Domain-Aware Multi-View Contrastive Representation Learning for Protein Subcellular Localization Prediction

Qiang Zhang[1], Feng Yang[1], Weihong Huang[1], Jing Feng[2], Juan Liu[2*]

[1] School of Computer Science, Wuhan University, Wuhan 430070, China
[2] School of Artificial Intelligence, Wuhan University, Wuhan 430070, China
{qiangzhang_, feng.yang, 2023102110011, gfeng, liujuan}@whu.edu.cn

Abstract

Protein subcellular localization prediction is essential for understanding protein function and cellular organization. However, existing methods exhibit two major limitations: (1) they overlook the critical role of evolutionarily conserved protein domains, which are fundamental functional and structural units that significantly influence functions and subcellular localization, and (2) they rarely learn residue order and backbone coordinates simultaneously, neglecting the complementary information inherent in multi-modal representations. In this paper, we propose a novel Domain-Aware Multi-View Contrastive Representation Learning for Protein Subcellular Localization prediction, named DMVCL. Firstly, it devises domain-sequence/structure attention modules, which identify functionally significant regions in protein structures/sequences that critically determine subcellular localization. Secondly, it introduces a multi-view contrastive learning framework that unites inter-view and intra-view objectives. Inter-view contrastive learning aligns protein sequences with their corresponding structures by maximizing mutual information, thereby capturing the consistency of protein residue order and backbone coordinates. Intra-view contrastive learning enhances the representation discriminability of each modality by explicitly separating proteins with no common location and attracting those with any shared localization. Extensive experiments demonstrate that DMVCL significantly outperforms existing baselines. Ablation studies and visualizations further highlight the contributions of domain-sequence/structure attention and multi-view contrastive learning in achieving superior predictive performance.

Introduction

Protein subcellular localization refers to the specific compartment or organelle within a cell where a protein resides and performs its biological function (Rajendran, Knölker, and Simons 2010). Proteins must be localized to the correct subcellular compartments to function properly (Zhang et al. 2025). Mislocalization is often associated with diseases, including cancer, neurodegenerative disorders, and metabolic syndromes (Ng et al. 2024). Understanding protein localization is crucial for elucidating protein function, cellular organization, and disease mechanisms (Wan, Mak, and Kung 2017). Traditional experimental methods, such as fluorescence microscopy and mass spectrometry, are accurate but time-consuming and expensive (Costanzo et al. 2016). With the exponential growth of protein sequences, computational methods have become indispensable for large-scale subcellular localization prediction.

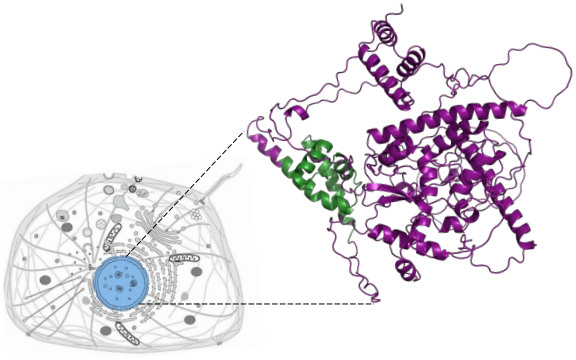

Figure 1: Example of protein "A0A0A0MQE0": the POU-specific domain (IPR000327, green-colored) determines that it resides in the nucleus (blue-colored).

Recent computational methods have significantly advanced protein subcellular localization prediction, leveraging inputs such as sequence, structure, or curated knowledge features (Wang et al. 2024a) (Jiang et al. 2021) (Liu et al. 2022). Domains are the evolutionarily conserved, modular units of proteins that affect molecular function and subcellular localization (Wang et al. 2025). Figure 1 demonstrates that the protein "A0A0A0MQE0" (Uniprot ID) contains the POU-specific domain (IPR000327: green-colored), which endows it with DNA-binding and transcription-factor activity (Sturm and Herr 1988). Consequently, it typically operates within the nucleus (blue-colored). Domains typically provide proteins with crucial cues for their subcellular localization (Jin et al. 2009). However, existing methods rarely incorporate domain information of proteins as informative features into their models. Due to overlooking the domain information, they forfeit a more in-depth biological insight for subcellular localization prediction.

Since proteins frequently localize to multiple subcellu-

*Corresponding authors.

lar compartments, predicting protein subcellular localization is inherently a multi-label classification problem (Almagro Armenteros et al. 2017). To address this problem, existing subcellular localization methods generally handle sequence (residue order) and structure (backbone coordinates) independently, without integrating these two critical aspects (Kaleel et al. 2020) (Wang et al. 2023) (Yuan et al. 2025). Despite mounting evidence proving that joint modal representation yields more expressive features, existing subcellular localization methods rarely learn both modalities simultaneously (Wang et al. 2024b) (Wang, Wang, and Zhang 2024). In addition, these methods often fail to explicitly capture the nuanced similarities and differences between protein representations, resulting in poorly discriminative representations.

