FusionProt: Fusing Sequence and Structural Information for Unified Protein Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Accurate protein representations that integrate sequence and three-dimensional (3D) structure are critical to many biological and biomedical tasks. Most existing models either ignore structure or combine it with sequence through a single, static fusion step. Here we present FusionProt, a unified model that learns representations via iterative, bidirectional fusion between a protein language model and a structure encoder. A single learnable token serves as a carrier, alternating between sequence attention and spatial message passing across layers. FusionProt is evaluated on Enzyme Commission (EC), Gene Ontology (GO), and mutation stability prediction tasks. It improves F_{max} by a median of +1.3 points (up to +2.0) across EC and GO benchmarks, and boosts AUROC by +3.6 points over the strongest baseline on mutation stability. Inference cost remains practical, with only $\sim$ 2–5% runtime overhead. Beyond state-of-the-art performance, we further demonstrate FusionProt's practical relevance through representative biological case studies, indicating that the model captures biologically relevant features.

1 Introduction

Proteins are essential for biological processes and for understanding complex mechanisms in living organisms. They comprise linear chains of amino acids that fold into a specific three-dimensional (3D) structure, which underscores their functional diversity and dynamic behaviors (Zhang et al., 2023b). An effective understanding of proteins is essential for understanding disease mechanisms and synthetic biology and for advancing drug development (Liu et al., 2024).

Current methodologies for protein representation primarily emphasize the exploration of proteins' one-dimensional (1D) structures, specifically the relationships between amino acids. These approaches, such as ProteinBERT (Brandes et al., 2021) and ESM (Rives et al., 2019; Lin et al., 2023), often utilize text-based techniques and transformer architectures (Vaswani et al., 2017), which are trained on extensive protein sequence datasets. These models take an amino acid sequence as input and typically produce protein representations by averaging the representations of individual amino acids. However, this narrow focus on amino acid sequences neglects crucial protein structure details, thereby limiting the effectiveness of these methods.

The intricate 3D structure of proteins is crucial, as the conformation plays a pivotal role in determining their activities (Zhang et al., 2023a). Proteins possess specific active sites for interactions with other molecules, which are defined by the 3D arrangement of amino acids. This 3D structure determines the specificity and affinity of binding interactions, aspects that a 1D representation cannot adequately capture (Kastritis & Bonvin, 2013; Yan et al., 2013). Furthermore, drug design relies on an understanding of 3D structures to identify binding sites and find molecules that modulate protein function (Luo, 2022; Liu et al., 2022). One of the state-of-the-art (SOTA) techniques today for 3D protein representation is GearNet (Zhang et al., 2023b). This approach converts the 3D structure of a protein into a graph that captures its biological characteristics. Subsequently, graph neural network techniques (Kipf & Welling, 2017; Schlichtkrull et al., 2017) are applied to this graph, facilitating the creation of comprehensive protein representations.

Recent research emphasizes the importance of comprehensive protein representation that includes both 1D and 3D structures to capture the protein's functional and interactional properties accurately. ESM-

GearNet (Zhang et al., 2023a) was one of the first approaches to integrate these modalities. Although the study explored various fusion strategies, empirical results showed that the most effective method is using a large protein language model (PLM) such as ESM (Lin et al., 2023) to generate representations, which were then used as context for a graph encoder like GearNet (Zhang et al., 2023b). Other approaches, such as SaProt (Su et al., 2024), leverage an AlphaFold-based model (van Kempen et al., 2022) to reduce the 3D structure to tokens and train them along with amino acid tokens using a traditional PLM. However, these approaches are limited as they reduce one modality into context for another model, which potentially leads to the loss of critical structure information.

In this study, we introduce FusionProt (See Figure 1), a novel approach designed to learn a unified representation of the 1D and 3D structures of proteins simultaneously. Despite the vast number of proteins in nature, the number of known 3D structures remains limited (Váradi et al., 2023). To address this, we leverage an AlphaFold model (Jumper et al., 2021) for accurate protein structure predictions. We introduce an innovative learnable fusion token that serves as an adaptive bridge, enabling an iterative exchange of information between a PLM and the protein's 3D structure graph. This token is integrated into the training process of both modalities, enabling seamless propagation of information and facilitating comprehensive representation through iterative learning cycles. In practice, this token is concatenated to the sequence, allowing attention mechanisms to query the unique fusion token alongside the amino acids. This process extracts and integrates valuable information, enhancing the learning of amino acid representations. Then, the fusion token is incorporated as an additional node in the graph representing the protein's 3D structure, connected to all nodes. A graph encoder (i.e., a structure model) processes this graph, generating a new representation for the fusion token, which is subsequently used in the PLM training over the amino acids. Through this iterative process, the model representations are combined to form a refined protein representation.

Unlike prior models such as ESM-GearNet (Zhang et al., 2023a), which perform static or one-shot fusion, such as simple concatenation after independent encoding, FusionProt introduces a dynamic mechanism where the fusion token continuously evolves through repeated interaction across layers. This design allows sequence and structure modalities to co-adapt throughout training, resulting in richer, functionally informed representations that better capture the complexity of protein behavior.

We perform an empirical evaluation over several protein tasks, spanning a broad spectrum of biological domains, comparing numerous methods of protein representations and achieving SOTA performance across various benchmarks, with statistical significance improvements. We present ablation tests to better study the performance of the algorithm.

The contributions of this study are threefold: (1) We introduce FusionProt, a novel approach that learns a unified representation for both the 1D and 3D structures of proteins simultaneously. Our main focus is on the fusion of 1D and 3D models in an effective manner. Our method enhances the accuracy of capturing functional and interactional properties of proteins, addressing limitations of previous methods that treated these structures separately; (2) We propose a novel fusion architecture that utilizes a specialized learnable fusion token, enabling an iterative exchange of information between a PLM and the protein's 3D structure graph. This token is integrated into the training process of both modalities, enabling seamless propagation of information and facilitating comprehensive representation through iterative learning cycles. The iterative process facilitates the exchange of contextually relevant structure and sequential features, improving the model's ability to capture both 1D and 3D protein characteristics; (3) We conduct an empirical evaluation of our work over several protein tasks establishing SOTA results and presenting biological case studies that further demonstrate the model's strengths. Also, we contribute our code to the community [1].

[1] `https://anonymous.4open.science/r/FusionProt-4554`

2 Related Work

2.1 Sequence-based Representation Learning

Proteins are comprised of sequences of amino acids, that establish a natural analogy to tokens in natural language processing. Recently, the adoption of unsupervised deep learning techniques has become prevalent in modeling protein sequence data.

The advent of Transformers (Vaswani et al., 2017) has led to the development of numerous PLMs such as MSA Transformer (Rao et al., 2021), ProteinBERT (Brandes et al., 2021), ProteinLM (Xiao et al., 2021), ProtBERT-BFD (Elnaggar et al., 2021), ProtTrans (Elnaggar et al., 2022), ESM-1b (Rives et al., 2019), and ESM-2 (Lin et al., 2023) which is considered as the SOTA PLM. These models were trained on data from UniRef (Suzek et al., 2007) which contains hundreds of billions of protein sequences, via masked language modeling (Devlin et al., 2019).

A protein's biological function hinges on its 3D native structure (Dill & MacCallum, 2012). However, many PLMs do not explicitly encode protein 3D structures, which are pivotal in understanding protein functions.

In this work, we aim to overcome this limitation by enhancing a PLM through novel fusion algorithms that integrate protein 3D structure models into the protein embedding. This approach seeks to capture both sequential and structure attributes of proteins, thereby advancing the capabilities of existing PLMs in biological research and applications.

2.2 Structure-based Representation Learning

The rising success of AlphaFold (Jumper et al., 2021; Senior et al., 2020) in predicting the 3D structure of proteins has led to deeper insights into their functional roles. Moreover, the release of more than 200 million protein structures in AlphaFoldDB (Váradi et al., 2023; 2021) has significantly advanced the development of large-scale protein structure models (van der Weg et al., 2025).

