Protein Language Modeling: Coding Life's Code

Bolutito Babatunde
Department of Mechanical Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
bbabatun@andrew.cmu.edu

Madeline Davis
Department of Biomedical Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
madelind@andrew.cmu.edu

Abstract

Just as human language is composed of words, protein sequences consist of amino acids, forming the basis of biological processes. Understanding the relationship between a protein's structure and function is pivotal for advancements in healthcare and biotechnology. However, the complexity of proteins, influenced by factors such as folding patterns and evolutionary variations, poses a significant challenge. Recent developments, especially in deep learning like AlphaFold, have revolutionized protein structure prediction, offering more accuracy even for proteins without known homologs. This work focuses on single-sequence-based prediction methods, leveraging SPOT-1D-LM for predicting protein structural properties. By integrating embeddings from pre-trained models like ESM-1b and ProteinBERT, we will enhance our understanding of proteins' structural and functional motifs, crucial for addressing the protein-function prediction problem.

See Github: Here

1 Introduction

In a manner analogous to human language, protein sequences can be depicted as strings of letters, each corresponding to the 20 standard amino acids, excluding rare and atypical types (Figure 1). As fundamental macromolecules, proteins are essential to all biological processes sustaining life. Anfinsen famously referred to fully understanding the link between protein structure and function as the 'Holy Grail' of structural biology. [1]. This understanding opens doors to advancements in disease research, drug discovery, and biotechnological development. Despite their ubiquity, our current grasp of protein structural characteristics remains limited due to their immense diversity and complexity. This complexity arises from various factors, including intricate folding patterns, evolutionary differences, dynamic behaviors, post-translational modifications, and interactions between different proteins [2]. In past two decades the application of deep-learning based approaches to predict protein structure have emerged as a viable alternatives to traditional methods (e.g. X-ray Crystallography [3], Cryo-Electron Microscopy [4]), which are constrained by low throughput, high resource demand and low resolution [2, 5, 6]. Specifically, utilizing Protein Language Models to infer structural properties from a protein's foundational amino acid sequence (sequence-based prediction) has presented a viable methodology to further our understanding of the relationship between protein sequence, structure, and function [2].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

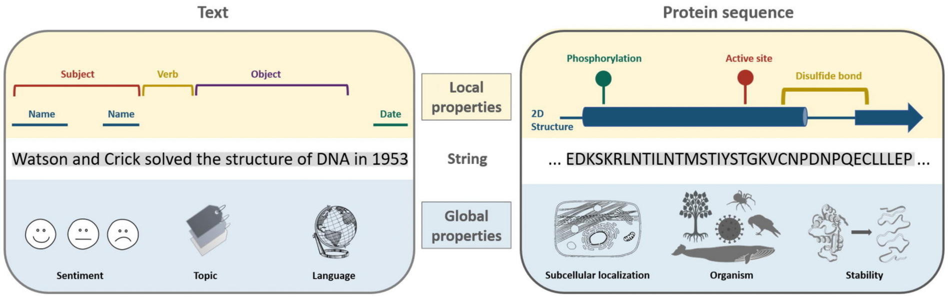


Figure 1: Protein sequences; natural language analogs for analyzing local and global protein properties. (left) Deconstruction of natural language phrase. (right) Deconstruction of protein sequence. (Sourced Ofer et al. [7])

Protein structure prediction employs two methods: template-based and template-free modeling. Template-based modeling predicts novel protein structures by utilizing existing protein information in the protein data bank (PDB). This approach makes predictions by relying on homologous protein sequences, which are protein sequences with common evolutionary origin. However, for a notable number of proteins, especially from less-studied organisms, there are few or no identifiable homologous sequences available. [8]. Recent advances in protein structure prediction, including AI-driven methods like AlphaFold [6, 9, 10] and RoseTTAFold [11], have significantly enhanced the accuracy of protein structure predictions without known homologs. The current state of the art protein language models rely on evolutionary-based methods such as multiple sequence alignemnt (MSA) and homology modeling (Figure 2) to exploit the features of a protein's evolutionary relatives [12, 13, 14, 15]. On the other hand, template-free modeling utilize non-evolutionary or single-sequence based methods [16]. The recent adoption of sequence-based prediction has vastly improved the efficacy of de-novo modeling methods for protein secondary structure prediction, contact map prediction and tertiary structure prediction, as observed in CASP-15 [6].

Single sequence-based prediction methods, which utilise deep learning to capture the biological structural and functional motifs ingrained within a protein's primary structure, are crucial to tackling the protein-function prediction problem. In this work, we will explore end-to-end protein structure prediction by integrating single-sequence-based methods with learned representations optimized specifically for the unique characteristics of protein. We employ an architecture based on the model utilized in the recently published SPOT-1D-LM [17] to generate predictions for 1D structural properties and protein secondary structure states based on definitions in the Dictionary of secondary structure predictions (DSSP) [18, 19]. In addition to one-hot encodings of amino-acid sequences, our input included the concatenated embeddings of two pre-trained models. Similarly to SPOT-1D-LM [17] we utilized embeddings from ESM-1b [20] a language model pre-trained on the Uniref50 Dataset [21]. However, in addition to this we leveraged the outputs of the novel model ProteinBERT to investigate utilizing an attention map from a deep language model designed for proteins, to perform downstream contact map prediction [22]. The concatenated inputs were then passed through SPOT-1D-Single [23], an ensemble of hybrid BiLSTM-CNNs.

2 Related Work

2.1 Overview of Protein Language Modelling

The basic primary structure of proteins consists of a polypeptide chain of amino acids that fold into a unique 3D secondary and tertiary structure [7]. Thus, in a similar manner to words in natural language, proteins can be conceptualized as a linear sequence of modular amino-acid building blocks that can be rearranged and assembled to produce structures with distinct functions [7]. Interestingly, protein amino-acid sequences exhibit characteristics such as hierarchical organization, context sensitivity and folding constraints that parallel linguistic syntax and grammar rules (Figure 1) [24].

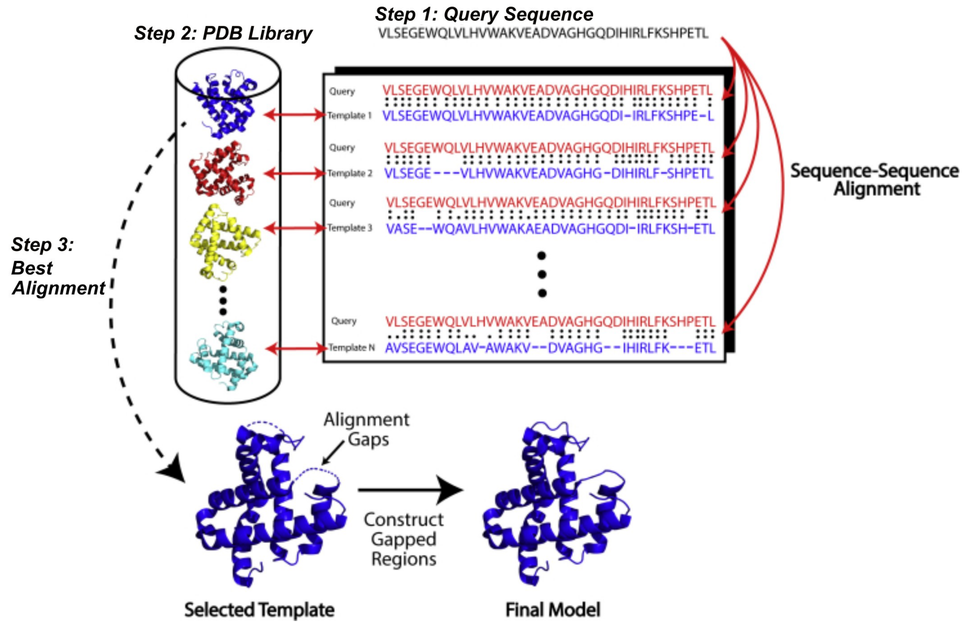

Step 1: Query Sequence
VLSEGEWQLVLHVWAKVEADVAGHGQDIHIRLFKSHPETL
Step 2: PDB Library
Query
VLSEGEWQLVLHVWAKVEADVAGHGQDIHIRLFKSHPETL
Template 1
VLSEGEWQLVLHVWAKVEADVAGHGQDI-IRLFKSHPE-L
Query
VLSEGEWQLVLHVWAKVEADVAGHGQDIHIRLFKSHPETL
Template 2
VLSEGE---VLHVWAKVEADVAGHG-DIHIRLF-SHPETL
Query
VLSEGEWQLVLHVWAKVEADVAGHGQDIHIRLFKSHPETL
Template 3
VASE--WQAVLHVWAKAEADVAGHGQDI-IRLFKSH-ETL
Query
VLSEGEWQLVLHVWAKVEADVAGHGQDIHIRLFKSHPETL
Template N
AVSEGEWQLAV-AWAKV--DVAGHG--IHIRLFK---ETL
Sequence-Sequence Alignment
Step 3: Best Alignment
Alignment Gaps
Construct Gapped Regions
Selected Template
Final Model

Figure 2: Homology-based modeling scheme: (1) Indentify templates with sequence alignment algorithms, (2) Copy structural framework of best templates, and (3) Generate final model by constructing unaligned regions. (Adapted from Pearce and Zhang [2])

2.1.1 Amino Acid Token Embedding

Protein language modeling regards individual amino acids as character-level tokens [24, 7].There are 20 standard tokens utilised to represent all known amino acids, as well as additional tokens to represent the beginning/end of a sequence, gaps in sequences, unknown amino acids and modified amino acids.

2.1.2 Structure Prediction Outputs

RoseTTAFold employs a unique three-track architecture. It consists of a 1D track for amino acid sequences, a 2D track for pairwise residue interactions, and a 3D track for three-dimensional structure prediction. It effectively integrates information from the sequence itself (1D), patterns of interaction between pairs of amino acids (2D), and the 3D spatial structure. The model iteratively refines its predictions, passing information between the 1D, 2D, and 3D tracks to progressively improve the accuracy of the protein structure prediction. By considering multiple aspects of protein structure (sequence, interactions, and 3D conformation), it achieves a more comprehensive and accurate prediction. The main output is the predicted three-dimensional structure of the protein. It also provides information on secondary structures and spatial relations between amino acid residues.

End-to-End methods streamline the protein structure prediction pipeline by utilizing deep learning to directly output three-dimensional properties of proteins as a function of amino acid sequence inputs[25]. Notable protein language models include RoseTTAFold [11], RaptorX [26], ProteinUnet [27], SPOT-1D-Single [23] and NetSurf [13] output secondary structure predictions.

2.2 State of the Art

2.2.1 Attention Based Learning for Protein Structure Prediction

The integration of self-attention based transformer architectures into protein language models constitutes the most influential innovation to the field in recent years [28, 29]. Self-attention mechanisms, in conjunction with parallel processing, have extended the capacity of models to process lengthy sequences of amino acids and effectively capture the long-range and complex dependencies between residues [2]. Notably DeepMind's AlphaFold2 represents the current industry benchmark, with a structure prediction accuracy that matches experimental methods [10] and outperformed contending models at CASP-14 [9]. AlphaFold2's innovative "evoformer" network conceptualizes a protein's 3D structure as a graph, in which residues constitute nodes and their spatial and evolutionary relationships are represented by edges. The network consists of 48 evoformer blocks joined by residual connections in which axial attention is used to iteratively refine and integrate evolutionary patterns from MSA representations, and pairwise interactions between amino acids [30].

Like other MSA-based methods AlphaFold2 displays diminished performance on proteins lacking co-evolutionary data, with accuracy scores of 20% for all metagenomic sequences and 11% for eukaryotic and viral proteins [31]. Additionally the requirement for database-searching in MSA methods constitutes a bottleneck that significantly reduces time efficiency and the ability of a model to process a large number of proteins [25].

2.2.2 Single-Sequence Prediction

In response to the shortcomings of evolutionary based methods, there is a growing interest in single-sequence-based prediction methods, which utilize the intrinsic properties of amino-acid sequences to determine protein structure [25, 28]. The physical process of protein folding is driven primarily by the interplay of physicochemical forces and interactions of it's chain, thus single-sequence prediction is more biologically plausible than evolutionary-based prediction, and is increasingly tractable with growing advances in deep learning [31, 32]. In the absence of a MSA, single-sequence prediction methods typically leverage statistical methods, energy minimisation, and most recently transfer learning, to gain additional context of a protein's structure-function relationship.