To address these problems, we propose a **D**omain-Aware **M**ulti-**V**iew **C**ontrastive Representation **L**earning for Protein Subcellular Localization Prediction, named **DMVCL**. The core of DMVCL is a domain-aware, multi-view contrastive learning architecture that simultaneously leverages sequence (residue order) and structure (backbone coordinates). First, multi-modal encoders transform the sequence and structure into dense embeddings, while the domain encoder then maps discrete domain annotations into dense embeddings. Secondly, we introduce domain-sequence attention and domain-structure attention modules to identify and emphasize regions critical for both function and subcellular localization within each modality. Thirdly, we devise an **Inter**-view **C**ontrastive **L**earning (**Inter-CL**) that maximizes the Mutual Information (MI) (Wang, Wang, and Zhang 2024) (Doquire and Verleysen 2011) between sequence and structure embeddings of the same protein to align the two modalities. We further introduce **Intra**-view **C**ontrastive **L**earning (**Intra-CL**), which explicitly pushes apart the embeddings of proteins from different subcellular compartments while drawing together those from the same compartments. The above modules lead to more stable and accurate predictions. Overall, the main contributions of this work are as follows:

- We introduce a novel domain-sequence/structure attention module that explicitly integrates domain information to precisely identify function-critical regions in both protein sequences and structures.
- We devise the Inter-CL module, which aligns two modality representations in latent space, yielding a robust and complementary representation.
- We devise the Intra-CL module. It explicitly pushes apart the embeddings of proteins from different subcellular compartments. Concurrently, it draws together the embeddings of proteins that share at least one common compartment. As a result, it enhances the discriminability of protein representations.

Related Work

Subcellular Localization

In recent years, computational methods for predicting protein subcellular localization have made significant progress, with methods broadly categorized into three classes: knowledge-based, sequence-based, and structure-based methods. Knowledge-based methods are among the earliest developed to predict protein subcellular localization (Liu et al. 2022) (Shen and Chou 2007). ML-locMLFE (Liu et al. 2022) employs various feature extraction modules to gather multi-source information, including pseudo amino acid composition, grouped weight encoding and Gene Ontology (GO) (Ashburner et al. 2000). Hum-mPLoc (Shen and Chou 2007) achieves higher accuracy for 10-12 compartments by leveraging hierarchical voting, but this method increases the overhead of feature engineering. Sequence-based methods focus on leveraging protein sequence information (Wang et al. 2023). SCLpred-EMS (Kaleel et al. 2020) employs an N-to-1 convolutional neural network to predict subcellular localization by processing vector representations derived from homologous sequence comparison results. MULocDeep (Jiang et al. 2021) utilizes a two-layer bidirectional long short-term memory (LSTM) network (Gal and Ghahramani 2016) to process amino acid embeddings and derive context matrices using multi-head self-attention layers. With the development of accurate protein structure prediction tools, structure-based methods have become increasingly prominent (Zhang et al. 2022) (Jumper et al. 2021). GPSFun (Yuan et al. 2024) utilizes a large language model to predict structural information and extract sequence features, subsequently updating protein features via a graph neural network. DeepMTC (Bai et al. 2024) employs a graph transformer to update embeddings and leverages GO information to enhance localization prediction accuracy. Despite recent advances, three critical gaps remain: (i) overlooking domain cues pivotal for subcellular localization, (ii) failing to align residue order and backbone coordinates in a joint latent space, thereby failing to effectively learn multi-modal complementary information, and (iii) inadequate modeling of the similarities and differences among protein representations, which yields representations with low discriminability.

Contrastive Learning

Contrastive learning has emerged as a powerful paradigm in self-supervised representation learning, enabling models to distill knowledge from data-derived pseudo-labels (Chen et al. 2020) (Xiong et al. 2023). The core principle of contrastive learning by contrasting positive and negative pairs can be traced back to early metric learning methods and has recently gained unprecedented momentum with the progress of deep learning innovations (Liu et al. 2021) (Zhou et al. 2025). Most graph contrastive learning methods are built upon the underlying graph structure. The key to graph contrastive learning generally lies in constructing positive and negative sample pairs, which is typically achieved through graph augmentation strategies (Du et al. 2025). These strategies include augmenting global views by perturbing graph structures or shuffling initial node features (You et al. 2020). However, such methods are not well-suited for proteins, as even minor perturbations can disrupt the precise spatial arrangement, which is critical for their biological function (Wu, Chang, and Zou 2024). An alternative paradigm

