Briefings in Bioinformatics, 2025, **26(1)**, bbae695
https://doi.org/10.1093/bib/bbae695
Problem Solving Protocol

ProteinF3S: boosting enzyme function prediction by fusing protein sequence, structure, and surface

Mingzhi Yuan[1,2,‡], Ao Shen[1,2,‡], Yingfan Ma[1,2], Jie Du[1,2], Bohan An[1,2], Manning Wang[1,2,*]

[1] Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, 131 Dong'an Road, 200032 Shanghai, China
[2] Shanghai Key Laboratory of Medical Image Computing and Computer Assisted Intervention, Fudan University, 131 Dong'an Road, 200032 Shanghai, China
[*] Corresponding author. Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China.
E-mail: mnwang@fudan.edu.cn
[‡] Mingzhi Yuan and Ao Shen have contributed equally to this work.

Abstract

Proteins can be represented in different data forms, including sequence, structure, and surface, each of which has unique advantages and certain limitations. It is promising to fuse the complementary information among them. In this work, we propose a framework called ProteinF3S for enzyme function prediction that fuses the complementary information across protein sequence, structure, and surface. To achieve more effective fusion, we propose a multi-scale bidirectional fusion strategy between protein structure and surface, in which the hierarchical features of a surface encoder and a structure encoder interact with each other bidirectionally. Based on these interactions, more distinctive features can be obtained. After that, we achieve further fusion by concatenating the sequence features with the features containing structure and surface information, so that better performance can be achieved. To validate our method, we conduct extensive experiments on tasks including enzyme reaction classification and enzyme commission number prediction. Our method achieves new state-of-the-art performance and shows that fusing different forms of data is effective in enzyme function prediction.

Keywords: enzyme function prediction; protein representation learning; information fusion

Introduction

Enzyme function prediction is essential for elucidating the roles of proteins in biochemical processes, which is vital for advancing disease diagnosis, therapeutic strategies [1, 2], and the design of novel pharmaceuticals [3, 4]. Traditionally, finding out proteins' enzyme function often involves costly wet experiments. In recent years, the explosion of available protein sequence and structure data fueled by high-throughput sequencing technologies [5], cryogenic electron microscopy [6], and algorithms [7] for protein structure prediction have laid the groundwork for learning-based proteins' enzyme function prediction. With the advancement of deep learning, a series of deep learning-based methods [8–10] has been applied to enzyme function prediction, yielding inspiring performance.

Proteins are composed of twenty different amino acids, which form peptide bonds through dehydration synthesis between amino acids and undergo complex folding to ultimately yield proteins with diverse functions. Consequently, numerous deep learning-based methods [11–16] have adopted different representation forms for proteins. Among them, the most intuitive one is the sequence form. This form directly models proteins as a string of discrete amino acids and employs natural language processing to understand the inherent patterns within the sequence [11, 12]. Leveraging language modeling, the encoded information of amino acid sequences is often concise, efficient, and encompasses global contextual information. However, relying solely on sequence modeling may lead to ambiguities. For instance, proteins with similar sequences may exhibit completely different structures, consequently resulting in different functions [17]. Correspondingly, protein structure, typically represented in the form of amino acid coordinates and processed by graph neural networks [13, 18], holds an advantage in protein spatial geometric modeling and has fewer ambiguities. Specifically, as depicted in Fig. 1, adjacent amino acids in Euclidean space may appear distant in sequence. Therefore, structure form can avoid sequence ambiguity to a certain extent and directly describe the protein geometric structure that is closely related to protein function. However, limited by the computational overhead, most structure-based methods often operate at the residue level rather than the atom level, leading to the neglect of some local physicochemical information. For example, the same amino acid can have different side chains, and minor disturbances in side chain angles can result in changes in protein function [14, 15]. Furthermore, many protein properties often have a relatively weak correlation with the internal structure of the protein. Therefore, the structure form is not a completely perfect representation for protein learning.

In addition to the two most common forms mentioned above, protein surface is often overlooked but complementary to the above forms. The surface of proteins directly engages in interactions with other molecules, focusing more on the external

Received: June 4, 2024. **Revised:** October 30, 2024. **Accepted:** December 18, 2024

© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

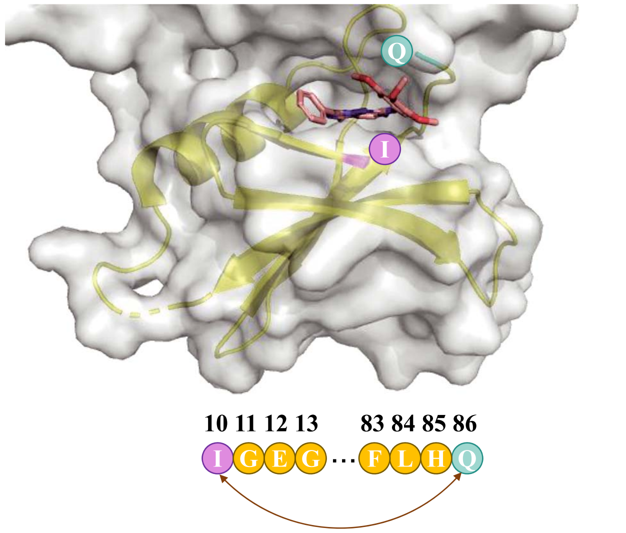


Figure 1. Illustration of the advantages of structure over sequence. Taking protein 3EOC as an example, the 10th and 86th amino acids are far apart in sequence but relatively close in actual Euclidean space, which may have a correlation in protein function. Based on sequence alone, it is difficult to infer the functional correlation between these two amino acids directly.

characteristics of proteins and providing clearer modeling of local properties at the atomic level. However, correspondingly, the protein surface overlooks the internal information of protein structure, which may lead to a lack of in-depth understanding of certain functions. For instance, many membrane proteins [16] have internal conduits that determine the specificity and selectivity of channels, thereby regulating the permeability of ions or molecules. In summary, each of the aforementioned forms has its unique advantages and limitations. Therefore, fusing the complementary information extracted from these different forms is a promising direction.

There are several existing works [19–22] involved in fusing complementary information extracted from different forms of proteins. For instance, LM-GVP [20] cascades an LSTM and a graph network to leverage the complementary information from protein sequence and structure. STEPS [23] proposes a bi-level optimization scheme for sequence and structure. HOLOPROT [21] extracts efficient protein representation by fusing features extracted from surface and structure, and proposes a superpixel algorithm to save computational overhead. Nevertheless, most existing methods merely focus on fusing two out of the three forms of protein representation (sequence, structure, and surface), rarely leveraging the complementary information from all three forms simultaneously.

