S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure

Duolin Wang[1], Usman L Abbas[2], Qing Shao[2]*, Jin Chen[3]* and Dong Xu[1]*

[1] Department of Electrical Engineering and Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

[2]Chemical & Materials Engineering, University of Kentucky, Lexington, KY, 40506, USA

[3]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY, 40536, USA

*To whom correspondence should be addressed. Emails: qshao@uky.edu; chen.jin@uky.edu; xudong@missouri.edu.

Abstract: Large protein language models (PLMs) have presented excellent potential to reshape protein research. The trained PLMs encode the amino acid sequence of a protein to a mathematical embedding that can be used for protein design or property prediction. It is recognized that protein 3D structure plays an important role in protein properties and functions. However, most PLMs are trained only on sequence data and lack protein 3D structure information. The lack of such crucial 3D structure information hampers the prediction capacity of PLMs in various applications, especially those heavily depending on the 3D structure. We utilize contrastive learning to develop a 3D structure-aware protein language model (S-PLM). The model encodes the sequence and 3D structure of proteins separately and deploys a multi-view contrastive loss function to enable the information exchange between the sequence and 3D structure embeddings. Our analysis shows that contrastive learning effectively incorporates 3D structure information into sequence-based embeddings. This implementation enhances the predictive performance of the sequence-based embedding in several downstream tasks.

A. Introduction

Proteins play a crucial role in diverse biomedical research endeavors, ranging from unraveling the mechanisms of biological systems to discovering novel therapeutics for various diseases. Deep learning technologies have significantly advanced in developing large protein language models (PLMs) that enable an in-depth exploration of intricate patterns within protein sequences [1]. By considering amino acid sequences of proteins as natural languages, several PLMs were developed based on the sequence data in an unsupervised or self-supervised learning manner. The current paradigm for PLM development and deployment includes two stages: (1) train a deep language model to convert the amino acid sequence into a mathematical representation (embedding) by means of the masked language modeling, where the model predicts masked amino acids in the sequence based on the surrounding context [2-4]; and (2) fine-tune the trained PLMs with the protein property data to perform specific protein prediction tasks. Most PLMs have been developed following this paradigm, such as the ProtBert [3] and the ESM2 [4] models. These PLMs have shown encouraging results for downstream protein property prediction tasks if well fine-tuned on specific data and demonstrated their potential for new knowledge discovery and analyses [5-7].

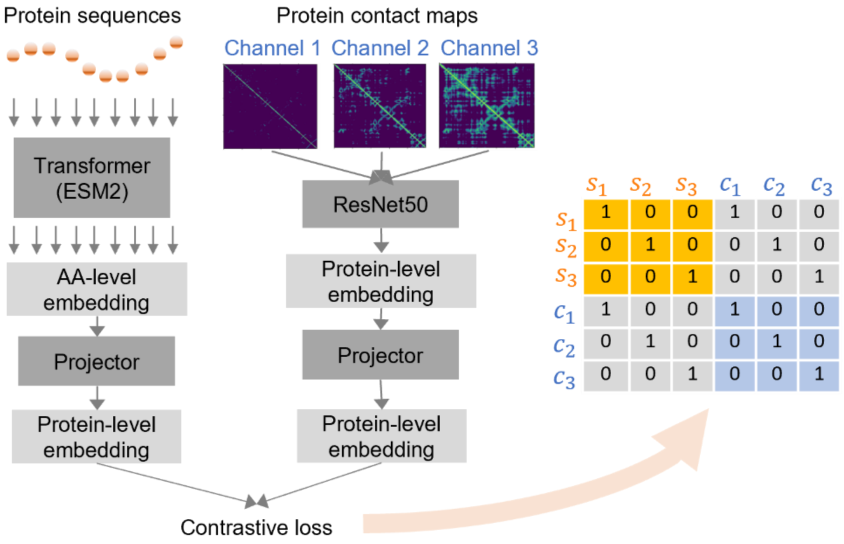
One challenge in developing PLMs is incorporating critical biophysical knowledge into the embeddings. It is well known that a protein's function relies on its 3D structure. However, most PLMs are trained solely on the amino acid sequences incorporating limited information regarding protein structures, thereby constraining their predictive capabilities, especially those heavily depending on the 3D structure of proteins. There have been efforts to integrate evolutionary and protein function knowledge into PLMs. Some studies incorporated multiple sequence alignments (MSAs) into PLMs, including AlphaFold's Evoformer [8] and MSA Transformer [9], which use the MSA directly as input to include the evolutionary information. ProteinBERT is another PLM that used a pre-training strategy similar to other masked language models but incorporated the gene ontology (GO) annotation prediction task in the pre-training scheme. It utilized a denoising autoencoder to pre-train on corrupted protein sequences and GO annotations and performed comparatively well in many protein tasks despite its relatively smaller size. From the algorithm perspective, such methods belong to cross-view representation learning. To incorporate structure information, some methods have been developed using joint-embedding approaches. These approaches aim to learn combined representations of both protein sequences and structures, typically requiring both sequence and structure as input in downstream tasks. For instance, Chen *et al.* proposed a self-supervised learning-based method for structure representation to leverage the available pre-trained PLMs [10]. However, for downstream tasks, their model requires reliable protein structures to provide reliable inference.