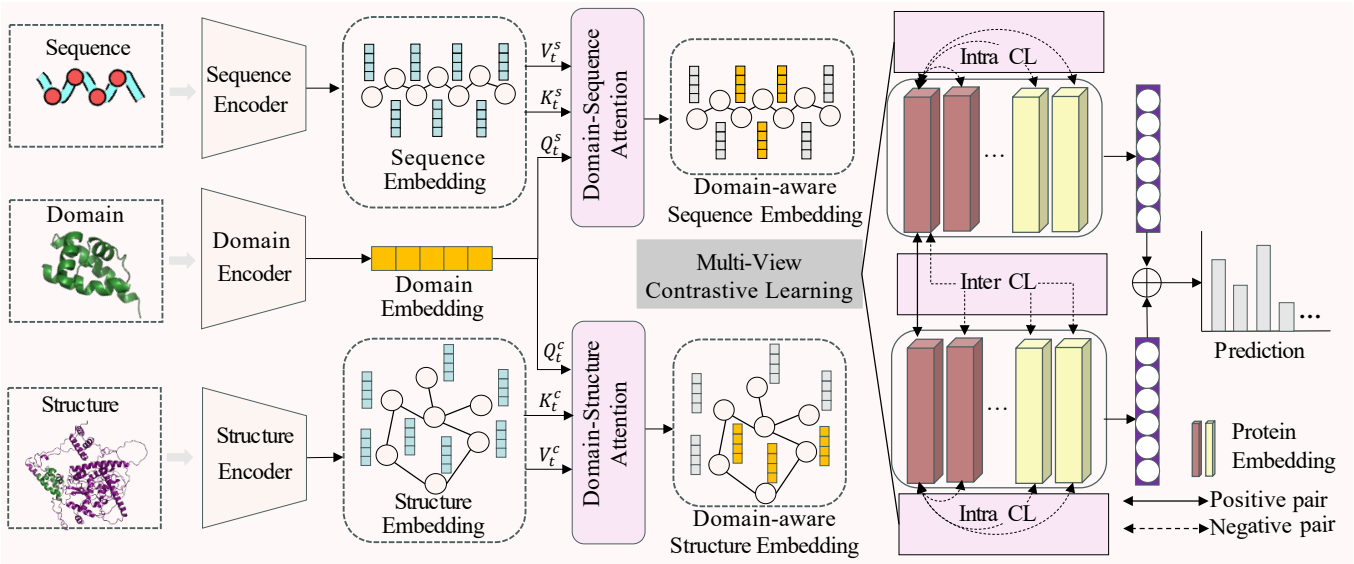

Figure 2: DMVCL pipeline. First, the multi-modal encoder converts sequences, structure and domain into dense embeddings. Next, domain-sequence/structure attention modules spotlight sequences or structure regions deemed functionally related. Finally, Inter-CL aligns the sequence-structure embeddings pair, while Intra-CL pushes apart proteins from different compartments and draws together those that share the same compartment.

is multi-modal contrastive learning, exemplified by CLIP (Radford et al. 2021) (Li et al. 2025), which aligns images with text captions, enabling zero-shot transfer. Despite its impressive success, the application of multi-modal contrastive learning in the protein field remains limited.

Method

In this section, we describe the components of our DMVCL framework, as shown in Figure 2.

Problem Formulation

Let $S = \{s_n\}_{n=1}^{N}$ be a protein sequence of N residues, where s_n denotes the n-th residue. Let $C = \{c_n \in R^3 \mid n = 1, 2, \ldots, N\}$ denote the corresponding 3D coordinate set of the protein atoms obtained by ESMFold (Zhang et al. 2022).

We formulate protein subcellular localization as a multi-label classification problem, with the goal of predicting the binary vector $\hat{y}_i \in \{0, 1\}^K$ for a protein i using its sequence and structure features, where K represents the number of subcellular compartments.

Multi-Modal Encoders

Sequence encoder. For the sequence of protein i, we initialize the sequence representation using one-hot encoding of amino-acid types. Formally, given a sequence of length N, we construct the input feature matrix $X_i^s \in \{0, 1\}^{N \times 20}$, where each row is a 20-dimensional one-hot vector indicating the amino-acid type. These features are processed by the sequence encoder F_s to obtain sequence embedding h_i^s,

$$h_i^s = F_s(X_i^s) \in R^{N \times l}, \tag{1}$$

where l denotes the embedding dimension. F_s consists of a pre-trained network (Lin et al. 2023) followed by a two-layer fully connected network (Zhao et al. 2025).

Structure encoder. We model the structure of protein i as a graph $G_i = (X_i^c, A_i)$, where each node corresponds to a residue. Following the standard graph-convolutional architecture (Fout et al. 2017), we update node embeddings h_i^c,

$$h_i^c = F_c(X_i^c, A_i) \in R^{N \times l}, \tag{2}$$

where F_c is a two-layer graph convolutional network. The node feature $X_i^c \in R^{N \times l'}$ is initialized with a matrix obtained by concatenating: (i) one-hot amino-acid encoding vector, (ii) DSSP secondary-structure vector (Kabsch and Sander 1983), (iii) backbone torsion angles vector (ϕ, ψ, ω) and bond angles vector (α, β, γ) (Yang et al. 2020) (Pereira et al. 2021), (iv) atomic distance vector, and (v) directional vector. These features ensure the rotation and translation invariance of the graph, and the detailed information is provided in Appendix A. Besides, $A_i \in \{0, 1\}^{N \times N}$ denotes the adjacency matrix. Each element of the adjacency matrix indicates whether the C_α distance between the corresponding residue pair is within 10 Å.

Domain encoder. For each protein, we extract functional-domain annotations using InterProScan (Jones et al. 2014). Let $X_i^d \in \{0, 1\}^{1 \times z}$ denote the resulting binary vector, where z is the total number of possible domain types and an entry 1 indicates the presence of the corresponding domain in the protein i. We feed this sparse vector into a two-layer fully connected network F_d to obtain a dense domain embedding h_i^d,

$$h_i^d = F_d(X_i^d) \in R^{1 \times l}, \tag{3}$$

where l denotes the hidden dimension.

Domain-Sequence/Structure Attention

Our proposed model leverages domain information to identify functionally essential sequence regions or structure regions. Inspired by (Vaswani et al. 2017) (Wang et al. 2025), we compute multi-head attention between the domain embedding $h_i^d \in R^{1\times l}$ and the sequence embeddings $h_i^s \in R^{N\times l}$. For each head $t = 1, \ldots, T$, the output of attention for head t is defined as:

$$Q_t^s = h_i^d\, W_t^{(Q,s)}, K_t^s = h_i^s\, W_t^{(K,s)}, V_t^s = h_i^s\, W_t^{(V,s)}, \quad (4)$$

$$H_{i,t}^s = \mathrm{Softmax}\left(\frac{Q_t^s K_t^{s\top}}{\sqrt{d}}\right) V_t^s \in R^{1\times e}, \quad (5)$$

where $W_t^{(Q,s)}, W_t^{(K,s)}, W_t^{(V,s)} \in R^{l\times e}$ are learnable weights and $e = l/T$ is the per-head dimension. Besides, $K_t^s, V_t^s \in R^{N\times e}$, $Q_t^s \in R^{1\times e}$. The softmax function is then applied along the row dimension.