To leverage these complementary information effectively, we propose a framework called **ProteinF3S** for enzyme function prediction that incorporates the complementary knowledge across protein **s**equence, **s**tructure, and **s**urface. Given a protein, we extract features for the sequence, structure, and surface using specialized networks, and then fuse them. It's crucial to note that the fusion strategy plays a pivotal role. Simple cascade or concatenation often fails to achieve optimal performance or even incurs negative effects. Therefore, we extensively explore fusion strategy. Ultimately, we propose a multi-scale bidirectional fusion strategy. Specifically, as both the structure and surface features of proteins are extracted hierarchically, we perform bidirectional fusion between structure and surface to achieve information complementary at multiple scales. Unlike unidirectional information propagation in cascade fusion, bidirectional fusion simultaneously enhances the performance of both the structure network and the surface network, thereby achieving better overall performance. In addition, we also use concatenation to fuse the protein sequence information after above multi-scale bidirectional fusion, so that the network can achieve further improvement in performance. To validate our method, we apply our ProteinF3S to protein enzyme reaction classification and enzyme commission number prediction. By simultaneously incorporating complementary information from sequence, structure, and surface, our method achieves state-of-the-art performance. Moreover, we also conduct a series of experiments to completely explore the role of fusion strategies.

Our contributions are as follows: (1) we propose a framework called ProteinF3S for enzyme function prediction, which incorporates domain knowledge across protein sequence, structure, and surface. By leveraging the complementary information across these three forms, our method gains a significant advantage in predicting proteins' enzyme function. (2) We propose a multi-scale bidirectional fusion strategy to fuse information between protein structure and surface. Based on this strategy, more effective fusion can be achieved. (3) We conduct extensive experiments on tasks including enzyme reaction classification and enzyme commission number prediction. Our method outperforms previous methods by a large margin and successfully achieves new state-of-the-art performance. (4) To explore the effectiveness of fusion strategies, we conduct a comprehensive comparison among many fusion strategies and empirically analyse their effects.

Related work

Protein representation learning

Protein representation learning aims to abstract a protein into a feature representation and use this representation to achieve a series of predictions. Typically, current learning-based enzyme function prediction tasks, such as enzyme reaction classification and enzyme commission number prediction, are primarily based on protein representation learning. Methods for protein representation learning can be categorized based on the form of input protein data. First, many methods [11, 12, 24] adopt the intuitive way to treat amino acid sequences as words and apply natural language processing techniques to understand them. For example, Protein-Bert [12] successfully applies BERT [25] to the field of protein representation learning. Since amino acid sequences are the most available form of protein data, sequence-based protein representation learning widely benefits from large-scale pre-training. For instance, the ESM [24] model used in our method has been pre-trained on a large dataset [26] with about 48 million protein sequences, thus exhibiting strong initial representation capabilities. Second, since the function of protein is directly determined by its structure, structure-based methods [9, 10, 27] have also attracted significant attention. Constrained by computational overhead, these methods typically model proteins as residue graphs and process them using graph neural networks. Additionally, there are works [8, 28] operating at the atom level. However, local characteristics and over-smoothing still hinder structure-based protein representation learning. Third, protein surface [29–33] is also widely utilized in protein representation learning because it directly encodes the external chemical and physical properties of proteins. However, many surface-based

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

Table 1. Comparison of some existing protein representation learning methods fusing multiple forms of proteins. Most of them fuse two out of the three forms of protein (sequence, structure, and surface).

	Sequence	**Structure**	**Surface**
HOLOPROT [21]		✓	✓
DeepFRI [19]	✓	✓	
LM-GVP [20]	✓	✓	
STEPS [23]	✓	✓	
GraphQA [13]	✓	✓	
CDConv [9]	✓	✓	
MASSA [22]	✓	✓	
ProteinF3S (Ours)	✓	✓	✓

methods [31, 32] are constrained by the surface building, as generating surfaces using software [34] often requires significant time. To address this challenge, dMaSIF [29] proposed a rapid method for online surface building. We utilize this method to generate protein surface in the form of point cloud. In addition to the aforementioned methods, there are many methods [19, 21, 22] that take more than one form of protein data as input. Typically, due to the complementary information within input data, these methods often achieve better performance.

Information fusion in protein representation learning

As analysed above, fusing information from different forms of protein is a promising direction. As shown in Table 1, we enumerate several methods [9, 13, 19–23] that fuse multiple forms of protein. Most methods fuse protein sequence and structure but neglect the surface. In our work, we reveal that surface also plays important roles in protein representation learning and fuse all these three forms. Moreover, the fusion strategies are rarely explored, despite their significant roles. Most methods follow a unidirectional fusion paradigm. For example, DeepFRI [19], LM-GVP [20], and HOLOPROT [21] follows a cascade paradigm, which transfer the knowledge from sequence/surface to structure. CDConv [22] implicitly encodes sequence information into the structure via a variable convolution kernel, which is determined by current regular displacements and continuous displacements. Unlike previous work, we propose a bidirectional fusion strategy in our ProteinF3S. Information is no longer propagated unidirectionally, allowing both surface features and structure features to benefit, ultimately leading to better performance.

Materials and methods

In this section, we first provide an overview of our ProteinF3S framework in Section Overview of ProteinF3S. Following that, we introduce specialized encoders for structure, surface, and sequence in Section Encoders. Subsequently, we provide the details of our fusion strategies in Section Fusion strategy. Finally, we introduce the loss and implementation details for tasks including enzyme reaction classification and enzyme commission number prediction.

Overview of ProteinF3S

As shown in Fig. 2, our ProteinF3S takes a raw protein in the form of a PDB file as input and converts it into structure, surface, and sequence. These three forms of data are separately input into three specialized encoders. Since both the structure and surface encoders follow hierarchical structures, we employ a multi-scale bidirectional fusion strategy between the structure encoder and the surface encoder. Through multi-scale bidirectional fusion, the complementary information from structure and surface can continuously enrich the features extracted by both encoders, thereby generating more powerful features at higher levels. As the sequence encoder adopts a global feature interaction rather than a hierarchical manner and truncates the sequence, it is difficult to align the sequence with the structure or surface. Therefore, we use concatenation to fuse the sequence feature $f^{\text{seq}} \in \mathbb{R}^d$ with the high-level feature $f^{\text{ss}} \in \mathbb{R}^d$ obtained from the previous multi-scale bidirectional fusion, resulting in the final feature representation $f^{\text{final}} \in \mathbb{R}^d$ used for tasks such as enzyme reaction classification. The feature representation is input into a classifier i.e. a multi-layer perceptron, which ultimately outputs the prediction result.