Protein structures are commonly represented as graphs, where amino acids serve as the nodes. Therefore, utilizing protein structure models on these graphs is a common practice. Models such as GVP (Jing et al., 2021), CDConv (Fan et al., 2023) and GearNet (Zhang et al., 2023b) have shown promising results, with different learning techniques. GearNet, which incorporates a Multiview Contrastive pre-training algorithm, is considered SOTA (Zhang et al., 2023b) and has outperformed the IEConv model (Hermosilla et al., 2021), which proposed to apply a learnable kernel function on edge features. Furthermore, CDConv (Fan et al., 2023) has outperformed HoloProt (Somnath et al., 2022) and ProNet (Wang et al., 2022a) in a recent study (Liu et al., 2023).

Foldseek (van Kempen et al., 2022) suggested a different approach, optimized for protein structural search, which utilizes a VQ-VAE (van den Oord et al., 2017) to encode protein structures into informative tokens. Then, SaProt (Su et al., 2024) integrates residue and structure tokens during training, derived from encoding 3D protein structures using Foldseek.

In this study, we enhance the structure-based representation learning approach by integrating sequential information, thus capturing both sequential and structure attributes of proteins.

2.3 Joint Representation Learning

The integration of protein sequence-based models with protein structure models has gained popularity in recent years (Quan et al., 2024; Ko et al., 2024; Wu et al., 2022; Li et al., 2024). Early efforts, such as LM-GVP (Wang et al., 2022b) or MIF-ST (Yang et al., 2022), aimed to combine PLMs with Graph Neural Networks (GNNs) (Scarselli et al., 2009). More recently, ESM-GearNet (Zhang et al., 2023a) proposed to incorporate sequential information into distinct residue-level models such as GearNet, GVP, and CDConv (Zhang et al., 2023a). SaProt-GearNet (Su et al., 2024) also integrates GearNet and achieves similar results. However, these approaches are limited by reducing one modality into context for another model, potentially losing critical structure information.

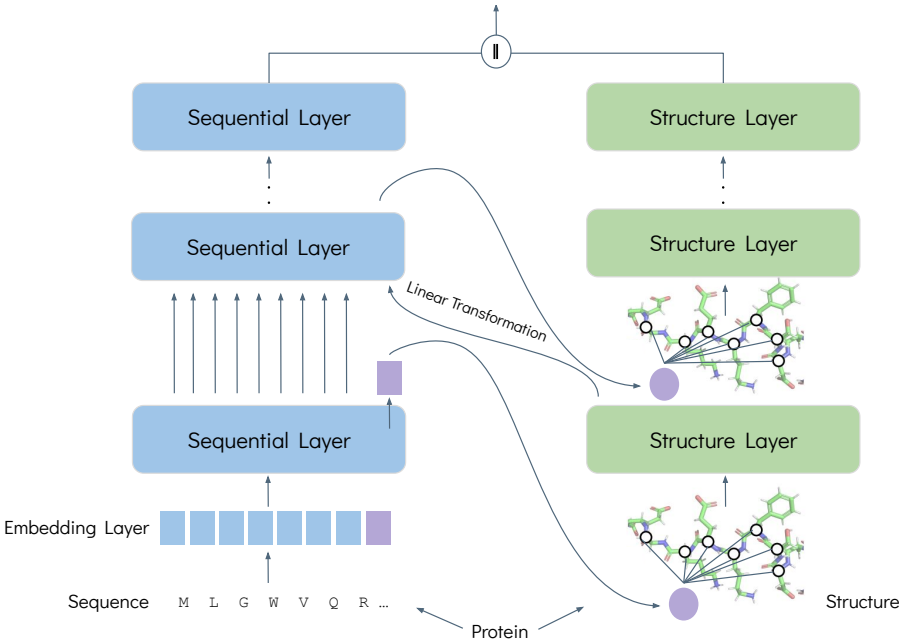

Figure 1: The **FusionProt** pre-training architecture. The model is trained on amino acid sequences and their corresponding 3D structures, utilizing the SOTA AlphaFold2 (Jumper et al., 2021) for accurate protein structure predictions. We introduce an innovative learnable fusion token that serves as an adaptive bridge, enabling an iterative exchange of information between a PLM and the protein's 3D structure graph. This token is concatenated to the protein sequence, allowing attention mechanisms to query the unique fusion token alongside the amino acids. Then, the fusion token is incorporated as an additional node in the graph representing the protein's 3D structure, connected to all nodes. A graph encoder (i.e., a structure model) processes this graph, generating a new representation for the fusion token. This representation is subsequently used in the PLM's sequential layers. A learnable linear transformation is applied between each sequential and structure layer pair to align and adapt their distinct modality spaces. Through this iterative process, the model representations are combined to form a refined protein representation.

Unlike earlier approaches, we introduce a novel fusion architecture that simultaneously learns a unified representation for both the 1D and 3D structures of proteins. To the best of our knowledge, no existing PLMs are based on an iterative fusion of structure and sequential models simultaneously.

3 Methods

We introduce the **FusionProt** model (See Figure 1). Given a protein $p = (S, R)$, which combines its amino acid sequence S and its 3D structure R, FusionProt simultaneously learns a unified representation for both the 1D and 3D structures of the protein. We propose a specialized learnable fusion token designed to enable an iterative exchange of information between a PLM and the protein's 3D structure graph.

In this section, we formally define sequential (1D) and structure (3D) protein layers, which are used in the model as part of the protein representation learning, and present the FusionProt algorithm. Information from both layers is fused simultaneously during learning throughout the proposed specialized fusion token.

3.1 Protein Sequential Layer

3.1.1 Protein Sequence

The simplest level of protein structure, known as the primary structure, is a 1D structure, consisting of a linear sequence of amino acids, referred to as residues. The protein sequence exhibits similarities with natural

language sequences, making the application of language models a common practice in this domain (Xiao et al., 2021; Brandes et al., 2021).

Given a protein sequence $S = [s_1, s_2, ..., s_n]$, where n denotes the number of residues, the sequential model aims to capture the essential features of the protein sequence and outputs a protein representation denoted by $z = [z_1, z_2, ..., z_n] \in \mathbb{R}^{n \times D}$, where D is the embedding dimension.

3.1.2 Sequential Layer Definition

We utilize ESM-2 (Lin et al., 2023) as our sequential model, leveraging its superior performance as a PLM. Alternatively, other advanced sequential models (i.e., PLMs) can be used.

Within the context of a given sequential layer denoted by l, we denote $z^{(l)}$ as the output representation of this layer, initialized with $z_i^{(0)} = \text{Embedding}(s_i) \in \mathbb{R}^D$, where D denotes the embedding representation dimension. The layer performs the following update:

$$z^{(l)} = \textit{Attention}\left(z_1^{(l-1)}, z_2^{(l-1)}, ..., z_n^{(l-1)}\right)$$

where n denotes the number of residues in the protein, and *Attention* is a multi-head self-attention layer (Vaswani et al., 2017).

3.2 Protein Structure Layer

3.2.1 Protein 3D Structure

A protein 3D structure is uniquely determined by its primary structure (amino acid sequences) (Dill & MacCallum, 2012). There are only 20 standard residue (amino acid) types, each containing multiple components connected to a central carbon atom known as alpha carbon. Following GearNet (Zhang et al., 2023b), we use only alpha carbons to represent the main backbone structure of each protein. Therefore, we define a protein 3D structure as $R = [r_1, r_2 ... r_n] \in \mathbb{R}^{n \times 3}$, where r_i represents the Cartesian coordinates of the $i - th$ alpha carbon atom of each amino acid, and n denotes the number of residues, which is the sequence length.

3.2.2 Protein Structure Graph

We represent proteins using a multi-relational residue graph $G = (V, E, T)$, where V is the set of residues, E is the set of edges, and T is the edge types. The set of edges E consists of three directed edge types, namely sequential edges, radius edges, and KNN edges:

$$\begin{aligned} E_{\text{seq}} &= \{(i,j) \mid i,j \in V, |j-i| < d_{\text{seq}}\} \\ E_{\text{radius}} &= \{(i,j) \mid i,j \in V, |r_j - r_i| < d_{\text{radius}}\} \\ E_{\text{KNN}} &= \{(i,j) \mid i,j \in V, j \in \text{KNN}(i)\} \\ E &= E_{\text{seq}} \cup E_{\text{radius}} \cup E_{\text{KNN}} \end{aligned}$$

where $d_{\text{seq}} = 3$ defines the sequential distance threshold, $d_{\text{radius}} = 10\text{Å} = 1$ [nm] defines the spatial distance threshold, and $\text{KNN}(i)$ indicates the K-nearest neighbors (Peterson, 2009) of node i with $k = 10$ (all parameters were set as reported in GearNet (Zhang et al., 2023b)).