2.2.3 Protein Representation Learning for Single Sequence Prediction

The accuracy of single-sequence methods has been enhanced with protein representation learning, the process of encoding proteins into a computationally tractable, generalisable form for downstream structure prediction tasks [24]. Protein Language Models ESM-1b [20] and ProtT5-XL-U50 (ProtTrans) [33] have generated rich contextual embeddings, to transfer information with regards to protein structure, the effect of sequence variation on function and inverse folding to subsequent applications.

The transformer based ESM-1b language model (outlined in Section 4.1.3) achieves SOTA benchmarks for secondary structure prediction, whilst significantly improving upon the computational costs of profile -based models such as AlphaFold2 and RoseTTAFold [34]. The ProtT5-XL-U50 model, based on the auto-encoder model T5, additionally approaches state of the art accuracy without the use of MSAs [33]. Like ESM-1b the model effectively generates essential biophysical amino acid features, however it is significantly larger and doesn't have a restriction on protein sequence length [35].

The use of learned representations from pre-trained language models such as ESM-1b and ProtT5-XL-U50 offers improved efficiency and computational performance, enables the use of large datasets and addresses the lack of existing labeled protein structures. Additionally, models such as SPOT-1D-Single [23] have demonstrated that passing embeddings from ESM-1b in addition to amino acid sequences through an Feed-Foward Neural Network, can lead to improved performance for secondary structure prediction. Not only was SPOT-1D-Single able to outperform alternative single-sequence prediction methods and evolutionary-based methods, but it provided preferable results to using ESM-1b alone.

3 Dataset

3.0.1 Trainining & Validation Datasets

Similar to SPOT-1D-LM [17], our study utilizes datasets from the SPOT-1D-Single model [23], which is sourced from ProteinNet [36] and caps sequence identity at 95% to enhance training data diversity. This means that sequences that are more that 95% identical to each other are removed. This provides 50,914 proteins registered in the PDB prior to 2016, with resolution < 2.5 Å, indicating a high level of detail. To prevent overfitting, 100 proteins were randomly selected for validation against the training set using Hidden Markov Model comparison, excluding similar sequences and those over 1024 amino acids in length, yielding 39,012 proteins with a 80-20 split between training and validation. We used a batch size of 10 for the training and a batch size of 5 for the validation.

3.0.2 Test Datasets

We also used the same test set from SPOT-1D-LM: TEST2018, CASP12-FM, CASP13-FM, CASP14-FM, TEST2020-HQ, and Neff1-2020. TEST2018 is the same test set used in SPOT-1D [12], which consists of 250 proteins released between January 01, 2018 and June 17, 2018 with resolution <2.5 Å and R-free <0.25, indicating high model-to-data agreement, with sequence identity at 25% using the BlastClust software [37]. R-free [38] is used to assess the quality of a protein structure acquired by X-ray crystallography. The CASP (Critical Assessment of Structure Prediction techniques) test sets—CASP12-FM [39], CASP13-FM [40], and CASP14-FM [41]—comprises of 22, 17, and 15 free modeling targets, which are particularly challenging proteins without known structural templates in the PDB at their release date. SPOT-1D-Single further acquires a hard data set TEST2020, which consists of proteins released between May 2018 and April 2020 that are additionally refined by removing close and remote homologs using Hidden Markov Model comparison, excluding similar sequences and those over 1024 amino acids in length to yield 671 proteins. Close homolgs have high percentage of sequence identity and therefore have direct evolutionary relationships, while remote homologs share less sequence identity with more distant evolutionary relationships. These proteins were further constrained to a resolution <2.5 Å and R-free <0.25, yielding 241 proteins for the TEST2020-HQ test set. Separating the proteins without homologs (Neff1) from TEST2020 yields the Neff1-2020 test set with 46 proteins. Proteins without homologs creates a much harder test set with no detectable sequence similarity to other proteins in the database and therefore eliminating bias.

3.1 Overview of Data Types

3.1.1 Input data

Protein sequence inputs are provided in FASTA format [42, 43], a standardised text-based format for representing peptide sequences in bioinformatics applications. FASTA encodes proteins as a string of letters in Table 1, each corresponding to one of the 20 amino acid codes.

Amino Acid Code	Meaning
A	Alanine
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine/Start codon
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
V	Valine
W	Tryptophan
Y	Tyrosine
X	Others

Table 1: Character Assignments for Amino Acids

3.1.2 Output Data

Protein secondary structures can be classified into the three-state (SS3) and the more detailed eight-state (SS8). These secondary states provide essential insights into their complex configurations. Our model primarily utilizes the SS8 labels for detailed structural analysis. SS3 broadly categorizes structures into helix (H), strand (E), and coil (C) formations (Figure 3 (a)). SS8 refines this by splitting helix (H) into 3_{10}helix (G), alpha helix (H), and pi helix (I); strand into beta strand (E) and beta bridge (B); and coil (C) into beta turn (T), high curvature loop (S), and irregular (C). Some of these eight state formations are illustrated in Figure 3 (b). The baseline paper [17] includes Dictionary of Secondary Structure of Proteins (DSSP) files that provide the eight-state secondary structure labels, each corresponding to an amino acid in a protein sequence.

In addition to classification tasks, protein language models can be used to predict other properties such as the accessible surface area (ASA), half-sphere exposure (HSE), coordination number (CN), and backbone angles (ψ, ϕ, θ, and τ). ASA, which is typically measured in square Angstroms, is a measure of the area of amino acid residue exposure to a solvent molecule [44]. HSE, represented by integer values, measures how buried an amino acid is in the overall protein structure. HSE-up and HSE-down splits this HSE value by separating the amino acid's contacting sphere into half-spheres. CN is the sum of HSE-up and HSE-down [45]. The dihedral angles ϕ and ψ are the protein backbone torsion angles describing the angles between atomic planes that typically range from $-180.0°$ to $180.0°$. As illustrated in Figure 4 (a), ϕ measures the counter-clockwise rotation around the nitrogen atom (N) and the alpha carbon atom ($C\alpha$) bond when viewed from the N-terminus to the C-terminus, while ψ measures the counter-clockwise rotation around the alpha carbon atom ($C\alpha$) and carbonyl carbon (C) bond when viewed from the $C\alpha$ end [46]. Figure 4 (b) illustrates τ, which is a dihedral angle measuring the counter-clockwise rotation between the two consecutive alpha bonds of neighboring amino acid residues ($C_{\alpha i} - C_{\alpha i+1}$) when viewed from the $C_{\alpha i+1}$ end, while θ measures the angle formed by the nitrogen atom (N), the alpha carbon atom ($C\alpha$), and the carbonyl carbon atom (C) of an amino acid residue [47].

The baseline paper [17] includes DSSP files with ASA, ϕ, and ψ values, $.\mathbf{h}$ files for HSE-up, HSE-down, and CN values, and $.\mathbf{t}$ files for τ and θ values. Each set of values corresponds to specific amino acids in a protein sequence. The properties and labels can be downloaded from the resources tab in the Zhou Laboratory website. While the baseline model predicts these properties, our project currently focuses on eight-state label training due to time constraints. It would be interesting to add on more prediction capabilities in the future.

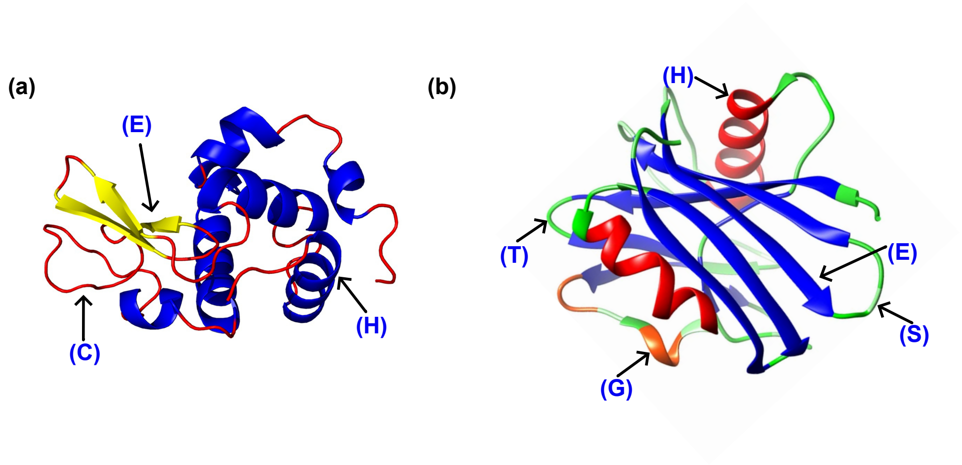


Figure 3: The protein structure is represented in two ways: as three states (SS3) and eight states (SS8); (a) In SS3, proteins are categorized into helix (H), coil (C), and strand (E). (Adapted from Duran et al. [48]). (b) For a more detailed representation, SS8 expands these categories. For example, the helix (H) from SS3 is further divided into 3_{10}helix (G) and alpha helix (H), and the strand (E) into beta-strand (E). Additional SS8 states include high-curvature loop (S), beta-turn (T), and a broader coil (C) category. (Adapted from Hong et al. [49]).

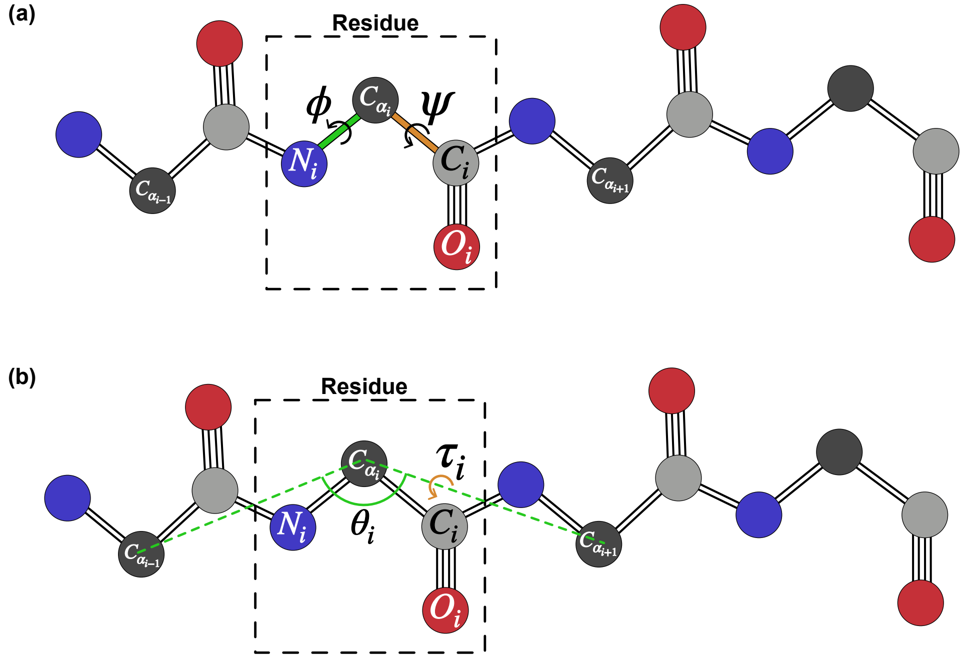


Figure 4: The protein backbone angles; (a) ϕ and ψ dihedral angle illustration, where ϕ is the counter-clockwise rotation along the $N - C\alpha$ bond from the N-terminus. (b) τ is the counter-clockwise rotation along the $(C_{\alpha i} - C_{\alpha i+1})$ from the $C_{\alpha i+1}$ end. θ is the angle formed by the nitrogen atom (N), the alpha carbon atom $(C\alpha)$, and the carbonyl carbon atom (C). (Adapted from Hong et al. [47]).

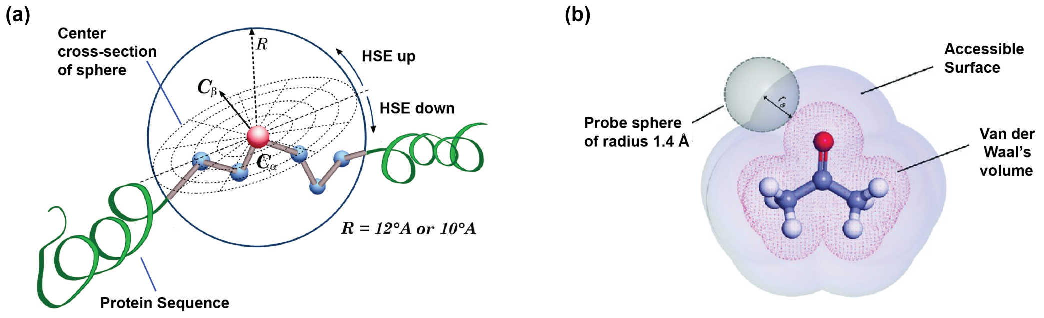


Figure 5: (a) Visualisation of the Accessible Surface Area, the surface of a biomolecule accessible to a solvent, in comparison to the Vander Waal's Surface, which provides a representation of the molecule's physical boundary (Adapted from Sharma et al. [50]). (b) Half Sphere Exposure provides a measure of the number of neighbouring amino acids located within two half-spheres (up and down) surrounding an amino acid of interest (Adapted from Jarray et al. [51]).