Encoders

To effectively extract feature representations for different forms of protein, we utilize three specialized encoders for protein structure, surface, and sequence, respectively. The structure and surface encoders adopt a hierarchical manner, and they are symmetric with respect to each other. The sequence encoder follows a Transformer [35] architecture, taking amino acid sequences as input and extracting global representations of the sequences. The details of each encoder is described below.

Structure encoder

We employ CDConv [9] as our structure encoder. The core of CDConv lies in its convolution kernel which is simultaneously determined by current regular displacements and continuous displacements. Specifically, given an amino acid, which occupies the tth position in the sequence, with its current feature at the lth scale denoted as $f_t^{\text{struct},l} \in \mathbb{R}^d$, CDConv employs a (3+1) convolution kernel to update its feature:

$$f_t^{\text{struct},l'} = \sum_{\substack{\|p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}\| \le r, \\ -\lfloor s/2 \rfloor \le \Delta \le \lfloor s/2 \rfloor}} g^{\text{struct}}\left(p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}; \theta_\Delta\right) f_{t+\Delta}^{\text{struct},l} \tag{1}$$

where $f_t^{\text{struct},l'}$ denotes the updated feature corresponding to the tth amino acid at the lth scale, $g^{\text{struct}}\left(p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}; \theta_\Delta\right)$ denotes the kernel function determined by the different displacements Δ and relative position coordinates $p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}$, $\{\theta_\Delta\}$ denotes the different weights for different displacements, r and s denote the 3D geometric radius and 1D sequential kernel size. To ensure the SE(3)-invariance in updating features, the kernel function $g^{\text{struct}}(\cdot)$ is specifically designed as follows:

$$\begin{aligned} g^{\text{struct}}\left(p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}; \theta_\Delta\right) = \theta_\Delta \cdot \Bigg(& \left\|p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}\right\|, \\ & O_t^T \cdot \frac{p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}}{\left\|p_{t+\Delta}^{\text{struct}} - p_t^{\text{struct}}\right\|},\ O_t^T \cdot O_{t+\Delta} \Bigg) \end{aligned} \tag{2}$$

where $O_t = \left(b_t^{\text{struct}}, n_t^{\text{struct}}, b_t^{\text{struct}} \times n_t^{\text{struct}}\right)$ denotes a local reference frame determined by local coordinates $\left\{p_{t-1}^{\text{struct}}, p_t^{\text{struct}}, p_{t+1}^{\text{struct}}\right\}$ and $u_t^{\text{struct}} = \frac{p_t^{\text{struct}} - p_{t-1}^{\text{struct}}}{\|p_t^{\text{struct}} - p_{t-1}^{\text{struct}}\|}$, $b_t^{\text{struct}} = \frac{u_t^{\text{struct}} - u_{t+1}^{\text{struct}}}{\|u_t^{\text{struct}} - u_{t+1}^{\text{struct}}\|}$, $n_t^{\text{struct}} = \frac{u_t^{\text{struct}} \times u_{t+1}^{\text{struct}}}{\|u_t^{\text{struct}} \times u_{t+1}^{\text{struct}}\|}$.

Based on the above convolution kernel and mean pooling, our encoder constructs a series of blocks to extract protein structure features at multiple scales.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

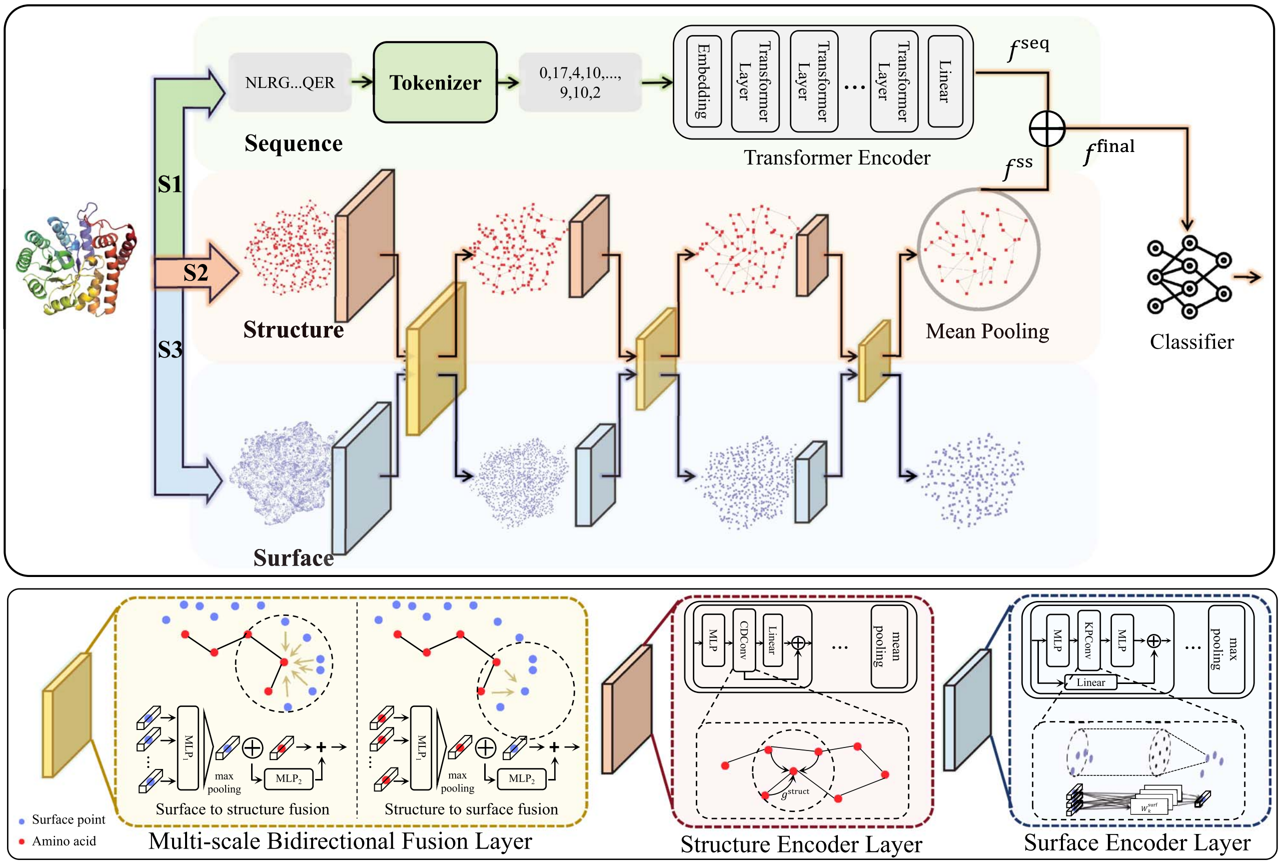


Figure 2. Overview of our ProteinF3S. The input protein is transformed into three distinct forms: sequence, structure, and surface, each processed by specialized encoders. During the encoding of structure and surface, a multi-scale bidirectional fusion is employed within the extracted hierarchical features, thereby allowing the features in the two encoders to benefit from complementary information simultaneously. Finally, the high-level feature containing structure and surface information is concatenated with the feature extracted by the sequence encoder, generating protein feature representation used for specific tasks such as classification.