3.2.3 Structure Layer Definition

We utilize GearNet (Zhang et al., 2023b) as the structure model, leveraging its contextual understanding of protein structures. Alternatively, other advanced structure models can be used in the algorithm (See Section 5.2).

Given a protein's 3D structure, R, we construct its corresponding protein structure graph $G = (V, E, T)$ (See Section 3.2.2). Then, the structure layer employs a relational message passing procedure, based on a relational graph convolutional neural network (Schlichtkrull et al., 2017).

Within the context of a given structure layer denoted by l, we denote $u^{(l)}$ as the output representations of this layer, initialized with $u_i^{(0)} = \text{Embedding}(v_i) \in \mathbb{R}^D$, where $v_i \in V$, and D denotes the embedding representation dimension. The layer performs the following update:

$$u_i^{(l)} = u_i^{(l-1)} + \sigma\left(\sum_{t \in T} W_t \sum_{j \in \mathcal{N}_t(i)} u_j^{(l-1)}\right)$$

where $\mathcal{N}_t(i)$ is the set of neighbors of i with edge type t, $\sigma(\cdot)$ is a ReLU activation function, and the weight matrix W_t is learned per edge type t. This approach allows the model to incorporate various types of relational information, thereby enhancing its ability to learn comprehensive protein representations.

3.3 FusionProt

3.3.1 Fusion Design

We introduce a specialized learnable fusion token, designed to enable an iterative exchange of information between a PLM and the protein's 3D structure graph. This fusion token serves as a dynamic bridge, enabling iterative information exchange between the two modalities. During training (See Figure 1), the specialized fusion token, holds information from both models, the sequential and the structure. In practice, the amino acids query the associated unique fusion token to extract and integrate valuable information for learning amino acid representations via attention mechanisms. Additionally, the fusion token is incorporated as an additional node in the graph representing the protein's 3D structure. We connect it to all nodes, with a new edge type, thus enabling the structure model to learn important features from the sequential model. Hence, the token enables a novel fusion between both models, leading to an enhanced protein representation.

The choice to connect the fusion token to all nodes in the 3D graph is informed by principles from GNNs (Scarselli et al., 2009), where nodes exchange information to capture global and local dependencies. The fusion token is not merely an addition to the sequence or structure; rather, it mediates the dynamic interaction between the two modalities, ensuring iterative refinement of the protein representation. This approach contrasts with previous methods that simply concatenate or independently process the sequence and structure information. This novel fusion mechanism leverages multi-modal learning and graph-based information propagation. By iteratively combining the 1D protein sequence and 3D structure features, FusionProt captures both local and global dependencies, enabling the model to learn richer and more holistic representations of proteins.

3.3.2 Fusion Algorithm

We provide a formal description of the FusionProt algorithm in Algorithm 1.

Given a protein $p = (S, R)$, which combines its amino acid sequence S and its 3D structure R, FusionProt concatenates the learnable fusion token to the sequence, resulting in a new sequence $S' = [s_1, s_2, ..., s_n, f]$ with a length of $n + 1$. Then, it is used to initialize the sequential embedding layer:

$$\forall s_i \in S' : z_i^{(0)} = \text{Embedding}(s_i)$$

Within the context of a given sequential layer denoted by l, the algorithm yields a sequence representation:

$$z^{(l)} = \textit{Attention}\left(z_1^{(l-1)}, z_2^{(l-1)}, ..., z_n^{(l-1)}, z_{n+1}^{(l-1)}\right)$$

where $z_{n+1}^{(l-1)}$ is the intermediate representation of the fusion token. Then, the fusion token representation is passed into the corresponding structure layer l, represented as node $n + 1$, while connecting to all nodes in the 3D structure graph, yielding a structure representation:

$$u_i^{(l)} = u_i^{(l-1)} + \sigma\left(\sum_{t \in T'} W_t \sum_{j \in \mathcal{N}_t(i)} u_j^{(l-1)}\right)$$

Algorithm 1 FusionProt Algorithm

Input: Protein $p = (S, R)$ with sequence $S = [s_1, ..., s_n]$, and 3D structure $R = [r_1, ..., r_n] \in \mathbb{R}^{n\times 3}$, Sequential encoder with L_1 layers, Structure encoder L_2 layers.
Output: Unified protein representation h

Create a learnable fusion token $s_{n+1} = f$
Augment sequence: $S' = [s_1, ..., s_n, s_{n+1}]$
Initialize sequential embeddings: $z_i^{(0)} = \text{Embedding}(s_i)\ \forall i \in [1, ..., n+1]$
Construct graph $G = (V, E, T)$ from R
Add node v_{n+1} to G, connect $v_{n+1} \leftrightarrow v_i$ with edge type $t_f\ \forall i \in [1, ..., n]$
Set $u_i^{(0)} = \text{Embedding}(v_i)$ for all $v_i \in V$
for $l = 1$ to L_2 **do**
    for $j = 1$ to $L_1 \div L_2$ **do**
        $l_j = j + (l-1) \cdot (L_1 \div L_2)$
        $z^{(l_j)} = \textit{Attention}(z_1^{(l_j-1)}, ..., z_n^{(l_j-1)}, z_{n+1}^{(l_j-1)})$
    end for
    $u_{n+1}^{(l-1)} = \text{Linear}_l(z_{n+1}^{(l_j)})$
    Update node v_{n+1} in graph G with $u_{n+1}^{(l-1)}$
    $u_i^{(l)} = u_i^{(l-1)} + \sigma\left(\sum_{t\in T} W_t \sum_{j\in\mathcal{N}_t(i)} u_j^{(l-1)}\right)\ \forall u_i \in V$
    $z_{n+1}^{(l_j)} = \text{Linear}'_l(u_{n+1}^{(l)})$
end for
Concatenate final embeddings: $h = [z_{1:n}^{(L_1)}; u_{1:n}^{(L_2)}]$
return h

where $u_{n+1}^{(l-1)} = \text{Linear}_l\left(z_{n+1}^{(l-1)}\right)$, $T' = T \cup \{t_f\}$, $\mathcal{N}_{t_f}(n+1) = V$, and $\forall i \in [1, ..., n] : \mathcal{N}_{t_f}(i) = \{n+1\}$, as we connected the fusion token to all residues in the protein structure graph with a new type of edge t_f (in both directions). To bridge the gap between the sequential and structure representation space, we utilize an affine transformation Linear_l (a linear layer) per each structure layer l to project these representations into the same space. Subsequently, we set the fusion token representation $u_{n+1}^{(l)}$ as the input for the following sequential layer $l+1$, resulting in:

$$z_{n+1}^{(l)} = \text{Linear}_{l'}\left(u_{n+1}^{(l)}\right)$$

where $\text{Linear}_{l'}$ is the affine transformation in the opposite direction (i.e., structure to sequential space or vice versa).

The protein representation undergoes transformations across the structure and sequential layers, where L_1 and L_2 are the number of sequential and structure layers, respectively, and $L_1 \geq L_2$ (note that L_1 need not be equal to L_2). Our aim is to uniformly integrate 3D structure information across the L_1 sequential and L_2 structure layers. Therefore, the fusion token is passed after $L_1 \div L_2$ sequential layers into the structure model. Then, the fusion token is passed back into the sequential model after every structure layer.

Lastly, the final hidden states of both the sequential and structure layers are concatenated to form the final protein output representation, which is:

$$h = \left[z^{(L_1)}, u^{(L_2)}\right]$$

3.4 Pre-training Objective

Following GearNet (Zhang et al., 2023b) and ESM-GearNet (Zhang et al., 2023a), we use Multiview Contrastive learning as our pre-training objective, for a fair comparison. Other pre-training algorithms could also be applied. We compare different pre-training algorithms in Section 5.3.

