CRYOPROT: A PROTEIN PRETRAINING FRAMEWORK WITH CROSS-BOX INTERACTIONS ON CRYO-EM DENSITY MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the growing availability of cryo-electron microscopy (cryo-EM) density maps, effectively leveraging them for protein representation remains challenging. First, current methods lack a general-purpose protein pretraining framework tailored for cryo-EM density maps, designed for protein-related property prediction. Second, existing approaches typically partition density maps into local box regions and model them independently, overlooking interactions across boxes which are essential for capturing global structural context in cryo-EM density map. To address these challenges, we propose **CryoProt**, a protein pretraining framework designed for cryo-EM density maps. CryoProt introduces a Map Encoder based on multi-head latent attention (MLA), where box-level representations interact through a shared latent space, enabling explicit modeling of cross-box dependencies within the density map. Furthermore, we adopt a multi-task pretraining strategy to learn generalizable representations that can be effectively transferred to diverse downstream tasks, such as protein flexibility prediction, where cryo-EM density maps are not required and can be inferred implicitly by the pretrained model. Experimental results demonstrate that CryoProt consistently outperforms existing state-of-the-art methods across multiple benchmarks, achieving up to 12% improvement over the best-performing baselines, highlighting the importance of modeling cross-box interactions in cryo-EM data. The source code is publicly available at `https://anonymous.4open.science/r/CryoProt`.

1 INTRODUCTION

Proteins carry out a wide range of essential biological processes (Truong Jr and Bepler, 2023; Madsen et al., 2025), and effective protein representation is fundamental for understanding and predicting their functional properties (Duan et al., 2025; Detlefsen et al., 2022). Existing protein representations primarily rely on either sequence information or structural information (Han et al., 2025). The former leverages large-scale amino acid sequences to capture evolutionary and contextual information (Rao et al., 2021), while the latter utilizes three-dimensional coordinates of residues to model spatial organization and geometric relationships (Yuan et al., 2025). Both paradigms rely on abstracted representations that are either indirectly inferred or computationally predicted, which may fail to fully capture the underlying physical properties of biomolecules.

In contrast, cryo-electron microscopy (cryo-EM) offers a more direct and physically grounded modality. Recent advances in cryo-EM have led to a rapid increase in the availability of high-resolution density maps (Kimanius et al., 2024; He et al., 2024), which approximate the electron scattering potential of biomolecules and provide a continuous three-dimensional view of their spatial organization (Raghu et al., 2025). Unlike sequence or discretized structural representations, cryo-EM density maps preserve rich low-level structural signals, capturing both local environments and global conformational variations (He et al., 2024; Sanchez-Garcia et al., 2021; Liu et al., 2025), thereby enabling the learning of generalizable protein representations. However, existing approaches leveraging cryo-EM density maps still face two critical challenges that limit their effectiveness in modeling proteins.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

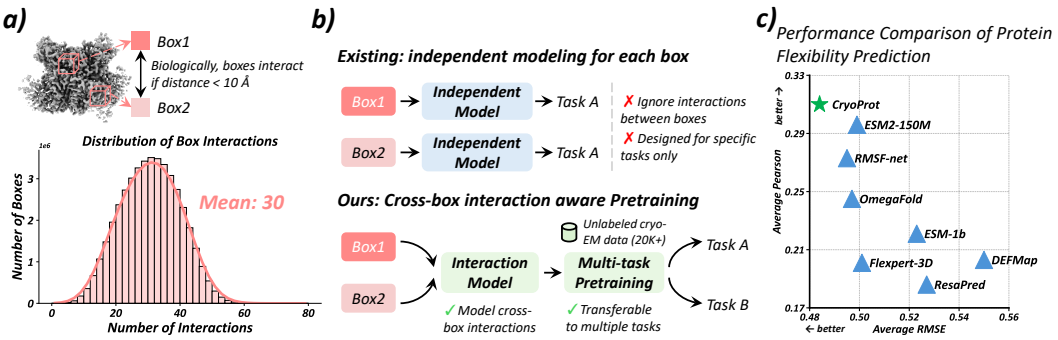


Figure 1: (a) The upper sub-figure illustrates how interactions between local box regions in a density map are determined. From a biological perspective, boxes are considered interacting if their distance is less than 10 Å. The lower sub-figure shows the distribution of the number of boxes under different interaction counts. (b) Existing methods model each box independently, whereas the proposed pre-trained CryoProt considers interactions across boxes. (c) Performance comparison between CryoProt and state-of-the-art methods on the protein flexibility prediction task.

Challenge 1: Lacking general-purpose protein pretraining models for cryo-EM density maps. Existing methods leveraging cryo-EM maps are often tailored to specific downstream applications, such as protein flexibility prediction (Matsumoto et al., 2021; Song et al., 2024; Vander Meersche et al., 2024). While these task-specific designs achieve promising performance within their respective domains, they lack a unified framework for learning transferable representations across diverse tasks. Consequently, the rich structural information in cryo-EM density maps is not fully utilized, and the resulting models often show limited generalization, making it difficult to develop scalable pretraining frameworks for cryo-EM data.

Challenge 2: Overlooking interactions across local boxes in density map. As shown in the upper panel of Fig. 1 (b), many existing methods partition density maps into local regions based on residue positions and process each box independently (Chen et al., 2024; Selvaraj et al., 2025; Zhou et al., 2024). While this design reduces computational cost, it overlooks interactions across boxes. In practice, protein structures are inherently interconnected, and spatial dependencies between regions are important for capturing structural and functional properties (Koren et al., 2023; Jumper et al., 2021). As illustrated in Fig. 1 (a), each box interacts with around 30 neighboring boxes on average. Neglecting such dependencies can lead to incomplete structural representations, as the model fails to capture global structural context, ultimately degrading downstream performance.

To address these challenges, we propose CryoProt, a novel general-purpose protein pretraining framework built directly upon cryo-EM density maps. CryoProt employs a Map Encoder based on multi-head latent attention (MLA) to effectively model density maps and capture complex interactions across local boxes in shared latent space. Furthermore, we design a multi-task pretraining strategy that enables the model to learn generalizable representations, which can be readily transferred to diverse downstream tasks. As illustrated in Fig. 1 (c), protein flexibility prediction is shown as a representative downstream task, where CryoProt achieves state-of-the-art performance. Across a wide range of downstream benchmarks, CryoProt consistently outperforms existing methods, demonstrating its strong generalization capability. Our contributions are as follows:

- We propose CryoProt, the first general-purpose protein pretraining framework that explicitly models cross-box interactions in cryo-EM density maps, enabling the learning of transferable representations from large unlabeled cryo-EM data for diverse downstream tasks.
- We design an MLA-based Map Encoder for explicit cross-box interaction modeling, where box-level representations interact through a shared latent space. Furthermore, we develop a multi-task pretraining strategy to enhance representation learning and generalization.
- CryoProt achieves state-of-the-art performance across multiple benchmarks, with up to 12% improvement over the best-performing baselines. Ablation studies and visualization analyses further validate the effectiveness of modeling cross-box interactions in cryo-EM data.

2 Related Work

Protein Pretraining Models. Sequence-based pretrained models such as ESM-1b (Rives et al., 2021), ESM-1v (Meier et al., 2021), and ESM-2 (Lin et al., 2022) use Transformers to capture evolutionary information. ProtBert (Elnaggar et al., 2021) enhances functional understanding by incorporating Gene Ontology (Hu et al., 2024). Beyond sequence-based approaches, structure-aware models are explored. GearNet (Zhang et al., 2022) introduces sparse edge-based message passing for structural encoding, while OmegaFold (Wu et al., 2022) employs memory-efficient self-attention without multiple sequence alignments (Zielezinski et al., 2025). Most methods rely on sequence or structure, with limited use of experimental properties such as cryo-EM density maps.

Protein-related Property Prediction. Numerous task-specific models have been proposed for predicting protein-related properties, including residue-level tasks (e.g., protein flexibility prediction and identification of active sites) and protein-level tasks (e.g., protein–protein binding affinity prediction and mutation-induced binding free energy change $\Delta\Delta G$ prediction). For flexibility prediction, ResaPred (Wang et al., 2025) employs residual networks with self-attention to model sequence features, while Flexpert-3D (Kouba et al., 2025) integrates sequence and structural information. For site identification, Deep-ProBind (Khan et al., 2025) combines transformer-based encoding with handcrafted features, whereas MMSite (Ouyang et al., 2024) and M^{3}Site (Ouyang et al., 2025) adopt multimodal frameworks to incorporate structural and functional context. For binding affinity prediction, Bind-ddG (Shan et al., 2022) adopts an attention-based geometric neural network, while GearBind (Cai et al., 2024) and PPIgraphomer (Xie et al., 2025) utilize geometric graph neural networks and graph Transformers, and Island (Abbasi et al., 2020) relies on sequence features for regression. For $\Delta\Delta G$ prediction, PPIformer (Bushuiev et al., 2023) introduces an SE(3)-equivariant architecture, ProBASS (Gurusinghe et al., 2025) leverages pretrained protein language models, DGCddG (Jiang et al., 2023) applies graph convolutional network to model mutation effects, and MFFN (Zhang et al., 2025a) employs multi-scale feature fusion with attention mechanisms.