3.2 Datasets Used for Pre-trained Models

3.2.1 ESM-1b

ESM-1b was pre-trained with the UniRef50 dataset, a set of clustered protein sequences provided by the UniProt Knowledge Base and UniParc Records. UniRef50 consists of 86 billion amino acids spanning 250 million protein sequences clustered at a sequence identity level of 50%.

3.2.2 ProteinBERT

ProteinBERT was pretrained using UniRef90 and UniProtKB. UniRef90 consists of a set of 106 million protein sequences clustered at least 90% sequence identity, whilst UniProtKB provides Gene Ontology annotations for selected protein sequences. From both UniRef90 and UniProtKB a subset of 46 million proteins and 8943 Gene Ontology annotations were selected for training.

3.3 Dataset Preparation and Preprocessing

To prepare the dataset, the following preprocessing steps were undertaken. Firstly, the protein backbone angles, which are inherently periodic were converted into their sine and cosine components. This accommodates for the circular nature of the input data, whilst enabling the model to efficiently process the angles. The resulting sin and cosine components were then normalised between 0 and 1. Additionally, the ASA values were normalised according to the maximum values observed for each amino acid type. Normalisation accounts for the varying sizes and solvent exposure properties of the different amino acids, standardising ASA features across the dataset. HSE values, for both HSE-u and HSE-d, were normalised by the maximum values of the entire dataset.

To handle the diversity of amino acid residues and to simplify the model's language comprehension, rare amino acids distinct from those outlined in 1 (e.g. U = Selenocysteine) are replaced with the token "X". This designation of "X" as a placeholder for undefined or less common residues helps to establish a common tokenization mechanism, streamlining the integration of embeddings and one-hot encodings in the model's architecture. This unification is crucial for the model to process and learn from the vast array of protein sequences effectively.

4 Baseline Model and Implementation

4.1 Baseline Selection

4.1.1 Spot-1D-LM

The SPOT-1D-LM Model proposed by Singh et. al [17] was selected as our baseline model. A comparative analysis in [6] demonstrates SPOT-1D-LM outperforms competing single-sequence-prediction methods PSIRED Single [52], SPIDER3-Single [16] and ProteinUnet [27] for structure prediction tasks. Moreover, SPOT-1D-LM has demonstrated higher prediction accuracy than SOTA profile-based tools such as NetSurfP-2.0 for proteins without homologous sequences. Exploring this architecture as our foundational benchmark can be beneficial for understanding the model's behaviour, verifying results and potentially extending its capabilities. The model is demarcated by its combination of three distinct of input features: one-hot encoded single sequence embeddings, ESM-1b encodings and Prot-Trans encodings. The input streams are concatenated together and passed through the SPOT-1D-Single Architecture, an ensemble of ResNet & LSTM architectures. Similarly to SPOT-1D-Single's training procedure, two separate ensembles were trained for classification (SS3 & SS8) and regression (ASA, HSE, CN & backbone angles) tasks. Both models were trained using a batch size of 10, and Cross-entropy loss and L1-Loss were utilised for the classification and regression ensembles respectively. The classification outputs were evaluated based on percentage accuracy. Pearson's Correlation Coefficient was used to assess ASA, HSE-u, HSE-d and CN and predictions, whilst Mean Absolute error was selected to evaluate protein backbone angle predictions.

4.1.2 SPOT-1D-Single Architecture

SPOT-1D-Single [23] is a novel architecture for single-sequence prediction that utilises an ensemble of 3 models with different variations of hybrid LSTM bidirectional neural networks and convolution neural networks. SPOT-1D-Single is an ensemble consisting of a 2-Layer BiLSTM, a multi-scale parallel ResNet (MS-ResNet) and a multi-scale parallel Res-LSTM (MS-Res-LSTM) as visualised in 7.

The first model is composed of 2 Bi-directional LSTM layers, each with a hidden dimensino of 1024 and a dropout layer of p=0.5. The LSTM outputs are then passed through 2 fully connected layers with a dimension of 2000. The second model consists of 3 parallel stacks of ResNets with identical compositions, excluding kernel size (k=7, k=9, k=15). All 3 streams contain a 15 1D ResNet blocks

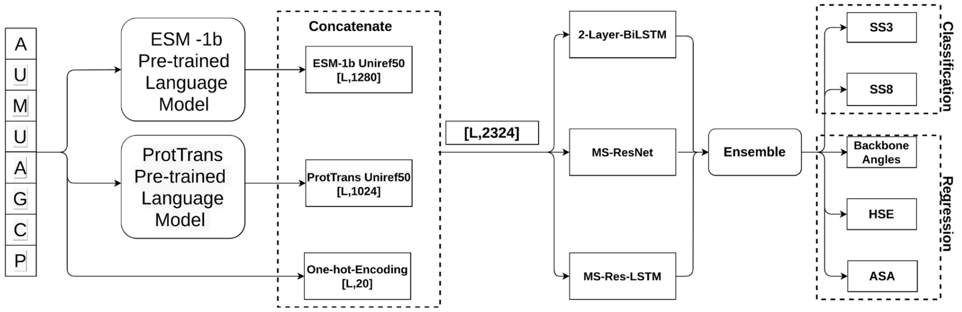


Figure 6: The SPOT-1D-LM architecture (Sourced from Singh et al. [17])

stacked sequentially, where the first 5 blocks use filter size = 64, the second 5 blocks use filter size = 128 and the final 5 blocks use filter size 256. All blocks are appended with batch normalisation and ReLU activation layers. Outputs of the parallel stacks are concatenated and passed through a linear output layer. The third model features a similar architecture to the MsResNet, but the first stack is replaced with a 4 layer BiLSTM with a hidden size of 128 and a dropout layer of p= 0.5.

All three models are used for both classification and regression tasks, and follow the training method outlined 4.1.1(SPOT-1D-LM). For the classification tasks, 11 output nodes (3 for SS3 and 8 for SS8) are used and a sigmoid output activation function is appended to each model to interpret model predictions as a distribution of class probabilities. For classification, SPOT-1D-Single averages the probability outputs from each model in the ensemble, while regression tasks involving properties like ASA, HSE-u, HSE-d, and CN use the mean ensemble of all models. For protein backbone angles predictions, SPOT-1D-Single employs the median ensemble to avoid generating unfeasible angles.

4.1.3 ESM-1b

Our selected Baseline architecture SPOT-1D-LM utilises pre-trained language model ESM-1b to generate embeddings for downstream structure prediction. ESM-1b, a protein language model based on the RoBERTa architecture and training procedure, has demonstrated state of the art performance on mutational effect, secondary structure prediction and contact prediction. RoBERTa (Robustly Optimized BERT Approach) [53], is a variant of the BERT architecture that has demonstrated improved convergence properties by adopting a larger training dataset, longer training times and large batch sizes. Moreover, RoBERTa applies dynamic masking during training and uses Byte-Pair Encoding (BPE) for tokenization over WordPiece to enhance generalisation and efficiency. Following the RoBERTa training approach, ESM-1b uses an unsupervised masked language modelling objective, masking 15 % of amino acids, to capture contextual and spatial patterns between residues [54].This results in a learned representation space that acutely reflects the biochemical properties and biological variations of amino acids, as well as the secondary and tertiary structure of proteins. Moreover, ESM-1b is trained on 86 billion amino acids across 250 million evolutionary diverse protein sequences (Uniref50 Dataset), inherently facilitating the learning of evolutionary dependencies without MSA [20]. The elimination of the computationally expensive search process for related proteins, as well as a simplified architecture, enables ESM 1-b to perform inference 60x faster than AlphaFold2 [34].

5 Model Description

Our selected learning pipeline was inspired by our Baseline Model SPOT-1D-LM's integration of ensemble architectures and transfer learning to achieve high performance in protein structure and contact map prediction. As illustrated in Figure 8, we first utilised language models ESM-1b and ProteinBERT to generate contextual embeddings for individual protein sequences. These embeddings are concatenated with the protein sequences, represented as one-hot vectors, and passed through through the SPOT-1D-Single Ensemble.

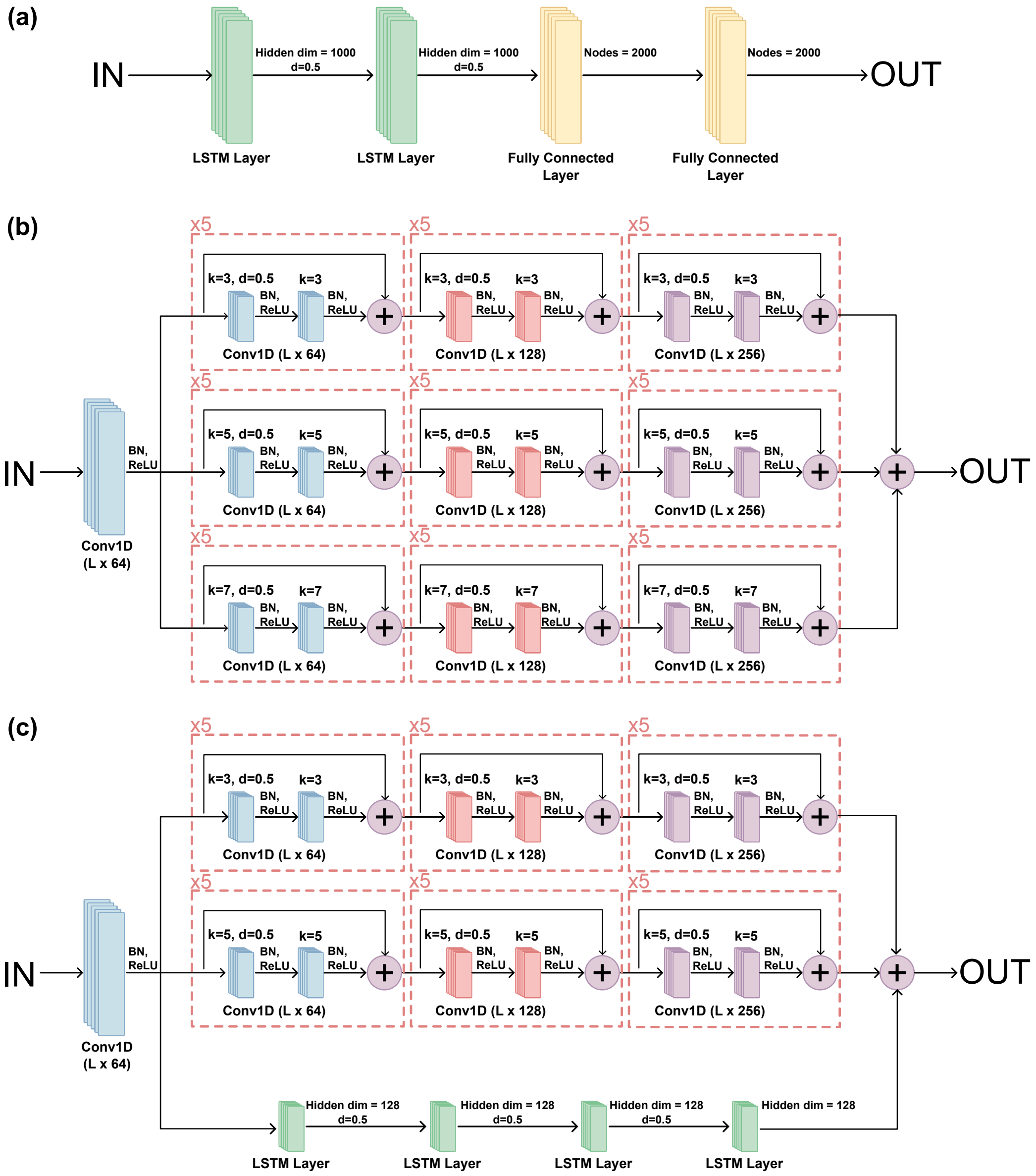


Figure 7: The SPOT-1D-Single Architecture (Adapted from Singh et al. [23])