Multi-view contrastive learning has emerged as a promising technique to reduce the dependence on protein structures at inference and exhibited favorable properties compared to other representation learning methods. By focusing on learning a coordinated embedding that integrates and leverages information from different views, this approach enables the model to capture rich and complementary features from diverse perspectives. During training, the contrastive loss function pushes the representations of similar views to be close together in the embedding space while simultaneously separating representations of dissimilar views further apart. Unlike the joint-embedding approaches, it does not require all the available views to be presented during inference after training. As a result, multi-view contrastive learning effectively captures underlying semantic structures and patterns in the data. From a head-to-head comparison, it has demonstrated more effective representations than the cross-view representation learning strategy [11]. Notably, ProtST has made notable contributions in this direction in protein research. They injected the biomedical texts into a PLM by aligning these two modalities through a contrastive loss [12]. After training, their PLM enhanced protein property information and demonstrated superiority over previous ones on diverse protein representations and classification benchmarks.

To this end, we propose S-PLM, a 3D structure-aware protein language model developed through multi-view contrastive learning. Unlike the joint-embedding-based methods that rely on both protein structure and sequence for inference, S-PLM encodes the sequence and 3D structure of proteins individually. This unique characteristic allows S-PLM to perform inference solely on sequences, significantly reducing the dependence on protein 3D structure, which is highly valuable in certain application scenarios. Although S-PLM has the capability to accept both structure and sequence as input, this article does not explicitly delve into this aspect. The main focus lies in demonstrating the benefits of S-PLM in encoding protein sequences and leveraging their structural information to improve sequence-based prediction tasks. The experimental results presented in this study underscore the effectiveness of S-PLM for various downstream tasks.

B. Results

B.1 Structure-aware protein language model (S-PLM)



	s_1	s_2	s_3	c_1	c_2	c_3
s_1	1	0	0	1	0	0
s_2	0	1	0	0	1	0
s_3	0	0	1	0	0	1
c_1	1	0	0	1	0	0
c_2	0	1	0	0	1	0
c_3	0	0	1	0	0	1

Figure 1. The architecture of S-PLM model. During training, the model inputs the amino acid sequences and backbone Cα contact maps. The protein sequences are transformed into amino-acid level embeddings (AA-level embedding) through a Transformer-based encoder (ESM2), while the contact maps are transformed into protein-level embeddings through ResNet50. Subsequently, the sequences and contact maps are further converted into separate protein-level embeddings by respective projector layers. Finally, the S-PLM model is trained using contrastive loss applied to the protein-level embeddings from each modality. The expected similarity matrix between the sequences and contact maps is shown on the right. Here, s_i denotes the sequence embedding from protein index *i*, and c_i denotes the contact map embedding from the same protein index *i*. The diagonal elements of the upper-right and lower-left corners represent the expected high similarity between the sequence and contact map embeddings of the same protein, while the off-diagonal elements of the upper-right and lower-left corners indicate expected dissimilarity between the sequence and contact map embeddings of different proteins. Additionally, the off-diagonal elements in the upper-left corner indicate expected dissimilarity between the embeddings of sequences from different proteins, and the off-diagonal elements in the lower-right corner indicate expected dissimilarity between the embeddings of contact maps from different proteins.

Figure 1 depicts the architecture of the S-PLM model. The S-PLM model consists of two encoders, one to encode protein sequence and the other to encode protein 3D structures. In this study, the one-letter amino acid sequences are utilized as the input of the protein sequence. The backbone C_α contact maps are used to represent the protein 3D structures because inter-residue distance contains comprehensive and essential information about protein structure. During training, the inputs of S-PLM are the amino acid sequences and backbone C_α contact maps. The protein sequence information is converted into amino-acid level embedding through a sequence encoder, while the contact map information is transformed into a protein-level embedding through a structure encoder. Then through each corresponding projector layer, sequences and contact maps are converted into separate protein-level embeddings. Finally, the S-PLM model is trained using contrastive learning as the contrastive loss is applied to a batch of sequences and contact maps. The detail of the encoders is presented in the Method section.

The objective of the S-PLM model is to maximize the alignment of the embeddings for the sequence and structure from the same protein and the de-alignment between the embeddings for the sequences and structures of the different proteins. Inspired by the SimCLR method [13], the expected similarity matrix between the sequences and contact maps is illustrated on the right (with a batch size=3). According to the expected similarity matrix, the sequence and contact-map embeddings of the same protein exhibit a high value (1), while those of different proteins show dissimilarity with low values (0). Additionally, embeddings of sequences and embeddings of contact maps from different proteins should also display dissimilarity.

