Do Protein Transformers Have Biological Intelligence?

Fudong Lin[1], Wanrou Du[2], Jinchan Liu[3], Tarikul Milon[4], Shelby Meche[4], Wu Xu[4], Xiaoqi Qin[2], and Xu Yuan[1] (✉)

[1] University of Delaware, Newark, DE 19716, USA
{fudong,xyuan}@udel.edu
[2] Beijing University of Posts and Telecommunications, Haidian, Beijing 100876, China
{wanroudu,xiaoqiqin}@bupt.edu.cn
[3] Yale University, New Haven, CT 06520, USA
jinchan.liu@yale.edu
[4] University of Louisiana at Lafayette, Lafayette, LA 70504, USA
{tarikul-islam.milon1,wu.xu}@louisiana.edu

Abstract. Deep neural networks, particularly Transformers, have been widely adopted for predicting the functional properties of proteins. In this work, we focus on exploring whether Protein Transformers can capture biological intelligence among protein sequences. To achieve our goal, we first introduce a protein function dataset, namely *Protein-FN*, providing over 9000 protein data with meaningful labels. Second, we devise a new Transformer architecture, namely *Sequence Protein Transformers (SPT)*, for computationally efficient protein function predictions. Third, we develop a novel Explainable Artificial Intelligence (XAI) technique called *Sequence Score*, which can efficiently interpret the decision-making processes of protein models, thereby overcoming the difficulty of deciphering biological intelligence bided in Protein Transformers. Remarkably, even our smallest SPT-Tiny model, which contains only 5.4M parameters, demonstrates impressive predictive accuracy, achieving 94.3% on the Antibiotic Resistance (AR) dataset and 99.6% on the Protein-FN dataset, all accomplished by training from scratch. Besides, our Sequence Score technique helps reveal that our SPT models can discover several meaningful patterns underlying the sequence structures of protein data, with these patterns aligning closely with the domain knowledge in the biology community. We have officially released our Protein-FN dataset on Hugging Face Datasets https://huggingface.co/datasets/Protein-FN/Protein-FN. Our code is available at https://github.com/fudong03/BioIntelligence.

Keywords: Protein Transformers · Explainable AI · AI for Science.

1 Introduction

Proteins serve as the architects of life, orchestrating an extraordinary range of functions that bring vitality and complexity to the biological world. Their roles encompass everything from catalyzing critical biochemical reactions to facilitating precise cellular communication. Decoding the intricate relationship between a protein's sequence, structure, and functional properties holds the key to unraveling these life-sustaining

mysteries. This endeavor is more than a scientific pursuit; it is a profound exploration of the fundamental processes that define life itself.

Since the intricate patterns of protein sequences are analogous to the syntactic and semantic structures found in human languages, existing state-of-the-art Protein Language Models (PLMs) [55,59,52,56,38,24,6,49,32,34] harness the advanced language models [67,20] to decipher how the intricate structures of protein sequences dictate their functional properties. However, these methods require pre-training on millions or even billions of protein sequences for satisfactory performance. The excessive computational demands of self-supervised pre-training render PLMs unattainable for resource-constrained research groups.

In this work, we develop a computationally efficient Transformer architecture, namely Sequence Protein Transformer (SPT), for unraveling the complex interplay between a protein's sequence and its functional property, by leveraging the Transformer architectures in the vision domain [21,66,45,30]. Specifically, our work focuses on answering the following research question: *Can our Protein Transformers learn biological intelligence underlined in protein sequences?*

To help answer this question, we first introduce a new protein function dataset called *Protein-FN*, offering over 9000 protein data, with each containing the protein's 1D amino acid sequences, 3D structures, as well as its functional properties annotated by biological experts of our team. Second, different from PLMs, where the protein sequence is naturally encoded by the letter abbreviation of amino acids (*e.g.*, with letter "A" for representing the amino acid "Alanine"), how to encode the protein data for new Transformer architecture remains unexplored, making the applications of existing Vision Transformers (ViT) variants [21,66,45,70,4,25,9,7,44,42,30,41,43,16] for protein function predictions technically infeasible. We develop the Sequence Protein Transformer (SPT) model to address this issue. Featuring an innovative embedding mechanism tailored for protein data, our SPT model excels in predicting the functional properties of proteins. Remarkably, it can achieve a superior prediction performance without relying on computationally extensive self-supervised pre-training. Third, Explainable Artificial Intelligence (XAI) techniques [61,28,27,63,8,36,23], especially those Transformer-specific solutions [1,68,9,12,11,71], can offer insightful perspectives into the decision-making mechanisms of deep neural networks (DNNs), making them suitable for deciphering biological intelligence resided in Protein Transformers. However, current XAI approaches face significant challenges when handling protein sequences that vary in the number of amino acids. They either fail to accommodate these variances or incur substantial computational burdens, *e.g.*, $\mathcal{O}(L^2 \cdot P^4)$ for Attention Flow [1], where L and P represent the model depth and the protein sequence length, respectively. Consequently, these methods prove impractical for analyzing the biological insight within protein sequences. In this study, we introduce the Sequence Score, a novel gradient-based XAI approach, tailored specifically to manage protein sequences of varying amino acid lengths. It advances existing Transformer-specific XAI solutions through its computational efficiency, which *scales linearly* with protein sequence length. This advancement facilitates a more efficient and effective interpretation of the biological intelligence bided in Protein Transformers.