Modeling Based on Cryo-EM Density Maps. Existing studies on cryo-EM density maps can be divided into two directions. The first focuses on map reconstruction and resolution enhancement, with methods such as CryoDRGN (Zhong et al., 2021), CryoNeFEN (Huang et al., 2024), and EModelX (Chen et al., 2024), often partitioning maps into local box regions to reduce computational cost. The second explores modeling protein-related properties from cryo-EM data, which remains less studied. Methods such as DEFMap (Matsumoto et al., 2021) extract residue-centered regions and apply 3D convolutional networks to predict flexibility, while RMSF-Net (Song et al., 2024) incorporates simulated density maps for improved performance. However, these approaches are task-specific and lack generalization. To address this limitation, we propose CryoProt to leverage cryo-EM density maps for general protein property prediction across diverse downstream tasks.

3 Method

3.1 CryoProt Framework

As illustrated in Fig. 2, the pretraining protein model CryoProt consists of key components (e.g., a Sequence Encoder for sequence-level representations, a Box Encoder for local structural contexts, and a Map Encoder for global cross-box interactions). Details are described below.

Sequence Encoder. Given a protein sequence $S = \{a_1, a_2, \ldots, a_L\}$, we first tokenize it into amino acid tokens and encode it using a frozen ESM2-150M model (Lin et al., 2022; Hsu et al., 2022; Ferruz and Höcker, 2022), yielding contextual residue-level representations $\mathbf{E}^{\text{seq}} \in \mathbb{R}^{L\times 640}$. To align feature dimensions with the cryo-EM Map Encoder, we apply a linear projection followed by a ReLU activation function

$$\tilde{\mathbf{E}}^{\text{seq}} = \text{ReLU}\left(\mathbf{W}_s\,\mathbf{E}^{\text{seq}} + \mathbf{b}_s\right), \tag{1}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are learnable parameters, with $\tilde{\mathbf{E}}^{\text{seq}} \in \mathbb{R}^{L\times 240}$. We set the maximum sequence length to 1024. Sequences exceeding this threshold are excluded during preprocessing, while shorter ones are zero-padded and accompanied by an attention mask for valid token-level computation.

Box Encoder. We first represent each cryo-EM density map as a set of residue-centered local subvolumes. Given a map representation N, we partition it according to residue positions into a set of local boxes $N \to \{\text{box}_1, \text{box}_2, \ldots, \text{box}_L\}$. Each $\text{box}_i \in \mathbb{R}^{10\times 10\times 10}$ encodes the local structural

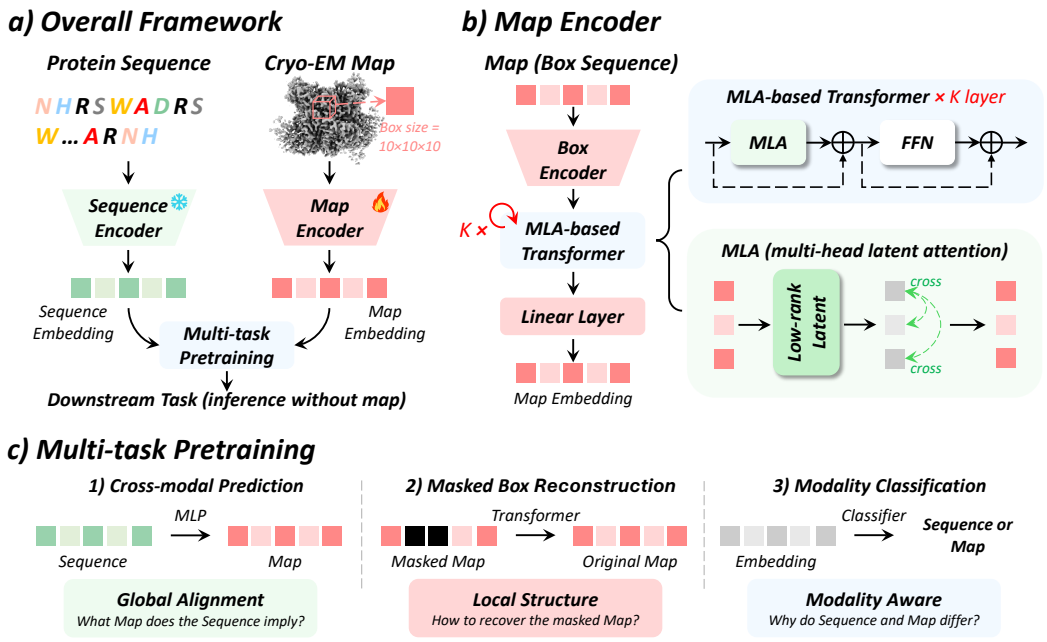


Figure 2: (a) Overview of the CryoProt framework, which employs an MLA-based Map Encoder to capture cross-box interactions and incorporates multi-task pretraining to learn transferable protein representations. (b) The proposed Map Encoder. (c) The designed multi-task pretraining strategy.

environment of residue i and is subsequently reshaped into a vector $\mathbf{E}_i^{\text{box}} \in \mathbb{R}^{1000}$. To incorporate spatial information, we encode the 3D coordinates of each box within the cryo-EM density map $\mathbf{C}_i = (x_i, y_i, z_i)$ using sinusoidal positional encoding (Vaswani et al., 2017)

$$\text{PE}(c_i, 2k) = \sin\left(\frac{c_i}{10000^{2k/d}}\right), \quad \text{PE}(c_i, 2k+1) = \cos\left(\frac{c_i}{10000^{2k/d}}\right), \tag{2}$$

where $c_i \in \{x_i, y_i, z_i\}$ and k indexes the embedding dimension. The structural and positional information are first fused, and then projected into a latent space via a learnable MLP

$$\tilde{\mathbf{E}}_i^{\text{box}} = \text{MLP}\left(\mathbf{E}_i^{\text{box}} + \text{PE}(\mathbf{C}_i)\right). \tag{3}$$

Map Encoder. After the box encoding stage, the cryo-EM density map is represented as a sequence of box embeddings $\mathbf{E}^{\text{map}} = \{\tilde{\mathbf{E}}_1^{\text{box}}, \tilde{\mathbf{E}}_2^{\text{box}}, \ldots, \tilde{\mathbf{E}}_L^{\text{box}}\}$. To model global dependencies and cross-box interactions, we design a Transformer-based architecture equipped with Multi-Head Latent Attention (MLA) (Liu et al., 2024). Unlike conventional attention mechanisms that compute pairwise interactions over all boxes, leading to quadratic complexity, MLA projects the input embeddings into a lower-dimensional latent space to obtain compressed representations $\mathbf{c}_i$. This design significantly reduces the computational cost while preserving essential structural information. More importantly, the latent-space formulation enables efficient modeling of long-range cross-box interactions by allowing each box to attend to globally aggregated latent representations rather than individual box features. This provides a compact yet expressive mechanism to capture global structural dependencies in cryo-EM density maps, which are often spatially sparse and noisy. Based on the latent representations, the attention output at position i is computed as

$$\mathbf{o}_i = \sum_{j=1}^{L} \text{softmax}_j\left(\frac{(\mathbf{W}_q \tilde{\mathbf{E}}_i^{\text{box}})^\top (\mathbf{W}_k \mathbf{c}_j)}{\sqrt{d}}\right)(\mathbf{W}_v \mathbf{c}_j), \tag{4}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable projection matrices. Final map representation is obtained as

$$\tilde{\mathbf{E}}^{\text{map}} = f_{\text{MLA}}(\mathbf{E}^{\text{map}}), \tag{5}$$

where $f_{\text{MLA}}(\cdot)$ denotes the proposed architecture, and $\tilde{\mathbf{E}}^{\text{map}} \in \mathbb{R}^{L \times 240}$. The detailed design of the MLA-based Transformer architecture is provided in Appendix A. This design enables efficient and scalable modeling of cross-box interactions in cryo-EM density maps.