During the inference stage, S-PLM has the flexibility to accept either sequences or contact maps as input and produces corresponding embeddings at various levels tailored to specific downstream tasks. This versatility allows S-PLM to adapt and provide suitable representations based on the specific input data and requirements of the given task. In this paper and the subsequent results, the S-PLM model primarily focuses on generating sequence embeddings incorporating structural information.

B.2 Contrastive learning rearranges the alignment between the sequence and structure embeddings

We first evaluated how contrastive learning rearranges the alignment between the sequence and structure embeddings. We randomly selected 32 proteins from an independent test set and projected their corresponding sequence and structure embeddings on a 2D contour after UMAP processing with and without contrastive learning. As shown in Fig. 2 left, without contrastive learning, the sequence-based embeddings (obtained from pre-trained ESM2-t33_650M_UR50D) were distributed in a wide range while the structure-based embeddings (obtained from pre-trained ResNet50) were clustered in several small regions. The difference in their distribution indicated no information exchange between the two embeddings. As shown in Fig. 2 right, with contrastive learning, the distributions of both the sequence and structure embeddings redistributed in the 2D contour, and the sequence embedding and structure embedding from the same protein were relatively close. The redistributions of the sequence and structure embeddings imply that the contrastive learning procedure has successfully implemented 3D structure information into sequence-based embedding. The redistribution of sequence-based embeddings indicates that they gain new information. The closer alignment of structure and sequence embeddings suggests that the new information is likely from protein 3D structure.

We also grouped the proteins in the test dataset into different bins according to the sequence length. We showed the same UMAP plots to display the alignment of structure and sequence embeddings for specific sequence length ranges (Supplementary Fig. S1). In summary, our contrastive learning approach has the ability to align sequence and structure embedding from the same protein closer while repelling other embeddings further apart.

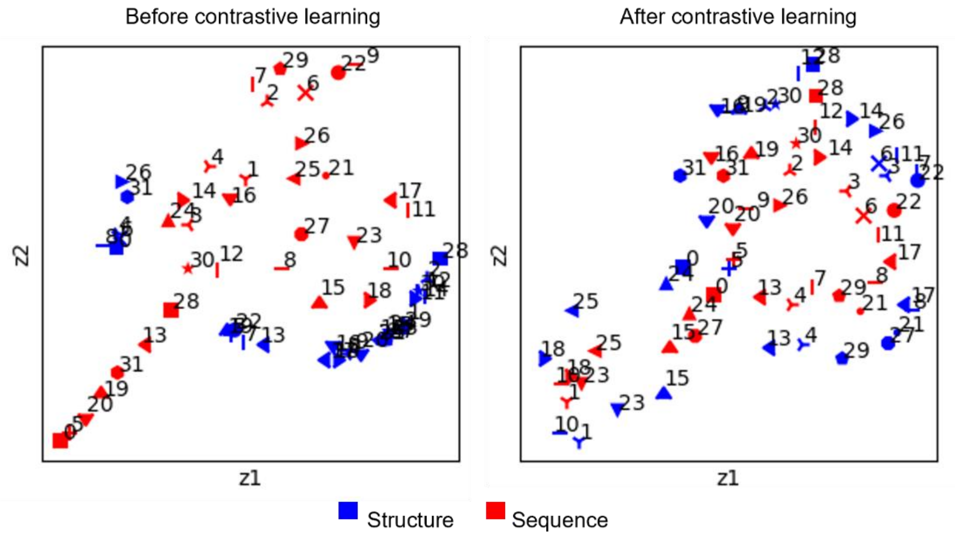


Figure 2. Comparison of structure and sequence embeddings with and without and after contrastive learning using UMAP. A node in blue indicates a structure-based 2D-UMAP embedding for a protein. A node in red indicates a sequence-based 2D-UMAP embedding for a protein. The number beside each node indicates the index of the protein. Different indices and shapes indicate different proteins. The structure-based and sequence-based embeddings from the same protein have the same shape and protein index.

B.3 Contrastive learning enhances the awareness of sequence-based embedding on structure

We further designed an experiment to analyze whether the contrastive learning refined sequence-based embeddings are aware of the protein structure information based on the CATH database [14], which categorizes proteins according to structure-based superfamilies. We selected 139 protein sequences (refer to Method for data preprocessing) with sequence lengths ranging from 200 to 400 AA. Then we represent these proteins with sequence-based embeddings obtained from two ESM2 models and two of our S-PLM models. They are the ESM2-t12_35M_UR50D model, which has the same number of parameters as our S-PLM (35M) model, and the ESM2-t33_650M_UR50D model that has the same number of parameters as our S-PLM (650M) model. Figure 3 shows the 2D T-SNE plot for the representations of these embeddings. Each dot represents a protein. The figures in the first-row color the individual dots based on the first digit of the CATH superfamily. The figures in the second-row color the dots based on the first two digits of the CATH superfamily (Supplementary Data 1).

