and G. Ghandour, "Comparative statistics for dna and protein sequences: single sequence analysis.," *Proceedings of the National Academy of Sciences*, vol. 82, no. 17, pp. 5800–5804, 1985.

[36] T. Bepler and B. Berger, "Learning protein sequence embeddings using information from structure," *arXiv preprint arXiv:1902.08661*, 2019.

[37] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (casp)—round x," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, pp. 1–6, 2014.

[38] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, and M. Zheng, "Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments," *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, 2020.

[39] A. Villegas-Morcillo, S. Makrodimitris, R. C. van Ham, A. M. Gomez, V. Sanchez, and M. J. Reinders, "Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function," *Bioinformatics*, vol. 37, no. 2, pp. 162–170, 2021.

[40] A.-S. Yang and B. Honig, "An integrated approach to the analysis and modeling of protein sequences and structures. iii. a comparative study of sequence conservation in protein structural families using multiple structural alignments," *Journal of molecular biology*, vol. 301, no. 3, pp. 691–711, 2000.

[41] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[42] W. L. DeLano, "The pymol molecular graphics system," *http://www. pymol. org/*, 2002.

[43] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures," *Nucleic acids research*, vol. 42, no. D1, pp. D304–D309, 2014.

[44] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, *et al.*, "Pfam: the protein families database," *Nucleic acids research*, vol. 42, no. D1, pp. D222–D230, 2014.

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

[46] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, *et al.*, "The swiss-prot protein knowledgebase and its supplement trembl in 2003," *Nucleic acids research*, vol. 31, no. 1, pp. 365–370, 2003.

[47] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium, "Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2015.

[48] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC bioinformatics*, vol. 20, pp. 1–17, 2019.

[49] D. Ofer, N. Brandes, and M. Linial, "The language of proteins: Nlp, machine learning & protein sequences," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1750–1758, 2021.

[50] A. Chandra, L. Tünnermann, T. Löfstedt, and R. Gratz, "Transformer-based deep learning for predicting protein properties in the life sciences," *Elife*, vol. 12, p. e82819, 2023.

[51] B. Kuhlman and P. Bradley, "Advances in protein structure prediction and design," *Nature reviews molecular cell biology*, vol. 20, no. 11, pp. 681–697, 2019.

[52] "Uniprot: the universal protein knowledgebase in 2023," *Nucleic acids research*, vol. 51, no. D1, pp. D523–D531, 2023.

[53] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. Sonnhammer, S. C. Tosatto, L. Paladin, S. Raj, L. J. Richardson, *et al.*, "Pfam: The protein families database in 2021," *Nucleic acids research*, vol. 49, no. D1, pp. D412–D419, 2021.

[54] "Bfd." Accessed on June 11, 2024.

[55] M. T. R. Laskar, X. Huang, and E. Hoque, "Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 5505–5514, 2020.

[56] M. Z. Atassi and E. Appella, *Methods in protein structure analysis*. Springer Science & Business Media, 2013.

[57] J. M. Walker, *The proteomics protocols handbook*. Springer, 2005.

[58] W. Wang, Z. Peng, and J. Yang, "Single-sequence protein structure prediction using supervised transformer protein language models," *Nature Computational Science*, vol. 2, no. 12, pp. 804–814, 2022.

[59] Z. Du, H. Su, W. Wang, L. Ye, H. Wei, Z. Peng, I. Anishchenko, D. Baker, and J. Yang, "The trrosetta server for fast and accurate protein structure prediction," *Nature protocols*, vol. 16, no. 12, pp. 5634–5651, 2021.

[60] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, *et al.*, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

[61] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "Proteinbert: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, 2022.

[62] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, "Msa transformer," in *International Conference on Machine Learning*, pp. 8844–8856, PMLR, 2021.

[63] P. Sturmfels, J. Vig, A. Madani, and N. F. Rajani, "Profile prediction: An alignment-based pre-training task for protein sequence models," *arXiv preprint arXiv:2012.00195*, 2020.

[64] Y. Fang, Y. Jiang, L. Wei, Q. Ma, Z. Ren, Q. Yuan, and D.-Q. Wei, "Deepprosite: structure-aware protein binding site prediction using esmfold and pretrained language model," *Bioinformatics*, vol. 39, no. 12, p. btad718, 2023.

[65] J. Singh, T. Litfin, K. Paliwal, J. Singh, A. K. Hanumanthappa, and Y. Zhou, "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*, vol. 37, no. 20, pp. 3464–3472, 2021.