3.2 MULTI-TASK PRETRAINING

As illustrated in Fig. 2 (c), we design three pretraining tasks to enable the model to learn generalizable representations from cryo-EM density maps and protein sequences, including cross-modal prediction, masked map reconstruction, and modality classification. These tasks are not independent but are designed to complement each other from different perspectives. Specifically, cross-modal prediction focuses on aligning global semantic representations between sequence and map modalities, masked map reconstruction emphasizes learning fine-grained local structural patterns, while modality classification enforces discriminative feature learning across modalities. Together, they provide a unified training objective that jointly captures global alignment, local structure, and modality-aware representations, leading to more robust and generalizable features. Detailed formulations of these pretraining tasks are provided in Appendix B.

Cross-modal Prediction. This task aims to align sequence and map representations while enabling the model to predict map features from sequence information alone. By learning this cross-modal mapping, the model can infer map representations from sequences, enabling downstream prediction without requiring cryo-EM density maps. This facilitates effective knowledge transfer between modalities and improves robustness when one modality is missing.

Masked Map Reconstruction. This task enhances the model's ability to capture local structural patterns by randomly masking a subset of input boxes and reconstructing the corresponding density representations within the learned feature space. By enforcing reconstruction from partial observations, the model is encouraged to learn fine-grained structural dependencies and contextual relationships across neighboring regions.

Modality Classification. Inspired by prior studies (Xiong et al., 2026; Ganin and Lempitsky, 2015; Wang et al., 2023), this task predicts the modality type (e.g., sequence or map) of each representation to improve modality-specific feature learning and enhance distinguishability between modalities. It further regularizes the shared representation space and prevents feature collapse across modalities.

We assign a balancing coefficient λ_i to each pretraining task to control the contribution of different loss terms, and adopt an uncertainty-based weighting scheme (Kendall et al., 2018; Kendall and Gal, 2017) to automatically determine these coefficients. This dynamic weighting strategy adaptively balances the learning difficulty of different tasks, ensuring stable optimization and preventing domination by any single objective. The overall training objective is defined as

$$\mathcal{L} = \sum_{i \in \{\text{cross,mask,class}\}} \left(\frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i \right), \tag{6}$$

where $\mathcal{L}$ denotes the total pretraining loss, and $\mathcal{L}_i$ represents the task-specific loss for the i-th pretraining objective. σ_i is a learnable parameter representing the uncertainty of the i-th task, and the corresponding coefficients satisfy $\lambda_i = \frac{1}{2\sigma_i^2}$.

4 EXPERIMENTS AND RESULTS

4.1 DATASET CONSTRUCTION

The pretraining dataset is constructed from the Electron Microscopy Data Bank (EMDB) (Lawson et al., 2016). Following prior studies (Matsumoto et al., 2021; Chen et al., 2024), we curate 20,530 high-quality cryo-EM protein samples, forming one of the largest curated datasets of this type compared with previous studies (Matsumoto et al., 2021; Dhakal et al., 2024; Zhang et al., 2025b). To evaluate generalizability, we consider four downstream tasks spanning residue- and protein-level prediction, namely protein flexibility prediction, active site identification, binding affinity prediction, and $\Delta\Delta G$ prediction. The corresponding datasets are RMSFNet (Song et al., 2024), ProTAD (Ouyang et al., 2024), HER2 (Shanehsazzadeh et al., 2023), and SKEMPI v2 (Jankauskaitė et al., 2019), respectively. Exact protein overlap between pretraining and downstream datasets is excluded, and an

Table 1: Summary of datasets in this study.

Dataset	Entries	Level
EMDB (filtered)	20,530	-
RMSFNet	335	Residue
ProTAD	6,318	Residue
HER2	422	Protein
SKEMPI v2	7,208	Protein

additional MMseqs2-based[1] analysis shows that sequence similarity between downstream proteins and the pretraining set is generally limited. Statistics are summarized in Table 1. All downstream experiments use five-fold cross-validation (AlQuraishi, 2019; Škrhák et al., 2025). Detailed dataset construction and similarity analysis are provided in Appendix C.

Table 2: Performance comparison across multiple benchmarks. Seq and Str denote sequence and structure respectively. CryoProt uses maps only for pretraining, downstream tasks require sequences only. The best results are in bold, and the second-best results are underlined.

A. Protein flexibility prediction (residue-level)						
Model	**Seq**	**Str**	**Map**	**RMSE ↓**	**Pearson ↑**	**Spearman ↑**
Domain-specific models						
DEFMap (Matsumoto et al., 2021)	✗	✗	✓	0.550 ± 0.055	0.203 ± 0.101	0.033 ± 0.102
RMSF-net (Song et al., 2024)	✗	✓	✓	0.495 ± 0.029	0.273 ± 0.050	0.222 ± 0.052
ResaPred (Wang et al., 2025)	✓	✗	✗	0.527 ± 0.046	0.186 ± 0.132	0.014 ± 0.132
Flexpert-3D (Kouba et al., 2025)	✓	✓	✗	0.501 ± 0.028	0.201 ± 0.032	0.208 ± 0.037
Pretrained models						
ESM-1b (Rives et al., 2021)	✓	✗	✗	0.523 ± 0.021	0.221 ± 0.047	0.230 ± 0.051
ESM2-150M (Lin et al., 2022)	✓	✗	✗	0.499 ± 0.019	0.296 ± 0.055	0.279 ± 0.062
OmegaFold (Wu et al., 2022)	✓	✓	✗	0.497 ± 0.024	0.245 ± 0.041	0.252 ± 0.044
CryoProt	✓	✗	✓	**0.484 ± 0.021**	**0.310 ± 0.041**	**0.315 ± 0.046**
B. Active site identification (residue-level)						
Model	**Seq**	**Str**	**Map**	**AUPRC ↑**	**MCC ↑**	**FPR ↓**
Domain-specific models						
Deep-ProBind (Khan et al., 2025)	✓	✓	✗	0.539 ± 0.010	0.537 ± 0.005	0.032 ± 0.002
MMSite (Ouyang et al., 2024)	✓	✗	✗	0.643 ± 0.010	0.592 ± 0.036	0.028 ± 0.002
M^{3}Site (Ouyang et al., 2025)	✓	✓	✗	0.735 ± 0.015	0.644 ± 0.021	0.024 ± 0.002
Pretrained models						
ESM-1b (Rives et al., 2021)	✓	✗	✗	0.672 ± 0.021	0.583 ± 0.024	0.029 ± 0.001
ESM2-150M (Lin et al., 2022)	✓	✗	✗	0.705 ± 0.025	0.612 ± 0.026	0.028 ± 0.001
OmegaFold (Wu et al., 2022)	✓	✓	✗	0.719 ± 0.019	0.628 ± 0.022	0.026 ± 0.002
ProtBert (Elnaggar et al., 2021)	✓	✗	✗	0.593 ± 0.011	0.561 ± 0.007	0.030 ± 0.008
CryoProt	✓	✗	✓	**0.746 ± 0.023**	**0.669 ± 0.019**	**0.021 ± 0.001**
C. Binding affinity prediction (protein-level)						
Model	**Seq**	**Str**	**Map**	**RMSE ↓**	**Pearson ↑**	**Spearman ↑**
Domain-specific models						
Bind-ddG (Shan et al., 2022)	✓	✓	✗	0.653 ± 0.027	0.392 ± 0.014	0.384 ± 0.016
GearBind (Cai et al., 2024)	✗	✓	✗	0.502 ± 0.022	0.495 ± 0.013	0.508 ± 0.015
PPIgraphomer (Xie et al., 2025)	✓	✓	✗	0.731 ± 0.029	0.252 ± 0.013	0.336 ± 0.017
Island (Abbasi et al., 2020)	✓	✗	✗	0.915 ± 0.034	0.108 ± 0.010	0.076 ± 0.012
Pretrained models						
ESM-1b (Rives et al., 2021)	✓	✗	✗	0.579 ± 0.028	0.352 ± 0.019	0.371 ± 0.021
ESM2-150M (Lin et al., 2022)	✓	✗	✗	0.546 ± 0.030	0.386 ± 0.021	0.402 ± 0.018
OmegaFold (Wu et al., 2022)	✓	✓	✗	0.521 ± 0.026	0.462 ± 0.018	0.471 ± 0.020
CryoProt	✓	✗	✓	**0.486 ± 0.039**	**0.524 ± 0.082**	**0.527 ± 0.087**
D. $\Delta\Delta G$ prediction (protein-level)						
Model	**Seq**	**Str**	**Map**	**RMSE ↓**	**Pearson ↑**	**Spearman ↑**
Domain-specific models						
PPIformer (Bushuiev et al., 2023)	✗	✓	✗	1.327 ± 0.044	0.598 ± 0.031	0.468 ± 0.029
ProBASS (Gurusinghe et al., 2025)	✓	✓	✗	1.183 ± 0.035	0.591 ± 0.032	0.433 ± 0.010
DGCddG (Jiang et al., 2023)	✗	✓	✗	1.463 ± 0.055	0.352 ± 0.035	0.309 ± 0.030
MFFN (Zhang et al., 2025a)	✓	✓	✗	1.518 ± 0.051	0.258 ± 0.026	0.334 ± 0.028
Pretrained models						
ESM-1b (Rives et al., 2021)	✓	✗	✗	1.302 ± 0.041	0.512 ± 0.024	0.341 ± 0.028
ESM2-150M (Lin et al., 2022)	✓	✗	✗	1.219 ± 0.039	0.548 ± 0.027	0.369 ± 0.032
OmegaFold (Wu et al., 2022)	✓	✓	✗	1.178 ± 0.038	0.571 ± 0.026	0.402 ± 0.029
CryoProt	✓	✗	✓	**1.036 ± 0.033**	**0.611 ± 0.025**	**0.488 ± 0.007**