Finally, we concatenate all heads to obtain the domain-aware sequence embedding $\hat{h}_i^s$,

$$\hat{h}_i^s = \mathrm{Concat}(H_{i,1}^s, \ldots, H_{i,T}^s) \in R^{1\times l}. \quad (6)$$

The $\mathrm{Concat}(\cdot,\cdot)$ denotes concatenate operator. Likewise, we compute Q_t^c, K_t^c, V_t^c and obtain the domain-aware structure embedding $\hat{h}_i^c \in R^{1\times l}$.

Multi-View Contrastive Learning

Both complementary information of the two-modal representation (residue order and backbone coordinates) and the discriminability of the each modal representation are essential for improving protein subcellular localization prediction. To this end, we have designed a multi-view contrastive learning framework. From one view, Inter-CL enhances the ability of the model to extract robust and complementary representations. It can capture complementary information by maximizing the MI between the two modalities across the batch. For another view, Intra-CL draws proteins within the shared subcellular compartment closer while pushing those from different compartments further apart within each modality. This process explicitly captures the nuanced similarities and differences between proteins, which forces the model to discern features critical for localization, thereby significantly enhancing the discriminative ability of the learned representations.

Inter-CL. We formulate a contrastive learning objective to align protein sequences and structures in a joint embedding space. For each protein in a mini-batch, the sequence representation serves as an anchor, with its corresponding structure as the positive pair and all other protein structures as negative pairs. This framework simultaneously maximizes the MI between sequence-structure positive pairs while minimizing the MI between negative pairs. To ensure bidirectional alignment, the structure representation alternately serves as the anchor, with its corresponding sequence representation as the positive pair and other sequences as negative pairs. The objective is optimized using a loss (Wang et al. 2024b):

$$\mathcal{L}_{\text{inter}} = -\frac{1}{2}\left[\log\frac{E(\hat{h}_i^s, \hat{h}_i^c)}{\sum_{j=1}^{M} E(\hat{h}_i^s, \hat{h}_j^c)} + \log\frac{E(\hat{h}_i^c, \hat{h}_i^s)}{\sum_{j=1}^{M} E(\hat{h}_i^c, \hat{h}_j^s)}\right], \quad (7)$$

where $\hat{h}_i^s \in R^{1\times l}$ and $\hat{h}_i^c \in R^{1\times l}$ denote domain-aware sequence and structure embeddings respectively, M represents the batch size, and $E(\cdot,\cdot)$ denotes the cosine similarity function.

Intra-CL. Within each mini-batch, we establish contrastive learning pairs by designating protein i as an anchor sample. Proteins that share at least partially identical subcellular compartments are treated as positive pairs, while those localized in distinct compartments serve as negative pairs. To enhance the discriminative capability of the representation, our objective is to maximize embedding similarity between positive pairs and minimize similarity between negative pairs. This yields the contrastive loss:

$$\mathcal{L}_{\text{intra}} = -\log\left[\frac{\sum_{(i,j)\in\mathcal{P}} E(\hat{h}_i^s, \hat{h}_j^s)}{\sum_{(i,j)\in\mathcal{P}} E(\hat{h}_i^s, \hat{h}_j^s) + \sum_{(i,j)\in\mathcal{N}} E(\hat{h}_i^s, \hat{h}_j^s)}\right] - \log\left[\frac{\sum_{(i,j)\in\mathcal{P}} E(\hat{h}_i^c, \hat{h}_j^c)}{\sum_{(i,j)\in\mathcal{P}} E(\hat{h}_i^c, \hat{h}_j^c) + \sum_{(i,j)\in\mathcal{N}} E(\hat{h}_i^c, \hat{h}_j^c)}\right], \quad (8)$$

where $\hat{h}_i^s \in R^{1\times l}$ and $\hat{h}_i^c \in R^{1\times l}$ denote sequence and structure embeddings respectively, $\mathcal{P}$ denotes the set of positive pairs (i,j) sharing at least one common class label, $\mathcal{N}$ represents the set of negative pairs (i,j) sharing no common class label, and $E(\cdot,\cdot)$ denotes the cosine similarity function.

Protein subcellular localization prediction. We concatenate sequence embedding $\hat{h}_i^s$ and and structure embedding $\hat{h}_i^c$ to obtain the final representation h,

$$h = \mathrm{Concat}(\hat{h}_i^s, \hat{h}_i^c) \in R^{1\times 2l}. \quad (9)$$

This combined representation is subsequently passed through a two-layer fully connected network to produce the final classification output probabilities.

Model training. The binary cross-entropy loss for the subcellular localization prediction task is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{M}\sum_{i=1}^{M}\Big[y_i \log\hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\Big], \quad (10)$$

where M denotes the batch size, y_i is the ground-truth label and $\hat{y}$ is the predicted subcellular localization probability for protein i.

The complete objective combines this supervised loss with the contrastive terms:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{BCE}} + \lambda_2\mathcal{L}_{\text{intra}} + \lambda_3\mathcal{L}_{\text{inter}}, \quad (11)$$

where positive scalars $\lambda_1, \lambda_2, \lambda_3$ balance the contributions of the BCE, intra-CL, and inter-CL losses.