[66] B. Chen, X. Cheng, P. Li, Y.-a. Geng, J. Gong, S. Li, Z. Bei, X. Tan, B. Wang, X. Zeng, *et al.*, "xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein," *arXiv preprint arXiv:2401.06199*, 2024.

[67] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[68] S. B. Pandit and J. Skolnick, "Fr-tm-align: a new protein structural alignment method based on fragment alignments and the tm-score," *BMC bioinformatics*, vol. 9, pp. 1–11, 2008.

[69] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[70] J. Singh, T. Litfin, J. Singh, K. Paliwal, and Y. Zhou, "Spot-contact-lm: improving single-sequence-based prediction of protein contact map using a transformer language model," *Bioinformatics*, vol. 38, no. 7, pp. 1888–1894, 2022.

[71] K. Karametsi, S. Kokkinidou, I. Ronningen, and D. G. Peterson, "Identification of bitter peptides in aged cheddar cheese," *Journal of agricultural and food chemistry*, vol. 62, no. 32, pp. 8034–8041, 2014.

[72] P. Charoenkwan, C. Nantasenamat, M. M. Hasan, B. Manavalan, and W. Shoombuatong, "Bert4bitter: a bidirectional encoder representations from transformers (bert)-based model for improving the prediction of bitter peptides," *Bioinformatics*, vol. 37, no. 17, pp. 2556–2562, 2021.

[73] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[74] R. D. Finn, J. Clements, and S. R. Eddy, "Hmmer web server: interactive sequence similarity searching," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W29–W37, 2011.

[75] T. Golubchik, M. J. Wise, S. Easteal, and L. S. Jermiin, "Mind the gaps: evidence of bias in estimates of multiple sequence alignments," *Molecular biology and evolution*, vol. 24, no. 11, pp. 2433–2442, 2007.

[76] T. M. Phuong, C. B. Do, R. C. Edgar, and S. Batzoglou, "Multiple alignment of protein sequences with repeats and rearrangements," *Nucleic acids research*, vol. 34, no. 20, pp. 5932–5942, 2006.

[77] T. Jiang, L. Fang, and K. Wang, "Deciphering the language of nature: A transformer-based language model for deleterious mutations in proteins," 2023.

[78] H. Yamaguchi and Y. Saito, "Evotuning protocols for Transformer-based variant effect prediction on multi-domain proteins," *Briefings in Bioinformatics*, vol. 22, p. bbab234, 06 2021.

[79] L. S. Tavares, C. d. Silva, V. C. Souza, V. L. Silva, C. G. Diniz, and M. D. Santos, "Strategies and molecular tools to fight antimicrobial resistance: resistome, transcriptome, and antimicrobial peptides," *Frontiers in Microbiology*, vol. 4, 2013.

[80] M. Steinegger, "Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, 2017.

[81] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. Tosatto, and R. D. Finn, "The Pfam protein families database in 2019," *Nucleic Acids Research*, vol. 47, pp. D427–D432, 10 2018.

[82] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 09 1997.

[83] G. W. Wilburn and S. R. Eddy, "Remote homology search with hidden potts models," *PLoS Comput. Biol.*, vol. 16, p. e1008085, Nov. 2020.

[84] A. Behjati, F. Zare-Mirakabad, S. S. Arab, and A. Nowzari-Dalini, "Protein sequence profile prediction using ProtAlbert transformer." Sept. 2021.

[85] M. Heinzinger, M. Littmann, I. Sillitoe, N. Bordin, C. Orengo, and B. Rost, "Contrastive learning on protein embeddings enlightens midnight zone," *NAR Genom. Bioinform.*, vol. 4, p. lqac043, June 2022.

[86] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "ProtTrans: Towards cracking the language of life's code through self-supervised learning." July 2020.

[87] J. Ingraham, V. Garg, R. Barzilay, and T. Jaakkola, "Generative models for graph-based protein design," *Advances in neural information processing systems*, vol. 32, 2019.

[88] X. Li, P. Han, G. Wang, W. Chen, S. Wang, and T. Song, "Sdnn-ppi: self-attention with deep neural network effect on protein-protein interaction prediction," *BMC genomics*, vol. 23, no. 1, p. 474, 2022.

[89] M. N. Asim, M. A. Ibrahim, M. I. Malik, A. Dengel, and S. Ahmed, "Adh-ppi: An attention-based deep hybrid model for protein-protein interaction prediction," *Iscience*, vol. 25, no. 10, 2022.