[1] https://github.com/soedinglab/MMseqs2

4.2 Experimental Setup

All experiments are conducted on a Linux server with 25 CPU cores and a single NVIDIA GeForce RTX 5090 GPU with 32 GB of VRAM. We first pretrain the CryoProt model on our pretraining cryo-EM dataset and then fine-tune it on four downstream tasks, with performance evaluated via five-fold cross-validation to ensure robustness and reliability. During pretraining, the model learns generalizable representations from unlabeled Cryo-EM data, which are subsequently adapted to task-specific objectives in the fine-tuning stage. Detailed training configurations, hyperparameter settings, fine-tuning procedures, baseline methods, and evaluation metrics are provided in Appendix D.

4.3 Comparison Results

The comparison results are shown in Table 2. Overall, CryoProt consistently outperforms baselines across four downstream tasks, achieving an average improvement of approximately 5% over the second-best baselines, demonstrating strong generalization ability for diverse prediction tasks. Specifically, for protein flexibility prediction, CryoProt outperforms existing protein pretrained models, including both sequence-based and structure-based approaches. This indicates that cryo-EM density maps provide complementary structural dynamics information that is beneficial for modeling protein flexibility. Moreover, compared with existing density map-based methods such as RMSF-net, our approach achieves superior performance, which may be attributed to the explicit modeling of cross-box interactions, enabling better capture of global structural dependencies. Notably, although CryoProt adopts ESM2-150M as its Sequence Encoder, it still significantly outperforms the variant using ESM2-150M alone. This suggests that the performance gain does not solely come from pretrained sequence embeddings, but also from effective learning of structural signals from cryo-EM density maps. These results further validate the effectiveness of the proposed Map Encoder.

4.4 Ablation Study

We conduct ablation studies on cross-box interactions, cryo-EM maps, local resolution, pretraining tasks, and sequence encoder variants. Selected results on protein flexibility prediction are shown in Table 3, while additional results and analyses are provided in Appendix E.

Table 3: Selected ablation study on protein flexibility prediction. The best results are in bold.

Ablation	Model	RMSE ↓	Pearson ↑	Spearman ↑
Box interactions	w/o MLA (No interactions)	0.546 ± 0.019	0.271 ± 0.020	0.278 ± 0.025
	w/ local-window attention	0.519 ± 0.018	0.291 ± 0.019	0.296 ± 0.023
	w/ self-attention	0.497 ± 0.017	0.302 ± 0.018	0.307 ± 0.021
Cryo-EM maps	w/ synthesized maps	0.521 ± 0.019	0.286 ± 0.021	0.293 ± 0.023
	w/ PDB structure	0.536 ± 0.020	0.276 ± 0.020	0.284 ± 0.022
	w/ sequence	0.551 ± 0.023	0.263 ± 0.024	0.270 ± 0.026
Pretraining tasks	w/o pretraining task2	0.498 ± 0.015	0.308 ± 0.016	0.309 ± 0.021
	w/o pretraining task3	0.508 ± 0.016	0.302 ± 0.017	0.306 ± 0.022
	w/o pretraining task2 + 3	0.532 ± 0.018	0.284 ± 0.019	0.291 ± 0.024
	CryoProt	**0.484 ± 0.021**	**0.310 ± 0.041**	**0.315 ± 0.046**

Box interactions. Removing MLA disables cross-box interactions and increases RMSE from 0.484 to 0.546. We also replace MLA with local-window attention or self-attention. The local window size is set to 30 based on the average number of interacting boxes. Both alternatives outperform the variant without cross-box interactions, but MLA achieves the best performance. These results suggest that MLA's latent global interactions help capture complex dependencies among boxes.

Cryo-EM maps. We replace experimental cryo-EM maps with ChimeraX-synthesized maps (Meng et al., 2023), PDB structures, or sequence representations while retaining the MLA architecture and pretraining objectives. The PDB-based variant uses ESM-IF1 (Hsu et al., 2022) residue embeddings, whereas the sequence-only variant uses ESM2 embeddings. These replacements increase RMSE from 0.484 to 0.521, 0.536, and 0.551, respectively. The consistent decline across all three metrics supports the contribution of experimental density information.

Pretraining tasks. We remove masked map reconstruction, modality classification, or both to assess their contributions. Cross-modal prediction is retained in all variants because it enables the model to infer map representations from sequences when density maps are unavailable at inference. Removing either auxiliary task degrades performance across all metrics, and removing both produces a larger decline. These results support the joint use of the two auxiliary objectives during pretraining.

4.5 Parameter Sensitivity Analysis

We analyze the sensitivity of box size and Transformer depth on protein flexibility prediction, as shown in Fig. 3. For box size, performance remains relatively stable across different settings, indicating that CryoProt is not highly sensitive to spatial granularity. The best overall performance is achieved at a box size of 10, while larger boxes substantially increase memory and computational costs, with a box size of 20 approaching the hardware limit. For Transformer depth, performance generally improves as the number of layers increases, suggesting that deeper architectures better capture cross-box dependencies. Due to computational constraints, we set the maximum depth to 6 layers. Additional sensitivity analyses for the attention heads and mask ratio are provided in Appendix F.

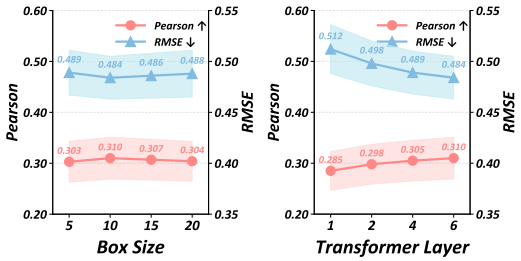


Figure 3: Parameter sensitivity analysis for box size and Transformer layers.

4.6 Case Study

Case on Top-L coverage. Due to the inherent resolution heterogeneity in cryo-EM density maps (Cushing et al., 2024; Vilas et al., 2020; Ramírez-Aportela et al., 2020), we evaluate whether CryoProt can capture structural relationships under varying resolution conditions. Specifically, we randomly select 40 proteins from each of three resolution ranges (2–4 Å, 4–6 Å, and 6–8 Å), yielding 120 proteins in total, and compute their ground-truth residue distance maps from the corresponding native structures. We then construct residue similarity maps based on embeddings from CryoProt and ESM2 and assess their alignment with the distance maps using the Top-L coverage metric (Li et al., 2019; Xiong et al., 2017). As summarized in Table 4, CryoProt consistently outper-