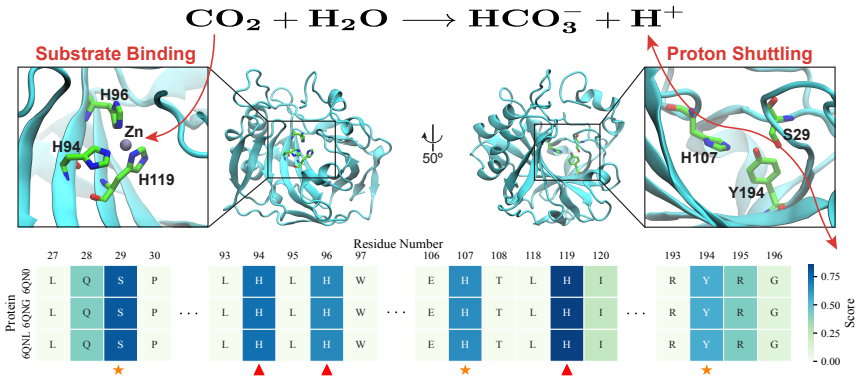


Fig. 1: Illustration of two key conserved motifs, *i.e.*, "His94-His96-His119" and "Ser29-His107-Tyr194", for the Carbonic Anhydrases, a vital enzyme class. Here, "His" (Histidine), "Ser" (Serine), and "Tyr" (Tyrosine) are the amino acids forming these motifs, abbreviated as "H", "S", and "Y", respectively. The residual numbers signify the positions of amino acids within the protein sequence. In the upper section, two figures, along with a corresponding equation, utilize the protein "6QN0" as an example to visualize the three-dimensional structures of the motifs and elucidate their roles in catalyzing the reaction. A heatmap in the lower section displays the importance scores generated by our approach, with the motifs "His94-His96-His119" and the "Ser29-His107-Tyr194" distinctly marked with red triangles and orange stars, respectively.

Our Sequence Score technique has successfully identified several meaningful biological patterns, showcasing its prowess in revealing the biological intelligence ingrained in Protein Transformers. For example, Carbonic Anhydrases (CAs) are a class of enzymes vital to many biological processes, such as respiration and acid-base balance in organisms. Central to the functionality of these enzymes are two highly conserved motifs: "His94-His96-His119" and "Ser29-His107-Tyr194". The "His94-His96-His119" pattern, known as the zinc-binding motif, plays a critical role in the catalytic activity of CAs. These histidine residues coordinate with a zinc ion, which is essential for the hydration of carbon dioxide, to realize a primary reaction catalyzed by CAs. This interaction is fundamental to maintain the enzyme's active site structure and its catalytic efficiency. On the other hand, the "Ser29-His107-Tyr194" motif is crucial for substrate specificity and the orientation of water molecules in the active site. This motif contributes to the positioning and polarization of water molecules, facilitating the transfer of protons and hence supporting the enzyme's catalytic mechanism. By applying our Sequence Score technique to interpret the prediction results of our SPT models, we discover that our models can capture the importance of the "His94-His96-His119" and the "Ser29-His107-Tyr194" motifs for the functional properties of CAs. Figure 1 (see its heatmap) shows the importance scores for the three proteins of the CA class, where our Sequence Score technique assigns very high scores on both the "His94-His96-His119" (marked by red triangles) and the "Ser29-His107-Tyr194" (marked by orange stars)

patterns. This exhibits that our SPT models have captured biological intelligence underlined in the protein sequence for predicting the functional properties of proteins.

2 Related Work

Protein Language Models. Protein Language Models (PLMs) have demonstrated remarkable performance across a spectrum of biological tasks. AlphaFold [38] is a well-known study in PLMs, and it uses multiple sequence alignment to predict the 3D structure of proteins. Recently, PLMs are also developed for predicting the functional properties of proteins, including single sequence-based methods, *i.e.*, TAPE [55], ESM-1b [59], and ESM-1v [52], multiple sequence alignment-based approaches, *i.e.*, MSA Transformer [56], and others [26,58,2,22,6,57,24,49,32,34]. Despite their effectiveness, PLMs require extensive pre-training on millions of protein data for satisfactory performance, making them computationally inaccessible for resource-limited groups. Different from prior studies, our SPT model can achieve superb performance in protein function predictions by training from scratch. Therefore, it advances previous PLMs by significantly reducing the computational overhead. We hope that the exceptionally computational efficiency of our SPT model can shed light on future work in adopting its model architecture for protein-relevant tasks.

XAI Techniques. Explainable Artificial Intelligence (XAI) methods provide valuable insights into the decision-making processes of deep neural networks (DNNs), making them well-suited for interpreting biological intelligence underlined in Protein Transformers. The mainstream XAI techniques targeting DNNs can be roughly grouped into two categories, *i.e.*, XAI for CNNs and XAI for Transformers. The former category is popularized by Grad-CAM [61], which weights the activation maps by global-average-pooled gradients flowing into the last convolutional layer. Subsequently, saliency-based [18,50,62], activation-based [73,39], perturbation-based [28,27,48,54,72], and gradient-based [63,8,64,36,23,10,29,37,69,19,15] XAI techniques are developed for deciphering the decision-making processes of CNNs. Despite their popularity, these methods face challenges when applied to Protein Transformers due to the structural differences between Transformers [67] and CNNs. Recently, Attention Rollout and Attention Flow [1], which map information flow using a Directed Acyclic Graph, has been proposed to interpret the decision-making processes in Transformer architectures, with its success inspiring a volume of Transformer-specific XAI techniques [14,68,71,9,12,11,13]. However, existing XAI solutions for Protein Transformers typically involve substantial computational demands, *e.g.*, $\mathcal{O}(L^2 \cdot P^4)$ for Attention Flow, with L and P respectively representing the depth of the model and the length of the protein sequence, making them infeasible to decipher the biological insight underlined in long protein sequences. This stems from their requirement to aggregate information from attention weights throughout every layer of the Transformer Encoders. In sharp contrast, our Sequence Score technique, while classified in the second category, revolutionizes the interpretation of decision-making processes in Transformers. This achievement stems from its linear complexity with respect to the protein sequence length. Therefore, our solution significantly advances previous Transformer-specific XAI techniques in computational