Surface encoder

Due to the considerable time for constructing protein surfaces in mesh format, we employ the point cloud-based protein surface construction method from dMaSIF [29]. This method circumvents complex computations such as surface electrostatics, retaining only the raw atom type encoding, and can achieve online surface construction based on smooth distance function.

For constructed protein surface point clouds, we utilize a KPConv-based network [36] to extract hierarchical features. This network has almost the same architecture as our structure encoder, except for the convolution kernel. Given a point with coordinate $p_i^{\mathrm{surf},l} \in \mathbb{R}^3$ and feature $f_i^{\mathrm{surf},l} \in \mathbb{R}^d$ in a constructed protein surface at the lth scale, the surface encoder employs the following convolution kernel to update its feature:

$$f_i^{\mathrm{surf},l'} = \sum_{\left\|p_{i+\Delta}^{\mathrm{surf}} - p_i^{\mathrm{surf}}\right\| \le r} g^{\mathrm{surf}}\left(p_{i+\Delta}^{\mathrm{surf}} - p_i^{\mathrm{surf}}\right) f_{i+\Delta}^{\mathrm{surf},l} \qquad (3)$$

where $f_i^{\mathrm{surf},l'}$ denotes the updated feature, $g^{\mathrm{surf}}(\cdot)$ denotes the kernel function, r denotes the neighboring radius of p_i^{surf}. Slightly different from the kernel function in the structure encoder, $g^{\mathrm{surf}}(\cdot)$ does not assign different weights for different displacements. Instead, it provides K^{surf} kernel points, each associated with a set of weights $\mathrm{W}_k^{\mathrm{surf}}$. Specifically, the kernel function is represented as follows:

$$g^{\mathrm{surf}}\left(p_{i+\Delta}^{\mathrm{surf}} - p_i^{\mathrm{surf}}\right) = \sum_{k<K^{\mathrm{surf}}} \max\left(0, 1 - \frac{\left\|p_{i+\Delta}^{\mathrm{surf}} - p_i^{\mathrm{surf}} - \tilde{p}_k^{\mathrm{surf}}\right\|}{\sigma}\right) \mathrm{W}_k^{\mathrm{surf}} \qquad (4)$$

where $\tilde{p}_k^{\mathrm{surf}}$ denotes kth the kernel point and $\mathrm{W}_k^{\mathrm{surf}}$ is its corresponding weight, σ is a parameter to control the influence distance of kernel points.

Consistent with the structure encoder, our surface encoder employs alternating blocks and downsampling to extract features at multiple scales. At each scale, we perform feature fusion between features from these two encoders, details will be provided in Section Fusion strategy.

Sequence encoder

Compared to protein structure and surface, protein sequence data are more accessible. Therefore, many sequence-based methods [12, 24] benefit from pre-training on large datasets of protein sequences. Previous studies [11, 24] have also indicated that these pre-trained sequence-based networks provide effective representations for proteins. Hence, we adopt ESM [24] as our sequence encoder. It is worth noting that the quadratic complexity of the self-attention in the Transformer necessitates truncation of the input protein sequences. Consequently, our sequence encoder cannot perfectly align with other encoders at the amino acid

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

level due to potential partial omissions. To address this issue, we directly utilize the global feature representation extracted by ESM rather than dense amino acid level representations. Additionally, due to the large parameter size of ESM, fine-tuning ESM requires significant memory overhead. Therefore, our sequence encoder utilizes ESM by the linear probe. The linear layer ultimately outputs the overall feature representation $f^{\text{seq}} \in \mathbb{R}^d$ of the protein sequence, which is used for subsequent fusion.

Fusion strategy

As shown in Fig. 2, we conducted two fusion processes. First, a multi-scale bidirectional fusion is performed between the structure encoder and the surface encoder. Subsequently, we further fuse the feature obtained from the previous fusion using concatenation with the feature extracted by our sequence encoder, generating the final feature for specific tasks.

Multi-scale bidirectional fusion

As previously analysed, protein structure and surface contain complementary information. Therefore, fusing them holds promise. To fully leverage the complementary information, we propose a multi-scale bidirectional fusion strategy. Unlike direct concatenation or cascade, we conduct bidirectional fusion at each scale. Through this bidirectional fusion, structure and surface features can acquire complementary information from each other and evolve into more distinctive features. Furthermore, conducting fusion at multiple scales allows both structure and surface features to be enhanced early on, thereby generating more prominent features in the final stage. Specifically, given an amino acid with coordinate $p_t^{\text{struct}} \in \mathbb{R}^3$ and feature $f_t^{\text{struct},l} \in \mathbb{R}^d$, and a surface point with coordinate $p_i^{\text{surf}} \in \mathbb{R}^3$ and feature $f_i^{\text{surf},l} \in \mathbb{R}^d$, we update the structure feature $f_t^{\text{struct},l}$ and surface feature f_i^{surf} at the lth scale according to the following formulas:

$$f_i^{\text{surf},l} \leftarrow f_i^{\text{surf},l} + MLP^l_{\text{struct2surf}_2} \left(\max \sum_{\left\| p_i^{\text{surf}} - p_j^{\text{struct}} \right\| \le r} MLP^l_{\text{struct2surf}_1} \left(f_i^{\text{surf},l} \oplus f_j^{\text{struct},l} \right) \right) \quad (5)$$

$$f_t^{\text{struct},l} \leftarrow f_t^{\text{struct},l} + MLP^l_{\text{surf2struct}_2} \left(\max \sum_{\left\| p_t^{\text{struct}} - p_j^{\text{surf}} \right\| \le r} MLP^l_{\text{surf2struct}_1} \left(f_t^{\text{struct},l} \oplus f_j^{\text{surf},l} \right) \right) \quad (6)$$

where $\oplus$ denotes the concatenation operation, r denotes the neighboring radius of the multi-scale bidirectional fusion, $MLP^l_{\text{struct2surf}_1}(\cdot)$ denotes a multilayer perceptron taking the concatenation of feature $f_i^{\text{surf},l}$ and its neighboring feature $f_j^{\text{struct},l}$ as input, and $MLP^l_{\text{surf2struct}_1}(\cdot)$ similarly, $\max(\cdot)$ denotes a max pooling layer to achieve permutation-invariant aggregation within neighbor, and $MLP^l_{\text{struct2surf}_2}(\cdot)$ and $MLP^l_{\text{surf2struct}_2}(\cdot)$ denote the multilayer perceptrons for aggregated features. As shown in Fig. 2, our bidirectional fusion thoroughly considers the spatial positions between amino acids and surface points, selecting those close amino acids and surface points for fusion. For certain internal amino acids that may lack spatially adjacent surface points, we substitute a learnable feature for neighboring features.