The Multiview Contrastive objective is to preserve the similarity between correlated protein subcomponents when mapped to a lower-dimensional latent space. Therefore, for a protein structure graph, we ran-

domly select consecutive subsequences and apply random edge masking to hide 15% of edges in the protein graph, generating diverse views. Then, we align their representations in the latent space with an InfoNCE loss (van den Oord et al., 2018).

4 Empirical Results

In this section, we present our empirical results. For each task, we report the mean and standard deviation of FusionProt's performance over five independent fine-tuning runs (with different random seeds) to ensure reproducibility. Additionally, we validate the statistical significance of performance differences, using a two-tailed paired t-test with a 95% confidence level ($p < 0.05$), comparing observations from the tested models. The normality of the paired differences was confirmed using the Shapiro–Wilk test (Shapiro & Wilk, 1965). Significant results are marked with an asterisk (*) in the tables.

Finally, we compute Cohen's d effect sizes (Cohen, 1969) to quantify the magnitude of FusionProt's improvements over baselines on each task. We observe large effects ($d > 0.8$) across all tasks, indicating that the improvements are not only statistically significant but also practically meaningful.

In Supplementary Section A.1, we outline our empirical setting, including the pre-training dataset, task details, baselines, implementation details, and a computational complexity analysis.

Table 1: Evaluation results on EC and GO prediction under various pre-trained baseline models. "PLM" and "Structure Info." indicate the usage of protein language models and structure information in the model, respectively. The $F_{\max}$ score is the evaluation metric. Statistically significant results ($p < 0.05$) using a paired t-test across proteins in the test set are marked with an asterisk (*). For FusionProt, we report the mean and standard deviation over 5 independent fine-tuning runs (different seeds; EC: $n = 1604$, GO: $n = 3350$). The best result is highlighted in bold.

Method	**PLM**	**Structure Info.**	**EC**	**GO-BP**	**GO-MF**	**GO-CC**
			$F_{\max}$	$F_{\max}$	$F_{\max}$	$F_{\max}$
ProtBERT-BFD (Elnaggar et al., 2021)	✓	✗	0.838	0.279	0.456	0.408
DeepFRI (Gligorijević et al., 2021)	✓	✗	0.631	0.399	0.465	0.460
ESM-1b (Rives et al., 2019)	✓	✗	0.859	0.320	0.661	0.392
ESM-2 (Lin et al., 2023)	✓	✗	0.877	0.345	0.668	0.411
GVP (Jing et al., 2021)	✗	✓	0.886	0.495	0.672	0.420
CDConv (Fan et al., 2023)	✗	✓	0.820	0.453	0.654	0.479
GearNet (Zhang et al., 2023b)	✗	✓	0.871	0.481	0.650	0.476
MIF-ST (Yang et al., 2022)	✓	✓	0.803	0.239	0.627	0.322
ESM-GearNet (Zhang et al., 2023a)	✓	✓	0.886	0.512	0.670	0.495
SaProt (Su et al., 2024)	✓	✓	0.884	0.486	0.678	0.479
SaProt-GearNet (Su et al., 2024)	✓	✓	0.886	0.512	0.672	0.504
FusionProt	✓	✓	**0.904***±0.003	**0.524***±0.004	**0.689***±0.002	**0.518***±0.004

4.1 Task 1: EC Number Prediction

In Table 1, we compared the performance of FusionProt with eleven baseline methods on the EC number prediction task. In particular, FusionProt significantly outperformed all baseline methods, achieving the highest $F_{\max}$ score.

PLMs such as ProtBERT-BFD (Elnaggar et al., 2021) or ESM-2 (Lin et al., 2023), which rely solely on sequence data, produced substantially lower $F_{\max}$ scores, compared to FusionProt. This highlights the benefit of incorporating 3D structure information into representation learning.

Similarly, FusionProt, which leverages both structure and sequential information, performed better than SOTA structure models such as GearNet (Zhang et al., 2023b) or CDConv (Fan et al., 2023), which trained only on 3D structure data. This indicates that while structure models are strong, they are not sufficient alone,

as they require PLMs for optimal performance, probably due to their effective self-attention layers (Vaswani et al., 2017).

While models such as MIF-ST (Yang et al., 2022), ESM-GearNet (Zhang et al., 2023a), and SaProt-GearNet (Su et al., 2024) also attempt to utilize both types of information, they reduce one modality into a context for another model, leading to a loss of critical structure information. In addition, the importance of the fusion technique is emphasized by the significant underperformance of MIF-ST compared to other fusion models. Similarly to ESM-GearNet, MIF-ST uses outputs from a pre-trained sequence-only PLM as input to a structure model.

By integrating 1D and 3D protein structure information synergistically, FusionProt attempts to capture subtle structure features that influence enzyme specificity and activity, which are crucial for EC prediction. This performance improvement is essential for real-world applications, such as the diagnosis of enzyme deficiency-related diseases (Li et al., 2017).

4.2 Task 2: GO Term Prediction

Table 1 presents the results of the GO term prediction tasks compared to eleven baseline methods, including PLMs, structure models, and ensemble models. FusionProt showed strong performance across the board, achieving the highest $F_{\max}$ scores in all tasks, with a statistical significance.

The results demonstrate that sequential or structure information alone is insufficient, similarly to the EC prediction task (See Section 4.1). This indicates that the novel fusion of FusionProt is critical for achieving superior performance.

In the GO-BP task, which involves predicting a protein's role in biological processes, FusionProt achieves an $F_{\max}$ score of 0.524, significantly outperforming the next-best method. Biological processes often rely on long-range structural interactions and cooperative protein functions, which makes this task particularly sensitive to accurate 3D structure representations. The substantial performance gap between SaProt and SaProt-GearNet further emphasizes the critical role of structure in this setting, a factor that FusionProt leverages more effectively through early integration.

In the GO-CC task, which predicts the cellular component of proteins, FusionProt achieves an $F_{\max}$ of 0.518, outperforming all baselines including ESM-GearNet (Zhang et al., 2023a), probably due to the more direct connection between 3D structure characteristics and subcellular localization (Gillani & Pollastri, 2024). The prediction of cellular components is strongly based on the spatial organization of proteins within the cell, which is well represented by detailed 3D structures (Song et al., 2022).

This reliance on structural information benefits models such as CDConv (Fan et al., 2023), although it exhibits weaker performance in EC prediction. In contrast, predicting molecular functions involves complex interactions and dynamic changes that are less directly captured by 3D structures (Saraç et al., 2010), resulting in a comparatively smaller performance gap than in the other tasks.

4.3 Task 3: Mutation Stability Prediction

Table 2 reports the results for the Mutation stability prediction (MSP) task, which evaluates a model's ability to assess the effect of amino acid substitutions on protein stability, a critical problem in protein engineering, variant effect prediction, and drug design.

Following the evaluation protocol of ESM-GearNet (Zhang et al., 2023a), we compare FusionProt on the MSP task against the SOTA structure-based method for this task, GVP (Jing et al., 2021) and the joint sequence–structure baseline ESM-GearNet (Zhang et al., 2023a). Sequence-only PLMs do not take structural inputs and therefore cannot address structure-dependent tasks such as MSP (Zhang et al., 2023a).

FusionProt achieves the highest AUROC among all evaluated methods, reaching 0.745 with statistical significance ($p < 0.05$), outperforming both structure-aware (GVP) and sequence-structure fused (ESM-GearNet) baselines. Notably, FusionProt improves upon ESM-GearNet by approximately 24.37%, and surpasses the current SOTA GVP by 5.1%. This highlights the effectiveness of our iterative fusion mechanism in pre-

serving and integrating long-range sequence and structural dependencies, which are especially important for modeling stability changes that may arise from distal or context-dependent mutations.

Compared to GVP, which is limited to local spatial interactions, and ESM-GearNet, which performs shallow one-shot fusion, FusionProt allows multi-layer cross-modal refinement via a learnable token and bidirectional interaction path. This enables a more expressive and biologically grounded representation, particularly for capturing distributed compensatory effects in allosteric regions, commonly overlooked in residue-centric models.

Table 2: Evaluation results on MSP prediction across different models. "PLM" and "Structure" indicate the usage of protein language models and structure information in the model, respectively. The AUROC score is used as the evaluation metric. Statistically significant results ($p < 0.05$) using a paired t-test across proteins in the test set are marked with an asterisk (*). For FusionProt, we report the mean and standard deviation over 5 independent fine-tuning runs (different seeds; MSP: $n = 347$). The best result is highlighted in bold.