Table 1: Overview of our Protein-FN dataset

Datasets	Samples	Classes					
		Protease	Kinase	Receptor	Carbonic Anhydrase	Phosphatase	Isomerase
Training	7211	2439	2003	1172	972	343	282
Test	1803	628	499	265	234	89	88
Total	9014	3067	2502	1437	1206	431	371

efficiency, making it well-suited for interpreting the biological intelligence resided in Protein Transformers.

3 The Protein-FN Dataset

We introduce our curated protein function dataset, namely *Protein-FN*, designed specifically for such biological tasks as protein function prediction [59,52], motif identification and discovery [40], *etc*. Table 1 presents the details of our ProFunc-9K dataset. This dataset, sourced from the Protein Data Bank (PDB) [53], provides diverse 1D amino acid sequences, 3D protein structures, functional properties of 9014 proteins (7211 and 1803 samples for the training and the test datasets, respectively). These proteins, after carefully examined by biological experts in our team, fall into six categories, *i.e*., protease, kinase, receptor, carbonic anhydrase, phosphatase, and isomerase. Notably, kinases, phosphatases, proteases, and receptors play essential roles in signal transduction. Most drugs act on proteins involved in signal transduction. Isomerases and carbonic anhydrases are two enzymes that are not directly involved in signal transduction pathways, but they catalyze critical reactions.

4 Our Approaches

To unveil the biological intelligence embedded within Protein Transformers, we have developed two key innovations: i) the Sequence Protein Transformer (SPT), designed for the efficient and effective prediction of protein functions, and ii) the Sequence Score, aimed at efficiently interpreting the decision-making processes of Protein Transformers.

4.1 Problem Statement

Given a protein dataset consisting of N samples, denoted as $\mathbb{X} = \{(\boldsymbol{x_i}, y_i) \mid i \in 1, 2, \cdots, N\}$, each sample $\boldsymbol{x} \in \mathbb{R}^{P\times 1}$ represents the primary structure (*i.e*., the sequence of amino acids) of the protein. Here, P denotes the sequence length, and $y \in [C]$ indicates the specific function of the protein, *e.g*., protease, kinase, receptor, *etc*. Notably, the length of the primary structure of proteins, *i.e*., the number of amino acids in a polypeptide chain, can vary widely in the real scenario. In this work, our goals are twofold. First, we aim to develop a simple, computation-efficient Protein Transformer (PT) $f_{\boldsymbol{\theta}} : \mathbb{R}^{P\times 1} \to [C]$ for accurate protein function predictions, where $\boldsymbol{\theta}$ denotes the

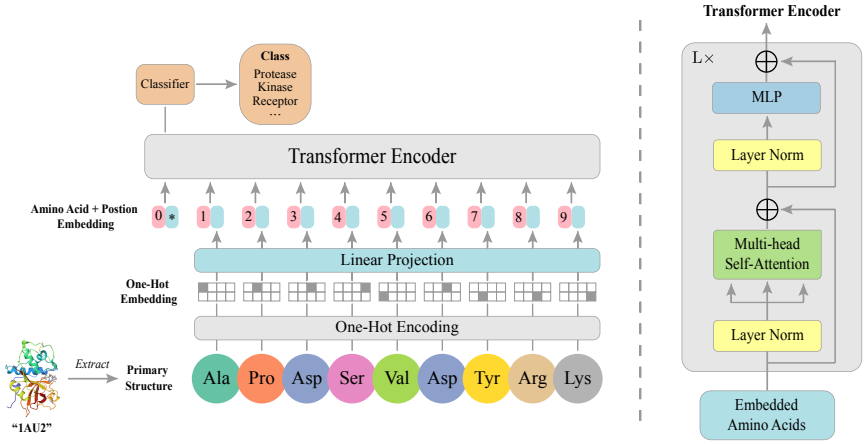


Fig. 2: The architecture of our Sequence Protein Transformers (SPT) model.

PT's hyperparameters. Second, we plan to propose a novel Explainable Artificial Intelligence (XAI) technique, which can efficiently interpret long protein sequences. As such, given a well-trained PT model $f_{\boldsymbol{\theta}}$, the proposed XAI technique $g : \mathbb{R}^{P\times 1} \rightarrow \mathbb{R}^{P\times 1}$ can decode its biological intelligence by deciphering its decision-making process.

4.2 Our Sequence Protein Transformer

To achieve our goal, we propose a simple, computation-efficient Transformer architecture, namely Sequence Protein Transformer (SPT), for predicting the functional properties of proteins. It can achieve superb prediction performance without relying on computation-intensive self-supervised pre-training.