Finally, after L layers of fusion, we input the features $\left\{ f_i^{\text{struct},L} \right\}$, which contain both structure and surface information, into a mean pooling layer to obtain a feature representation $f^{\text{ss}} \in \mathbb{R}^d$. This feature will be fused with the sequence representation $f^{\text{seq}} \in \mathbb{R}^d$ and utilized for subsequent specific tasks.

Concatenation fusion

Having obtained the feature representation f^{seq} for sequence and the feature representation f^{ss} for structure and surface, we employ a concatenation operation to fuse them and obtain the final protein representation $f^{\text{final}} = f^{\text{seq}} \oplus f^{\text{ss}}$. For the selected tasks including enzyme reaction classification and enzyme commission number prediction, this feature representation f^{final} will be input into a classifier to generate the final prediction.

Loss and implementation

Since the subsequent tasks, enzyme reaction classification and enzyme commission number prediction are both classification tasks, we utilize cross-entropy loss as the loss function [37]. Our ProteinF3S is implemented in PyTorch and PyTorch-Geometric. All the experiments are conducted on a single A100 graphic card. For more details, please refer to the supplementary materials.

Results and discussion

In this section, we conduct experiments on two tasks including enzyme reaction classification and enzyme commission number prediction to verify the effectiveness of our method. Furthermore, we also conduct comprehensive ablation studies to demonstrate the effectiveness of protein representation fusion and our proposed fusion strategy.

Enzyme reaction classification

Dataset. The enzyme reaction classification is valuable for bioinformatics and machine learning research. It facilitates the development of algorithms for the automatic classification of enzyme reactions based on enzyme and substrate characteristics. With information on reaction types and associated Enzyme Commission (EC) numbers, the dataset [38] enables the training of models to predict enzyme function and reaction specificity. We follow the split in [8].

Evaluation metric. As enzyme reaction classification is a single label classification task with 384 classes, we follow [9] to utilize accuracy as the evaluation metric.

Competitors. For a comprehensive evaluation, we compare a series of competitors, including several state-of-the-art methods. Some methods [8, 10, 27, 28, 39–43] take protein sequences or structures as input. The remaining competitors [9, 13, 20, 21] utilize two of the sequence, structure, and surface as inputs. Unlike them, our ProteinF3S takes all three simultaneously as inputs.

Results. Table 2 illustrates the results of enzyme reaction classification. It is evident that our method surpasses all others, achieving the best performance. This is attributed to our fusion, which fully leverages the complementary information. Further discussions regarding inputs and fusion strategies will be conducted in detail in the ablation studies.

Enzyme commission number prediction

Dataset. Enzyme commission number prediction is a task in bioinformatics aiming to predict the EC numbers for enzymes based on their sequences or structures. It involves using machine learning and bioinformatics techniques to classify enzymes into different EC number categories, facilitating the understanding of their functions and catalytic activities. This task is crucial for elucidating enzyme functionalities, metabolic pathways, and aiding

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

Table 2. Enzyme reaction classification results. We utilize bold number to indicate the best results.

Method	Accuracy (%)
CNN [39]	51.7
ResNet [40]	24.1
LSTM [40]	11.0
Transformer [40]	26.6
GCN [41]	67.3
GAT [42]	55.6
3D CNN [43]	72.2
GraphQA [13]	60.8
GVP-GNN [27]	65.5
IEConv [8]	87.2
HOLOPROT [21]	78.9
ProtNet [28]	86.4
GearNet [10]	79.4
GearNet-IEConv [10]	83.7
GearNet-Edge [10]	86.6
GearNet-Edge-IEConv [10]	85.3
CDConv [9]	88.5
Ours	**89.2**

Table 3. Enzyme commission number prediction results (< 95% sequence identity to the training set). We utilize bold number to indicate the best results.

Method	F_{max}
CNN [39]	54.5
ResNet [40]	60.5
LSTM [40]	42.5
Transformer [40]	23.8
GCN [41]	32.0
GAT [42]	36.8
3D CNN [43]	7.7
GraphQA [13]	50.9
GVP-GNN [27]	48.9
IEConv [8]	73.5
HOLOPROT [21]	–
ProtNet [28]	–
GearNet [10]	73.0
GearNet-IEConv [10]	80.0
GearNet-Edge [10]	81.0
GearNet-Edge-IEConv [10]	81.0
CDConv [9]	82.0
Ours	**87.3**

in various biotechnological and pharmaceutical applications. We utilize the dataset and cutoff splits in [18].

Evaluation metric. Since enzyme commission number prediction is a multi-label classification problem, which can be regarded as 538 binary classification problems, we adopt the Protein-centric maximum F-Score [19] i.e. $F_{\max}$ as the evaluation metric, following the settings in [19]. $F_{\max}$ is a metric designed for measuring the accuracy of multi-label classification. It is defined as follows:

$$F_{\max} = \max_{\lambda \in [0,1]} \left\{ \frac{2 \times \mathbf{precision}(\lambda) \times \mathbf{recall}(\lambda)}{\mathbf{precision}(\lambda) + \mathbf{recall}(\lambda)} \right\} \tag{7}$$

where λ is a decision threshold, $\mathbf{precision}(\lambda)$ and $\mathbf{recall}(\lambda)$ are the average precision and recall for all binary classifications based on the decision threshold. More details on the definitions can be found in the supplementary material.

Competitors. We use the same competitors as for the enzyme reaction classification task and mainly report their results with no more than 95% sequence identity to the training set.

Results. Since the 'no more than 95% sequence identity' condition has been paid the most attention in previous comparisons, we first compared our method with various other methods under this setting. As shown in Table 3, our method significantly outperforms others and achieves state-of-the-art performance. Additionally, we investigated performance under different cutoffs and compared it with competitive methods like CDConv [9] and GearNet [10] in Table 4. Our method achieves superior performance across all cutoffs, demonstrating its effectiveness.