Method	**PLM**	**Structure**	**MSP (AUROC)**
ESM-GearNet (Zhang et al., 2023a)	✓	✓	0.599
GVP (Jing et al., 2021) (SOTA)	✗	✓	0.709
FusionProt	✓	✓	$\mathbf{0.745}^{*} \pm 0.006$

5 Ablation and Analysis

We conduct a series of ablation and diagnostic studies to evaluate the key design decisions behind FusionProt. These include varying the fusion-injection frequency, comparing different structure encoders, and testing alternative pre-training objectives. We also assess the model's robustness to noise in predicted 3D structures. Finally, we present biological case studies that highlight FusionProt's practical utility in real-world tasks such as drug discovery and disease research.

5.1 Ablation on Fusion-Injection Frequency

We aimed to determine the optimal number of fusion injections in our FusionProt model, which fused information between the structure and sequential models using a specialized learnable fusion token. In our standard method, given a sequential model with L_1 layers and a structure model with L_2 layers, where $L_1 \geq L_2$, the fusion token is passed after $L_1 \div L_2$ sequential layers into the structure model (See Section 3.3.2). Then, the fusion token is passed back into the sequential model after every structure layer. In this ablation, we experimented with different injection frequencies: half the regular number (every $2 \cdot (L_1 \div L_2)$ sequential layers and two structure layers) and one-third the regular number (every $3 \cdot (L_1 \div L_2)$ sequential layers and three structure layers). The results in Table 3 indicate that the standard fusion injection frequency consistently provides the best performance across tasks. Reducing fusion injections negatively impacted performance, highlighting the importance of the fusion token. By decreasing the frequency of injections, the model's ability to integrate these two types of information is compromised, leading to suboptimal predictions.

Table 3: Ablation test for the number of fusion injections in FusionProt. Statistically significant results with $p < 0.05$ using a t-test are marked with an asterisk (*). The best result is highlighted in bold.

Number of Fusion	**EC**	**GO-BP**	**GO-MF**	**GO-CC**
	$F_{\max}$	$F_{\max}$	$F_{\max}$	$F_{\max}$
Few (One-Third)	0.876	0.511	0.659	0.498
Medium (Half)	0.882	0.514	0.669	0.504
Full (Standard)	$\mathbf{0.904}^{*}$	$\mathbf{0.524}^{*}$	$\mathbf{0.689}^{*}$	$\mathbf{0.518}^{*}$

5.2 3D Structure Models Comparison

FusionProt is designed to be flexible and modular, allowing the structure encoder component to be replaced with alternative architectures. The ablation study for various SOTA 3D structure models is detailed in Table 4. We evaluated several structure models, including GearNet (Zhang et al., 2023b), GVP (Jing et al., 2021), and CDConv (Fan et al., 2023). For each structure model, we applied our fusion technique and compared its performance with FusionProt, which utilizes the GearNet (Zhang et al., 2023b) model (See Section 3.2.3). The GVP model replaces standard MLPs (Murtagh, 1991) in GNN (Scarselli et al., 2009) layers with generalized vector perceptrons, which handle scalar and geometric features as vectors that adapt to spatial rotations. In contrast, CDConv utilizes GearNet's multi-type message passing to capture sequential and spatial interactions among residues. For consistency, we follow the model selection protocol of ESM-GearNet (Zhang et al., 2023a). Our results demonstrate that the FusionProt model, utilizing the GearNet structure, consistently achieves superior performance compared to other models. This aligns with previous studies (Zhang et al., 2023a), which have also established GearNet as a SOTA approach for joint representation learning.

Table 4: Ablation test for various structure models used by FusionProt. For all models, the PLM is the ESM-2 (Lin et al., 2023). Statistically significant results with $p < 0.05$ using a t-test are marked with an asterisk (*). The best result is highlighted in bold.

Method	**EC**	**GO-BP**	**GO-MF**	**GO-CC**
	$F_{\max}$	$F_{\max}$	$F_{\max}$	$F_{\max}$
FusionProt (GVP)	0.889	0.507	0.677	0.462
FusionProt (CDConv)	0.837	0.478	0.665	0.497
FusionProt (GearNet)	**0.904***	**0.524***	**0.689***	**0.518***

5.3 Pre-training Algorithm Ablation

Table 5 compares FusionProt, trained with Multiview Contrastive learning (Section 3.4), with other pre-training algorithms. First, we tested self-prediction methods (Zhang et al., 2023b), which aim to predict one part of the protein given the remaining context. Specifically, we utilized the Residue Type Prediction method, a self-supervised task that performs masked prediction on individual residues. Next, we evaluated diffusion-based methods (Zhang et al., 2023c), inspired by the success of diffusion models in capturing the relationship between sequences and structures. During training, noise levels are added to structures and sequences, with higher noise levels indicating more distortion. We tested the SiamDiff objective, which refines structures via torsional adjustments and noise reduction. The results show that the Multiview Contrastive approach outperformed both Residue Type Prediction (using only sequence data) and SiamDiff (which integrates sequence and structure separately). Unlike methods that treat sequence and structure separately, Multiview Contrast integrates both, aligning subsequence representations from the same protein to capture interrelated sequences and structure patterns.

Table 5: Ablation test for the FusionProt model with different pre-training algorithms. Statistically significant results with $p < 0.05$ using a t-test are marked with an asterisk (*). The best result is highlighted in bold.

Method	**EC**	**GO-BP**	**GO-MF**	**GO-CC**
	$F_{\max}$	$F_{\max}$	$F_{\max}$	$F_{\max}$
Residue Type	0.890	0.524	0.673	0.506
SiamDiff	0.874	0.509	0.645	0.512
Multiview Contrast	**0.904***	**0.524***	**0.689***	**0.518***

5.4 Robustness to Noise in 3D Structures

Robustness in protein modeling often refers to a model's ability to tolerate minor perturbations (Cho et al., 2024)—an important property given that predicted structures (e.g., from AlphaFold2) may contain local inaccuracies, especially in flexible regions. To evaluate FusionProt's robustness to such noise, we injected Gaussian perturbations into the alpha-carbon backbone coordinates of AlphaFold2-predicted structures. The noise was zero-mean, with standard deviations ranging from 0.1 Å to 1.5 Å, simulating local structure deviations commonly observed in predicted or low-confidence regions. FusionProt maintained stable performance up to a noise level of 0.9 Å, with only a modest drop of 3–5 $F_{\max}$ points observed in the EC and GO-BP tasks beyond this threshold. These results demonstrate that FusionProt is resilient to realistic structure noise and remains effective even when relying on imperfect 3D structures, as commonly encountered in high-throughput or large-scale prediction settings.

5.5 Biological Case Studies

FusionProt consistently outperforms SOTA baselines across all downstream tasks (Table 1, Table 2). When such performance gains are driven by biologically meaningful inputs, such as detailed 3D structural features, they can indicate that the model is capturing relationships of functional relevance.

To better understand the nature of these improved learned protein representations, we evaluated a model's ability to predict EC numbers, as outlined in Section A.1.2, using pre-trained representations without any fine-tuning. This approach enables an assessment of the inherent quality of the learned embeddings, independent of task-specific training. We then compared FusionProt's predictions with those of ESM-GearNet, focusing on the cases with the largest discrepancies.

In Supplementary Table 6, we present these proteins into distinct biological groups based on shared mechanistic characteristics. We focus on two common mechanisms and provide a representative example for each group (See Supplementary Section A.2).

6 Conclusions

In this paper, we introduced FusionProt, a novel architecture designed to learn a unified representation of the 1D and 3D structures of proteins simultaneously. FusionProt incorporates a specialized learnable fusion token that enables an iterative exchange of information between a PLM and the protein's 3D structure graph, facilitating a more comprehensive representation through iterative learning cycles between a PLM and a structure model.

We propose a novel fusion algorithm that enables effective propagation of information between a PLM and a structure model. This fusion technique outperforms previous methods, which often convert one modality into context for another model, potentially leading to the loss of crucial structural information.