Figure 2 shows an overview of our model architecture. We start by extracting the primary structure $\boldsymbol{x} \in \mathbb{R}^{P\times 1}$, *i.e.*, a sequence of amino acids, from a protein (*e.g.*, "1AU2"). Then, the one-hot encoding is utilized to encode the sequence of amino acids. As such, each amino acid is represented by a binary vector of length d. Note that we set $d = 20$ in this study as there are 20 types of amino acids. Next, a linear projection layer Proj: $\mathbb{R}^{P\times d} \rightarrow \mathbb{R}^{P\times D}$ is used to project the low dimentional one-hot embedding $\boldsymbol{E}_{\text{oh}} \in \mathbb{R}^{P\times d}$ to the high dimentional amino acid embedding $\boldsymbol{E}_{\text{ami}} \in \mathbb{R}^{P\times D}$, *i.e.*,

$$\boldsymbol{E}_{\text{ami}} = \text{Proj}\left(\text{OH}(\boldsymbol{x})\right). \tag{1}$$

Here, D is the hidden size of the Transformer Encoder, and OH represents one-hot encoding. Similar to prior Transformer variants [21,66,45], our model prepends a learnable classification token $\boldsymbol{E}_{\text{cls}} \in \mathbb{R}^{1\times D}$ to the embedding sequence. As such, the input sequence of the Transformer Encoder can be obtained by summing up the amino acid embedding and the positional embedding $\boldsymbol{E}_{\text{pos}} \in \mathbb{R}^{(P+1)\times D}$. Finally, the head of the output sequence $\boldsymbol{z} \in \mathbb{R}^{1\times D}$, encoded by a stack of Transformer blocks, is fed to a linear

classifier for protein function predictions, *i.e.*, $\hat{y} = \boldsymbol{W}^T \boldsymbol{z} + \boldsymbol{b}$, with $\boldsymbol{W}$ and $\boldsymbol{b}$ respectively representing the weights and the bias of the classifier, and $\hat{y} \in [C]$ indicating the predicted protein function.

The rightmost chart of Figure 2 depicts the architecture of the Transformer Encoder, where each Transformer block consists of a Multi-head Self-Attention (MSA) block [67] and an MLP block. Each of the two blocks includes a Layer Normalization [5] before the block and the residual connection [31] after the block. Similar to previous studies [21,45,70,66], the MLP block is a two-layer neural network with a GELU non-linearity. Mathematically, our Transformer Encoder can be expressed as below,

$$\begin{aligned} \boldsymbol{E}_0 &= [\boldsymbol{E}_{\text{cls}}; \boldsymbol{E}^1_{\text{ami}}; \boldsymbol{E}^2_{\text{ami}}; \cdots; \boldsymbol{E}^P_{\text{ami}}] + \boldsymbol{E}_{\text{pos}}, \\ \boldsymbol{E}'_\ell &= \text{MSA}\left(\text{Norm}(\boldsymbol{E}_{\ell-1})\right) + \boldsymbol{E}_{\ell-1}, \\ \boldsymbol{E}_\ell &= \text{MLP}\left(\text{Norm}(\boldsymbol{E}'_\ell)\right) + \boldsymbol{E}'_\ell, \\ \boldsymbol{z} &= \text{Norm}(\boldsymbol{E}^0_L). \end{aligned} \tag{2}$$

Here, $\ell = 1, 2, \cdots, L$ indicates the ℓ-th block of Transformer Encoder. Different from prior Transformer variants, our Transformer Encoder has a flexible number of positional embeddings, enabling the SPT to address the primary structure of proteins, whose sequence lengths vary significantly. Inspired by the Multi-head Self-Attention (MSA) mechnism [67], our MSA block here is devised to capture the global protein representation by learning the dependency among a sequence of amino acids, *i.e.*,

$$\begin{aligned} \text{MSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= \text{Concat}(\text{head}_1, \cdots, \text{head}_h)\boldsymbol{W}^{\text{O}}, \\ \text{head}_i &= \text{Softmax}(\boldsymbol{Q}_i \boldsymbol{K}_i^T / \sqrt{d_k})\boldsymbol{V}_i, \\ \text{where } \boldsymbol{Q}_i &= \boldsymbol{Q}\boldsymbol{W}_i^Q,\ \boldsymbol{K}_i = \boldsymbol{K}\boldsymbol{W}_i^K,\ \boldsymbol{V}_i = \boldsymbol{V}\boldsymbol{W}_i^V. \end{aligned} \tag{3}$$

Here, Concat indicates the operation of feature concatenation, h is the number of attention heads, and $d_k = D/h$ represents the dimension for queries, keys, and values of the Attention mechanism. Like those in prior studies [67,21,60,41], $\boldsymbol{W}_i^Q \in \mathbb{R}^{D \times d_k}$, $\boldsymbol{W}_i^K \in \mathbb{R}^{D \times d_k}$, $\boldsymbol{W}_i^V \in \mathbb{R}^{D \times d_k}$, and $\boldsymbol{W}^O \in \mathbb{R}^{D \times D}$ are four learnable projection matrices.

4.3 Our Sequence Score

Interpreting the biological intelligence encoded in Protein Transformers (PT) demands an XAI technique capable of handling long protein sequences composed of a substantial number of amino acids. Previous Transformer-specific XAI methods [1,68,71,12,11] have proven effective in elucidating the decision-making processes of various Transformer models [21,45]. However, their use is predominantly limited to analyzing shorter token sequences, due to their substantial computational overhead involved in aggregating attention weights across all layers of Transformer Encoders. This limitation is particularly problematic for interpreting Protein Transformers (PTs), which analyze amino acid sequences that often feature extensive lengths, thus obstructing their potential to unlock the decision-making processes within PTs. To address this challenge, we introduce the Sequence Score, an innovative XAI method characterized by its time complexity *growing linearly* with the protein sequence. This feature renders it exception-

ally suitable for decoding the biological intelligence usually embedded within protein sequences.