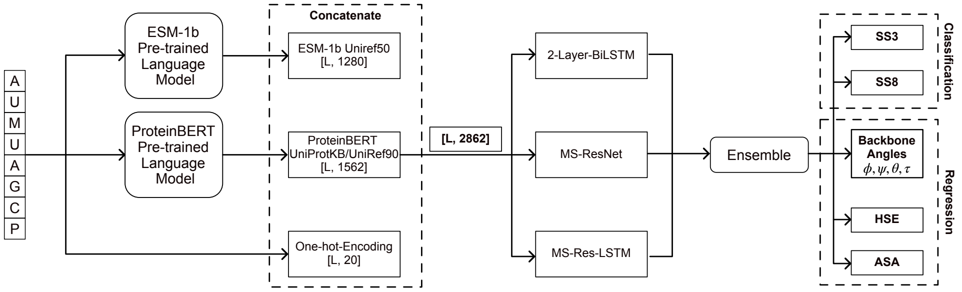


Figure 8: Proposed Model Architecture for Protein Structure Prediction

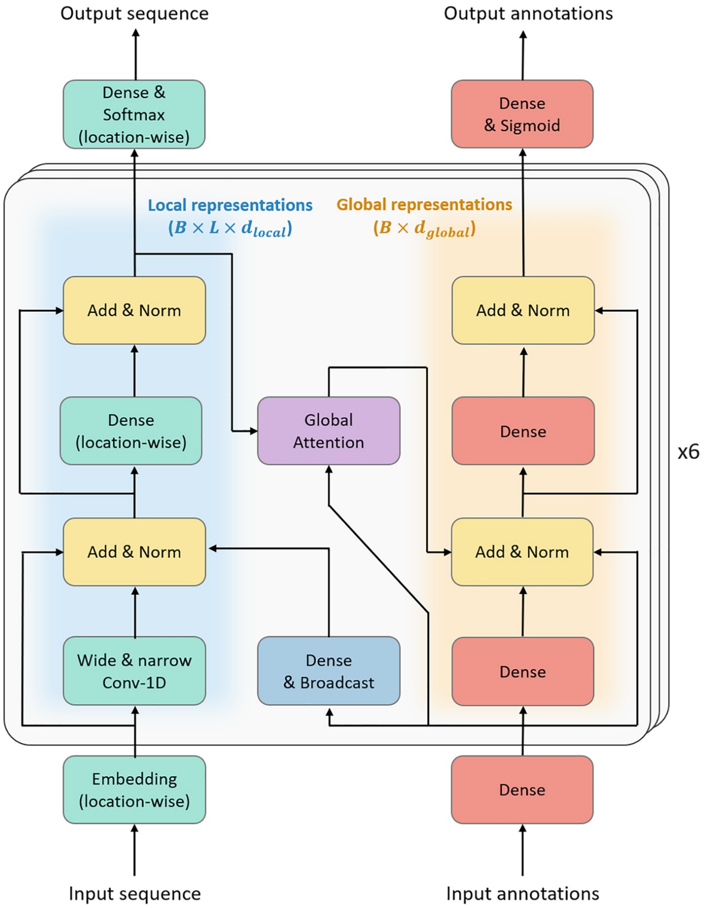


Figure 9: The proteinBERT architecture (Sourced from Brandes et al. [22])

5.1 Generating Embeddings - ProteinBERT Language Model

To adapt our selected baseline implementation we propose leveraging the contextual embeddings from ProteinBERT in place of the Prot-Trans Language Model encodings used in SPOT-1D-LM. Pre-existing language model architectures adapted for a protein dataset, such as ESM-1b and Prot-Trans are inherently constrained by their basis in natural language. Although proteins and natural language share some similar characteristics, protein sequences exhibit several fundamental differences rooted in their biological nature. Specifically, protein sequences have a lower entropy than the English language; proteins don't have an analog to punctuation to distinguish between structural features; protein functional units often overlap; protein sequences have greater variability in length and due to their structure residues often form long-range dependencies. In contrast, ProteinBERT is a deep language model distinguished by its comprehensive, protein-orientated pre-training task, and an adaptation of the Transformer/BERT architecture that enables the distinction between local and global representations [22] (Figure 9). These distinctions enable ProteinBERT to outperform existing protein language models in terms of speed and size, whilst approaching or exceeding state-of-the-art performance for structure prediction benchmarks

5.1.1 Pre-Training Tasks

The pre-training procedure for ProteinBERT can be segmented into two distinct tasks that are performed concurrently; Masked language modeling and Gene Ontology (GO) annotation prediction. Gene Ontology annotations provide up-to-date analogues of a protein's molecular function, cellular component and involvement in biological processes [55]. Integrating protein language models with the rich, standardized biological information provided by Gene Ontology enables ProteinBERT to bridge the gap between raw sequence data and functional biological information.

Task 1: Masked Language Modelling Model learns a bidirectional representation of an amino acid sequence

- *Learns Local Features*- protein sequences represent local features, detailing the micro-level, immediate structure of proteins that determines their physical configuration and direct molecular interactions
- *Input Type* - Protein sequences encoded as sequences of 26 unique integer tokens (20 amino acids, START, END, PAD, UNDEFINED, OTHER, selenocysteine U)
- *Loss Function* - Categorical Cross Entropy over the protein sequences

$$\mathcal{L} = -\sum_{i=1}^{l} log(\hat{S}_{i,S_i})$$

where l is the sequence length, $S_i \in \{1, ...26\}$ is the sequence's true token at position i, and $\hat{S}_{i,S_k} \in [0, 1]$ is the probability that the token k (for any $k \in \{1, ...26\}$)is predicted at position$i$

Task 2: Gene Ontology Annotation Prediction Model Learns GO annotation classifications for specific proteins

- *Learns Global Features* - Gene Ontology annotations represent global features as they provide a macro-level view of a protein's role and function within the broader context of biological systems and processes
- *Input Type* - Binary sequence of length 8943 representing attributes of a given protein
- *Loss Function* - Binary Cross Entropy over GO annotations

$$\mathcal{L} = -\sum_{j=1}^{8943} (A_j \cdot log(\hat{A}_j)) + (1 - A_j) \cdot log(1 - \hat{A}_j)$$

where $A_j \in 0, 1$ is the true value of the annotation at position j (for any $j \in \{1, ...8943\}$), and $\hat{A}_j \in 0, 1$ is the predicted probability that the protein has annotation j

5.1.2 Model Architecture

ProteinBERT is a denoising autoencoder consisting of two parallel paths that distinctly process the local amino acid inputs and the global GO annotation inputs. The model consists of 6 transformer blocks, with 4 global attention heads per block and skip connections between hidden layers. Within the blocks, local representations are transformed by 1D convolutional layers and global representations are transformed by 2 FC layers, avoiding the use of recurrent connections to improve efficiency and stability with respect to sequence length. Information between both paths is integrated via broadcast fully-connected layers (from global to local representations) and global attention layers (from local to global representations).

5.1.3 Global Attention Layer

ProteinBERT's Global Attention Layers constitute a novel architectural component based on self-attention, with reduced resource consumption. The objective of global attention is to determine the local positions of an input sequence with respect to to an additional fixed-size "global" input vector. Each single-head global attention layer takes in a global representation vector $x \in \mathbb{R}^{d_{global}}$ and a set of L positional representation vectors $s_1, ...s_L \in \mathbb{R}^{d_{local}}$, and outputs a global vector $y \in \mathbb{R}^{d_{value}}$. Similarly to self-attention, values and key vectors are calculated with respect to each position $i \in \{1, ...L\}$, however the query vector is derived from the global input, as elucidated below:

Table 1: Global Attention Functions

Values	$v_i = \sigma(W_i s_i)$	$\in \mathbb{R}^{d_{value}}$
Keys	$k_i = tanh(W_k s_i)$	$\in \mathbb{R}^{d_{key}}$
Query	$q = tanh(W_q x)$	$\in \mathbb{R}^{d_{key}}$
Attention Weights	$z_1, ...z_L = softmax\{\frac{\langle q,k_i \rangle}{\sqrt{d_{key}}}\})$	$\in [0, 1]$
Output	$y = \sum_{i=1}^{L} z_i v_i$	$\in \mathbb{R}^{d_{value}}$

where W_v, W_k and W_q are trainable parameter matrices and σ is the GELU activation function.

5.2 Loss Functions

5.2.1 Loss for Classification Tasks

We directly adapted the loss function utilised for the classification task in the Baseline model, Cross Entropy Loss. This is as cross entropy loss is well suited for multi-class classification and provides a robust measure for comparing model's output predicted probability distribution with the ground truth for the corresponding SS8 classification. Moreover, cross entropy loss penalizes high-confidence incorrect prediction, enhancing the reliability and precision of our ensemble outputs. A definition is provided below, where $n = 8$ is the number of output 8 state secondary structure classes, y_i is the ground truth label, and p_i is the probability distribution for the ith class.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{8} y_i log(p_i)$$

5.2.2 Loss for Regression Tasks

Similarly, we adapt the loss function for regression utilised in the baseline, L1-Loss. Specifically L1 loss is beneficial for determining the numerical values of ASA, HSE and protein backbone angles as it is robust to outliers and promotes stable convergence. For the definition below, y_i is the ground truth value and $\hat{y_i}$ is the output prediction from the model

$$\mathcal{L}_{L1} = -\sum_{i=1}^{n} |y_i - \hat{y_i}|$$

5.2.3 Masking Loss

Datasets utilised for protein structure prediction often contain unlabeled or partially labelled data, as a consequence of several challenges inherent to the field of protenomics. Experimental methods utilised to determine protein structures (e.g. X-ray crystallography, cryo-electron microscopy) may incomplete data due to limitations in resolution or difficulties in resolving certain regions in a protein [3]. These unresolved regions lead to gaps in the structural data, resulting in missing labels for certain amino acids. Additionally discrepancies in experimental techniques or errors in data annotation can also contribute to missing labels.

To ensure the learning is not influenced by missing amino acid data, the loss function utilised in the Baseline will be adapted to mask out amino acids with missing label data. By masking out the amino acids with missing labels, we ensure that the model focuses only on the reliable, well-characterized portions of the protein sequences during training. This approach prevents the model from learning from potentially erroneous data, thereby enhancing the accuracy and reliability of the structure predictions it generates. Furthermore, it aligns the model's training process more closely with the reality of available protein structure data, which often includes these gaps and uncertainties.

6 Evaluation Metrics

6.1 Metrics for Classification

6.1.1 Prediction Accuracy

The model's performance in terms of residue-level class predictions for SS8 configurations is measured in terms of accuracy.

$$Accuracy = \frac{Total\ Correct\ Class\ Predictions}{Total\ Samples}$$

6.2 Metrics for Regression

6.2.1 Mean Absolute Error

Mean absolute error (MAE) is used as a performance metric for the protein backbone angles ψ, ϕ, θ, τ, as it appropriately handles the cyclical nature of the data.

$$MAE = \frac{1}{n} \sum i = 1n|y_i - \hat{y_i}|$$

where n is the protein sequence length, y_i is the predicted value, and $\hat{y}_i$ is the actual value

6.2.2 Pearson's Correlation Coefficient

Pearson's correlation coefficient (r) is utilised to evaluate linear correlations between model predictions and target values for ASA, HSE-up, HSE-down and contact number.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

where x_i and y_i are the individual samples of the two variables, and $\overline{x}$ and $\overline{x}$ are the means of the x and y variables, respectively.

7 Implemented Extensions and Experiments

We conducted a series of experiments to assess the performance of our baseline model with our ProteinBERT module extension. Following the approach detailed in SPOT-1D-LM [17], experimentation was performed using the classification ensemble only to reduce computation time and resource consumption. Evaluation was based on the accuracy metric for SS8 classification defined in Section 6.1.1, as well as stability and efficiency of a model's convergence for the validation dataset. Additionally, training batch size was kept at a constant value of 10. Experimentation was directed towards three primary objectives: determining a masking procedure for loss calculation, exploring architecture variations for the SPOT-1D-Single Ensemble and assessing the effectiveness of different regularisation methods.

7.1 Integration of Masking for Loss Calculations

As outlined in Section 5.2.3, to account for unlabelled or partially labelled points in our dataset during loss calculation we selected to implement masking. Our initial implementation strategy involved masking out amino acids with missing label data during loss calculation (labelled "Post Masking" in Table 7.1), to prevent these values from to the computed value. However, as depicted in Table 7.1, when implementing this methodology for classification tasks we observed low values for both training (24.53%) and validation (31.41%), due to vanishing gradients accrued when passing the masked values through the cross entropy loss function.

To address this issue, we experimented with an alternative technique we have termed "pre-masking". This involved masking missing labels during preprocessing and embedding, before the data is fed through the model. We selected to explore this alternate method of masking, as the MLM learning procedure used by both ESM-1b and ProteinBERT similarly, masks out input data during a preprocessing stage to enhance the efficiency of representational learning. As demonstrated in Table 7.1, we found adopting this strategy greatly increased the prediction for out model performance. Both the "post-masking" and "pre-masking" trials summarised in Table 7.1, utilised the optimizer Adam, an initial learning rate of 1e-3 and a Cosine Annealing learning rate scheduler.