We use the Calinski-Harabasz index [15] to quantify the ability of the four sequence-based embeddings to differentiate the proteins belonging to distinct superfamilies. The Calinski-Harabasz index for the α-dominated vs. β-dominated clusters (Figure 3, second row) is 1.89 for ESM2_t12_35M_UR50D embedding, 423.22 for our S-PLM (35M) embedding, as well as 2.33 for ESM2_t33_650M_UR50D embedding, and 46.73 for our S-PLM (659M) embedding. The Calinski-Harabasz index for the two digits of the CATH superfamily is 2.33, 58.93, 1.27, and 6.32 for the four embeddings. Given that the CATH superfamily was established using protein structures, this analysis strongly suggests that the sequence-based embedding from the developed S-PLM exhibits an inherent awareness of protein structures, surpassing the other two sequence-only PLMs in effectively distinguishing proteins with diverse structural characteristics. The poor performance of the sequence-only PLMs also indicated their limitations in explicitly acquiring protein structure knowledge. Surprisingly, S-PLM with smaller parameters (S-PLM 35M) for this case showed significantly better performance then S-PLM with larger parameters (S-PLM 650M), which might be because the larger model is harder to train than the smaller model.

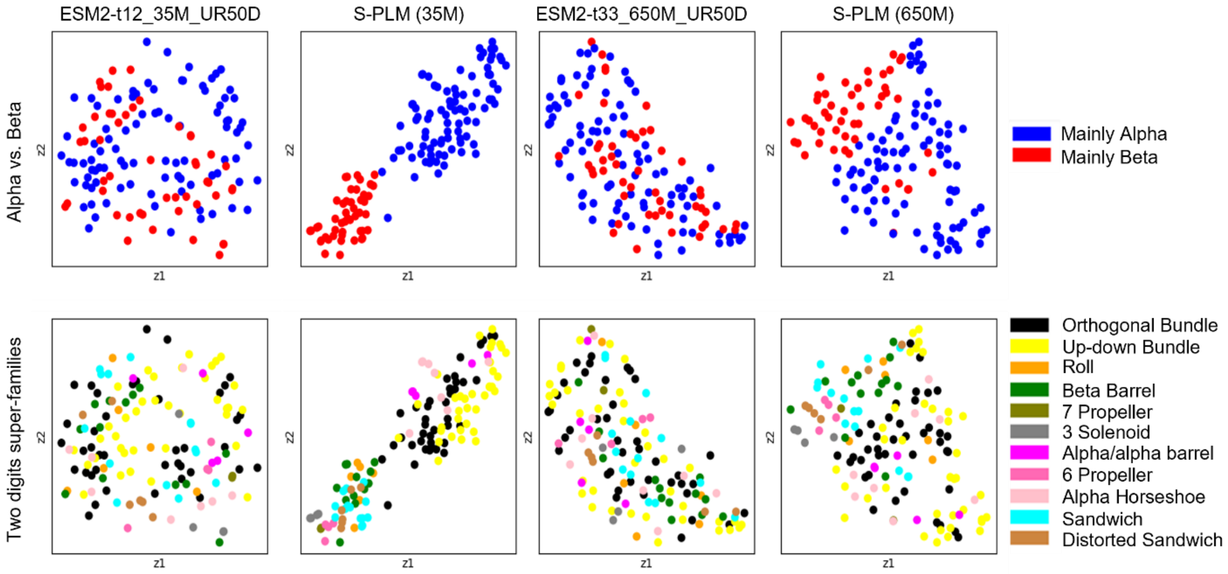


Figure 3. T-SNE plots for different sequence-based embeddings. Each column shows a pair of T-SNE plots for one sequence-based embedding method. The dots in each T-SNE plot that share the same color indicate a cluster of sequences that belong to the same CATH superfamily. The colors in the first row are determined by the first digits of the CATH superfamily, which represent the alpha-dominant family (red) and the beta-dominant family (blue). The colors in the second row are determined by the first two digits of the CATH superfamily.

We also conducted the same experiments using the contact map calculated from the known protein structure provided by the CATH database and encoded them into 2048D vectors by ResNet50 pre-trained on imagenet1k_v1 dataset. The corresponding results are shown in the supplementary Fig. S2. The results using the embeddings from known contact maps served as a benchmark. Based on the results in Fig. 3, Fig. S2, and the Calinski-Harabasz index for the embeddings, our structure-aware sequence embedding demonstrated the most effective distribution in distinguishing structure-based superfamilies.