Ablation study

In this section, we conduct comprehensive ablation studies to verify the effectiveness of our fusion strategy. First, we conduct an ablation on the input forms of protein to illustrate the importance of fusion. In this ablation, we focus on the network's performance when using different combinations of protein forms as inputs, with each combination adopting the optimal fusion strategy. Specifically, we utilize concatenation for fusing sequence with any other form, utilize multi-scale bidirectional fusion for fusing surface and structure, and adopt ProteinF3S as the fusion strategy for the combination of all forms. Additionally, we also conduct an ablation on the fusion strategies to demonstrate the significance of fusion strategies. We compare our multi-scale bidirectional fusion strategy with other fusion strategies such as cascade and concatenation. All ablation studies were conducted on the enzyme reaction classification task.

Table 4. Enzyme commission number prediction results under different cutoffs. We utilize bold number to indicate the best results.

Method\Cutoff	30%	40%	50%	70%	95%
CNN [39]	36.6	36.1	37.2	42.9	54.5
ResNet [40]	40.9	41.2	45.0	52.6	60.5
LSTM [40]	24.7	24.9	27.0	33.3	42.5
Transformer [40]	16.7	17.3	17.5	19.7	23.8
GCN [41]	24.5	24.6	24.6	28.0	32.0
GearNet [10]	55.7	57.0	61.5	69.3	73.0
GearNet-Edge [10]	62.5	64.6	69.4	75.7	81.0
CDConv [9]	63.4	65.9	70.2	76.8	82.0
Ours	**75.1**	**77.2**	**80.7**	**84.6**	**87.3**

As previously analysed, protein sequence, structure, and surface contain complementary information. Therefore, fusing different forms of proteins has potential. To validate this, we test the performance of sequence, structure, and surface encoder, as well as the performance of their combinations. The results are shown in Table 5. Both the sequence encoder and structure encoder perform remarkably well. The former benefits from pre-training on large-scale data, while the latter achieves good modeling of geometric structures. The performance of the surface encoder is relatively limited, consistent with previous results in [21]. Notably, any combination of two encoders yields better performance than before, demonstrating that different forms of protein representations contain complementary information. Leveraging the complementary information from all three forms simultaneously yields the best performance.

We also present a series of specific proteins to illustrate the importance of fusion. First, as shown in Fig. 4(a), for protein

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

Table 5. Influence on input forms of protein. We utilize bold number to indicate the best results.

Input	Accuracy (%)
Sequence	87.6
Structure	88.5
Surface	72.6
Sequence + Structure	88.9
Sequence + Surface	87.9
Structure + Surface	88.9
Sequence + Structure + Surface	**89.2**

Table 6. Comparison on fusion strategies between protein structure and surface. CAT denotes fusion using concatenation. MSF denotes multi-scale fusion. Struct2surf denotes unidirectional fusion from the structure encoder to the surface encoder. Surf2Struct denotes unidirectional fusion from the surface encoder to the structure encoder. We utilize bold number to indicate the best results. Wo Res denotes update fused features without residue design.

Fusion Strategy	Input	Accuracy (%)
–	Structure	88.5
–	Surface	72.6
CAT	Structure + Surface	87.9
Cascade (HOLOPROT [21])	Structure + Surface	88.4
MSF + Struct2Surf	Structure + Surface	86.9
MSF + Surf2Struct	Structure + Surface	88.3
MSF + Bidirectional (wo Res)	Structure + Surface	88.4
MSF + Bidirectional	Structure + Surface	**88.9**

Table 7. Time cost on enzyme reaction classification task. We utilize bold number to indicate the best results.

Input	Accuracy (%)	Time (ms/batch)
Structure	88.5	**43**
Sequence	87.6	298
Surface	72.6	190
Structure + Sequence + Surface	**89.2**	309

surface, we take protein 1a0i_a as an example, which is a part of a protease responsible for cleaving specific peptide bonds. Its active site is typically deeply buried within the protein, relying on its structure to form a precise catalytic pocket. The enzyme's function is closely linked to the internal structure of the substrate binding site, making surface features insufficient to capture the key attributes of its active center. Thus, the surface encoder fails to predict its enzyme reaction class accurately. Similarly, many enzymes exhibit characteristics tied to internal and external protein structures, which further explains why the surface encoder underperforms in this task. Second, for protein structure, we take protein 1o6z_C in Fig. 4(b) as an example. Its highly complex fold, particularly with auxiliary structures like helices or loops irrelevant to the reaction, can introduce additional noise into the classification process. This complexity hampers the structure encoder from capturing the core functional characteristics of the enzyme's reaction. In contrast, sequence and surface forms offer more concise information. As a result, both surface and sequence encoders succeed in prediction, whereas the structure encoder fails. Although the structure form might appear to offer the best overall performance, it still exhibits certain shortcomings. Finally, for protein sequence, we refer to protein 5jis_d in Fig. 4(c), which participates in the glycolysis pathway. The enzyme's active site relies on the three-dimensional fold (with red representing α-helices and yellow representing β-sheets). Since the sequence alone cannot directly reflect the protein's fold and spatial structure, predictions based on sequence features are limited in this case. Overall, each form of protein representation has its own strengths and weaknesses, and fusing multiple protein forms proves both reasonable and effective.

Furthermore, we conducted a comprehensive ablation on fusion strategies, as there are many fusion strategies available. Here, we focus on the fusion between structure and surface, as the sequence encoder extracts a global representation of the protein, allowing only concatenation fusion and no finer-grained fusion. We compared concatenation, cascade, and unidirectional multi-scale fusion, with the data flow during fusion illustrated in Fig. 3. The results are shown in Table 6. It can be observed that conventional cascade and concatenation fusion do not yield positive effects. This may be because the performance of the surface encoder in this task is relatively limited, dragging down the structure encoder. The performance of surf2struct outperforming struct2surf in unidirectional fusion also confirms this point. However, our multi-scale bidirectional fusion achieves positive effects, demonstrating the significance of fusion strategies. Additionally, the bottom two rows in Table 6 also demonstrate the necessity of residual design during fusion. The design of residuals enables selective fusion, allowing the transmission of complementary information at reasonable scales and directions.