Masking Type	**Training Accuracy (%)**	**Validation Accuracy (%)**	**Convergence Epoch**
Post-Masking	31.41	24.53	16
Pre-Masking	94.71	77.16	5

Table 2: Accuracy and Convergence Epoch Results for Masking Objectives

7.2 Exploration of Alternate Ensemble Architectures

During initial iterations of training the baseline model with our ProteinBERT extension, we observed characteristics indicative of overfitting. Specifically, consistently across ablations after only a small number of epochs, there was a marked increase from validation loss, causing divergence from the training loss, as visualised in Figure 10. We hypothesised with a reduced dataset due to the masking, the complexity of the ensemble model resulting in overfitting. Specifically we theorised the large number of parameters present in the BiLSTM layers contribute to a large learning capacity that poorly correlates to the task. To investigate this notion, we replaced the entire ensemble with just the MSResNet architecture (Model 2), multiscale ResNet used in SPOT-1D-Single's Model 2 (MSResNet) and performed further tuning. Utilising MSResNet alone resulting a two-fold increase in the number of epochs performed before a marked divergence of validation loss. Despite this across all trials an increase in validation loss was observed, suggesting the need for the integration of explicit regularisation measures.

Model	**Scheduler Parameters**	**Initial LR**	**Training Accuracy (%)**	**Validation Accuracy (%)**	**Divergence Epoch**
Baseline Ensemble	factor=0.5 patience=5	1.00e-03	92.72	76.90	4
Baseline Ensemble	factor=0.5 patience=5	1.00e-04	85.80	76.92	5
Baseline Ensemble	factor=0.5 patience=5	2.00e-04	85.80	76.92	3
MSResNet Only	factor=0.5 patience=2	1.00e-04	85.90	75.22	11
MSResNet Only	factor=0.5 patience=1 threshold=1e-4	1.00e-03	84.02	75.89	9
MSResNet Only	factor=0.5 patience=1 threshold=1e-3	1.00e-03	86.48	75.00	13

Table 3: Hyperparameter Tuning for Ensemble Architecture

7.3 Addition of Regularization Techniques

To explore solutions to the overfitting observed during training, we ran an ablation study evaluating a number of different regularization methods with different optimizers and architectures. Adding weight decay to our optimizer demonstrated a significant reduction in overfitting. Integrating a weight decay term of 1e-5 increased the divergence epoch for the "MSResNet Only" Architecture from the maximum value of 13 achieved without regularization to 20 (Table 3 & 4). Furthermore, number of epochs until validation loss increased was lifted to 43 when a weight decay value of 1e-4 was used.

Additionally, we tested a variation of the MSResNet architecture that added an 5 Resnet blocks with a convolutional filter size of 512 to the end of all three stacks. The kernel sizes were additionally changed from k=7 to k=3 for the first stack, k=9 to k=5 for the second stack and k=15 to k=7 for the third stack. When this model was trained alone with a Adam and weight decay term of 1e-4, similar results to the original MsResNet model were observed. To further asses the validity of this variation we tested replacing the MsResNet in the SPOT-1D-Single ensemble with the Adapted ResNet. The MsResLSTM model in the ensemble was also adapted by appending the 5 Blocks with a filter size = 512 to the end of the two parallel ResNet Stacks. Although weight decay was added, similarly to the tests run in Section 7.2, the increased complexity of the ensemble model bolstered the effect of overfitting causing the validation and training losses to diverge at the 5th epoch. We initially attempted to resolve this by increasing the value of the weight decay from 1e-4 to 1-e2, however no marked change in results was observed, as demonstrated in Table 4. This prompted us to integrate L1 regularization to prevent overfitting. Defined Below, L1 Regularization enhances model generalization by adding a penalty value equal to the absolute magnitude of coefficients, promoting sparsity.

$$\mathcal{L}_{L1reg} = Loss + \lambda \sum_{i=1}^{N} |w_i|$$

Model	**Optimizer**	**Training Accuracy (%)**	**Validation Accuracy (%)**	**Convergence / Divergence Epoch**
MsResNet + Weight Decay (1e-5)	Adam	81.03	74.99	20
MsResNet + Weight Decay (1e-4)	Adam	77.26	74.71	43
MsResNet + Weight Decay (1e-4)	SGD	82.54	74.75	22
MsResNet Variation + Weight Decay (1e-4)	Adam	74.90	72.70	37
Ensemble Variation + Weight Decay (1e-4)	AdamW	84.17	77.36	5
Ensemble Variation + Weight Decay (1e-2)	AdamW	84.13	77.54	5
Ensemble Variation + Weight Decay (1e-2) + L1 Regularization	AdamW	74.47	73.41	50

Table 4: Hyperparameter Tuning with Regularization

8 Results

Figure 10 captures the training behavior across the last 12 experiments, utilizing the same average percentage accuracy evaluation as our baseline, achieved by concatenating all proteins. Figures 10(a) & (c) depict validation and training accuracy, respectively, while Figure 10 (b) & (d) present validation and training loss. Lastly, Figure 10 (e) & (f) highlight the accuracy and loss of our best-performing model during its training phase.

Following training, we evaluated our top-performing model using unseen test sets: TEST2018, TEST2020-HQ, Neff1-2020, and CASP12-FM through CASP14-FM. Figure 11 displays a comparative analysis of SS8 label prediction accuracy across these six test sets for each model, including SPOT-1D-LM (baseline), our SPOT-1D-LM implementation, and our extended version. Additionally, Figure 12 illustrates the congruence of predicted SS8 labels with ground truth for each test set in a confusion matrix, highlighting accuracy along the diagonal. For baseline model comparison, Figure 13 is provided. Figures 14 to 19 further contrast the baseline's SS8 prediction accuracy for the 20 standard amino acids with that of our extended model across all test sets.

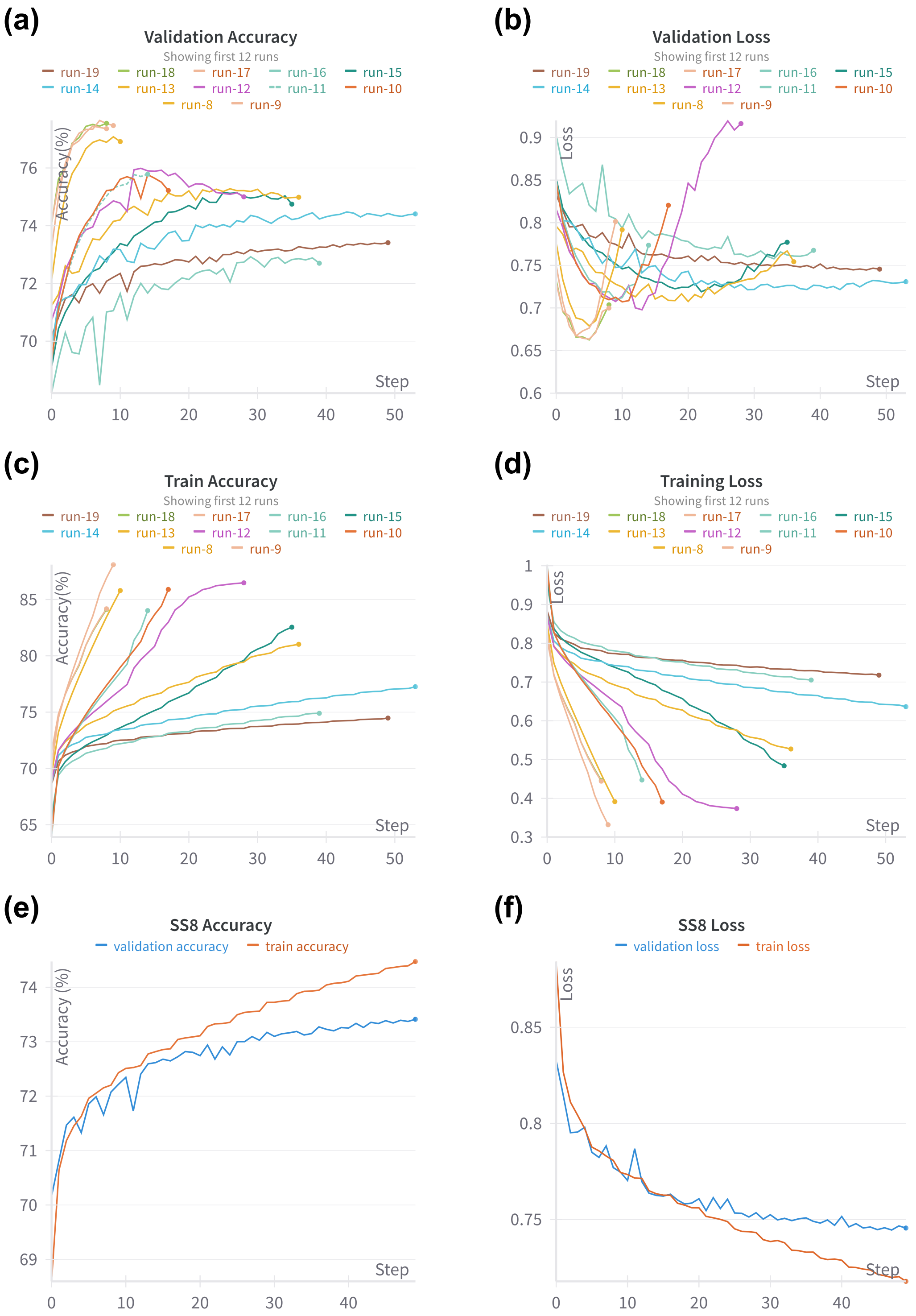


Figure 10: Overview of the Extended Architecture's Performance During Training; (a) Validation accuracy across 12 selected experiments. (b) Validation loss trends for the same 12 experiments. (c) Training accuracy for the same 12 selected experiments. (d) Training loss trends for the same 12 experiments. Detailed view of run-19 among the 12 experiments: (e) Training vs. validation accuracy. (b) Training vs. validation loss.

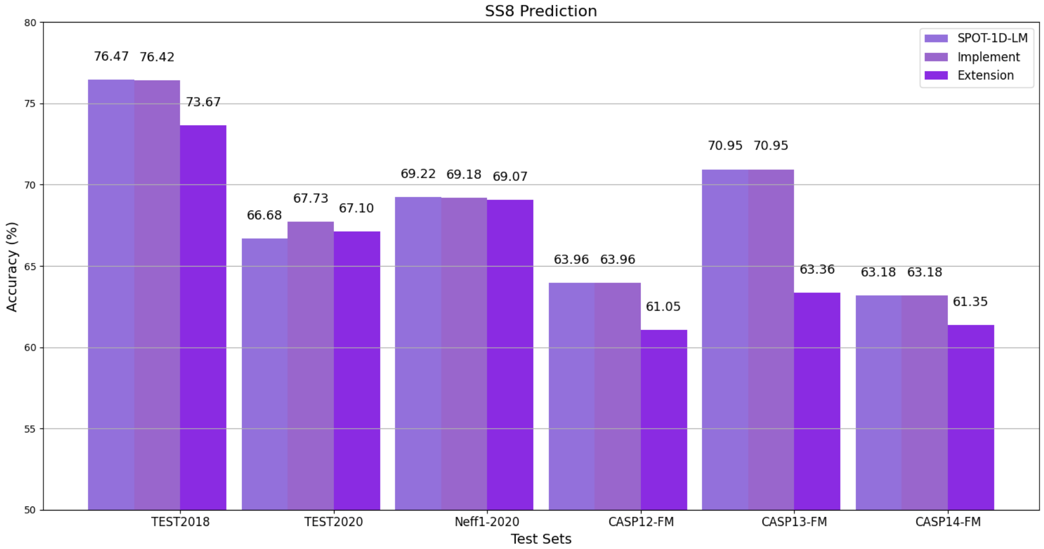


Figure 11: Secondary structure prediction accuracy across six test sets. This comparison includes SPOT-1D-LM (baseline), our implementation of SPOT-1D-LM, and our extended version of SPOT-1D-LM. Test sets include TEST2018, TEST2020, Neff1-2020, CASP12-FM, CASP13-FM, and CASP14-FM, focusing on eight-state (SS8) secondary structure predictions.