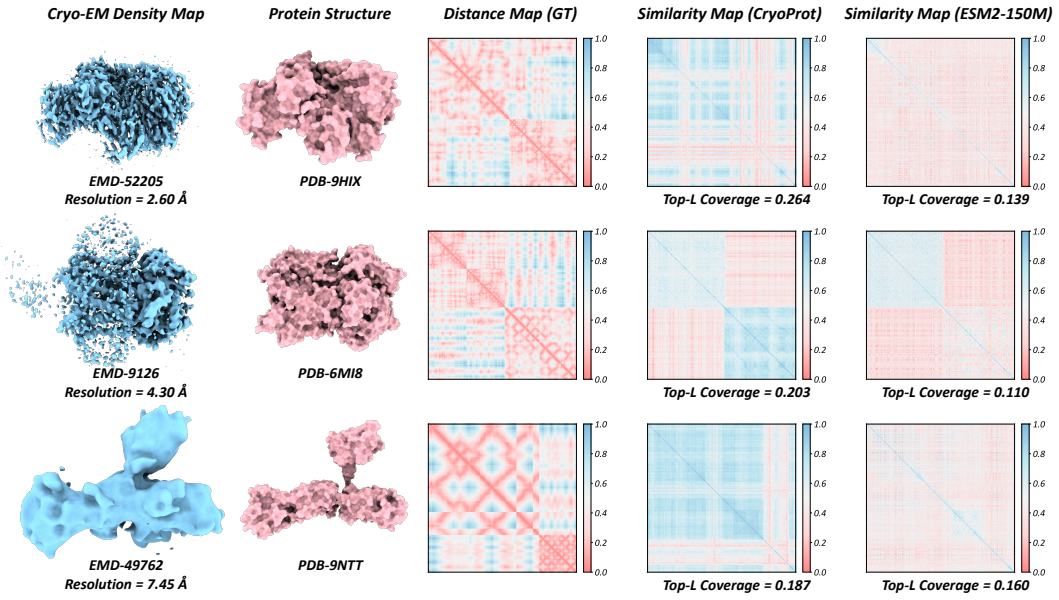


Figure 4: Visualization of residue distance maps and embedding similarity maps.

forms ESM2 across all resolution ranges, indicating that it can implicitly capture spatial relationships without explicit structural supervision and remains robust under low-resolution conditions. As shown in Fig. 4, representative proteins from different resolution ranges consistently show that the similarity maps produced by CryoProt better align with the ground-truth residue distance maps than those produced by ESM2, further demonstrating the ability of CryoProt to capture residue-level spatial relationships from cryo-EM density information. Detailed definitions and interpretation of the Top-L coverage metric, together with per-protein results, are provided in Appendix G.

Table 4: Top-L coverage comparison. The best results are in bold, and the second-best results are underlined.

Model	2–4 Å	4–6 Å	6–8 Å	Mean
ESM2-8M	0.135	0.120	0.101	0.119
ESM2-35M	0.141	0.127	0.108	0.125
ESM2-150M	<u>0.147</u>	<u>0.132</u>	<u>0.112</u>	<u>0.130</u>
CryoProt	**0.166**	**0.159**	**0.128**	**0.151**

Case on protein flexibility prediction. We further present qualitative results on two representative proteins, identified by their PDB IDs 7BP3 (EMD-30143, *human monocarboxylate transporter*) and 5K10 (EMD-9182, *isocitrate dehydrogenase*), which are associated with disease-related biological processes. The ground-truth and predicted flexibility are visualized on the 3D structures using color mapping with ChimeraX software[2], as shown in Fig. 5. CryoProt achieves RMSE values of 0.486 and 0.357 on two proteins, respectively, and the predicted patterns exhibit good consistency with the ground-truth distributions. These results demonstrate that our model can effectively capture residue-level flexibility and generalize well to downstream tasks. Additional visualization results are provided in Appendix I.

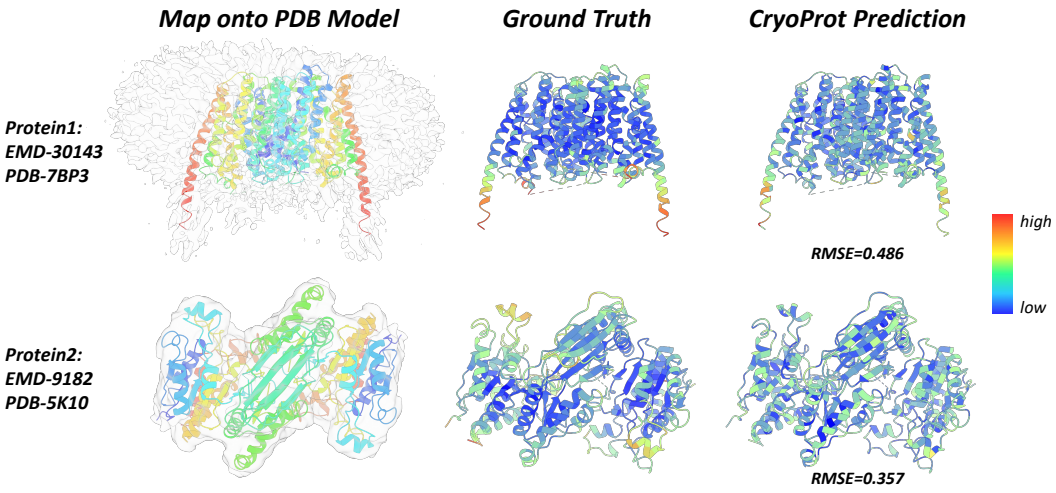


Figure 5: Case study on protein flexibility prediction.

5 Discussion

In this work, we propose CryoProt, a pretraining framework that leverages cryo-EM density maps to learn protein representations. By introducing an MLA-based Transformer architecture to model interactions across local boxes, together with a multi-task pretraining strategy, CryoProt consistently outperforms state-of-the-art methods across multiple downstream tasks. Through case studies on Top-L coverage and visualization, we further demonstrate that CryoProt maintains strong performance across proteins with varying resolution levels. Notably, the model can implicitly capture protein spatial relationships using only sequence information and cryo-EM density maps, without requiring explicit structural supervision. However, CryoProt is currently limited to protein density maps for pretraining. Future work will explore extending this framework to other biomolecular systems, such as RNA and viral proteins. Moreover, incorporating generative modeling to jointly learn protein structures and cryo-EM density maps may provide a promising direction to better address resolution heterogeneity and noise in density data.

[2] https://www.cgl.ucsf.edu/chimerax

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing, grammar correction, and improving the clarity and readability of the manuscript. They were also used for limited assistance in code debugging and visualization-related scripting. Generative AI tools were not used to generate experimental results, fabricate data, conduct statistical analyses, or make scientific conclusions. All AI-assisted text and code were carefully reviewed, verified, and revised by the authors. The experimental results, analyses, interpretations, and scientific claims were independently validated by the authors. We take full responsibility for the final content of this work, including any text or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete source code at `https://anonymous.4open.science/r/CryoProt`. The repository includes scripts and instructions for dataset downloading and preprocessing, implementation of the model architecture and pretraining framework, downstream task evaluation, and visualization. We also provide the configurations and processing pipelines required to reproduce the experiments reported in this work. Detailed descriptions of the datasets, model architecture, training settings, and evaluation protocols are provided in the main text and appendix.

REFERENCES

Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. In *Advances in Neural Information Processing Systems*, volume 36, pages 77379–77415, 2023.

Christian D Madsen, Agnese Barbensi, Stephen Y Zhang, Lucy Ham, Alessia David, Douglas EV Pires, and Michael PH Stumpf. The topological properties of the protein universe. *Nature Communications*, 16(1):7503, 2025.

Haonan Duan, Marta Skreta, Leonardo Cotta, Ella Miray Rajaonson, Nikita Dhawan, Alán Aspuru-Guzik, and Chris J Maddison. Boosting the predictive power of protein representations with a corpus of text annotations. *Nature Machine Intelligence*, 7(9):1403–1413, 2025.

Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

Rong Han, Xiaohong Liu, Tong Pan, Jing Xu, Xiaoyu Wang, Wuyang Lan, Zhenyu Li, Zixuan Wang, Jiangning Song, Guangyu Wang, et al. Copra: Bridging cross-domain pretrained sequence models with complex structures for protein-rna binding affinity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 246–254, 2025.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

Xinyu Yuan, Zichen Wang, Marcus D Collins, and Huzefa Rangwala. Protein structure tokenization: Benchmarking and new recipe. In *International Conference on Machine Learning*, pages 73645–73670. PMLR, 2025.

Dari Kimanius, Kiarash Jamali, Max E Wilkinson, Sofia Lövestam, Vaithish Velazhahan, Takanori Nakane, and Sjors HW Scheres. Data-driven regularization lowers the size barrier of cryo-em structure determination. *Nature Methods*, 21(7):1216–1221, 2024.

Bintao He, Fa Zhang, Chenjie Feng, Jianyi Yang, Xin Gao, and Renmin Han. Accurate global and local 3d alignment of cryo-em density maps using local spatial structural features. *Nature Communications*, 15(1):1593, 2024.

Rishwanth Raghu, Axel Levy, Gordon Wetzstein, and Ellen D Zhong. Multiscale guidance of protein structure prediction with heterogeneous cryo-em data. *arXiv preprint arXiv:2506.04490*, 2025.