Experiments

Experimental Settings

Datasets. To evaluate our proposed DMVCL model, we utilize two benchmark datasets: Thumuluri's dataset (Thumuluri et al. 2022) and Bai's dataset (Bai et al. 2024). Both datasets encompass proteins from 10 distinct subcellular localizations. For data preparation, we employ ESMFold (Zhang et al. 2022) to generate structure PDB files for these proteins and InterProScan (Jones et al. 2014) to acquire their InterPro annotations. Proteins without InterPro records are excluded from the analysis. The original dataset splitting method for the Thumuluri dataset (Yuan et al. 2024) results in the test set that does not contain all subcellular localization classes, thereby limiting a comprehensive evaluation of model performance. To overcome this limitation, we redivide the dataset using a random split to create new training and testing sets. We guarantee that the test set includes all classes while maintaining the sample size of the test set consistent with that reported in the reference (Yuan et al. 2024). For Bai's dataset, we randomly divide it into training, testing, and validation sets in an 8:1:1 ratio. For detailed information about the datasets, please refer to Appendix B.

Baselines. We benchmark our model against several baseline methods. DeepMTC (Bai et al. 2024) is a multi-task deep learning framework that jointly predicts subcellular localization and protein function. It leverages a graph transformer for structure feature extraction, and it also models task correlations to achieve joint optimization. ESM2 (Lin et al. 2023) is a large-scale protein language model trained on millions of sequences to generate contextual embeddings. It can be the pre-training model of protein tasks, making different downstream task predictions. DeepLoc 2.0 (Thumuluri et al. 2022) predicts protein subcellular localization using pre-trained protein language models and a transformer-based module. GPSFun (Yuan et al. 2024) is a geometric graph neural network (GNN) that encodes protein structures as residue-level graphs to capture structural features for predicting protein subcellular localizations. To further enhance the comprehensiveness of our comparison, we introduce two additional methods originally designed for protein function prediction, both of which have demonstrated strong performance in their respective field. We reproduce the two methods and apply them to the task of protein subcellular localization. Specifically, GAT-GO (Lai and Xu 2022) is a graph attention network-based method, while DeepGOPlus (Kulmanov and Hoehndorf 2020) is a sequence-based method that employs 21-1D convolution layers to predict protein functions. By comparing these methods, we aim to provide a more comprehensive evaluation of protein subcellular localization prediction methods. For detailed descriptions of the baseline methods, please refer to Appendix C.

Evaluation metrics. Given the complexity of multi-class classification of subcellular localization prediction, we adopt a comprehensive set of evaluation metrics:

- Accuracy and Jaccard similarity focus on the matching and overlap of predicted and actual localizations, respectively.
- Micro-AUC, Micro-AUPR, and Micro-F1 aggregate contributions from all samples to assess overall effectiveness, particularly useful for imbalanced datasets.
- Macro-AUC, Macro-AUPR, and Macro-F1 evaluate the model's average performance across all classes, providing insight into its ability to handle different classes.

This dual macro/micro perspective ensures balanced evaluation of both per-class and global predictive performance. Please refer to Appendix D for specific details of evaluation metrics.

Implementation details. Our model is trained for 50 epochs with a batch size of 32 and a learning rate lr of 0.0001, using the Adam optimizer on a single GTX 3090 Ti GPU card. Please refer to Appendix E for more detailed network architecture parameters. Our code, data, and appendix are available on GitHub (https://github.com/Qzhangyx/DMVCL)

Performance Comparison

Comparison with baselines. Table 1 provides a comparative analysis of DMVCL with various baseline models across two datasets. The results consistently indicate that DMVCL achieves superior performance compared to other models on both datasets in all evaluation metrics. Notably, DMVCL demonstrates a 5.59% increase in accuracy on Thumuluri's dataset and a 2.01% improvement on Bai's dataset. This phenomenon may be primarily attributed to DMVCL's use of domain information to accurately identify key regions of structures and sequences associated with protein function and subcellular localization. The performance advantages of DMVCL over unimodal methods are evident, which support the notion that the integration of sequential and structural information through inter-CL effectively captures the complementarity across modalities, thereby generating a stable and comprehensive representation. Furthermore, DMVCL exhibits significant improvements with enhancements of 1.30% in Macro-AUC, 2.46% in Macro-AUPR, 1.82% in Macro-F1, and 3.00% in Jaccard Similarity in Thumuluri's dataset and with enhancements of 1.64% in Macro-AUC, 6.96% in Macro-AUPR, 3.48% in Macro-F1, and 1.84% in Jaccard Similarity in Bai's dataset. DMVCL employs intra-CL to explicitly separate protein embeddings from entirely different subcellular compartments while pulling together those that share compartments. This enhances the model's sensitivity to proteins with different subcellular localization, leading to more discriminative representations and more robust predictions, consequently.

Visualization of localization results for each class. To further demonstrate the effectiveness of the DMVCL method, Figure 3 provides a detailed comparative analysis of its performance against other baselines across various subcellular compartments in two datasets. Each box plot illustrates the distribution of evaluation scores, with the central box representing the interquartile range, the line within the box indicating the median, and the whiskers extending to the most extreme data points not considered outliers. Overall,