Given a decision of interest (*e.g.*, the protease class), our Sequence Score, with respect to the gradients of a well-trained PT model, can generate a sequence of important scores, based on the primary structure of proteins. Next, we introduce the details of our Sequence Score technique. Consider a decision interest of class $c \in [C]$, our Sequence Score technique first calculates the gradient of the logit for the class c, with respect to feature maps $\boldsymbol{A} \in \mathbb{R}^{P \times D}$ of any Transformer block (*e.g.*, the last block of Transformer Encoders), *i.e.*, $\frac{\partial y^c}{\partial \boldsymbol{A}}$. Here, the term "logit" refers to the classification score before being passed through a sigmoid (or softmax) function to produce a probability distribution over the classes. Then, the neuron importance weight $\boldsymbol{w}^c \in \mathbb{R}^D$, is obtained by performing global average pooling over the sequence length (indexed by j), *i.e.*,

$$\boldsymbol{w}^c = \frac{1}{P} \sum_{j=1}^{P} \frac{\partial y^c}{\partial \boldsymbol{A}_j}. \tag{4}$$

Next, we arrive at the importance score for the j-th amino acid S_j^c by summing up a weighted combination of the feature map activations $\boldsymbol{A}$, *i.e.*,

$$S_j^c = \sum_{k=1}^{D} \boldsymbol{w}_k^c \boldsymbol{A}^k, \quad j = 1, 2, \cdots, P. \tag{5}$$

Similar to prior XAI techniques [61,10], attention is paid solely to features that positively affect the prediction of interest. In other words, the negative importance scores should be dropped. Meanwhile, our preliminary experimental results indicate that if the primary structure of proteins is too long, the importance score for each amino acid will be small (*i.e.*, < 0.001). We develop a novel trick to address the two issues simultaneously, expressed as follows,

$$S_j^c = \frac{\max(0,\ S_j^c)}{\max(\boldsymbol{S}^c)}, \quad j = 1, 2, \cdots, P. \tag{6}$$

Here, the numerator and the denominator of Eq. (6) serve for dropping the negative scores and normalizing the positive scores, respectively. As such, our Sequence Score technique can interpret biological intelligence underlined in PT models by revealing their decision-making processes when predicting the functional properties of proteins. It is noteworthy that the computation of our Sequence Score technique achieves linear time complexity, denoted as $\mathcal{O}(D \cdot P)$, where D represents the hidden dimension of the Transformer Encoders and P denotes the length of the protein sequences. This efficiency underscores its suitability for analyzing the intricate biological intelligence embedded in protein structures.

5 Experiments and Results

5.1 Experimental Settings

Datasets. We conduct experiments across three benchmarks: i) **Protein-FN**, having $9,014$ protein data with their 1D amino acid sequences, 3D protein structures, and func-

Table 2: Model variants of our Sequence Protein Transformers (SPT), with their model details listed below

Model	Layers	Hidden Size D	Head	MLP Size	Parameters
SPT-Tiny	12	192	4	768	5.4M
SPT-Small	12	384	6	1536	21.5M
SPT-Base	12	768	12	3072	85.5M

tional properties; ii) **Antibiotic Resistance (AR)** [51], containing $3,416$ protein samples, each associated with its antibiotic type; and iii) **Metal Ion Binding (MIB)** [33], offering $7,332$ single protein sequences, collected from PDB with annotation as metal ion binding.

Model Variants. We set our Sequence Protein Transformers (SPT) configurations based on the Transformer settings reported in previous studies [21,66]. Three model variants, *i.e.*, SPT-Tiny, SPT-Small, and SPT-Base, are developed, tailored for protein function predictions across different scales of data. Table 2 presents the model details of those SPT variants. Specifically, all SPT variants are composed of 12 layers of Transformer blocks, with their hidden sizes set to 192, 384, and 768, and their numbers of heads set to 4, 6, and 12, respectively for the SPT-Tiny, the SPT-Small, and the SPT-Base models. The MLP sizes are fixed to four times of their corresponding hidden sizes.

Compared Approaches. As our SPT models belong to the single sequence-based Protein Transformers, we consider three prominent single sequence-based Protein Language Models (PLMs), *i.e.*, **TAPE** [55], **ESM-1b** [59], and **ESM-1v** [52], for baseline comparison. The hyperparameters for PLM counterparts, if not specified, are set as reported in their original literature.

Hyperparamters. In sharp contrast to prior PLMs [55,59,52], our SPT models do not require computationally extensive self-supervised pre-training. Instead, they are all trained from scratch by employing the AdamW [47] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.05. The training epochs for SPT variants are set to 100, including 5 warmup epochs. We utilize the cosine decay learning rate schedule [46], with a base learning rate of $1e-3$ and a layer-wise learning rate decay [17] of 0.75. Following [30], we also apply the label smoothing [65] and the path dropping [35], with their values both set to 0.1. As such, our SPT models are superb computation-efficient. All experiments were conducted on a lab computer with an RTX 4090 GPU, having its memory usage consistently $\leq 9.6\%$.

5.2 Comparisons to Protein Language Models

Experiments on the Protein-FN Dataset. We conducted experiments on the *Protein-FN* dataset to evaluate the performance of our SPT models. Three state-of-the-art PLMs mentioned in Section 5.1 are taken into account as baselines for comparison. Table 3 presents experimental results, where two notations, *i.e.*, "pre-trained" and "scratch", respectively indicate the counterparts with and without the self-supervised pre-training[5].

[5] ESM-1v (scratch) is the same as ESM-1b (scratch) as they use the same structure but are pre-trained on different datasets.