[90] M. Tang, L. Wu, X. Yu, Z. Chu, S. Jin, and J. Liu, "Prediction of protein–protein interaction sites based on stratified attentional mechanisms," *Frontiers in Genetics*, vol. 12, p. 784863, 2021.

[91] I. Ieremie, R. M. Ewing, and M. Niranjan, "Transformergo: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms," *Bioinformatics*, vol. 38, no. 8, pp. 2269–2277, 2022.

[92] J. Ma, Z. Zhao, T. Li, Y. Liu, J. Ma, and R. Zhang, "Graphsformercpi: Graph transformer for compound–protein interaction prediction," *Interdisciplinary Sciences: Computational Life Sciences*, pp. 1–17, 2024.

[93] Z. Wu, M. Guo, X. Jin, J. Chen, and B. Liu, "Cfago: cross-fusion of network and attributes based on attention mechanism for protein function prediction," *Bioinformatics*, vol. 39, no. 3, p. btad123, 2023.

[94] M. Kong, Y. Zhang, D. Xu, W. Chen, and M. Dehmer, "Fctp-wsrc: protein–protein interactions prediction via weighted sparse representation based classification," *Frontiers in genetics*, vol. 11, p. 18, 2020.

[95] A. Nambiar, S. Liu, M. Heflin, J. M. Forsyth, S. Maslov, M. Hopkins, and A. Ritz, "Transformer neural networks for protein family and interaction prediction tasks," *Journal of Computational Biology*, vol. 30, no. 1, pp. 95–111, 2023.

[96] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, "Deepppi: boosting prediction of protein–protein interactions with deep neural networks," *Journal of chemical information and modeling*, vol. 57, no. 6, pp. 1499–1510, 2017.

[97] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, *et al.*, "The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic acids research*, p. gkw937, 2016.

[98] F. Zhu, F. Li, L. Deng, F. Meng, and Z. Liang, "Protein interaction network reconstruction with a structural gated attention deep model by incorporating network structure information," *Journal of Chemical Information and Modeling*, vol. 62, no. 2, pp. 258–273, 2022.

[99] Y. Li, Y. Chen, Y. Qin, Y. Hu, R. Huang, and Q. Zheng, "Protein-protein interaction relation extraction based on multigranularity semantic fusion," *Journal of Biomedical Informatics*, vol. 123, p. 103931, 2021.

[100] H. Zhang and M. Xu, "Graph neural networks with multiple kernel ensemble attention," *Knowledge-Based Systems*, vol. 229, p. 107299, 2021.

[101] N. Warikoo, Y.-C. Chang, and W.-L. Hsu, "Lbert: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations," *Bioinformatics*, vol. 37, no. 3, pp. 404–412, 2021.

[102] Q. Yuan, S. Chen, J. Rao, S. Zheng, H. Zhao, and Y. Yang, "Alphafold2-aware protein–dna binding site prediction using graph transformer," *Briefings in bioinformatics*, vol. 23, no. 2, p. bbab564, 2022.

[103] W. Shi, M. Singha, L. Pu, G. Srivastava, J. Ramanujam, and M. Brylinski, "Graphsite: ligand binding site classification with deep graph learning," *Biomolecules*, vol. 12, no. 8, p. 1053, 2022.

[104] Z. Wang, X. Tan, B. Li, Y. Liu, Q. Shao, Z. Li, Y. Yang, and Y. Zhang, "Bindtransnet: A transferable transformer-based architecture for cross-cell type dna-protein binding sites prediction," in *International Symposium on Bioinformatics Research and Applications*, pp. 203–214, Springer, 2021.

[105] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *British journal of pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011.

[106] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve r&d productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery*, vol. 9, no. 3, pp. 203–214, 2010.

[107] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: deep drug–target binding affinity prediction. bioinformatics 34: i821–i829," 2018.

[108] N. R. Monteiro, J. L. Oliveira, and J. P. Arrais, "Dtitr: End-to-end drug–target binding affinity prediction with transformers," *Computers in Biology and Medicine*, vol. 147, p. 105772, 2022.

[109] K. Huang, C. Xiao, L. M. Glass, and J. Sun, "Moltrans: molecular interaction transformer for drug–target interaction prediction," *Bioinformatics*, vol. 37, no. 6, pp. 830–836, 2021.