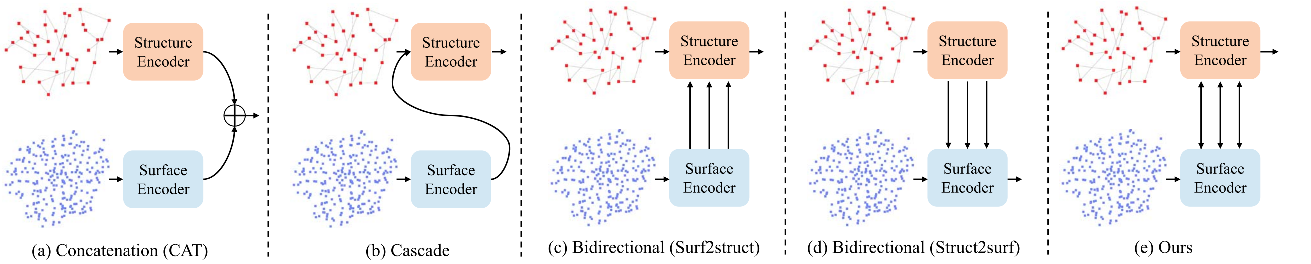
Discussion

Despite achieving state-of-the-art performance, ProteinF3S still faces some limitations. First, the model requires three different forms of protein data as inputs simultaneously. However, there may be instances of data corruption or absence. For example, in attempting other tasks such as gene ontology term prediction, data corruption impedes the construction of protein surfaces, making it difficult for ProteinF3S to accomplish the task. Solutions for handling incomplete inputs still require further exploration. Second, fusing different forms of protein inevitably increases computational cost. We record the inference times of the structure encoder, sequence encoder, surface encoder, and our ProteinF3S in Table 7. As shown, although ProteinF3S excels in accuracy, it struggle to strike a better balance between accuracy and computational cost than the structure encoder. Thus, in scenarios with limited computational resources, our ProteinF3S may not be the most suitable, but it is better suited for scenarios that prioritize accuracy. Third, as fine-tuning ESM incurs significant computational overhead, ProteinF3S utilizes a linear probe strategy for transfer learning, yielding commendable results. However, fine-tuning or other transfer strategies [44] still merit further exploration.

Conclusion

In this work, we analyse the complementary information existing among protein sequence, structure, and surface and propose a framework which fuses these three protein representation forms.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025



(a) Concatenation (CAT)

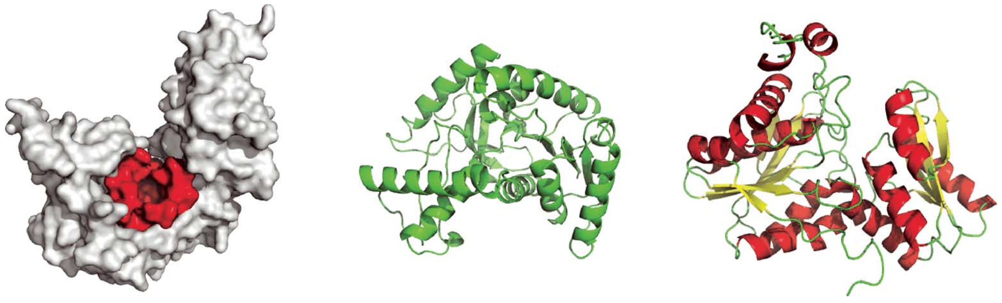
(b) Cascade

(c) Bidirectional (Surf2struct)

(d) Bidirectional (Struct2surf)

(e) Ours

Figure 3. Illustrations of different fusion strategies between protein structure and surface.



(a) 1a0i_a

(b) 1o6z_c

(c) 5jis_d

Figure 4. Misclassified cases in enzyme reaction classification. (a) The surface encoder fails, while both structure and sequence encoders, as well as our ProteinF3S, predict correctly. We color the catalytic pocket in red; (b) The structure encoder fails, but the other encoders and ProteinF3S succeed; (c) The sequence encoder fails, but the other encoders and ProteinF3S succeed. We color the α-helices in red and β-sheets in yellow.

Different from most previous works, our ProteinF3S does not neglect any of these three representation forms. Moreover, we also propose a multi-scale fusion strategy for protein structure and surface to further leverage the complementary information among them. The extensive experiments demonstrate that fusing different forms of proteins can indeed improve performance and that the fusion strategy also plays a vital role. Based on the complementary information and our effective multi-scale bidirectional fusion strategy, our ProteinF3S successfully achieves new state-of-the-art performance on enzyme reaction classification and enzyme commission number prediction tasks. Furthermore, we hope to extend our ProteinF3S to more protein tasks and generalize the fusion strategy to molecular representation learning [45] in the future.

Key Points

- We propose a framework called ProteinF3S for enzyme function prediction, which incorporates domain knowledge across protein sequence, structure, and surface. By leveraging the complementary information across these three forms, our method gains a significant advantage in predicting proteins' enzyme function.
- We propose a multi-scale bidirectional fusion strategy to fuse information between protein structure and surface. Based on this strategy, more effective fusion can be achieved.
- We conduct extensive experiments on tasks including enzyme reaction classification and enzyme commission number prediction. Our method outperforms the previous methods by a large margin and successfully achieves new state-of-the-art performance.
- To explore the effectiveness of fusion strategies, we conduct a comprehensive comparison among many fusion strategies and empirically analyse their effects.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the Science and Technology Innovation Plan Of Shanghai Science and Technology Commission 23S41900400 and AI for Science Foundation of Fudan University.

Data availability

Our code and data will be made public released at https://github.com/phdymz/ProteinF3S.

Author contributions statement

Mingzhi Yuan proposed the main idea and completed the main experiments and main writing. Ao Shen assisted at the dataset processing stage. Yingfan Ma provided help with image polishing and typesetting. Jie Du and Bohan An completed the proofreading of the paper. Manning Wang polished the paper and provided funding.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