Dataset	Method	Micro			Macro			Acc	Jaccard
		AUC	AUPR	F1	AUC	AUPR	F1		
Thumuluri's dataset	DeepGOplus	0.908	0.706	0.648	0.866	0.575	0.506	0.459	0.565
	ESM2	0.955	0.828	0.745	0.922	0.693	0.522	0.582	0.683
	GAT-GO	0.945	0.796	0.736	0.903	0.676	0.662	0.587	0.710
	Deeploc 2.0	0.945	0.818	0.768	0.917	0.732	0.714	0.604	0.733
	GPSFun	0.948	0.815	0.750	0.917	0.716	0.684	0.608	0.729
	DMVCL	**0.959**	**0.844**	**0.783**	**0.934**	**0.750**	**0.727**	**0.642**	**0.755**
	Up Ratio	0.42% ↑	1.93% ↑	1.95% ↑	1.30% ↑	2.46% ↑	1.82% ↑	5.59% ↑	3.00% ↑
Bai's dataset	DeepMTC	0.886	0.734	0.641	0.773	0.451	0.335	0.319	0.581
	DeepGOplus	0.895	0.775	0.666	0.805	0.557	0.464	0.370	0.581
	ESM2	0.919	0.814	0.691	0.856	0.589	0.386	0.405	0.638
	GAT-GO	0.898	0.777	0.722	0.847	0.563	0.531	0.441	0.671
	Deeploc 2.0	0.915	0.818	0.757	0.856	0.618	0.603	0.497	0.705
	GPSFun	0.896	0.793	0.706	0.833	0.581	0.538	0.415	0.656
	DMVCL	**0.920**	**0.833**	**0.761**	**0.870**	**0.661**	**0.624**	**0.507**	**0.718**
	Up Ratio	0.11% ↑	1.83% ↑	0.53% ↑	1.64% ↑	6.96% ↑	3.48% ↑	2.01% ↑	1.84% ↑

Table 1: Comparative performance of DMVCL against baseline methods on two subcellular localization datasets. Best results are highlighted in bold. The second-best performing results are underlined.

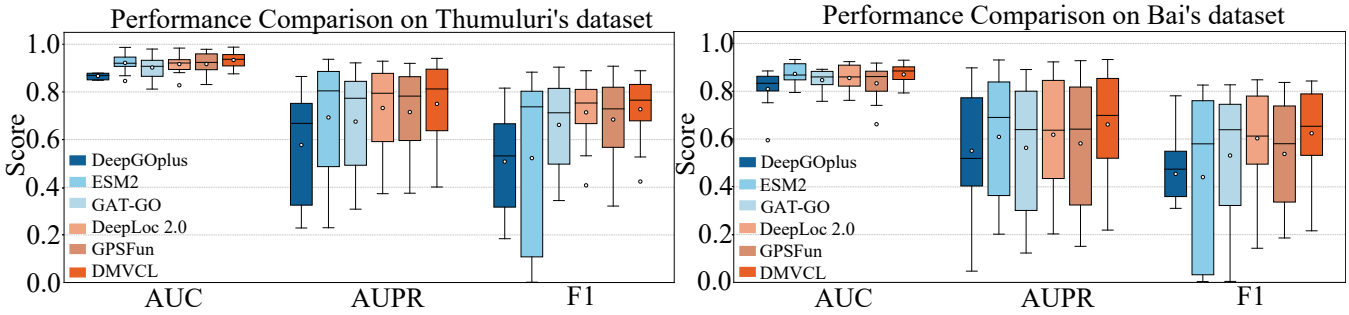

Figure 3: Boxplots of DMVCL and baseline models' performance across different subcellular compartments.

our method exhibits more robust performance across different classes compared to other baselines. This suggests that by integrating domain information, DMVCL can more effectively identify the key sequence and structure regions relevant to subcellular localization. Moreover, multi-view contrastive learning enables the model to learn more comprehensive and discriminative representations, thereby enhancing the prediction robustness and stability.

Ablation Study

To evaluate the significance of each component within the DMVCL framework for the subcellular localization prediction, we conduct comparative analyses between the full DMVCL model and its variants: (1) DMVCL without the domain-sequence/structure Attention module (w/o domain), (2) DMVCL without the Intra-CL module (w/o Intra-CL), and (3) DMVCL without the Inter-CL module (w/o Inter-CL).

The results in Figure 4 indicate that removing the domain-sequence/structure Attention module leads to a decline in performance across all metrics for subcellular localization prediction. This highlights the importance of domain features in helping the model learn key sequential and structural features related to protein subcellular localization. Removing the Inter-CL module also shows a decrease in performance, indicating that learning a comprehensive representation based on sequences and structures is crucial for accurate predictions. The decline in performance when the Intra-CL module is removed indicates that this module contributes to enhancing the discriminability of protein representation, which leads to more stable and accurate predictions.

To further validate the impact of each module within the DMVCL model on subcellular localization tasks. We conducted a thorough analysis using heatmaps, as shown in Figure 5. The number of corresponding samples for each class decreases from left to right. Please refer to Appendix B for detailed information on the sample size for each class of datasets. The model w/o Domain shows a significant perfor-

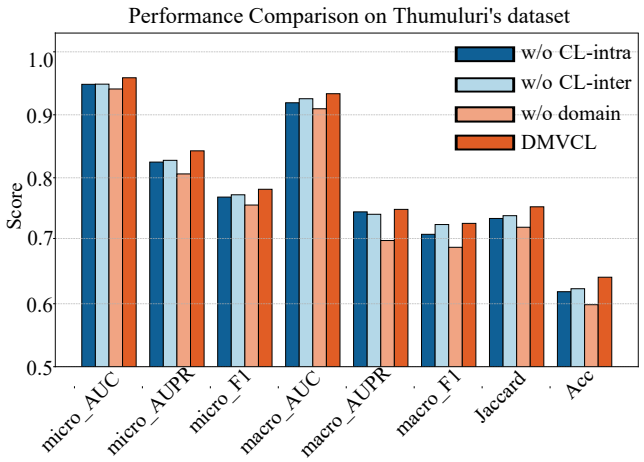

Figure 4: Comparative performance of ablation models using different metrics.