B.4 Clustering enzymes via S-PLM

Recently, Huang et al. introduced a structure-based protein clustering approach for discovering deaminase functions and identifying novel deaminase families [16]. Their work applied AlphaFold2 to predict and subsequently clustered the entire deaminase protein family based on the predicted structure similarities through structure alignment. They discovered new functions of the deaminase proteins and found new deaminases; such findings cannot be obtained by mining amino acid sequences. In their study, the authors compared the proposed structure-based clustering approach (MSTA) with a multiple-sequence-alignment (MSA) based method. The findings demonstrated that the MSTA-based method significantly outperformed the MSA approach in clustering the deaminase protein families. In this study, we investigated the effectiveness of our S-PLM model in clustering the deaminase family. We utilized the identical sequence data of this study to generate representations for each query protein sequence. The family annotations based on the InterPro database were used as the reference. The 242 sequences and their corresponding annotations were made available either in their paper or by requesting them directly from the authors. For this task, we utilized the S-PLM (650M) and obtained a 1280-dimensional representation vector for each protein. Subsequently, we employed T-SNE to reduce the vector to a two-dimensional representation. Based on the comparison (Fig. 4) between our method (Calinski-Harabasz index=112.47) and the MSA (Calinski-Harabasz index=19.50) and MSTA methods (Calinski-Harabasz index=282.43), our structure-aware sequence-based method performed worse than the structure-based method but significantly outperformed the traditional sequence-based method.

In Huang’s study, they evaluated one deaminase clade SCP1.201 and discovered that certain deaminases from the SCP1.201 clade exhibited the ability to deaminate single-stranded DNA substrates (ssDNA), despite being previously labeled as double-stranded DNA deaminase (dsDNA) toxin A-like (DddA-like) deaminases in the InterPro database (PF14428). Their evaluation was based on 332 SCP1.201 deaminases, where 10 proteins clustered within a specific subclade. Through experimental validation, 8 of these proteins were confirmed to perform dsDNA base editing (Ddd). When we tested our S-PLM model on these 332 SCP1.201 deaminases, our results (Fig. 4B) showed that all 8 Ddd proteins clustered together. This aligns with their findings, and additionally, the newly identified ssDNA proteins were found across all SCP2.201 sub-clades.

Kinases as a specific type of enzymes that facilitate the transfer of phosphate groups to proteins in a process known as phosphorylation, which plays a critical role in many biology processes. In this study, we utilized kinases from established kinase groups and assessed the sequence embeddings generated by our S-PLM model for clustering purposes. We extracted 336 kinases, categorized into 9 kinase groups, along with their respective kinase domain sequences from GPS5.0 [17]. Subsequently, sequence embeddings were generated for each kinase using its corresponding kinase domain sequences. For comparison, we also obtained sequence embeddings from another PLM model, ESM2-t33_650M_UR50D, which shares the same architecture as our S-PLM (650M) model but was pre-trained solely on sequence data. From the T-SNE plots (Fig. 4C), our structure-aware S-PLM model's sequence embeddings showed superior division of kinase groups with a Calinski-Harabasz index of 330.80 compared to 14.29 for the ESM2 model. This is likely because the S-PLM model incorporates protein structure information essential for determining kinase groups.