We evaluated our proposed method on several protein-level tasks to assess its effectiveness in protein representation learning. These tasks include protein annotations, which are essential for real-world medical applications. FusionProt significantly outperformed all baseline methods, including SOTA models, across all tasks, with statistically significant improvements.

We perform ablation tests to examine the contribution of our novel approach of learning unified representations of both 1D and 3D structures of proteins simultaneously. The tests confirm that our approach significantly enhances FusionProt's ability to capture the intricate relationships between protein sequences and their 3D structures, leading to improved performance across various real-world tasks.

Although the proposed fusion token mechanism is tailored for bidirectional interaction between sequence and structure modalities, it is conceptually extensible to more than two information channels. For example, additional modalities such as ligand descriptors, protein dynamics, or expression context could, in principle, be integrated into the framework by assigning separate fusion tokens or adopting a shared token passed through each encoder block in sequence. We leave the systematic evaluation of such multi-modal extensions to future work.

References

Nadav Brandes, Dan Ofer, Yam Peleg, and et al. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38:2102 – 2110, 2021.

Yehlin Cho, Sergey Ovchinnikov, and Christopher Frank. Enhancing protein design robustness through noise-informed sequence design. In *ICML 2024 AI for Science Workshop*, 2024. URL https://openreview.net/forum?id=1tUTPheba8.

Jacob Cohen. Statistical power analysis for the behavioral sciences. *The SAGE Encyclopedia of Research Design*, 1969. URL https://api.semanticscholar.org/CorpusID:123217261.

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*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338:1042 – 1046, 2012. URL https://api.semanticscholar.org/CorpusID:5756068.

Ahmed Elnaggar, Michael Heinzinger, and et al. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. doi: 10.1109/TPAMI.2021.3095381.

Ahmed Elnaggar, Michael Heinzinger, and et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 7112–7127, 2022.

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2023.

Maryam Gillani and Gianluca Pollastri. Protein subcellular localization prediction tools. *Computational and Structural Biotechnology Journal*, 23:1796–1807, 2024. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2024.04.032. URL https://www.sciencedirect.com/science/article/pii/S2001037024001156.

Vladimir Gligorijević, P. Douglas Renfrew, and et al. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12, 2021. URL https://api.semanticscholar.org/CorpusID:235218243.

Midori A. Harris, Jennifer I. Clark, Amelia Ireland, Jane Lomax, Michael Ashburner, and et al. Gene ontology consortium: The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32 Database issue:D258–61, 2004. URL https://api.semanticscholar.org/CorpusID:22565487.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, and et al. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *International Conference on Learning Representations*, 2021.

Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING/IJCLCLP*, 1997. URL https://api.semanticscholar.org/CorpusID:1359050.

Bowen Jing, Stephan Eismann, Pratham N. Soni, and et al. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, and et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

Panagiotis L. Kastritis and Alexandre M.J.J. Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society Interface*, 10, 2013. URL https://api.semanticscholar.org/CorpusID:14659814.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Eunji Ko, Seul Lee, Minseon Kim, Dongki Kim, and Sung Ju Hwang. Protein representation learning by capturing protein sequence-structure-function relationship. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024. URL https://openreview.net/forum?id=1YBM58iaNt.

Juha Kurkela, Julia Fredman, Tiina A Salminen, and Taina Tyystjärvi. Revealing secrets of the enigmatic omega subunit of bacterial rna polymerase. *Molecular Microbiology*, 115(1):1–11, 2021. doi: https://doi.org/10.1111/mmi.14603. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.14603.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4Z7RZixpJQ.

Yu Li, Sheng Wang, Ramzan Umarov, and et al. Deepre: sequence-based enzyme ec number prediction by deep learning. *Bioinformatics*, 2017. URL https://api.semanticscholar.org/CorpusID:3790226.

Zeming Lin, Halil Akin, Roshan Rao, and et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.

Zeming Lin, Halil Akin, Roshan Rao, and et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Jiajia Liu, Mengyuan Yang, Yankai Yu, Haixia Xu, Kang Li, and Xiaobo Zhou. Large language models in bioinformatics: applications and perspectives. *ArXiv*, 2024. URL https://api.semanticscholar.org/CorpusID:266899789.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. In *International Conference on Machine Learning*, 2022.

Shengchao Liu, weitao Du, Yanjing Li, and et al. Symmetry-informed geometric representation for molecules, proteins, and crystalline materials. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=ygXSNrIU1p.

Shitong Luo. A 3d generative model for structure-based drug design. In *Neural Information Processing Systems*, 2022. URL https://api.semanticscholar.org/CorpusID:244906742.

Renjith Mathew and Dipankar Chatterji. The evolving story of the omega subunit of bacterial rna polymerase. *Trends in microbiology*, 14 10:450–5, 2006. URL https://api.semanticscholar.org/CorpusID:27701631.

Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2:183–197, 1991. URL https://api.semanticscholar.org/CorpusID:7949799.

Catia Pesquita, Daniel Faria, Hugo Bastos, André Falcao, and Francisco Couto. Evaluating go-based semantic similarity measures. In *Proc. 10th Annual Bio-Ontologies Meeting*, volume 37, pp. 38, 2007.

Leif E. Peterson. K-nearest neighbor. *Scholarpedia*, 4:1883, 2009. URL https://api.semanticscholar.org/CorpusID:29611121.

Ruijie Quan, Wenguan Wang, Fan Ma, Hehe Fan, and Yi Yang. Clustering for protein representation learning, 2024. URL https://openreview.net/forum?id=IWfqn6Ythj.

Roshan M Rao, Jason Liu, Robert Verkuil, and et al. Msa transformer. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Alexander Rives, Siddharth Goyal, Joshua Meier, and et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118, 2019.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, 116:13996 – 14001, 2019. URL https://api.semanticscholar.org/CorpusID:195192023.

Hira Saleem, Usman Ali Ashfaq, Habibullah Nadeem, Muhammad Zubair, Muhammad Hussnain Siddique, and Ijaz Rasul. Subtractive genomics and molecular docking approach to identify drug targets against stenotrophomonas maltophilia. *PLOS ONE*, 16(12):1–17, 12 2021. doi: 10.1371/journal.pone.0261111. URL https://doi.org/10.1371/journal.pone.0261111.

Ömer Sinan Saraç, Volkan Atalay, and Rengul Cetin-Atalay. Gopred: Go molecular function prediction by combined classifiers. *PLoS ONE*, 5, 2010. URL https://api.semanticscholar.org/CorpusID:1258725.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20:61–80, 2009. URL https://api.semanticscholar.org/CorpusID:206756462.

M. Schlichtkrull, Thomas Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *Extended Semantic Web Conference*, 2017. URL https://api.semanticscholar.org/CorpusID:5458500.

Andrew W. Senior, Richard Evans, John M. Jumper, and et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710, 2020.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52 (3-4):591–611, dec 1965. doi: 10.1093/biomet/52.3-4.591. URL https://doi.org/10.1093/biomet/52.3-4.591.

Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-scale representation learning on proteins. In *Neural Information Processing Systems*, 2022. URL https://api.semanticscholar.org/CorpusID:236317531.

Bosheng Song, Xiaoyan Luo, Xiaoli Luo, Yuansheng Liu, Zhangming Niu, and Xiangxiang Zeng. Learning spatial structures of proteins improves protein–protein interaction prediction. *Briefings in Bioinformatics*, 23(2):bbab558, 01 2022. ISSN 1477-4054. doi: 10.1093/bib/bbab558. URL https://doi.org/10.1093/bib/bbab558.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6MRm3G4NiU.

Baris E. Suzek, Hongzhan Huang, Peter B. McGarvey, Raja Mazumder, and Cathy H. Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23 10:1282–8, 2007.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

Karel van der Weg, Erinc Merdivan, Marie Piraud, and Holger Gohlke. Topec: prediction of enzyme commission classes by 3d graph neural networks and localized 3d protein descriptor. *Nature Communications*, 16, 2025. URL https://api.semanticscholar.org/CorpusID:277148295.

Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, and et al. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42:243 – 246, 2022. URL https://api.semanticscholar.org/CorpusID:257837735.

Mihály Váradi, Stephen Anyango, Mandar S. Deshpande, and et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50:D439 – D444, 2021. URL https://api.semanticscholar.org/CorpusID:245770129.