References

1. Chou K-C, Shen H-B. *et al.* Recent advances in developing web-servers for predicting protein attributes. *Nat Sci* 2009;**01**:63–92. https://doi.org/10.4236/ns.2009.12011.
2. Sharma M, Garg P. Computational approaches for enzyme functional class prediction: a review. *Curr Proteom* 2014;**11**:17–22. https://doi.org/10.2174/1570164611666140415225013.
3. Shaker B, Ahmad S, Lee J. *et al.* In silico methods and tools for drug discovery. *Comput Biol Med* 2021;**137**:104851. https://doi.org/10.1016/j.compbiomed.2021.104851.
4. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. *Brief Bioinform* 2009;**10**:579–91. https://doi.org/10.1093/bib/bbp023.
5. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010;**11**:31–46. https://doi.org/10.1038/nrg2626.
6. Dubochet J, Adrian M, Chang J-J. *et al.* Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys* 1988;**21**:129–228. https://doi.org/10.1017/S0033583500004297.
7. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
8. Hermosilla P, Schäfer M, Lang M. *et al.* Intrinsic-extrinsic convolution and pooling for learning on 3D protein structures. *International Conference on Learning Representations*. Virtual Event: OpenReview.net, 2021.
9. Fan H, Wang Z, Yang Y. *et al.* Continuous-discrete convolution for geometry-sequence modeling in proteins. In: *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview, 2022.
10. Zhang Z, Xu M, Jamasb A. *et al.* Protein representation learning by geometric structure pretraining. *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview, 2023.
11. Rives A, Meier J, Sercu T. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**:e2016239118. https://doi.org/10.1073/pnas.2016239118.
12. Brandes N, Ofer D, Peleg Y. *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10. https://doi.org/10.1093/bioinformatics/btac020.
13. Baldassarre F, Hurtado DM, Elofsson A. *et al.* GraphQA: protein model quality assessment using graph convolutional networks. *Bioinformatics* 2021;**37**:360–6. https://doi.org/10.1093/bioinformatics/btaa714.
14. Jacobson MP, Friesner RA, Xiang Z. *et al.* On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol* 2002;**320**:597–608. https://doi.org/10.1016/S0022-2836(02)00470-9.
15. Agrawal V, Kishan RKV. Functional evolution of two subtly different (similar) folds. *BMC Struct Biol* 2001;**1**:1–6. https://doi.org/10.1186/1472-6807-1-5.
16. Catterall WA. Voltage-gated sodium channels at 60: structure, function and pathophysiology. *J Physiol* 2012;**590**:2577–89. https://doi.org/10.1113/jphysiol.2011.224204.
17. Alexander PA, He Y, Chen Y. *et al.* A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci* 2009;**106**:21149–54. https://doi.org/10.1073/pnas.0906408106.
18. Jing B, Eismann S, Suriana P. *et al.* Learning from protein structure with geometric vector perceptrons. In: *International Conference on Learning Representations*. Virtual Event: OpenReview.net, 2020.
19. Vladimir, Gligorijević P, Renfrew D, Kosciolek T. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:1–14. https://doi.org/10.1038/s41467-021-23303-9.
20. Wang Z, Combs SA, Brand R. *et al.* LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. *Sci Rep* 2022;**12**:6832. https://doi.org/10.1038/s41598-022-10775-y.
21. Somnath VR, Bunne C, Krause A. Multi-scale representation learning on proteins. *Adv Neural Inf Process Syst* 2021;**34**:25244–55.
22. Fan H, Yishen H, Zhang W. *et al.* A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks. *Adv Sci* 2023;**10**:2301223. https://doi.org/10.1002/advs.202301223.
23. Chen C, Zhou J, Wang F. *et al.* Structure-aware protein self-supervised learning. *Bioinformatics* 2023;**39**:btad189. https://doi.org/10.1093/bioinformatics/btad189.
24. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574.
25. Devlin J, Chang M-W, Lee K. *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of naacL-HLT*, vol. **1**, pp. 2. Minneapolis, USA: Association for Computational Linguistics, 2018.
26. Suzek BE, Wang Y, Huang H. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;**31**:926–32. https://doi.org/10.1093/bioinformatics/btu739.
27. Jing B, Eismann S, Suriana P. *et al.* Learning from protein structure with geometric vector perceptrons. In: *International Conference on Learning Representations*. Virtual Event: OpenReview.net, 2020.
28. Wang L, Liu H, Liu Y. *et al.* Learning hierarchical protein representations via complete 3D graph networks. *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview, 2022.
29. Sverrisson F, Feydy J, Correia BE. *et al.* Fast end-to-end learning on protein surfaces. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15272–81. Virtual Event: IEEE, 2021.
30. Yuan M, Shen A, Kexue F. *et al.* ProteinMAE: masked autoencoder for protein surface self-supervised learning. *Bioinformatics* 2023;**39**:btad724. https://doi.org/10.1093/bioinformatics/btad724.
31. Gainza P, Sverrisson F, Monti F. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184–92. https://doi.org/10.1038/s41592-019-0666-6.
32. Wang Y, Shen Y, Chen S. *et al.* Learning harmonic molecular representations on riemannian manifold. *The Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview, 2023.
33. Shen A, Yuan M, Ma Y. *et al.* SS-PRO: a simplified siamese contrastive learning approach for protein surface representation. *Front Comp Sci* 2024;**18**:185910. https://doi.org/10.1007/s11704-024-3806-9.
34. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 2010;**26**:2064–5. https://doi.org/10.1093/bioinformatics/btq322.
35. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

36. Thomas H, Qi CR, Deschaud J-E. *et al.* KPConv: flexible and deformable convolution for point clouds. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6411–20. Long Beach, California, USA: IEEE, 2019.
37. Bishop CM. *Pattern Recognition and Machine Learning*, vol. **4**, pp. 1122–8. Springer Google Scholar, 2006.
38. Loening KL. The terminology of biotechnology: a multidisciplinary problem. In: *Proceedings of the 1989 International Chemical Congress of Pacific Basin Societies PACIFICHEM*, Vol. **89**. Berlin: Springer, 1990.
39. Shanehsazzadeh A, Belanger D, Dohan D. Is transfer learning necessary for protein landscape prediction? arXiv preprint arXiv:2011.03443. 2020.
40. Rao R, Bhattacharya N, Thomas N. *et al.* Evaluating protein transfer learning with tape. *Adv Neural Inf Process Syst* 2019;**32**: 9689–701.
41. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*. Toulon, France: OpenReview, 2016.
42. Velickovic P, Cucurull G, Casanova A. *et al.* Graph attention networks. *Stat* 2017;**1050**:10–48550.
43. Derevyanko G, Grudinin S, Bengio Y. *et al.* Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* 2018;**34**:4046–53. https://doi.org/10.1093/bioinformatics/bty494.
44. Hu EJ, Shen Y, Wallis P. *et al.* LoRA: low-rank adaptation of large language models. *International Conference on Learning Representations*. Virtual Event: OpenReview.net, 2022.
45. Shen A, Yuan M, Ma Y. *et al.* Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction. *Brief Bioinform* 2024;**25**:bbae256. https://doi.org/10.1093/bib/bbae256.

Downloaded from https://academic.oup.com/bib/article/26/1/bbae695/7942219 by guest on 25 April 2025

© The Author(s) 2025. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Briefings in Bioinformatics, 2025, **26(1)**, bbae695
https://doi.org/10.1093/bib/bbae695
Problem Solving Protocol