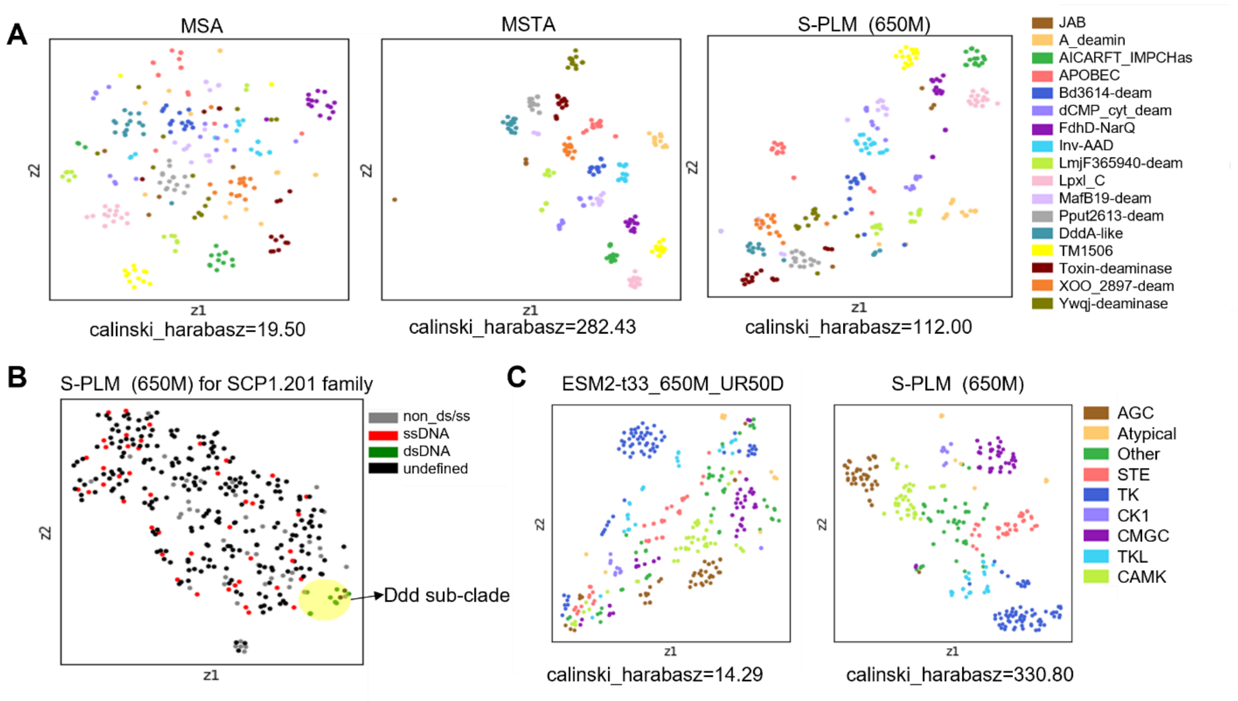


Figure 4. Clustering enzymes via S-PLM. (A) The clustering results for 242 deaminases proteins displayed by T-SNE plots of sequence-based (MSA), structure-based (MSTA) methods, and sequence-embeddings from our S-PLM (650M). Different family proteins are distinguished by different colors. MSA denotes the multiple-sequence-alignment method and MSTA denotes the multiple-structure-alignment method. (B) The clustering of SCP1.201 deaminases via S-PLM. The tested deaminases are shown in green (dsDNA, double-stranded editing), red (dsDNA, single-stranded editing), gray (non_ds/ss, no editing was detected), and black (undefined deaminases). (C) The clustering results for kinase groups displayed by T-SNE plots of sequence embeddings via a pre-trained PLM model (ESM22-t33_650M_UR50D) and our S-PLM (650M). Different kinase groups are distinguished by different colors.

B.5 S-PLM improves downstream protein prediction tasks

In this section, we demonstrate the superiority of our structure-aware S-PLM (650M) model for downstream protein prediction tasks. To assess the performance of our method, we utilized the PEER benchmark [18], a comprehensive and multi-task benchmark for protein sequence understanding, and compared it with the best PLM model presented in the PEER paper. The benefit of using PEER is that it provides a unified platform for training and evaluating protein prediction models. This makes it easy to train and assess large-scale PLMs, as all of the necessary code and data are provided. To ensure a fair comparison and to exclude any impact from model architecture, we selectively chose tasks on which the ESM-1b model, which has the same architecture as our S-PLM, achieved the best performance in the PEER paper.

Both ESM-1b and S-PLM were used as protein sequence encoders. The final output layer aggregates the representations of different residues into a protein-level representation for task prediction, utilizing a corresponding prediction head consisting of a 2-layer MLP and a ReLU activation function. In the PEER paper, they made two evaluation settings by either (1) learning the prediction head with PLM parameters frozen or (2) fine-tuning the PLM along with the prediction head. In our study, we specifically compared with the settings that achieved the best performance in their paper for a particular task and applied the same settings for the S-PLM model. For model training, we modified the PEER codes by substituting the ESM-1b model with our S-PLM (650M) as the sequence encoder, while retaining other training parameters such as learning rate, batch size, and loss function for each task.

From the comparison results in Table 1, we can clearly see the advantages of the fused structure information from our structure-aware S-PLM model. Especially for tasks like Betalactamase and secondary structure predictions, the results are significantly improved, which may be because structure features are more useful for these problems.

Table 1. Comparison of S-PLM and ESM-1b via PEER benchmarks.

Tasks	Metrics	S-PLM	PEER paper ESM-1b
$Betalactamase$ $(\beta - lac)$	Spearmanr	**0.90 (0.002)**	0.84 (0.053)
Solubility (Sol)	Accuracy	**72.09 (0.002)**	70.23 (0.75)
Subcellular localization (Sub)	Accuracy	**79.84* (0.001)**	79.82* (0.18)
Secondary structure (SSP)	Accuracy	**86.88* (0.001)**	83.14* (0.10)

* Used as a feature extractor with the pre-trained PLM weights frozen. The task names used in the PEER paper (Table 3 [18]) are indicated in parentheses.

C. Conclusion

In this work, we proposed a structure-aware protein language model via contrastive learning of sequence and structure. Distinguishing from joint-embedding-based methods, S-PLM independently encodes protein sequences and 3D structures. This unique feature allows S-PLM to make predictions using only the protein sequences, reducing the dependence on the protein's 3D structure. This aspect is particularly important when obtaining protein structure is difficult or time-consuming.

The results of our experiments in this study further highlight the efficacy of S-PLM across a range of downstream tasks. Firstly, S-PLM demonstrates the ability to align sequence and structure embeddings of the same protein effectively while keeping other embeddings from other proteins further apart. Secondly, compared to other PLMs that only use sequences, S-PLM shows impressive awareness of protein structures, as evidenced by its superior performance in distinguishing proteins from different CATH protein superfamilies. Thirdly, in enzyme clustering tasks, S-PLM outperforms both traditional sequence-based methods and other PLM methods significantly. Finally, our experiments on the PEER benchmarks reveal that S-PLM enhances downstream protein prediction tasks, further validating its effectiveness. These findings collectively highlight the potential of S-PLM as a powerful tool for protein analysis and prediction tasks.