Ruben Sanchez-Garcia, Josue Gomez-Blanco, Ana Cuervo, Jose Maria Carazo, Carlos Oscar S Sorzano, and Javier Vargas. Deepemhancer: a deep learning solution for cryo-em volume post-processing. *Communications biology*, 4(1):874, 2021.

Zhe Liu, Bintao He, Tian Zhang, Chenjie Feng, Fa Zhang, Zhongjun Yang, and Renmin Han. Cryoalign2: efficient global and local cryo-em map retrieval based on parallel-accelerated local spatial structural features. *Bioinformatics*, 41(5):btaf296, 2025.

Shigeyuki Matsumoto, Shoichi Ishida, Mitsugu Araki, Takayuki Kato, Kei Terayama, and Yasushi Okuno. Extraction of protein dynamics information from cryo-em maps using deep learning. *Nature Machine Intelligence*, 3(2):153–160, 2021.

Xintao Song, Lei Bao, Chenjie Feng, Qiang Huang, Fa Zhang, Xin Gao, and Renmin Han. Accurate prediction of protein structural flexibility by deep learning integrating intricate atomic structures and cryo-em density information. *Nature Communications*, 15(1):5538, 2024.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

Sheng Chen, Sen Zhang, Xiaoyu Fang, Liang Lin, Huiying Zhao, and Yuedong Yang. Protein complex structure modeling by cross-modal alignment between cryo-em maps and protein sequences. *Nature Communications*, 15(1):8808, 2024.

Joel Selvaraj, Liguo Wang, and Jianlin Cheng. Cryoten: efficiently enhancing cryo-em density maps using transformers. *Bioinformatics*, 41(3):btaf092, 2025.

Yi Zhou, Yilai Li, Jing Yuan, and Quanquan Gu. Cryofm: A flow-based foundation model for cryo-em densities. *arXiv preprint arXiv:2410.08631*, 2024.

Gil Koren, Sagi Meir, Lennard Holschuh, Haydyn DT Mertens, Tamara Ehm, Nadav Yahalom, Adina Golombek, Tal Schwartz, Dmitri I Svergun, Omar A Saleh, et al. Intramolecular structural heterogeneity altered by long-range contacts in an intrinsically disordered protein. *Proceedings of the National Academy of Sciences*, 120(30):e2220180120, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Bozhen Hu, Cheng Tan, Yongjie Xu, Zhangyang Gao, Jun Xia, Lirong Wu, and Stan Z Li. Protgo: Function-guided protein modeling for unified representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 88581–88604, 2024.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07, 2022.

Andrzej Zielezinski, Adam Gudyś, Jakub Barylski, Krzysztof Siminski, Piotr Rozwalak, Bas E Dutilh, and Sebastian Deorowicz. Ultrafast and accurate sequence alignment and clustering of viral genomes. *Nature Methods*, 22(6):1191–1194, 2025.

Wei Wang, Shitong Wan, Hu Jin, Dong Liu, Hongjun Zhang, Yun Zhou, and Xianfang Wang. Resapred: A deep residual network with self-attention to predict protein flexibility. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(1):216–227, 2025.

Petr Kouba et al. Learning to engineer protein flexibility. In *International Conference on Learning Representations*, 2025.

Salman Khan, Sumaiya Noor, Hamid Hussain Awan, Shehryar Iqbal, Salman A AlQahtani, Naqqash Dilshad, and Nijad Ahmad. Deep-probind: binding protein prediction with transformer-based deep learning model. *BMC bioinformatics*, 26(1):88, 2025.

Song Ouyang, Huiyu Cai, Yong Luo, Kehua Su, Lefei Zhang, and Bo Du. Mmsite: a multi-modal framework for the identification of active sites in proteins. In *Advances in Neural Information Processing Systems*, volume 37, pages 45819–45849, 2024.

Song Ouyang, Yong Luo, Huiyu Cai, Kehua Su, Fei Liao, Na Zhan, Huangxuan Zhao, Tailang Yin, Lin Zhao, and Dongjing Shan. M3site: multiclass multimodal learning for protein active site identification and classification. *Briefings in Bioinformatics*, 26(6):bbaf590, 2025.

Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, et al. Deep learning guided optimization of human antibody against sars-cov-2 variants with broad neutralization. *Proceedings of the National Academy of Sciences*, 119(11): e2122954119, 2022.

Huiyu Cai, Zuobai Zhang, Mingkai Wang, Bozitao Zhong, Quanxiao Li, Yuxuan Zhong, Yanling Wu, Tianlei Ying, and Jian Tang. Pretrainable geometric graph neural network for antibody affinity maturation. *Nature Communications*, 15(1):7785, 2024.

Jun Xie, Youli Zhang, Ziyang Wang, Xiaocheng Jin, Xiaoli Lu, Shengxiang Ge, and Xiaoping Min. Ppi-graphomer: enhanced protein-protein affinity prediction using pretrained and graph transformer models. *BMC bioinformatics*, 26(1):116, 2025.

Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas. Island: in-silico proteins binding affinity prediction using sequence information. *BioData Mining*, 13(1):20, 2020.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Marketa Gabrielova, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, et al. Learning to design protein-protein interactions with enhanced generalization. *arXiv preprint arXiv:2310.18515*, 2023.

Sagara NS Gurusinghe, Yibing Wu, William DeGrado, and Julia M Shifman. Probass—a language model with sequence and structural features for predicting the effect of mutations on binding affinity. *Bioinformatics*, 41(5):btaf270, 2025.

Yelu Jiang, Lijun Quan, Kailong Li, Yan Li, Yiting Zhou, Tingfang Wu, and Qiang Lyu. Dgcddg: deep graph convolution for predicting protein-protein binding affinity changes upon mutations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(3):2089–2100, 2023.

Hao Zhang, Yang Liu, Limin Yu, Zejie Wang, Yifei Liu, and Maozu Guo. Multi-scale feature fusion network for the prediction of protein-protein binding affinity changes upon mutations. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 218–223. IEEE, 2025a.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, 18(2):176–185, 2021.

Yue Huang, Chengguang Zhu, Xiaokang Yang, and Manhua Liu. High-resolution real-space reconstruction of cryo-em structures using a neural field network. *Nature Machine Intelligence*, 6(8): 892–903, 2024.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pages 8946–8970. PMLR, 2022.

Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Zhankun Xiong, Ziyan Wang, Feng Huang, Minyao Qiu, Shuyan Fang, Liuqing Yang, Xionghui Zhou, Shichao Liu, Ping Zhang, and Wen Zhang. Multi-to-uni modal knowledge transfer pretraining for molecular representation learning. *Nature Communications*, 2026.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15878–15887, 2023.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Catherine L Lawson, Ardan Patwardhan, Matthew L Baker, Corey Hryc, Eduardo Sanz Garcia, Brian P Hudson, Ingvar Lagerstedt, Steven J Ludtke, Grigore Pintilie, Raul Sala, et al. Emdatabank unified data resource for 3dem. *Nucleic acids research*, 44(D1):D396–D403, 2016.

Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. Cryotransformer: a transformer model for picking protein particles from cryo-em micrographs. *Bioinformatics*, 40(3):btae109, 2024.

Ziying Zhang, Liang Xu, Shuai Zhang, Chunxiang Peng, Guijun Zhang, and Xiaogen Zhou. Emol: modeling protein-nucleic acid complex structures from cryo-em maps by coupling chain assembly with map segmentation. *Nucleic acids research*, 53(W1):W228–W237, 2025b.

Amir Shanehsazzadeh, Sharrol Bachas, Matt McPartlon, George Kasun, John M Sutton, Andrea K Steiger, Richard Shuai, Christa Kohnert, Goran Rakocevic, Jahir M Gutierrez, et al. Unlocking de novo antibody design with generative artificial intelligence. *BioRxiv*, pages 2023–01, 2023.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.

Mohammed AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC bioinformatics*, 20(1):311, 2019.

Vít Škrhák, Marian Novotnỳ, Christos P Feidakis, Radoslav Krivák, and David Hoksza. Cryptobench: cryptic protein–ligand binding sites dataset and benchmark. *Bioinformatics*, 41(1): btae745, 2025.

Elaine C Meng, Thomas D Goddard, Eric F Pettersen, Greg S Couch, Zach J Pearson, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023.

Victoria I Cushing, Adrian F Koh, Junjie Feng, Kaste Jurgaityte, Alexander Bondke, Sebastian HB Kroll, Marion Barbazanges, Bodo Scheiper, Ash K Bahl, Anthony GM Barrett, et al. High-resolution cryo-em of the human cdk-activating kinase for structure-based drug design. *Nature Communications*, 15(1):2265, 2024.