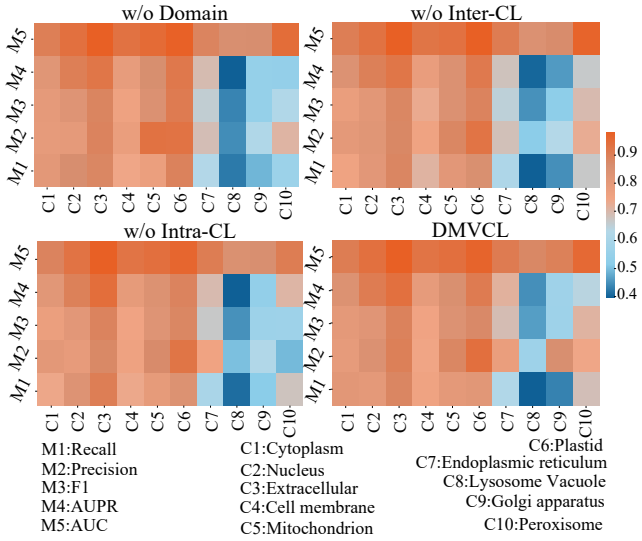

Figure 5: Comparative performance of ablation models across different subcellular compartments in Thumuluri's datasets.

mance drop, especially for less common classes, indicating its importance in recognizing key features for subcellular localization. The w/o Intra-CL model shows a decline in performance for rare classes, which is attributed Intra-CL module to effectively distinguish proteins across different subcellular compartments. Lastly, the model without Inter-CL demonstrates a decrease in performance across all classes, emphasizing the crucial role of learning the consistency of multimodal information in protein subcellular localization prediction.

Overall, these results demonstrate that the full DMVCL model achieves superior performance in predicting subcellular localizations, and the removal of any modules compromises its predictive power.

Hyper-parameter Sensitivity Analysis

In this section, we conduct two sets of parameter experiments. The first experiment investigates how different values of the three coefficients (λ_1, λ_2, and λ_3) affect model training by scaling the individual loss components in Eq.(11). The second experiment investigates the effect of varying the learning rate lr.

Loss weighting sensitivity study. To optimize the loss weights (λ_1, λ_2, and λ_3) in our model, extensive experiments are conducted on Thumuluri's dataset and Bai's dataset, with detailed results provided in Appendix F. For Thumuluri's dataset, the best performance is achieved with $\lambda_1 = 1$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.4$. For Bai's dataset, the optimal settings are $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.5$. The setting of loss weights ensures that the classification objective remains dominant, preventing the model from overemphasizing representation learning at the expense of localization accuracy. The differing optimal weights reflect the unique characteristics of each dataset, emphasizing the need for dataset-specific tuning to achieve optimal performance.

Learning rate sensitivity analysis. Further experiments are conducted by varying the learning rate lr across values of 0.00005, 0.0001, 0.0002, 0.0003, 0.0004, and 0.0005. The detailed results, presented in Appendix F, indicate that DMVCL achieves peak performance at $lr = 0.0001$. Consequently, we have adopted 0.0001 as the default learning rate, which optimally balances between convergence speed and model performance.

Conclusion

We propose DMVCL, a novel model that effectively tackles two critical limitations in existing methods for protein subcellular localization prediction. First, it introduces a domain-sequence/structure attention module that addresses the neglect of evolutionarily conserved protein domains, leveraging these domains to pinpoint key functional regions of protein sequences and structures. Second, it designs a multi-view contrastive learning strategy to mitigate the issues of limited representation comprehensiveness and discriminability. Specifically, Inter-CL aligns the two modality representations in the latent space to produce a robust and complementary representation. Meanwhile, Intra-CL enhances the discriminability of representation within each modality by separating proteins with no common localization and attracting those sharing at least one localization. Extensive benchmarking demonstrates that our model achieves state-of-the-art accuracy and robustness, and ablation studies confirm the substantial contribution of both domain information and contrastive learning to the performance gains.

Acknowledgments

This work is supported by the National Key R&D Program of China (No.2019YFA0904303).

References

Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; and Winther, O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21): 3387–3395.

Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25–29.

Bai, P.; Li, G.; Luo, J.; and Liang, C. 2024. Deep learning model for protein multi-label subcellular localization and function prediction based on multi-task collaborative training. *Briefings in Bioinformatics*, 25(6): 1–14.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607.

Costanzo, M.; VanderSluis, B.; Koch, E. N.; Baryshnikova, A.; Pons, C.; Tan, G.; Wang, W.; Usaj, M.; Hanchard, J.; Lee, S. D.; et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306): 1–34.

Doquire, G.; and Verleysen, M. 2011. Feature selection for multi-label classification problems. In *International Workconference on Artificial Neural Networks*, 9–16.

Du, K.; Zhang, J.; Cao, L.; Guo, Y.; and Sun, W. 2025. A facial structure sampling contrastive learning method for sketch facial synthesis. *Scientific Reports*, 15(1): 16056.

Fout, A.; Byrd, J.; Shariat, B.; and Ben-Hur, A. 2017. Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems*, 30.

Gal, Y.; and Ghahramani, Z. 2016. A theoretically grounded application of dropout in recurrent neural networks. *Advances in Neural Information Processing Systems*, 1–9.

Jiang, Y.; Wang, D.; Yao, Y.; Eubel, H.; Künzler, P.; Møller, I. M.; and Xu, D. 2021. MULocDeep: a deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Computational and structural biotechnology journal*, 19: 4825–4839.

Jin, J.; Xie, X.; Chen, C.; Park, J. G.; Stark, C.; James, D. A.; Olhovsky, M.; Linding, R.; Mao, Y.; and Pawson, T. 2009. Eukaryotic protein domains as functional units of cellular evolution. *Science Signaling*, 2(98): 1–18.

Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9): 1236–1240.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kabsch, W.; and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12): 2577–2637.

Kaleel, M.; Zheng, Y.; Chen, J.; Feng, X.; Simpson, J. C.; Pollastri, G.; and Mooney, C. 2020. SCLpred-EMS: Subcellular localization prediction of endomembrane system and secretory pathway proteins by deep N-to-1 convolutional neural networks. *Bioinformatics*, 36(11): 3343–3349.

Kulmanov, M.; and Hoehndorf, R. 2020. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2): 422–429.

Lai, B.; and Xu, J. 2022. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings in Bioinformatics*, 23(1): 1–9.

Li, J.; Li, Y.; Fu, Y.; Liu, J.; Liu, Y.; Yang, M.; and King, I. 2025. CLIP-powered domain generalization and domain adaptation: a comprehensive survey. *arXiv preprint arXiv:2504.14280*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; and Tang, J. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 857–876.

Liu, Y.; Jin, S.; Gao, H.; Wang, X.; Wang, C.; Zhou, W.; and Yu, B. 2022. Predicting the multi-label protein subcellular localization through multi-information fusion and MLSI dimensionality reduction based on MLFE classifier. *Bioinformatics*, 38(5): 1223–1230.

Ng, C. S.; Liu, A.; Cui, B.; and Banik, S. M. 2024. Targeted protein relocalization via protein transport coupling. *Nature*, 633(8031): 941–951.

Pereira, J.; Simpkin, A. J.; Hartmann, M. D.; Rigden, D. J.; Keegan, R. M.; and Lupas, A. N. 2021. High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, 89(12): 1687–1699.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. 8748–8763.

Rajendran, L.; Knölker, H.-J.; and Simons, K. 2010. Subcellular targeting strategies for drug design and delivery. *Nature Reviews Drug Discovery*, 9(1): 29–42.

Shen, H.-B.; and Chou, K.-C. 2007. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochemical and Biophysical Research Communications*, 355(4): 1006–1011.

Sturm, R. A.; and Herr, W. 1988. The POU domain is a bipartite DNA-binding structure. *Nature*, 336(6199): 601–604.

Thumuluri, V.; Almagro Armenteros, J. J.; Johansen, A. R.; Nielsen, H.; and Winther, O. 2022. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(1): 228–234.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30: 1–11.

Wan, S.; Mak, M.-W.; and Kung, S.-Y. 2017. FUELmLoc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms. *Bioinformatics*, 33(5): 749–750.

Wang, C.; Wang, Y.; Ding, P.; Li, S.; Yu, X.; and Yu, B. 2024a. ML-FGAT: Identification of multi-label protein subcellular localization by interpretable graph attention networks and feature-generative adversarial networks. *Computers in Biology and Medicine*, 170: 107944.

Wang, W.; Shuai, Y.; Zeng, M.; Fan, W.; and Li, M. 2025. DPFunc: accurately predicting protein function via deep learning with domain-guided structure information. *Nature communications*, 16(1): 1–13.

Wang, X.; Han, L.; Wang, R.; and Chen, H. 2023. DaDL-SChlo: protein subchloroplast localization prediction based on generative adversarial networks and pre-trained protein language model. *Briefings in Bioinformatics*, 24(3): 1–10.

Wang, Y.; Liu, X.; Huang, F.; Xiong, Z.; and Zhang, W. 2024b. A multi-modal contrastive diffusion model for therapeutic peptide generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1–11.

Wang, Z.; Wang, Y.; and Zhang, W. 2024. Improving paratope and epitope prediction by multi-modal contrastive learning and interaction informativeness estimation. *arXiv preprint arXiv:2405.20668*.

Wu, K. E.; Chang, H.; and Zou, J. 2024. Proteinclip: enhancing protein language models with natural language.

Xiong, Z.; Liu, S.; Huang, F.; Wang, Z.; Liu, X.; Zhang, Z.; and Zhang, W. 2023. Multi-relational contrastive learning graph neural network for drug-drug interaction event prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5339–5347.

Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; and Baker, D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3): 1496–1503.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33: 5812–5823.

Yuan, G.-H.; Li, J.; Yang, Z.; Chen, Y.-Q.; Yuan, Z.; Chen, T.; Ouyang, W.; Dong, N.; and Yang, L. 2025. Deep generative model for protein subcellular localization prediction. *Briefings in Bioinformatics*, 26(2): 1–11.

Yuan, Q.; Tian, C.; Song, Y.; Ou, P.; Zhu, M.; Zhao, H.; and Yang, Y. 2024. GPSFun: geometry-aware protein sequence function predictions with language models. *Nucleic Acids Research*, 52(W1): W248–W255.

Zhang, X.; Tseo, Y.; Bai, Y.; Chen, F.; and Uhler, C. 2025. Prediction of protein subcellular localization in single cells. *Nature Methods*, 1–11.

Zhang, Z.; Xu, M.; Jamasb, A.; Chenthamarakshan, V.; Lozano, A.; Das, P.; and Tang, J. 2022. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*.

Zhao, Q.; Wang, F.; Wang, W.; Zhang, T.; Wu, H.; and Ning, W. 2025. Research on intrusion detection model based on improved MLP algorithm. *Scientific reports*, 15(1): 1–11.

Zhou, H.; Yin, M.; Wu, W.; Li, M.; Fu, K.; Chen, J.; Wu, J.; and Wang, Z. 2025. ProtCLIP: Function-informed protein multi-modal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22937–22945.