This work does have certain limitations. A significant one is that the current model lacks control over the information flow of each modality. To address this limitation, future research could consider incorporating two additional losses. Firstly, for sequence embedding, utilizing a mask language model would allow fine-tuning and refining the sequence encoding process, enhancing the representation of protein sequences. Secondly, for structure embedding, employing diffusion models to reconstruct a protein contact map could facilitate better control over the structure encoding process, enabling more precise capturing of structural information within proteins. Implementing these two additional losses would significantly enhance the overall performance of S-PLM, providing improved control over the information flow of each modality and unlocking its full potential in various protein-related tasks. Another challenge lies in the scale of S-PLM, like other large PLMs, requiring substantial amounts of training data for prediction tasks. It is essential to devise more efficient approaches for training downstream prediction models based on S-PLM to address overfitting and make the process suitable for small-sample learning. These areas present promising avenues for further exploration and improvement in the future.

Methods

Sequence encoder

Our sequence encoder was fine-tuned based on the pre-trained ESM2 model [4]. We have tried two ESM2 pretrained models with similar transformer-based architectures but different parameter scales. One is the ESM2-t12_35M_UR50D model, which has 35 million parameters. The other one is ESM2-t33_650M_UR50D which has 650 million parameters. In particular, the input protein sequence was firstly tokenized by one-hot encoded for each amino acid, and then 12 (12 for ESM2-t12_35M_UR50D and 33 for ESM2-t33_650M_UR50D) layers of transformer encoders were applied, and the embedding dimension for each position was 480 for ESM2-t12_35M_UR50D and 1280 for ESM2-t33_650M_UR50D. In this procedure, a BEGIN token ("<cls>") and END token ("<eos>") will be added to the sequence and going through the transformer together with the amino acid tokens, and a "<pad>" token was used for padding sequences. Through the transformer layers, the output was 480D (or 1280D) vectors for each amino acid, the begin and end tokens, as well as the padding sequences. The embedding of the BEGIN token was then used to represent the protein-level embedding. Then, two projector layers were applied to the protein-level embedding that transformed the dimension into the final output protein-level embedding, which is 196D. The project layers were two linear layers. To retrain the sequence information obtained from the pre-trained PLMs and mitigate the training burden as well, we only fine-tuned the last four transformer layers. There were 11 million trainable parameters in total for the sequence encoder pre-trained based on ESM2-t12_35M_UR50D while 80 million trainable parameters for the one based on ESM2-t33_650M_UR50D.

Structure encoder

In our structure encoder, we specifically encoded the protein contact maps as they were utilized to represent the protein's 3D structure. Since the contact map representation resembles an image, ResNet50 architecture was employed as the encoder, a widely adopted model in image-related tasks. Its application enables effective feature extraction from the contact map representation. To meet the requirement of the ResNet50 model, which expects three input channels, we transformed our contact map into a representation with three channels. Given one sequence, the raw contact map was generated by calculating the coordinate distance between the C_α atoms for each amino acid for one sequence. In general, a contact map is a binary matrix with a value of 1 if the pairwise distance is within a chosen threshold, indicating contact between the residues; otherwise, assign a value of 0. In our case, we considered three distance thresholds (d_i: [10Å, 20Å, 30Å]), and used a continuous representation for the contact map. Specifically, we converted the raw contact map into a similarity matrix whose value ranges from 0 to 1, with 1 indicating the shortest pairwise distance and 0 indicating the longest distance. The final contact map can be famulated as the following:

$$(d_i - CLIP(C_{seq}, 0, d_i))/d_i \tag{1}$$

where CLIP is a function that can change the C_{seq} values that are higher than d_i into d_i.

Multi-view contrastive learning

The objective of contrastive learning in this study is to attract the sequence embeddings and structure embeddings from the same protein closer and repel all the embeddings from different proteins further apart in the latent space. To achieve this, we applied a multi-view contrastive loss function to the protein-level embeddings obtained from the last projection layer of the sequence and structure encoders. The multi-view contrastive loss function was modified based on the *NT-Xent* (the normalized temperature-scaled cross entropy loss) in SimCLR [13]. In contrast to the SimCLR paper, the positive pair is only defined for sequence embedding (E_{seq}^i) and structure embedding (E_{str}^i) from the same protein (*i*) in our approach. Then the multi-view contrastive loss function for the positive pair (E_{seq}^i,E_{str}^i) was defined in the following equation.