[110] M. Gao, D. Zhang, Y. Chen, Y. Zhang, Z. Wang, X. Wang, S. Li, Y. Guo, G. I. Webb, A. T. Nguyen, *et al.*, "Graphormerdti: a graph transformer-based approach for drug-target interaction prediction," *Computers in Biology and Medicine*, vol. 173, p. 108339, 2024.

[111] P. Zhang, Z. Wei, C. Che, and B. Jin, "Deepmgt-dti: Transformer network incorporating multilayer graph information for drug–target interaction prediction," *Computers in biology and medicine*, vol. 142, p. 105214, 2022.

[112] N. R. Monteiro, J. L. Oliveira, and J. P. Arrais, "Tag-dta: Binding-region-guided strategy to predict drug-target affinity using transformers," *Expert Systems with Applications*, vol. 238, p. 122334, 2024.

[113] Y. Liu, L. Xing, L. Zhang, H. Cai, and M. Guo, "Geformerdta: drug target affinity prediction based on transformer graph for early fusion," *Scientific Reports*, vol. 14, no. 1, p. 7416, 2024.

[114] L. Jiang, C. Jiang, X. Yu, R. Fu, S. Jin, and X. Liu, "Deeptta: a transformer-based model for predicting cancer drug response," *Briefings in bioinformatics*, vol. 23, no. 3, p. bbac100, 2022.

[115] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar, "Comprehensive analysis of kinase inhibitor selectivity," *Nature biotechnology*, vol. 29, no. 11, pp. 1046–1051, 2011.

[116] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, *et al.*, "Pubchem in 2021: new data content and improved web interfaces," *Nucleic acids research*, vol. 49, no. D1, pp. D1388–D1395, 2021.

[117] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, and A. Masoudi-Nejad, "Deepcda: deep cross-domain compound–protein affinity prediction through lstm and convolutional neural networks," *Bioinformatics*, vol. 36, no. 17, pp. 4633–4642, 2020.

[118] M. Kanehisa, "The kegg database," in *'In silico'simulation of biological processes: Novartis Foundation Symposium 247*, vol. 247, pp. 91–103, Wiley Online Library, 2002.

[119] K.-C. Chou, "Progresses in predicting post-translational modification," *International Journal of Peptide Research and Therapeutics*, vol. 26, no. 2, pp. 873–888, 2020.

[120] J.-W. Seo and K.-J. Lee, "Post-translational modifications and their biological functions: proteomic analysis and systematic approaches," *BMB Reports*, vol. 37, no. 1, pp. 35–44, 2004.

[121] M. Krassowski, M. Paczkowska, K. Cullion, T. Huang, I. Dzneladze, B. F. F. Ouellette, J. T. Yamada, A. Fradet-Turcotte, and J. Reimand, "Activedriverdb: human disease mutations and genome variation in post-translational modification sites of proteins," *Nucleic acids research*, vol. 46, no. D1, pp. D901–D910, 2018.

[122] C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto Jr, "Protein posttranslational modifications: the chemistry of proteome diversifications," *Angewandte Chemie International Edition*, vol. 44, no. 45, pp. 7342–7372, 2005.

[123] A. J. Bannister, E. A. Miska, D. Görlich, and T. Kouzarides, "Acetylation of importin-α nuclear import factors by cbp/p300," *Current Biology*, vol. 10, no. 8, pp. 467–470, 2000.

[124] Y. Qiao, X. Zhu, and H. Gong, "Bert-kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained bert models," *Bioinformatics*, vol. 38, no. 3, pp. 648–654, 2022.

[125] L. Meng, X. Chen, K. Cheng, N. Chen, Z. Zheng, F. Wang, H. Sun, and K.-C. Wong, "Transptm: a transformer-based model for non-histone acetylation site prediction," *Briefings in Bioinformatics*, vol. 25, no. 3, p. bbae219, 2024.

[126] N. Kalebic, S. Sorrentino, E. Perlas, G. Bolasco, C. Martinez, and P. A. Heppenstall, "αtat1 is the major α-tubulin acetyltransferase in mice," *Nature communications*, vol. 4, no. 1, p. 1962, 2013.

[127] Q. Wang and H. Gao, "Prediction of protein post-translational modifications in rice based on multi-head self-attention," in *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 1–6, IEEE, 2022.

[128] Z. Xu, H. Zhong, B. He, X. Wang, and T. Lu, "Ptransips: Identification of phosphorylation sites enhanced by protein plm embeddings," *IEEE Journal of Biomedical and Health Informatics*, 2024.