Jose Luis Vilas, Hemant D Tagare, Javier Vargas, Jose Maria Carazo, and Carlos Oscar S Sorzano. Measuring local-directional resolution and local anisotropy in cryo-em maps. *Nature Communications*, 11(1):55, 2020.

Erney Ramírez-Aportela, Jose Luis Vilas, Alisa Glukhova, Roberto Melero, Pablo Conesa, Marta Martínez, David Maluenda, Javier Mota, Amaya Jiménez, Javier Vargas, et al. Automatic local resolution-based sharpening of cryo-em maps. *Bioinformatics*, 36(3):765–772, 2020.

Yang Li, Jun Hu, Chengxin Zhang, Dong-Jun Yu, and Yang Zhang. Respre: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, 35(22):4647–4655, 2019.

Dapeng Xiong, Jianyang Zeng, and Haipeng Gong. A deep learning framework for improving long-range residue–residue contact prediction using a hierarchical strategy. *Bioinformatics*, 33(17): 2675–2683, 2017.

APPENDIX

A MLA-BASED TRANSFORMER ARCHITECTURE

The cryo-EM density map is represented as a sequence of box embeddings

$$\mathbf{E}^{\text{map}} = \{\tilde{\mathbf{E}}_1^{\text{box}}, \tilde{\mathbf{E}}_2^{\text{box}}, \ldots, \tilde{\mathbf{E}}_L^{\text{box}}\}, \tag{7}$$

where $\tilde{\mathbf{E}}_i^{\text{box}} \in \mathbb{R}^d$ denotes the embedding of the i-th box, and L denotes the sequence length, which is equal to the number of boxes. The architecture consists of L_{layer} stacked Transformer layers, each composed of a Multi-Head Latent Attention (MLA) module followed by a feed-forward network (FFN). At the l-th layer, the input representations are denoted as $\mathbf{H}^{(l)} = \{\mathbf{h}_1^{(l)}, \ldots, \mathbf{h}_L^{(l)}\}$. For each token, MLA first projects the input into a lower-dimensional latent space

$$\mathbf{c}_i^{(l)} = \mathbf{W}_D^{(l)} \mathbf{h}_i^{(l)}, \tag{8}$$

where $\mathbf{W}_D^{(l)} \in \mathbb{R}^{d_c \times d}$ is a learnable projection matrix and $d_c \ll d$. This latent projection compresses the input representations, enabling more efficient attention computation while retaining essential structural information. The latent representations are used to construct keys and values, while queries are obtained from the original input

$$\mathbf{q}_i^{(l)} = \mathbf{W}_Q^{(l)} \mathbf{h}_i^{(l)}, \quad \mathbf{k}_i^{(l)} = \mathbf{W}_K^{(l)} \mathbf{c}_i^{(l)}, \quad \mathbf{v}_i^{(l)} = \mathbf{W}_V^{(l)} \mathbf{c}_i^{(l)}, \tag{9}$$

where $\mathbf{W}_Q^{(l)}$, $\mathbf{W}_K^{(l)}$, and $\mathbf{W}_V^{(l)}$ are learnable parameters. In practice, multiple attention heads are used to capture diverse interaction patterns, and the outputs from different heads are concatenated and linearly projected. The attention output at position i is computed by aggregating information from all tokens in the sequence

$$\mathbf{o}_i^{(l)} = \sum_{j=1}^{L} \text{softmax}_j \left(\frac{(\mathbf{q}_i^{(l)})^\top \mathbf{k}_j^{(l)}}{\sqrt{d_c}} \right) \mathbf{v}_j^{(l)}, \tag{10}$$

which enables each box to attend to global contextual information through the latent representations. Compared with standard attention, this formulation avoids directly computing pairwise interactions in the original feature space, significantly reducing computational overhead. A residual connection is applied to combine the input and attention output

$$\hat{\mathbf{H}}^{(l)} = \mathbf{H}^{(l)} + \mathbf{o}^{(l)}, \tag{11}$$

where $\mathbf{o}^{(l)} = \{\mathbf{o}_1^{(l)}, \ldots, \mathbf{o}_L^{(l)}\}$. The representations are then fed into a feed-forward network (FFN), followed by another residual connection

$$\mathbf{H}^{(l+1)} = \hat{\mathbf{H}}^{(l)} + \text{FFN}(\hat{\mathbf{H}}^{(l)}), \tag{12}$$

where the FFN is implemented as a two-layer MLP with non-linear activation. After L_{layer} layers, a linear projection is applied to align the feature dimension with the Sequence Encoder

$$\tilde{\mathbf{E}}^{\text{map}} = \mathbf{H}^{(L_{\text{layer}})} \mathbf{W}_o, \tag{13}$$

where $\mathbf{W}_o$ is a learnable projection matrix, resulting in $\tilde{\mathbf{E}}^{\text{map}} \in \mathbb{R}^{L \times 240}$.

Compared with standard multi-head attention, MLA reduces computational complexity by operating in a compressed latent space, effectively decoupling attention computation from the original feature dimension. This design preserves the ability to model long-range dependencies while improving scalability for long sequences. Such a property is particularly suitable for cryo-EM density maps, where modeling cross-box interactions across spatially distant regions is essential but computationally challenging due to the large number of boxes and the sparsity of meaningful signals.

B MULTI-TASK PRETRAINING OBJECTIVES

Cross-modal Prediction. Given the sequence embedding $\tilde{\mathbf{E}}^{\text{seq}} \in \mathbb{R}^{L \times d}$ and map embedding $\tilde{\mathbf{E}}^{\text{map}} \in \mathbb{R}^{L \times d}$, we employ a learnable projection function to predict map representations from sequence

features. Specifically, each sequence token embedding is mapped into the map representation space as

$$\hat{\mathbf{E}}_i^{\text{map}} = f_{\text{proj}}(\tilde{\mathbf{E}}_i^{\text{seq}}), \tag{14}$$

where $f_{\text{proj}}(\cdot)$ denotes a learnable mapping function implemented as a multi-layer perceptron (MLP). This task encourages alignment between sequence and map modalities by minimizing the discrepancy between predicted and ground-truth map embeddings. The cross-modal prediction loss is defined as

$$\mathcal{L}_{\text{cross}} = \frac{1}{L}\sum_{i=1}^{L}\left\|\hat{\mathbf{E}}_i^{\text{map}} - \tilde{\mathbf{E}}_i^{\text{map}}\right\|_2^2. \tag{15}$$

This objective enables the model to infer structural representations from sequence information, facilitating downstream prediction when cryo-EM data is unavailable.

Masked Map Reconstruction. To enhance the model's ability to capture local structural dependencies, we randomly mask a subset of box representations during training. Let $\mathcal{M}$ denote the set of masked positions sampled from the input sequence, where each position is masked with a fixed probability. The model is then trained to reconstruct the original box embeddings based on the remaining visible context. The reconstruction loss is defined as

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\left\|\hat{\mathbf{E}}_i^{\text{box}} - \tilde{\mathbf{E}}_i^{\text{box}}\right\|_2^2, \tag{16}$$

where $\hat{\mathbf{E}}_i^{\text{box}}$ denotes the reconstructed representation obtained from the model. This objective encourages the model to learn fine-grained local structures and contextual relationships across neighboring regions.

Modality Classification. To enhance modality-specific feature learning, we introduce a token-level binary classifier that distinguishes between the two modalities. A lightweight classification head outputs the probability of each token belonging to one of the two modalities using a sigmoid function. The classification loss is

$$\mathcal{L}_{\text{class}} = -\frac{1}{L}\sum_{i=1}^{L}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right], \tag{17}$$

where $y_i \in 0, 1$ denotes the modality label of token i, and p_i is the predicted probability that token i belongs to the positive modality. This objective encourages the learned representations to retain modality-specific information.