$$L_{E_{seq}^i,E_{str}^i} = -\log \frac{\exp\left(\frac{sim\left(E_{seq}^i,E_{str}^i\right)}{\tau}\right)}{\sum_{k=1}^{N} 1_{k\neq i}\exp\left(\frac{sim\left(E_{seq}^i,E_{str}^k\right)}{\tau}\right)+1_{k\neq i}\exp\left(\frac{sim\left(E_{seq}^i,E_{seq}^k\right)}{\tau}\right)+1_{k\neq i}\exp\left(\frac{sim\left(E_{str}^i,E_{str}^k\right)}{\tau}\right))} \tag{2}$$

where $1_{k\neq i} \in \{0,1\}$ is an indicator function evaluating to 1 if $k \neq i$ and τ denotes a temperature parameter; *N* is the batch size. The $sim(x,y)$is a function to quantify the similarity between embeddings x and y, defined as follows:

$$sim(x,y) = x^T y/\|x\|\|y\| \tag{3}$$

This loss function helped to maximize the alignment of the embeddings for protein *i* in the two views (E_{seq}^i ,E_{str}^i) and minimize the alignment between protein *i* and the other proteins ($k \neq i$). The final contrastive loss was calculated for all positive pairs within a mini batch of samples.

Preparation of training dataset for contrastive learning and training process

We prepared the training database based on the proteins in the Swiss-Prot library. The amino acid sequences of the proteins were gained from the Swiss-Prot library and saved in the FASTA format. The 3D structures of the proteins were gained from the alphaFold2 repository. The C_α-C_α contact maps for individual proteins were determined using in-house Python code based on the alphaFold2 predicted 3D structures. We randomly selected 500 K proteins from the Swiss-Prot library for training and 41.5 K proteins for validation. The training was performed on a single V100 using the computational facility of the University of Kentucky. The batch size is 12 due to the GPU RAM limit. Each S-PLM was trained for one week.

Data preprocessing for CATH superfamily clustering

The sequence data with CATH superfamily information were downloaded from the CATH database (release v4_3_0) [14]. We only kept sequences that have records of known protein structures in the PDB database. To eliminate the effect of sequence length, we only remained sequences with lengths ranging from 200 to 400, and for each of the four digits of a superfamily, only one sequence was randomly selected. Finally, 139 protein sequences were used.

Acknowledgments

D.W. and D.X. are partially supported by the National Institutes of Health (grant R35GM126985). U.A., Q.S., and J.C. would like to thank AI in Medicine (AIM), University of Kentucky (NCATS UL1TR001998, NCI P30 CA177558). U.A. and Q.S. would also like to thank the financial support from the Start-up Fund of the University of Kentucky. We thank the University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for their support and use of the Morgan Compute Cluster and associated research computing resources. This work also used Delta-GPU at NCSA through allocation CIS230053 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1. Ofer, D., N. Brandes, and M. Linial, *The language of proteins: NLP, machine learning & protein sequences.* Comput Struct Biotechnol J, 2021. **19**: p. 1750-1758.
2. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
3. *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning.* bioRxiv, 2021.
4. Rives, A., et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.* Proc Natl Acad Sci U S A, 2021. **118**(15).
5. Vu, M.H., et al., *Linguistically inspired roadmap for building biologically reliable protein language models.* Nature Machine Intelligence, 2023. **5**(5): p. 485-496.
6. Wang, B., et al., *Pre-trained language models in biomedical domain: A systematic survey.* arXiv preprint arXiv:2110.05006, 2021.
7. Bepler, T. and B. Berger, *Learning the protein language: Evolution, structure, and function.* Cell Syst, 2021. **12**(6): p. 654-669.e3.
8. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

9. Rao, R.M., et al., *MSA Transformer*, in *Proceedings of the 38th International Conference on Machine Learning*, M. Marina and Z. Tong, Editors. 2021, PMLR: Proceedings of Machine Learning Research. p. 8844--8856.
10. Chen, C., et al., *Structure-aware protein self-supervised learning.* Bioinformatics, 2023. **39**(4).
11. Tian, Y., D. Krishnan, and P. Isola, *Contrastive multiview coding.* European conference on computer vision, 2020: p. 776-794.
12. Xu, M., et al., *ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts.* 2023.
13. Chen, T., et al., *A simple framework for contrastive learning of visual representations.* International conference on machine learning, 2020: p. 1597-1607.
14. Sillitoe, I., et al., *CATH: increased structural coverage of functional space.* Nucleic Acids Research, 2020. **49**(D1): p. D266-D273.
15. Caliński, T. and J. Harabasz, *A dendrite method for cluster analysis.* Communications in Statistics-theory and Methods, 1974. **3**(1): p. 1-27.
16. Huang, J., et al., *Discovery of deaminase functions by structure-based protein clustering.* Cell, 2023: p. S0092867423005937.
17. Wang, C., et al., *GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins.* Genomics Proteomics Bioinformatics, 2020. **18**(1): p. 72-80.
18. Xu, M., et al., *PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding.* 2022.