Table 3: Comparisons to state-of-the-art PLM counterparts on the Protein-FN dataset, where the last two columns respectively report the error rates on the training and the test sets, with the best results shown in bold

Methods	Parameters	GFLOPs	Traning Error (%)	Test Error (%)
TAPE (scratch)	92.4M	21.4	2.39	11.59
TAPE (pre-trained)			0.34	0.55
ESM-1b (scratch)	650M	160	1.11	11.73
ESM-1b (pre-trained)			1.33	1.86
ESM-1v (pre-trained)	650M	160	0.55	0.58
SPT-Tiny	**5.4M**	**1.4**	0.39	0.41
SPT-Small	21.5M	5.1	0.22	0.38
SPT-Base	85.5M	19.4	**0.11**	**0.31**

Here, we consider both computational complexity and model performance. The former is measured by the number of parameters and GFLOPs[6] calculated under a sequence of amino acids, while the latter is characterized by the top-1 error rates on the training and the test sets.

We have three observations. First, our SPT-Tiny model (containing only 5.4M parameters) achieves an exceptionally low test set error rate of 0.41%, outperforming all competitors, in terms of both computational efficiency and prediction accuracy. Moreover, its impressive capabilities in protein function predictions are achieved with the GFLOPs value of just 1.4. This represents a computational demand at least $15.2\times$ lower than that of its competitors, underscoring the model's remarkable balance between efficiency and accuracy. Second, our SPT-Base model achieves the best error rate of 0.11% and 0.31% respectively on the training and the test sets. This notable performance underscores a key finding: scaling up our model's size corresponds to a marked enhancement in its capabilities. Third, training PLMs from scratch may lead to substantial overfitting. This issue is evident in the significant discrepancies observed between training and test set performance outcomes, *e.g.*, the error rate of 2.39% *v.s.* 11.59% for the TAPE (scratch) model and of 1.11% *v.s.* 11.73% for the ESM-1b (scratch) model. In sharp contrast, our SPT models, despite trained from scratch, show excellent generalization abilities. This can be attributed to our SPT's design in the amino acid embedding, which lifts the requirement of self-supervised pre-training to learn meaningful protein representations.

Experiments on the AR and MIB Datasets. Here, we conducted experiments on the two widely-used benchmarks, *i.e.*, Antibiotic Resistance (AR) and Metal Ion Binding (MIB), for further exhibiting the effectiveness of our SPT models for protein function predictions. Table 4 presents experimental results. It is observed that our SPT-Base model beats all PLM counterparts on both datasets, with the best test error rates of 4.3% and 32.1% on the AR and MIB datasets. In addition, our SPT models stand out for their computational efficiency. Taking the the AR dataset for instance, our SPT-Tiny

[6] GFLOPs, or Giga Floating Point Operations, is a metric that quantifies a model's computational complexity. It indicates the number of billion floating-point operations needed by a model per second.

Table 4: Overall comparisons to PLMs on the AR and MIB datasets, with best results shown in bold

Methods	AR		MIB	
	GFLOPs	Test Error (%)	GFLOPs	Test Error (%)
TAPE (scratch)	106.3	11.8	65.7	39.8
TAPE (pre-trained)		7.3		34.2
ESM-1b (scratch)	172.1	11.1	79.1	40.1
ESM-1b (pre-trained)		8.6		35.9
ESM-1v (pre-trained)	172.1	9.8	79.1	33.3
SPT-Tiny	**19.2**	5.7	**5.6**	37.1
SPT-Small	31.2	4.5	18.5	32.7
SPT-Base	105.6	**4.3**	65.7	**32.1**

model not only achieves a low test error rate of 5.7% but did so with just 19.2 GFLOPs of computational demand. This efficiency makes it at least 5.5 times faster than previous PLMs, marking a significant advancement in processing speed and energy consumption. Finally, without self-supervised pre-training, previous PLMs suffer from a significant performance degradation (See 3rd *v.s.* 4th rows and 5th *v.s.* 6th rows). On the contrary, our SPT models, though trained from scratch, achieves superb prediction performance outcomes. This can be attributed to the effectiveness of our novel protein embedding mechanism.

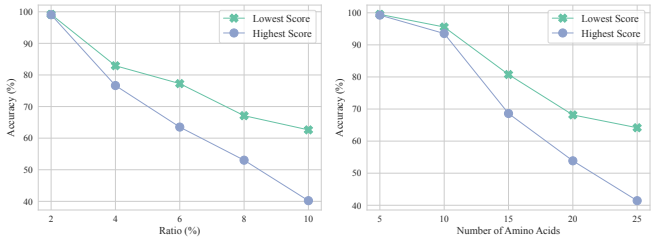
We've further evaluated our SPT models, comparing them with traditional bioinformatics approaches and conducting detailed ablation studies, with corresponding experimental results deferred to Appendices A.1 and A.2, respectively.

5.3 Evaluation on Our Sequence Score Technique

We consider two metrics proposed in the prior study [3], *i.e.*, **Faithfulness** and **Stability**, to evaluate the efficacy of our Sequence Score technique for explaining Protein Transformers. In this section, we present the experimental results regarding the *faithfulness* metric. The evaluation of the *stability* of our Sequence Score technique is detailed in Appendix B.1 of the supplementary materials for conserving space. In particular, *faithfulness* measures the degree to which the importance values which are attributed to amino acids, are aligned with their actual impact on the final prediction, expecting that amino acids with substantial effects will receive correspondingly high importance scores.