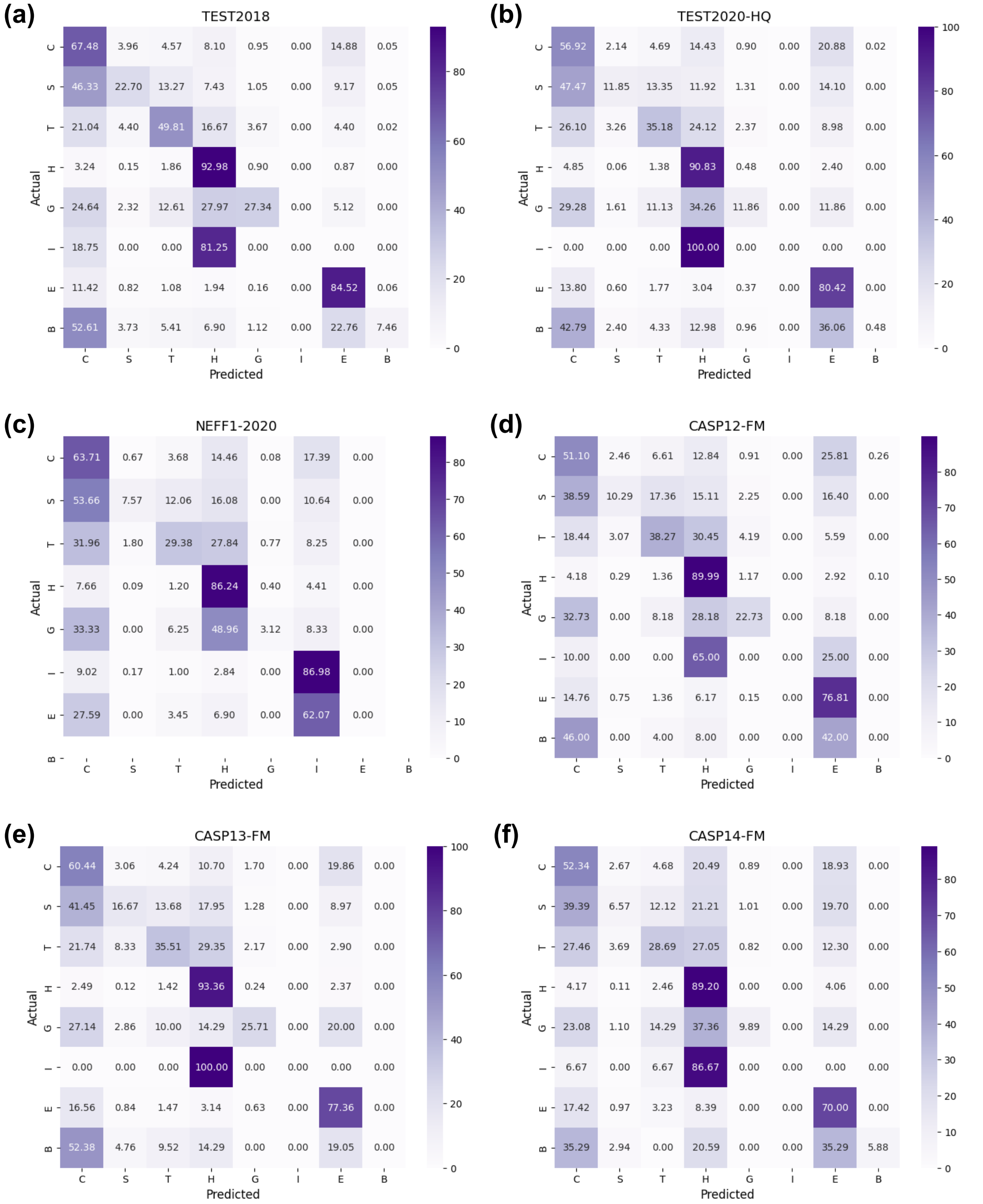
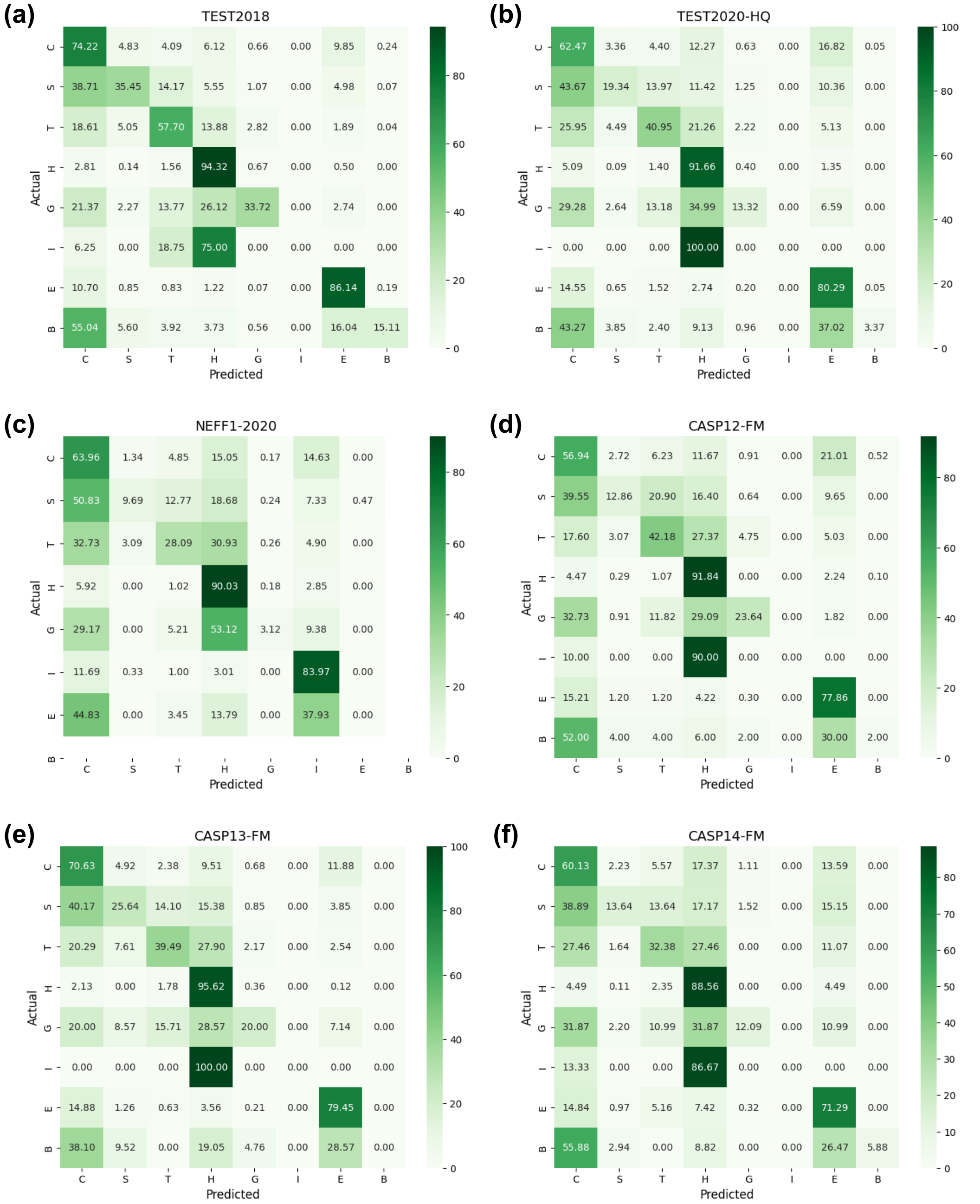


Figure 12: Confusion Matrix showcasing the model's performance, with each cell illustrating the percentage of actual versus predicted classifications. The diagonal entries represent accurate predictions highlighted by a dark purple hue, while off-diagonal entries indicate misclassifications highlighted by a light purple hue.



(a) TEST2018

	C	S	T	H	G	I	E	B
C	74.22	4.83	4.09	6.12	0.66	0.00	9.85	0.24
S	38.71	35.45	14.17	5.55	1.07	0.00	4.98	0.07
T	18.61	5.05	57.70	13.88	2.82	0.00	1.89	0.04
H	2.81	0.14	1.56	94.32	0.67	0.00	0.50	0.00
G	21.37	2.27	13.77	26.12	33.72	0.00	2.74	0.00
I	6.25	0.00	18.75	75.00	0.00	0.00	0.00	0.00
E	10.70	0.85	0.83	1.22	0.07	0.00	86.14	0.19
B	55.04	5.60	3.92	3.73	0.56	0.00	16.04	15.11

(b) TEST2020-HQ

	C	S	T	H	G	I	E	B
C	62.47	3.36	4.40	12.27	0.63	0.00	16.82	0.05
S	43.67	19.34	13.97	11.42	1.25	0.00	10.36	0.00
T	25.95	4.49	40.95	21.26	2.22	0.00	5.13	0.00
H	5.09	0.09	1.40	91.66	0.40	0.00	1.35	0.00
G	29.28	2.64	13.18	34.99	13.32	0.00	6.59	0.00
I	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
E	14.55	0.65	1.52	2.74	0.20	0.00	80.29	0.05
B	43.27	3.85	2.40	9.13	0.96	0.00	37.02	3.37

(c) NEFF1-2020

	C	S	T	H	G	I	E	B
C	63.96	1.34	4.85	15.05	0.17	14.63	0.00	
S	50.83	9.69	12.77	18.68	0.24	7.33	0.47	
T	32.73	3.09	28.09	30.93	0.26	4.90	0.00	
H	5.92	0.00	1.02	90.03	0.18	2.85	0.00	
G	29.17	0.00	5.21	53.12	3.12	9.38	0.00	
I	11.69	0.33	1.00	3.01	0.00	83.97	0.00	
E	44.83	0.00	3.45	13.79	0.00	37.93	0.00	
B								

(d) CASP12-FM

	C	S	T	H	G	I	E	B
C	56.94	2.72	6.23	11.67	0.91	0.00	21.01	0.52
S	39.55	12.86	20.90	16.40	0.64	0.00	9.65	0.00
T	17.60	3.07	42.18	27.37	4.75	0.00	5.03	0.00
H	4.47	0.29	1.07	91.84	0.00	0.00	2.24	0.10
G	32.73	0.91	11.82	29.09	23.64	0.00	1.82	0.00
I	10.00	0.00	0.00	90.00	0.00	0.00	0.00	0.00
E	15.21	1.20	1.20	4.22	0.30	0.00	77.86	0.00
B	52.00	4.00	4.00	6.00	2.00	0.00	30.00	2.00

(e) CASP13-FM

	C	S	T	H	G	I	E	B
C	70.63	4.92	2.38	9.51	0.68	0.00	11.88	0.00
S	40.17	25.64	14.10	15.38	0.85	0.00	3.85	0.00
T	20.29	7.61	39.49	27.90	2.17	0.00	2.54	0.00
H	2.13	0.00	1.78	95.62	0.36	0.00	0.12	0.00
G	20.00	8.57	15.71	28.57	20.00	0.00	7.14	0.00
I	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
E	14.88	1.26	0.63	3.56	0.21	0.00	79.45	0.00
B	38.10	9.52	0.00	19.05	4.76	0.00	28.57	0.00

(f) CASP14-FM

	C	S	T	H	G	I	E	B
C	60.13	2.23	5.57	17.37	1.11	0.00	13.59	0.00
S	38.89	13.64	13.64	17.17	1.52	0.00	15.15	0.00
T	27.46	1.64	32.38	27.46	0.00	0.00	11.07	0.00
H	4.49	0.11	2.35	88.56	0.00	0.00	4.49	0.00
G	31.87	2.20	10.99	31.87	12.09	0.00	10.99	0.00
I	13.33	0.00	0.00	86.67	0.00	0.00	0.00	0.00
E	14.84	0.97	5.16	7.42	0.32	0.00	71.29	0.00
B	55.88	2.94	0.00	8.82	0.00	0.00	26.47	5.88

Figure 13: Confusion Matrix showcasing the SPOT-1D-LM (baseline) performance, with each cell illustrating the percentage of actual versus predicted classifications. The diagonal entries represent accurate predictions highlighted by a dark green hue, while off-diagonal entries indicate misclassifications highlighted by a light green hue.

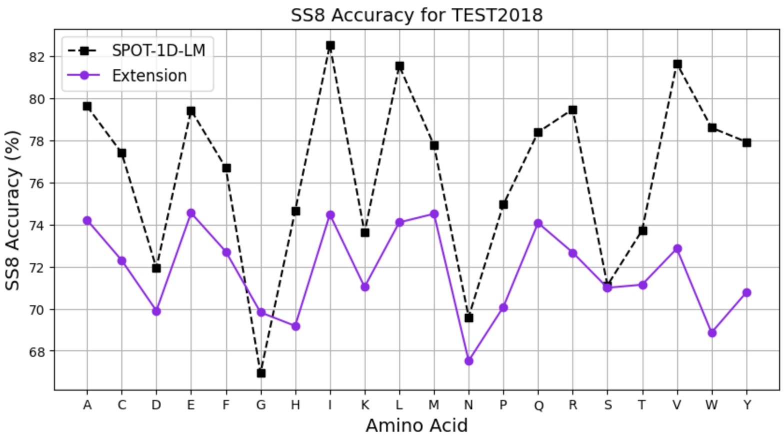


Figure 14: The accuracy of eight-state secondary structure prediction per the 20 standard amino acids for the TEST2018 (250 proteins) set given by SPOT-1D-LM and our extended model.