Mihály Váradi, Damian Bertoni, Paulyna Magaña, and et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52:D368 – D375, 2023. URL https://api.semanticscholar.org/CorpusID:265042036.

Ashish Vaswani, Noam M. 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*, 2017. URL https://api.semanticscholar.org/CorpusID:13756489.

Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3d graph networks. In *International Conference on Learning Representations*, 2022a. URL https://api.semanticscholar.org/CorpusID:257364758.

Zichen Wang, Steven A. Combs, Ryan Brand, and et al. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports*, 12, 2022b. URL https://api.semanticscholar.org/CorpusID:248415326.

Edwin C. Webb. Enzyme nomenclature 1992. recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. 1992. URL https://api.semanticscholar.org/CorpusID:83964422.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999.

Yijia Xiao, Jiezhong Qiu, Ziang Li, Chang-Yu Hsieh, and Jie Tang. Modeling protein using large-scale pretrain language model. *ArXiv*, abs/2108.07435, 2021.

Zhiqiang Yan, Liyong Guo, Liang Hu, and Jin Wang. Specificity and affinity quantification of protein-protein interactions. *Bioinformatics*, 29 9:1127–33, 2013. URL https://api.semanticscholar.org/CorpusID:1339389.

Kevin K Yang, Niccolò Zanichelli, and Hugh Yeh. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 10 2022. ISSN 1741-0126. doi: 10.1093/protein/gzad015. URL https://doi.org/10.1093/protein/gzad015.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, and et al. A systematic study of joint representation learning on protein sequences and structures. 2023a. URL https://api.semanticscholar.org/CorpusID:257496831.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, and et al. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=to3qCB3tOh9.

Zuobai Zhang, Minghao Xu, Aurélie Lozano, and et al. Pre-training protein encoder via siamese sequence-structure diffusion trajectory prediction. In *Advances in Neural Information Processing Systems*, 2023c.

A Appendix

A.1 Empirical Evaluation

In this section, we outline our empirical setting. We performed analysis on several downstream tasks, demonstrating the value of the representation we provide in real-world bioinformatics tasks.

A.1.1 Protein Structure Dataset

To reduce reliance on the availability and quality of external structural data, we leverage AlphaFold2 (Jumper et al., 2021), a well-known SOTA model, which enables us to generate 3D structures and eliminate the need for external sources. Furthermore, many recent high-performing methods, such as SaProt (Su et al., 2024) and ESM-GearNet (Zhang et al., 2023a), also used it for structure generation, ensuring consistency in comparisons. Therefore, we use the AlphaFold protein structure database (Váradi et al., 2023; 2021) for pre-training, which contains 805K protein structures predicted by AlphaFold2 (Jumper et al., 2021). Alternatively, other models, such as ESMFold (Lin et al., 2022), can be applied. Since AlphaFold predictions may include local inaccuracies, we assess FusionProt's robustness to structure noise in Section 5.4.

A.1.2 Tasks

We evaluated our proposed method on several common downstream tasks, selected based on SOTA works of GearNet (Zhang et al., 2023a) and SaProt (Su et al., 2024), to assess its effectiveness for protein representation learning. These tasks include protein annotations such as Enzyme Commission (EC) prediction, Gene Ontology (GO) prediction, and Mutation Stability Prediction. The GO prediction task includes three subtasks: predicting a protein's biological processes (BP), molecular functions (MF), and cellular components (CC).

To ensure the validity of our results, all downstream task splits are constructed so that sequences in one set share no more than 30% Needleman–Wunsch sequence identity with any sequence in the other sets. In addition, the hold-out sets for each task share no more than 30% identity with any protein in the AlphaFold2 pre-training corpus Jumper et al. (2021).

Enzyme Commission Number Prediction Annotation of enzyme function has a wide range of real-world applications, including metagenomics, diagnosis of enzyme-deficiency-related diseases (Li et al., 2017), and cellular metabolism (Ryu et al., 2019). This task focuses on determining enzyme function by predicting EC numbers, which characterize a protein's catalytic activity in biochemical reactions. It involves 538 binary classification problems derived from the third and fourth levels of the EC classification tree (Webb, 1992). We used dataset splits from DeepFRI (Gligorijević et al., 2021), and the evaluation metric is the $F_{\max}$ score.

Gene Ontology Term Prediction The GO knowledgebase (Harris et al., 2004) provides a set of structured and controlled terms describing gene products and their molecular properties. Many real-world biological applications, such as predictions of protein-protein interactions, rely on GO term-protein annotations (Pesquita et al., 2007; Jiang & Conrath, 1997). This benchmark includes three tasks: predicting a protein's biological processes, molecular functions, and cellular components. Each task involves multiple binary classification problems based on GO term annotations. We used dataset splits from DeepFRI (Gligorijević et al., 2021), and the evaluation metric is the $F_{\max}$ score.

Mutation Stability Prediction Mutation stability prediction (MSP) is crucial in bioinformatics for understanding how genetic mutations affect protein stability, which plays a key role in disease mechanisms and drug development. This task aims to predict whether a mutation enhances a protein complex's stability. We utilize the datasets and hyperparameters from ESM-GearNet (Zhang et al., 2023a). Evaluation is based on AUROC.

A.1.3 Baselines

We compare with numerous baseline models, including PLMs and structure models. We used ProtBERT-BFD (Elnaggar et al., 2021), DeepFRI (Gligorijević et al., 2021), ESM-1b (Rives et al., 2019), and ESM-2 (Lin et al., 2023) as SOTA PLM sequential models. GearNet (Zhang et al., 2023b), GVP (Jing et al., 2021), and CDConv (Fan et al., 2023) are used as SOTA structure models. Furthermore, we include MIF-ST (Yang et al., 2022), ESM-GearNet (Zhang et al., 2023a), SaProt (Su et al., 2024) and SaProt-GearNet (Su et al., 2024) as SOTA in joint learning of structure models with sequential models. For consistency, when using ESM-2 we used the ESM-2-650M variant (Lin et al., 2023). In addition, GearNet (Zhang et al., 2023b), ESM-GearNet (Zhang et al., 2023a), and FusionProt are pre-trained with the same objective, which is the Multiview Contrast (Zhang et al., 2023b) objective, as prior work showed its superior performance.

A.1.4 Implementation Details

Pre-Training Phase We follow ESM-GearNet Zhang et al. (2023a) in adopting the same training objective and model selection strategy. Our model uses a pre-trained ESM-2-650M Lin et al. (2023) as the base PLM, with 33 layers ($L_1 = 33$), and GearNet Zhang et al. (2023b) with 6 layers ($L_2 = 6$) and 512 hidden dimensions as the structure encoder. The embedding dimension D is set to 1280. Multiview Contrast Zhang et al. (2023b) is used as the pre-training objective. Hyperparameters were tuned using the same search procedure and ranges reported in ESM-2 Lin et al. (2023) and ESM-GearNet Zhang et al. (2023a), with the best configuration selected based on validation performance. For FusionProt, this configuration corresponded to training for 50 epochs with a learning rate of 2e-4 and a global batch size of 256 proteins. To accommodate long sequences, inputs are truncated to a maximum of 1,024 tokens. All implementations use the TorchDrug library.

Fine-Tuning Phase During inference, we incorporate task-specific classification heads to generate predictions for each downstream task. Following the protocol of the recent SOTA model SaProt Su et al. (2024), we evaluated our model and baselines under a consistent hyperparameter tuning procedure to ensure fair comparison. For all methods, we performed tuning within the hyperparameter ranges reported in SaProt Su et al. (2024), with the best configuration for each model selected based on validation performance. We use AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, treating learning rate, weight decay, and batch size as tunable; default initial values and exact ranges are provided in our GitHub repository. All models were trained to convergence, and the final checkpoint was chosen by the highest validation score (task-specific primary metric).

A.1.5 Computational Complexity

FusionProt was trained using 4× NVIDIA A100 80GB GPUs over 48 hours, totaling 192 GPU hours. It retains the underlying architecture of the base sequence and structure encoders (e.g., ESM-2 and GearNet), applying each layer only once, similar to prior baselines such as ESM-GearNet (Zhang et al., 2023a). The proposed fusion mechanism introduces only lightweight additions and arithmetic operations via a token exchange module between encoders, contributing a bounded $\mathcal{O}(1)$ overhead per layer. At inference time, FusionProt incurs a marginal runtime increase of approximately 2–5% relative to SOTA baselines, with latency rising from approximately 0.012 seconds to 0.014 seconds per 1,000 residues. Consequently, the overall computational complexity of FusionProt remains comparable to existing methods in both theoretical and empirical terms.