Here, we adopt the *deletion* method [54] to evaluate the *faithfulness* of our Sequence Score technique. Its key idea is to observe accuracy degradation incurred by masking a certain ratio (or number) of amino acids, with masked amino acids chosen based on their importance scores, *i.e.*, those with the highest scores *v.s.* those with the lowest scores. A larger drop in prediction accuracy, resulted from masking amino acids with high importance scores than with low scores, indicates that the assigned scores are well aligned with amino acids' actual significance to the final predictions.

Figure 3a illustrates the results of our experiments that involve selectively masking amino acids at various ratios. Our findings reveal a consistent pattern: masking amino



(a) Ratio (b) Number

Fig. 3: Comparisons of prediction performance by masking amino acids with the highest and the lowest importance scores under (a) masking a certain ratio of amino acids and (b) masking a specific number of amino acids.

acids identified as highly important invariably leads to a larger decline in prediction performance, versus masking amino acids with lower importance scores. For example, when the masking ratio is set to 10%, masking amino acids that have the highest importance scores results in performance degradation to be 22.41% greater than that resulted from masking those with the lowest importance scores. Additionally, lifting the masking ratio from 2% to 10% yields a noticeably faster decline in prediction performance when masking amino acids with the highest scores (see the blue line) than with the lowest scores (see the green line). This demonstrates a direct relationship between the importance scores assigned to the amino acids and their actual influence levels on the predictive accuracy of the model.

Complementing these findings, Figure 3b illustrates similar trends under a different experimental setting, where various numbers of amino acids are masked. As the number of masked amino acids increases from 5 to 25, prediction performance is observed to degrade significantly faster when masking amino acids with the highest scores than with the lowest scores. Specifically, a masking number of 25 leads to a substantial larger performance decline, *i.e.*, by 58.16%, when masking amino acids with the highest importance scores than with the lowest scores, *i.e.*, only by 35.41%. These statistical observations confirm that our Sequence Score technique adheres to the principle of *faithfulness* when interpreting Protein Transformers.

We have also conducted experiments to assess the *faithfulness* of our Sequence Score technique by simulating protein mutations, with their results deferred to Appendix B.2 of supplementary materials.

5.4 Discovery of Catalytic Triad in Serine Proteases

This section further interprets biological intelligence resided in Protein Transformers by unveiling its discovery of the catalytic triad in serine proteases. Serine proteases, a group of proteases, are crucial enzymes involved in a myriad of biological functions, including digestion, immune response, and blood coagulation. At the heart of their catalytic mechanism lies the catalytic triad, a set of three coordinated amino acids, usually following the pattern of "His57-Asp102-Ser195". This triad forms a potent synergis-

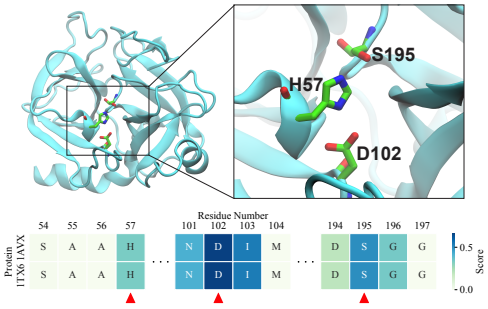


Fig. 4: Illustration of the catalytic triad, *i.e.*, "His57-Asp102-Ser195", in the two proteins of serine proteases. Here, "His", "Asp", and "Ser" are abbreviated as "H", "D" and "S", respectively. The figure in the upper section utilizes the protein "1TX6" as an example to show the 3D structure of the catalytic triad, while the heatmap in the lower section visualizes the importance scores generated by our Sequence Score technique, with amino acids corresponding to the catalytic triad distinctly marked by red triangles.

tic unit essential for the enzyme's function. Figure 4 depicts importance scores for the two proteins of serine proteases. It is obvious that our SPT models can identify the significance of the catalytic triad (marked by red triangles) for serine proteases. These results further confirm that our Protein Transformers can capture biological intelligence inherent within protein sequences.

It is worth noticing that the catalytic triad's importance extends beyond its biochemical role; its evolutionary conservation across various serine proteases underscores a fundamental mechanism critical to many physiological processes. Moreover, the detailed understanding of this triad, including the specific residue numbers, has been instrumental in the design of targeted pharmaceutical inhibitors to modulate serine protease activity in treating various diseases, such as cancer, inflammatory disorders, and coagulopathies. Hopefully, the capability of our SPT models to discover the catalytic triad of serine proteases can shed light on future studies, aided by the Protein Transformers for understanding the biochemical, physiological, and pharmaceutical processes.

6 Conclusion

This work has explored the capabilities of Protein Transformers in capturing biological intelligence resided in protein sequences. To achieve our goal, we first introduced the *Protein-FN* dataset, offering over 9000 protein sequence data as well as their functional properties created laboriously by biological experts. Then, we developed the Sequence Protein Transformers (SPT), a computationally efficient Transformer architecture, able to precisely predict the functional properties of proteins by leveraging their primary structures. Thanks to its novel protein embedding mechanism, our SPT models can achieve superb prediction performance without the requirement of self-supervised pretraining. Finally, we have developed the Sequence Score, a novel Explainable Artificial Intelligence (XAI) technique that advances beyond current Transformer-specific XAI

solutions in terms of computational efficiency. This efficiency increases linearly with the length of protein sequences, making it well-suited for analyzing the complex biological intelligence encoded within Protein Transformers. Extensive experimental results exhibited that our SPT models are efficient and effective in predicting the functional properties of proteins. Moreover, the devised Sequence Score technique helps reveal that our SPT models can capture important patterns underlying protein sequences, with these patterns aligning closely with the domain knowledge in the biology community. This demonstrates the capabilities of our Protein Transformers in capturing biological intelligence resided in protein sequences.