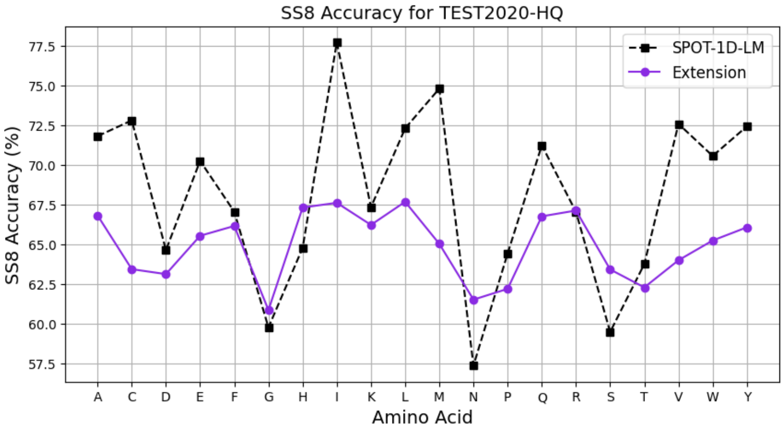


Figure 15: The accuracy of eight-state secondary structure prediction per the 20 standard amino acids for the TEST2020-HQ (241 proteins) set given by SPOT-1D-LM and our extended model.

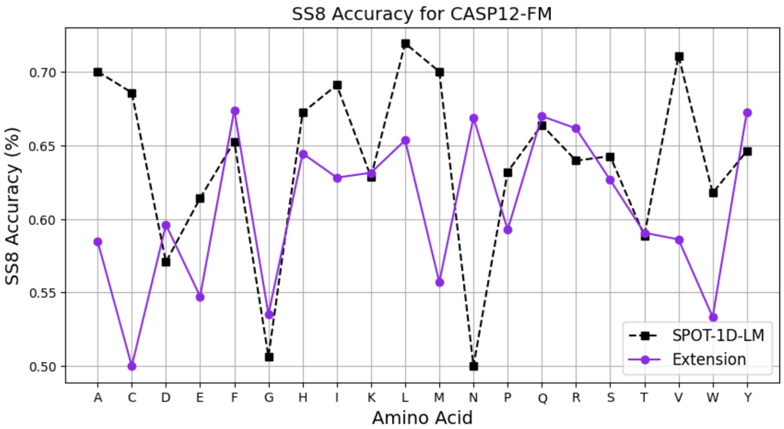


Figure 16: The accuracy of eight-state secondary structure prediction per the 20 standard amino acids for the CASP12-FM (22 proteins) set given by SPOT-1D-LM and our extended model.

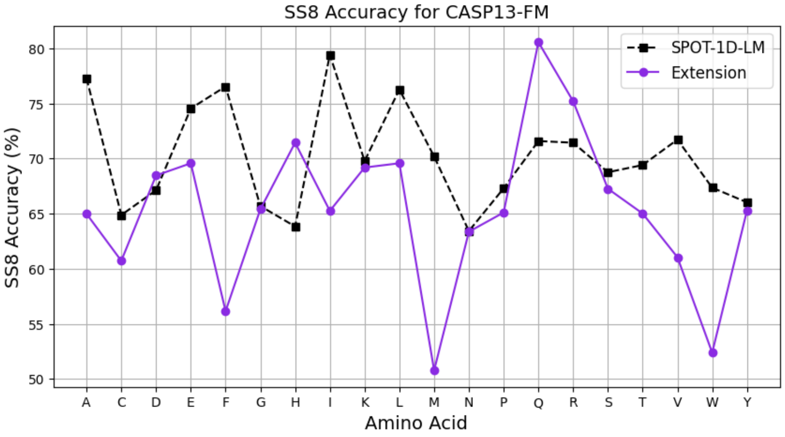


Figure 17: The accuracy of eight-state secondary structure prediction per the 20 standard amino acids for the CASP13-FM (17 proteins) set given by SPOT-1D-LM and our extended model.

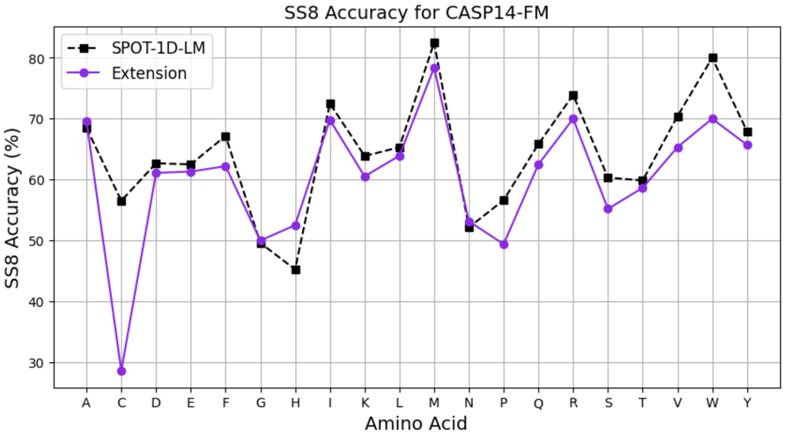


Figure 18: The accuracy of eight-state secondary structure prediction per the 20 standard amino acids for the CASP14-FM (15 proteins) set given by SPOT-1D-LM and our extended model.

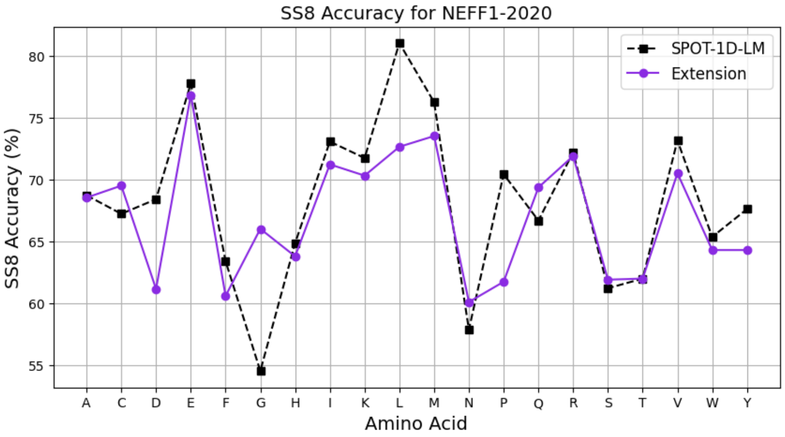


Figure 19: The accuracy of eight-state secondary structure prediction per the 20 standard amino acids for the Neff1-2020 (46 proteins) set given by SPOT-1D-LM and our extended model.

9 Discussion

Our extended model, with approximately 160M parameters, faced challenges with overfitting, evident from increasing training accuracy but decreasing validation accuracy and rising validation loss, as illustrated in Figure 10(a-d). Initially, we addressed this in our ms-resnet ensemble (runs 10 to 15 in Figure 10) by enhancing weight decay using the Adam optimizer. This change delayed the onset of validation loss increase from epoch 3 to epoch 43 (run 14 in Figure 10). Applying similar weight decay to the entire ensemble architecture resulted in validation loss starting to increase at epoch 5 (run 17 in Figure 10). In our final attempt to combat overfitting, we introduced an L1 coefficient (run 19 in Figure 10). This proved effective, as depicted in Figure 10(e-f), where validation loss stabilized by epoch 50. Despite this success, time constraints limited further training duration.

Figure 11 highlights the varied performance of our extended model, labeled "Extension" when compared to the original SPOT-1D-LM and our own implementation of it, referred to as "Implement." Figure 11 confirms the accuracy of our replication of the original SPOT-1D-LM, showcasing that our implementation closely matches the published results of the baseline model. Notably, in the TEST2018 set, the Extension model exhibits a marginal drop in accuracy against the original SPOT-1D-LM. While the Extension model surpasses TEST2020 our Implement, the Extension model for both TEST2020 and Neff1-2020 sets also delivers comparable results to SPOT-1D-LM, suggesting that our enhancements may be capturing some aspects of protein structure prediction more effectively. However, in the rigorous CASP12-FM, CASP13-FM, and CASP14-FM challenges, Extension's performance drops below both SPOT-1D-LM and Implement. This indicates that while our model has made strides in certain areas, it may require further refinement to consistently outperform the baseline in more complex scenarios. Overall, the Extension model shows promise, yet it highlights the complexity of protein structure prediction and the need for continued optimization.

Figures 12 and 13 offer a side-by-side performance comparison of our Extension model and the SPOT-1D-LM model, respectively, against ground truth labels across the same test sets. An analysis of Figures 12 and 13 reveals that both models exhibit strong predictive abilities for SS8 labels, particularly for irregular (C), beta turn (T), alpha helix (H), and beta strand (E). A common misclassification occurs with pi helix (I) being predicted as alpha helix (H) and beta bridge (B) as irregular (C) by both models. Notably, our Extension model has a tendency to confuse beta bridge (B) with beta strand (E) more frequently compared to SPOT-1D-LM. Despite this, the overall accuracy of SPOT-1D-LM does surpass that of our Extension model, which nevertheless delivers competitive results.

Figures 14 to 19 assess the prediction accuracy for the standard 20 amino acids, contrasting SPOT-1D-LM with our Extension model. For TEST2018 (Figure 14), our Extension model trails SPOT-1D-LM slightly but shows a 4.13% higher accuracy for Glycine (G) and matches SPOT-1D-LM for Serine (S). SPOT-1D-LM's highest advantage over our model is 14.14% for Cysteine (C). In TEST2020 (Figure 15), our Extension model outperforms SPOT-1D-LM for Glycine (G) by 1.80%, Histidine (H) by 3.79%, Asparagine (N) by 6.76%, Serine (S) by 6.21%, with Arginine (N) being nearly equal. SPOT-1D-LM's biggest lead is 14.96% for Cysteine (C). For the CASP12-FM and CASP13-FM datasets, SPOT-1D-LM's accuracy is generally higher. CASP12-FM (Figure 16) sees our model ahead for Aspartic acid (D) by 4.24%, Phenylalanine (F) by 3.13%, Glycine (G) by 5.37%, Asparagine (N) by an impressive 25.23%, Arginine (R) by 3.31%, and Tyrosine (Y) by 3.95%, while nearly matching for Lysine (K), Glutamine (Q), and Threonine (T). Here, SPOT-1D-LM's largest lead is 37.47% for Cysteine (C).

In CASP13-FM (Figure 17), our model shows better accuracy for Aspartic acid (D) by 1.96%, Glycine (G) by 5.37%, Histidine (H) by 10.64%, Glutamine (Q) by 11.14%, and Arginine (R) by 5.00%, with approximate parity for Lysine and Asparagine. SPOT-1D-LM's maximum advantage rises to 38.22% for Methionine/Start codon (M). The difference in SS8 accuracy for our Extension model for CASP14-FM and Neff1-2020 is much closer. For CASP14-FM (Figure 18), our model surpasses SPOT-1D-LM's accuracy for Alanine (A) by 1.56%, Histidine (H) by 13.83%, and Asparagine (N) by 1.90%, with Glycine (G) roughly equal. SPOT-1D-LM's greatest lead is a substantial 97.83% for Cysteine (C). In Neff1-2020 (Figure 19), our Extension model exceeds SPOT-1D-LM for Cysteine (C) by 3.28%, Glycine (G) by 17.41%, Asparagine (N) by 3.72%, Glutamine (Q) by 3.86%, and Serine (S) by 1.12%. The improvements, especially for Glycine, are substantial.