A.2 Biological Insights

FusionProt consistently outperforms SOTA baselines across all downstream tasks (Table 1, Table 2). When such performance gains are driven by biologically meaningful inputs, such as detailed 3D structural features, they can indicate that the model is capturing relationships of functional relevance. However, we note that one must avoid conflating improved prediction with causation: a model's ability to predict biological outcomes with more precision does not in itself support causal inferences. Nevertheless, significant increases in specific

classes of predictions can provide valuable hints that frame hypotheses, guide experimental questions, and motivate follow-up studies.

For FusionProt, the largest improvements are often observed in proteins whose function, stability, or interactions depend strongly on complex structural organization or long-range spatial effects. In contrast to methods such as ESM-GearNet (Zhang et al., 2023a) or SaProt (Su et al., 2024), which integrate sequence and structure in a single pass, FusionProt employs an iterative fusion token mechanism that enables multiple rounds of bidirectional information exchange between sequence and structure encoders. This repeated refinement allows the model to amplify subtle, spatially localized features that might otherwise be diluted in one-shot fusion.

To better understand the nature of these improved learned protein representations, we evaluated a model's ability to predict EC numbers, as outlined in Section A.1.2, using pre-trained representations without any fine-tuning. This approach enables an assessment of the inherent quality of the learned embeddings, independent of task-specific training. We then compared FusionProt's predictions with those of ESM-GearNet, focusing on the cases with the largest discrepancies.

We then organized these proteins into distinct biological groups based on shared mechanistic characteristics. We focus on two common mechanisms (Table 6) and provide a representative example for each group.

Table 6: Mechanisms highlighted by our case studies, with generalizable structural signals, a representative example, and EC-probe confidence.

Biological mechanism	**Insight / structural signal**	**Example protein**	**Confidence improvement**
Assembly interface	Interface fingerprint at subunit contacts (shape/electrostatics/hydrophobicity); gains when function depends on quaternary context rather than active-site chemistry.	RNAP ω–β' interface	0.90 vs 0.21
Loop-gated pocket	Dynamic loops that gate ligand/catalytic sites; loop residues occupy broader allowed (ϕ, ψ) basins (flexibility) supporting access and specificity.	D-Ala–D-Ala ligase (ATP-grasp)	0.88 vs 0.62

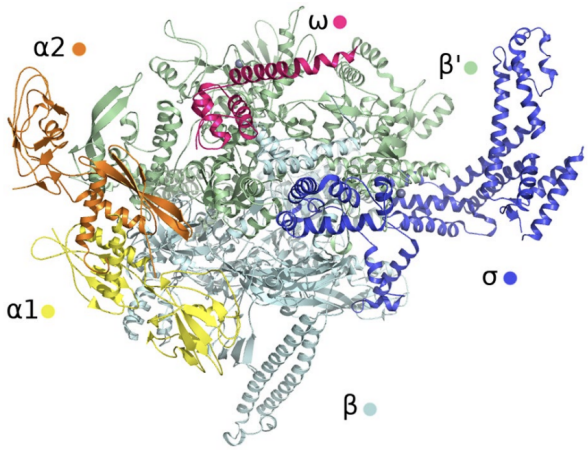

Figure 2: DNA-directed RNA polymerase highlighting the ω–β' assembly interface (ω in magenta, β' in light green) (Kurkela et al., 2021). Although ω is small and poorly conserved, it is essential for recruiting and stabilizing β' during holoenzyme assembly. FusionProt captured this interface-specific structural fingerprint and classified the correct complex EC labels with high confidence, while ESM-GearNet failed.

Structural subunits for macromolecular assembly. The bacterial RNA polymerase ω subunit (Figure 2) is small and poorly conserved, yet plays a key structural role in holoenzyme assembly by recruiting and stabilizing the β' subunit (Mathew & Chatterji, 2006). Its short length and weak sequence signal make it difficult for sequence-only models to classify correctly. FusionProt's iterative sequence–structure fusion appears to preserve features on the β' contact surface that are tied to assembly and stability, while one-pass fusion may reduce such localized signals. Consistent with this interpretation, FusionProt assigned the correct complex-level EC labels with high confidence (mean of 0.90), while the baseline ESM-GearNet (Zhang et al., 2023a) did not (mean of 0.21). Because ω is noncatalytic and the EC number reflects holoenzyme activity, we hypothesize that FusionProt recognizes the context of the quaternary structure, specifically an interface-centric fingerprint of local shape, charge, and hydrophobic patterning at the ω–β' surface, rather than active-site chemistry. Figure 2 highlights this interface (magenta: ω; light green: β'), where the model appears to localize signal more strongly than the baseline. This implies that weakening these interface features would reduce assembly propensity and lower model confidence, providing a concrete avenue for validation.

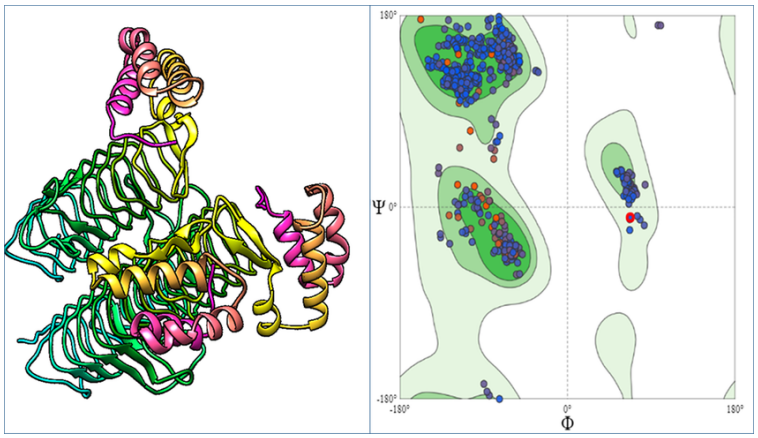

Figure 3: D-alanine–D-alanine ligase. Left: 3D structure showing the ATP-binding cleft and the surrounding pocket-adjacent loops. Right: Ramachandran plot (Saleem et al., 2021) of the dihedral angles (ϕ and ψ angles) of backbone for the same protein; shaded contours denote sterically allowed conformations. The broader spread within allowed basins is characteristic of flexible loop regions that gate the pocket, supporting our interpretation that FusionProt captures this loop-centric fingerprint and yields higher EC prediction confidence than ESM-GearNet.

Conserved ATP-dependent peptide ligases. D-alanine–D-alanine ligase (Figure 3) is a well-characterized member of the ATPgrasp enzyme superfamily, which catalyzes the formation of the D-Ala-D-Ala dipeptide, an essential step in bacterial peptidoglycan biosynthesis. Its conserved 3D fold encloses an ATP-binding site and supports relatively simple, well-understood catalytic chemistry. Both FusionProt and ESM-GearNet correctly predicted its EC classes, but FusionProt did so with higher confidence (mean of 0.88 vs. 0.62). We hypothesize that the iterative fusion mechanism amplifies local structural cues from the mobile loops surrounding the ligand-binding pocket, which mediate substrate recognition and catalysis, beyond what the global fold alone conveys. Consistent with this, the Ramachandran plot (Saleem et al., 2021) for the same structure (Figure 3) shows a broader distribution of the dihedral angles (ϕ and ψ angles) of the backbone for the loop residues, within allowed regions: hallmarks of conformational flexibility such as pocket gating and substrate specificity. This suggests that repeated sequence–structure refinement can sharpen flexible, function-critical motifs even in enzymes with highly conserved catalytic cores, implying that perturbations to these loops could influence both enzymatic activity and model confidence.

Across categories, FusionProt's unified embeddings capture structural determinants often missed in less iterative fusion frameworks, such as allosteric loops, oligomerization interfaces, and convergent active-site geometries. These enriched representations promise advances in function prediction, mutagenesis design, drug discovery, and the annotation of orphan proteins from metagenomic datasets.