Acknowledgments

This work was supported in part by NSF under Grants 2019511 and 2425812. Any opinions and findings expressed in the paper are those of the authors and do not necessarily reflect the views of funding agencies.

References

1. Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
2. Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 2019.
3. David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Neural Information Processing Systems 2018 (NeurIPS)*, 2018.
4. Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*, 2021.
5. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
6. Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021.
7. Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.
8. Oren Barkan, Omri Armstrong, Amir Hertz, Avi Caciularu, Ori Katz, Itzik Malkiel, and Noam Koenigstein. GAM: explainable visual similarity and classification via gradient activation maps. In *International Conference on Information and Knowledge Management (CIKM)*, 2021.
9. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.
10. Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018.

11. Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *International Conference on Computer Vision (ICCV)*, 2021.
12. Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
13. Ziheng Chen, Fabrizio Silvestri, Gabriele Tolomei, Jia Wang, He Zhu, and Hongshik Ahn. Explain the explainer: Interpreting model-agnostic counterfactual explanations of a deep reinforcement learning agent. *IEEE Transactions on Artificial Intelligence*, 5(4):1443–1457, 2022.
14. Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, Zhenhua Huang, Hongshik Ahn, and Gabriele Tolomei. Grease: Generate factual and counterfactual explanations for gnn-based recommendations. *arXiv preprint arXiv:2208.04222*, 2022.
15. Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, and Gabriele Tolomei. The dark side of explanations: Poisoning recommender systems with counterfactual examples. In *Proceedings of the 46th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 2426–2430, 2023.
16. Tiankuo Chu, Fudong Lin, Shubo Wang, Jason Jiang, Wiley Jia-Wei Gong, Xu Yuan, and Liyun Wang. Bonemet: An open large-scale multi-modal murine dataset for breast cancer bone metastasis diagnosis and prognosis. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
17. Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.
18. Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Neural Information Processing Systems (NeurIPS)*, 2017.
19. Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020.
20. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.
21. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
22. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
23. Alexandre Englebert, Olivier Cornu, and Christophe De Vleeschouwer. Backward recursive class activation map refinement for high resolution saliency map. In *International Conference on Pattern Recognition (ICPR)*, 2022.
24. Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, 2021.
25. Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

26. Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. Pfam: the protein families database. *Nucleic acids research*, 2014.
27. Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *International Conference on Computer Vision (ICCV)*, 2019.
28. Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision (ICCV)*, 2017.
29. Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *British Machine Vision Conference (BMVC)*, 2020.
30. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
31. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
32. Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning (ICML)*, 2022.
33. Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware & -free protein language models as protein function predictors. In *Neural Information Processing Systems (NeurIPS)*, 2022.
34. Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware &-free protein language models as protein function predictors. 2022.
35. Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.
36. Mohammad A. A. K. Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. CAMERAS: enhanced resolution and sanity preserving class activation mapping for image saliency. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
37. Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.*, 30:5875–5888, 2021.
38. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
39. Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.
40. Sarika Kondra, Titli Sarkar, Vijay Raghavan, and Wu Xu. Development of a tsr-based method for protein 3-d structural comparison with its applications to protein classification and motif discovery. *Frontiers in Chemistry*, 8:602291, 2021.
41. Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, Drew Gholson, Nicolas Ashwell, Tri Setiyono, Brenda Tubana, Lu Peng, Magdy Bayoumi, and Nian-Feng Tzeng. Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *International Conference on Computer Vision (ICCV)*, 2023.

42. Fudong Lin, Kaleb Guillot, Summer Crawford, Yihe Zhang, Xu Yuan, and Nian-Feng Tzeng. An open and large-scale dataset for multi-modal climate change-aware crop yield predictions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5375–5386, 2024.
43. Fudong Lin, Jiadong Lou, Xu Yuan, and Nian-Feng Tzeng. Towards robust vision transformer via masked adaptive ensemble. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1389–1399, 2024.
44. Fudong Lin, Xu Yuan, Yihe Zhang, Purushottam Sigdel, Li Chen, Lu Peng, and Nian-Feng Tzeng. Comprehensive transformer-based model architecture for real-world storm prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 54–71, 2023.
45. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.
46. Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2016.
47. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
48. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems (NeurIPS)*, 2017.
49. Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. 2022.
50. Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis. (IJCV)*, 120(3):233–255, 2016.
51. Andrew G McArthur, Nicholas Waglechner, Fazmin Nizam, Austin Yan, Marisa A Azad, Alison J Baylay, Kirandeep Bhullar, Marc J Canova, Gianfranco De Pascale, Linda Ejim, et al. The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7):3348–3357, 2013.
52. Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. 2021.
53. PDB. Protein data bank (pdb), 2023.
54. Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, page 151, 2018.
55. Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Neural Information Processing Systems (NeurIPS)*, 2019.
56. Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning (ICML)*, 2021.
57. Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations (ICLR)*, 2021.
58. Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 2018.
59. Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

60. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.
61. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, 2017.
62. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshop (ICLRW)*, 2014.
63. Suraj Srinivas and Francois Fleuret. Full-gradient representation for neural network visualization. In *Neural Information Processing Systems (NeurIPS)*, 2019.
64. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.
65. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
66. Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning (ICML)*, 2021.
67. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017.
68. Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
69. Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2020.
70. Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision (ICCV)*, 2021.
71. Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L. Zhang. Vit-cx: Causal explanation of vision transformers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.
72. Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.
73. Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.