Across various test sets, there is a nuanced performance landscape for our Extension model in comparison to SPOT-1D-LM. Glycine (G) emerges as a consistent strong suit for the Extension model, with a standout 17.41% higher accuracy in the Neff1-2020 test set. When predicting Asparagine (N), the Extension model notably surpasses SPOT-1D-LM, especially in the CASP12-FM set, where it achieves a remarkable 25.23% greater accuracy. For Serine (S), the Extension model demonstrates competitive accuracy, equaling or exceeding that of SPOT-1D-LM across most test sets. Conversely, SPOT-1D-LM maintains a substantial lead in predicting Cysteine (C), particularly evident in the TEST2018 and CASP14-FM sets, where it significantly outperforms the Extension model. The Extension model's performance displays variability across datasets. While it shows strong results in TEST2018 and TEST2020, SPOT-1D-LM outshines in the more challenging CASP12-FM and CASP13-FM datasets. However, this performance gap narrows in the CASP14-FM and Neff1-2020 sets, indicating the Extension model's potential for refinement and improvement. Histidine (H), Arginine (R), and Glutamine (Q) are among other amino acids where the Extension model frequently shows improved accuracy over SPOT-1D-LM. These trends suggest that with further training, the Extension model holds the promise of not only closing the current performance gap but potentially surpassing SPOT-1D-LM, particularly given its predictive strengths with amino acids like Glycine and Asparagine.

10 Future Works

We encountered several challenges and constraints, particularly relating to extensive training durations and persistent overfitting, which limited our ability to conduct comprehensive ablation studies aimed at surpassing the SPOT-1D-LM baseline model's performance. Given more time, a key area of interest for us would have been to experiment with substituting our current ResNet stacks with the advanced version proposed by Duta et al. [56]. This improved ResNet model has already shown promising results in the SPOT-Contact-Single [57] by Singh et al., the same team behind SPOT-1D-LM. In their work on SPOT-Contact-Single, they successfully integrated outer-concatenated one-hot encodings with pre-trained embeddings from both SPOT-1D-Single and ESM-1b, utilizing these as input for the enhanced ResNet blocks, ultimately achieving the same outputs as seen in SPOT-1D-LM. Their approach yielded a notable increase in F1-score for protein structures with Neff<1, which indicates the absence of homologs. The improved ResNet model is known for its superior performance in comparison to standard ResNet architectures, particularly in applications involving image and video data, suggesting a potential increase in accuracy for our Extension model in protein structure prediction tasks.

Our research faced time constraints that prevented us from training a regression model to predict key protein properties such as Solvent Accessible Surface Area (ASA), Half-Sphere Exposure (HSE), and protein backbone torsion angles (ϕ, ψ, θ, and τ). In the development of SPOT-1D-LM, the authors also encountered challenges with lengthy training times. Consequently, they conducted a focused hyperparameter ablation study specifically for prediction tasks and then applied the same parameters to the regression task. Moving forward, after exploring the integration of the improved ResNet blocks, we plan to apply these same hyperparameters from the classification tasks to the regression model. This approach will streamline our efforts to optimize the model's predictive capabilities in both classification and regression tasks for protein structure analysis.

11 Conclusion

This study explored protein structure prediction using deep learning, focusing on enhancing accuracy with the SPOT-1D-LM framework and the ProteinBERT language model. We tackled overfitting through weight decay and L1 regularization, improving performance and validation loss stability. Our model excelled in predicting Glycine and Asparagine and showed promise in difficult tests like Neff1-2020, though SPOT-1D-LM outperformed in some areas. This highlights the model's potential for further improvement. We identified the ongoing need for optimization in protein prediction models. Future work could include advanced ResNet blocks for better accuracy, especially for proteins without homologs, and extending the model to regression tasks for a more comprehensive protein analysis. Overall, this study underlines the importance of continued research in bioinformatics. Our results pave the way for more sophisticated models, enhancing our understanding of the complex protein structures vital to biological life.

12 Division of Work

- Bolutito Babatunde (Tito)
  - Ran Experiments/Ablations for the Extension model
  - Ran analysis on SPOT-1D-LM vs. Extension model
  - Wrote Dataset, Results, Discussion, Future Works, and Conclusion report sections
  - Generated all report Figures
  - Supported writing of script for presentation
  - Generated Results Figures for presentation
  - Supported literature review
- Madeline Davis (Madi)
  - Performed literature review
  - Produced Slides and script for presentation
  - Generated Figures for Presentation & Report
  - Supported ProteinBERT embeddings integration
  - Wrote Abstract to Implemented Extensions/Experiments report sections
  - Wrote supporting code included as Jupyter Notebook in referenced repository (note - did not have access to sufficient computational resources to perform ablations)

13 References

References

[1] Christian B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973. Publisher: American Association for the Advancement of Science.

[2] Robin Pearce and Yang Zhang. Toward the solution of the protein structure prediction problem. *The Journal of Biological Chemistry*, 297(1):100870, June 2021.

[3] J. P. Glusker. X-Ray Crystallography of Proteins. In *Methods of Biochemical Analysis*, pages 1–71. John Wiley & Sons, Ltd, 1993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470110584.ch1.

[4] John Cavanagh, Arthur G. Palmer III, Nicholas J. Skelton, Wayne J. Fairbrother, Mark Rance, Arthur G. Palmer III, John Cavanagh, Nicholas J. Skelton, Wayne J. Fairbrother, and Mark Rance. *Protein NMR Spectroscopy: Principles and Practice*. Elsevier Science & Technology, San Diego, UNITED STATES, 2006.

[5] K. Ravi Acharya and Matthew D. Lloyd. The advantages and limitations of protein crystal structures. *Trends in Pharmacological Sciences*, 26(1):10–14, January 2005.

[6] Arne Elofsson. Progress at protein structure prediction, as seen in CASP15. *Current Opinion in Structural Biology*, 80:102594, June 2023.

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

[8] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein Structure Determination using Metagenome sequence data. *Science (New York, N.Y.)*, 355(6322):294–298, January 2017.

[9] Letícia M. F. Bertoline, Angélica N. Lima, Jose E. Krieger, and Samantha K. Teixeira. Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, 3, 2023.

[10] Mehmet Akdel, Douglas E. V. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A. Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H. M. Rodrigues, Alistair S. Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Jérôme Basquin, Kresten Lindorff-Larsen, Alex Bateman, Andrey V. Kajava, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B. Ascher, Janet M. Thornton, Norman E. Davey, Amelie Stein, Arne Elofsson, Tristan I. Croll, and Pedro Beltrao. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11):1056–1067, November 2022. Number: 11 Publisher: Nature Publishing Group.

[11] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. Publisher: American Association for the Advancement of Science.

[12] Jack Hanson, Kuldip Paliwal, Thomas Litfin, Yuedong Yang, and Yaoqi Zhou. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 35(14):2403–2410, July 2019.

[13] Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjærgaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Sønderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, and Paolo Marcatili. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25674.

[14] Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Reports*, 6(1):18962, January 2016. Number: 1 Publisher: Nature Publishing Group.

[15] Gang Xu, Qinghua Wang, and Jianpeng Ma. OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics*, 36(20):5021–5026, December 2020.

[16] Rhys Heffernan, Kuldip Paliwal, James Lyons, Jaswinder Singh, Yuedong Yang, and Yaoqi Zhou. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *Journal of Computational Chemistry*, 39(26):2210–2216, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.25534.

[17] Jaspreet Singh, Kuldip Paliwal, Thomas Litfin, Jaswinder Singh, and Yaoqi Zhou. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Scientific Reports*, 12(1):7607, May 2022. Number: 1 Publisher: Nature Publishing Group.

[18] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[19] Claus A. F. Andersen, Arthur G. Palmer, Søren Brunak, and Burkhard Rost. Continuum Secondary Structure Captures Protein Flexibility. *Structure*, 10(2):175–184, February 2002.

[20] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners, December 2020. Pages: 2020.12.15.422761 Section: New Results.

[21] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007.

[22] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, April 2022.

[23] Jaspreet Singh, Thomas Litfin, Kuldip Paliwal, Jaswinder Singh, Anil Kumar Hanumanthappa, and Yaoqi 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*, 37(20):3464–3472, October 2021.

[24] Bozhen Hu, Jun Xia, Jiangbin Zheng, Cheng Tan, Yufei Huang, Yongjie Xu, and Stan Z. Li. Protein Language Models and Structure Prediction: Connection and Progression, November 2022. arXiv:2211.16742 [cs, q-bio].

[25] Shaun M. Kandathil, Andy M. Lau, and David T. Jones. Machine learning methods for predicting protein structure from single sequences. *Current Opinion in Structural Biology*, 81:102627, August 2023.

[26] Sheng Wang, Wei Li, Shiwang Liu, and Jinbo Xu. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Research*, 44(Web Server issue):W430–W435, July 2016.

[27] Krzysztof Kotowski, Tomasz Smolarczyk, Irena Roterman-Konieczna, and Katarzyna Stapor. ProteinUnet—An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. *Journal of Computational Chemistry*, 42(1):50–59, January 2021.

[28] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, June 2021.

[29] Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the Boundaries of Protein Language Models, June 2022. arXiv:2206.13517 [cs, q-bio].

[30] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Number: 7873 Publisher: Nature Publishing Group.

[31] Mihaly Varadi et al. Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nature Biotechnology*, 40(6):765–774, 2022.

[32] Liam J McGuffin et al. The psipred protein structure prediction server. *BMC Bioinformatics*, 7(1):178, 2006.

[33] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[34] Fang Wu, Lirong Wu, Dragomir Radev, Jinbo Xu, and Stan Li. Integration of pre-trained protein language models into geometric deep learning networks. *Communications Biology*, 6, 08 2023.

[35] Abel Chandra, Laura Tünnermann, Tommy Löfstedt, and Regina Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12:e82819, jan 2023.

[36] Mohammed AlQuraishi. Proteinnet: A standardized data set for machine learning of protein structure. *BMC bioinformatics*, 20(1):311–311, 2019.

[37] S.F Altschul, T.L Madden, A.A Schaffer, J.H 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*, 25(17):3389–3402, 1997.

[38] AT BRUNGER. Free r value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature (London)*, 355(6359):472–475, 1992.

[39] Joerg Schaarschmidt, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre M.J.J. Bonvin. Assessment of contact predictions in casp12: Co-evolution and deep learning coming of age. *Proteins, structure, function, and bioinformatics*, 86(S1):51–66, 2018.

[40] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins, structure, function, and bioinformatics*, 87(12):1011–1020, 2019.

[41] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins, structure, function, and bioinformatics*, 89(12):1607–1617, 2021.

[42] William R. Pearson and David J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences - PNAS*, 85(8):2444–2448, 1988.

[43] D.J Lipman and W.R Pearson. Rapid and sensitive protein similarity searches. *Science (American Association for the Advancement of Science)*, 227(4693):1435–1441, 1985.

[44] Cyrus Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248(5446):338–339, 03 1974.

[45] Rhys Heffernan, Abdollah Dehzangi, James Lyons, Kuldip Paliwal, Alok Sharma, Jihua Wang, Abdul Sattar, Yaoqi Zhou, and Yuedong Yang. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics*, 32(6):843–849, 11 2015.

[46] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13(3):289–302, Mar 1999.

[47] James Lyons, Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, Kuldip Paliwal, Abdul Sattar, Yaoqi Zhou, and Yuedong Yang. Predicting backbone c angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of computational chemistry*, 35(28):2040–2046, 2014.

[48] Tibo Duran, Bruna Minatovicz, Ryan Bellucci, Jun Bai, and Bodhisattwa Chaudhuri. Molecular dynamics modeling based investigation of the effect of freezing rate on lysozyme stability. *Pharmaceutical Research*, 39(10):2585–2596, 10 2022.

[49] Youngsuk Hong, Juyoung Song, Jaewoo Ko, et al. S-pred: protein structural property prediction using msa transformer. *Scientific Reports*, 12:13891, 2022.

[50] Alok Sharma, Artem Lysenko, Yosvany López, Abdollah Dehzangi, Ronesh Sharma, Hamendra Reddy, Abdul Sattar, and Tatsuhiko Tsunoda. Hsesumo: Sumoylation site prediction using half-sphere exposures of amino acids residues. *BMC Genomics*, 19(9):982, 2019.

[51] Ahmed Jarray, Herman Wijshoff, Jurriaan Luiken, and Wouter den Otter. Systematic approach for wettability prediction using molecular dynamics simulations. *Soft Matter*, 16:4299–4310, 04 2020.

[52] Liam J. McGuffin, Kevin Bryson, and David T. Jones. The PSIPRED protein structure prediction server . *Bioinformatics*, 16(4):404–405, 04 2000.

[53] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[54] Zeming Lin. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. PMID: 36927031.

[55] Gene Ontology Resource.

[56] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Improved residual networks for image and video recognition. *arXiv preprint arXiv:2004.04989*, 2020.

[57] Jaspreet Singh, Thomas Litfin, Jaswinder Singh, Kuldip Paliwal, and Yaoqi Zhou. Spot-contact-single: Improving single-sequence-based prediction of protein contact map using a transformer language model. *bioRxiv*, 2021.