C Dataset Details

Pretraining dataset. The pretraining data are collected from EMDB (Lawson et al., 2016), which provides a large number of experimentally determined cryo-EM density maps. Following the data selection protocol of DEFMap (Matsumoto et al., 2021), we retain protein density maps with resolutions ranging from 2 Å to 4 Å, where relatively high-resolution maps provide more reliable structural details for representation learning. In addition, we restrict the protein sequence length to be less than 1024, following the setting of ESM2, to ensure compatibility with the Sequence Encoder. To reduce the variability across different experimental conditions (e.g., microscope settings, noise levels, and local resolution), we adopt a normalization strategy inspired by EModelX (Chen et al., 2024). Given a raw density map $M \in \mathbb{R}^{w\times h\times d}$, we first align its coordinate system with the corresponding PDB structure and resample it to a unified voxel size of $1\times 1\times 1$ Å via trilinear interpolation, resulting in a processed map $M' \in \mathbb{R}^{w'\times h'\times d'}$. The normalized map N is then computed as

$$N_{xyz} = \begin{cases} 0, & M'_{xyz} < M'_{\text{med}} \\ \dfrac{M'_{xyz} - M'_{\text{med}}}{M'_{\text{top1}} - M'_{\text{med}}}, & M'_{\text{med}} \le M'_{xyz} < M'_{\text{top1}} \\ 1, & M'_{xyz} \ge M'_{\text{top1}} \end{cases} \tag{18}$$

where (x, y, z) denotes the voxel coordinate, M'_{med} is the median density value, and M'_{top1} denotes the top 1% density value. This normalization constrains voxel intensities to $[0, 1]$, reducing noise effects and improving training stability.

Downstream datasets. We evaluate CryoProt on four representative downstream tasks. The protein flexibility prediction dataset is obtained from RMSFNet (Song et al., 2024), which provides residue-level flexibility annotations. The active site identification dataset is derived from ProTAD (Ouyang et al., 2024), containing annotated catalytic residues. The protein–protein binding affinity prediction task uses the HER2 dataset (Shanehsazzadeh et al., 2023), while $\Delta\Delta G$ prediction is evaluated on the SKEMPI v2 dataset (Jankauskaitė et al., 2019). Following the pretraining setting, we further filter all downstream datasets by restricting protein sequence lengths to less than 1024, and only retain samples that satisfy this criterion. In addition, we remove any proteins that overlap with the pretraining dataset to prevent potential data leakage and ensure a fair evaluation. These tasks span both residue-level and protein-level prediction settings, where the first two tasks are defined at the residue level and the latter two are at the protein level, enabling a comprehensive evaluation of the learned representations.

Similarity analysis. Although proteins with exact overlap between the pretraining and downstream datasets are removed, proteins with similar sequences may still exist across the datasets. To further assess potential sequence-level data leakage, we perform an additional similarity analysis using MMseqs2. Specifically, each of the four downstream datasets is independently searched against the same pretraining set at sequence identity thresholds of 30%, 50%, 70%, and 90%, with at least 80% alignment coverage for both query and target sequences. For reproducibility, an example of the MMseqs2 commands used in our analysis is provided below:

```
for dataset in RMSFNet ProTAD HER2 SKEMPIv2; do
    for identity in 0.30 0.50 0.70 0.90; do
        mmseqs easy-search \
            ${dataset}.fasta \
            pretrain.fasta \
            ${dataset}_${identity}.m8 \
            tmp \
            --min-seq-id ${identity} \
            -c 0.80 \
            --cov-mode 0 \
            --format-output query,target,fident,qcov,tcov
    done
done
```

The results are summarized in Table 5. Overall, only a small fraction of downstream proteins exhibit high sequence similarity to proteins in the pretraining set. At the 90% identity threshold, the proportions are only 0.30%, 0.33%, 0.00%, and 2.77% for RMSFNet, ProTAD, HER2, and SKEMPI v2, respectively. These results suggest that the observed downstream performance is unlikely to be primarily driven by sequence-level overlap with the pretraining data. We focus on sequence similarity rather than exhaustive structure-level comparison. Structural similarity is harder to define consistently because proteins may differ in conformation, chain composition, and structural coverage, and simple metrics such as RMSD or TM-score may not directly indicate data leakage. We therefore adopt sequence similarity as a standardized and scalable criterion for assessing potential data leakage.

Table 5: Sequence similarity between downstream and pretraining datasets.

Dataset	Entries	≥30%	≥50%	≥70%	≥90%
RMSFNet	335	4 (1.19%)	3 (0.90%)	1 (0.30%)	1 (0.30%)
ProTAD	6,318	63 (1.00%)	43 (0.68%)	33 (0.52%)	21 (0.33%)
HER2	422	5 (1.18%)	2 (0.47%)	0 (0.00%)	0 (0.00%)
SKEMPI v2	7,208	275 (3.82%)	274 (3.80%)	272 (3.77%)	200 (2.77%)

Data splitting. To avoid potential protein data leakage (AlQuraishi, 2019; Škrhák et al., 2025), all datasets are split at the protein level rather than at the residue level (i.e., residues from the same protein are kept within the same fold rather than being split across folds), and a five-fold cross-validation strategy is employed for robust evaluation.

D More Experiment Details

D.1 Hyperparameter settings

The detailed hyperparameter settings used in both pretraining and fine-tuning are summarized in Table 6. The listed values of λ_1, λ_2, and λ_3 are their initial values, and these weights are subsequently learned through an uncertainty-based weighting scheme to balance the training objectives.

Table 6: Training hyperparameters of CryoProt.

Parameter	Value	Parameter	Value	Parameter	Value
Pretraining epochs	10	Batch size	48	Learning rate	1×10^{-4}
Weight decay	1×10^{-2}	Max length	1024	Embedding dimension	240
Transformer layers	6	Attention heads	8	Dropout rate	0.1
Box size	10	latent dimension	128	Mask ratio	0.15
λ_1	1.0	λ_2	0.5	λ_3	0.3

D.2 Fine-tuning on Downstream Tasks

After pretraining, CryoProt uses the protein sequence to generate sequence representations and infer map representations through the cross-modal prediction module. Thus, downstream inference does not require a cryo-EM density map. We concatenate the residue-level sequence representations with the sequence-inferred map representations along the feature dimension. For residue-level tasks, a task-specific prediction head operates on each concatenated residue representation. For protein-level tasks, masked mean pooling over valid residue positions produces a protein-level representation, which is then passed to a task-specific prediction head.

D.3 Introduction of Baseline Models

D.3.1 Domain-specific Models

Protein Flexibility Prediction

DEFMap (Matsumoto et al., 2021) leverages cryo-EM density maps to predict residue-level protein flexibility. It extracts local 3D regions centered at residues from density maps and employs 3D convolutional networks to learn structural patterns associated with flexibility. **RMSF-net** (Song et al., 2024) enhances flexibility prediction by combining real cryo-EM density maps with simulated density maps derived from protein structures. It further designs a tailored 3D CNN architecture to capture multi-scale structural features for improved prediction performance. **ResaPred** (Wang et al., 2025) is a deep learning-based method that predicts protein flexibility by integrating diverse sequence-derived features, including secondary structure, torsion angles, and solvent accessibility. It adopts a residual network architecture combined with self-attention mechanisms to effectively capture key patterns related to flexibility. **Flexpert-3D** (Kouba et al., 2025) is a protein flexibility prediction model that incorporates both sequential and structural information. It addresses data limitations and improves performance by capturing flexibility-related patterns from multiple protein feature sources.

Active Site Identification

Deep-ProBind (Khan et al., 2025) is a hybrid learning model for identifying protein binding sites by combining sequence-derived and evolutionary features with structural information. It integrates transformer-based encoding with handcrafted descriptors, and employs feature selection together with a neural classifier to perform residue-level binding site prediction. **MMSite** (Ouyang et al., 2024) is a multimodal framework that enhances protein language models by incorporating biomedical language representations. It leverages complementary information from protein sequences and textual biological knowledge to improve residue-level prediction performance. **M^{3}Site** (Ouyang et al., 2025) is a multimodal approach for residue-level active site prediction. It integrates protein sequence embeddings, structural graph representations, and functional textual annotations to jointly model biochemical context and improve multiclass active site identification.

Protein–Protein Binding Affinity Prediction

Bind-ddG (Shan et al., 2022) adopts an attention-based geometric neural network to model the effects of mutations on protein–protein interactions using 3D complex structures. As the training code is not publicly available, we directly use the released pretrained model for inference in our experiments. **GearBind** (Cai et al., 2024) represents protein–protein interfaces as all-atom graphs and models interactions through multi-level geometric message passing, enabling effective characterization of complex intermolecular interactions. **Island** (Abbasi et al., 2020) is a sequence-driven approach for binding affinity prediction. It utilizes a variety of features derived from protein sequences and applies regression-based learning to capture relationships associated with binding strength. **PPIgraphomer** (Xie et al., 2025) is a graph Transformer-based framework for binding affinity prediction. It incorporates pretrained protein representations and models interaction patterns at binding interfaces through refined graph construction and attention mechanisms.

$\Delta\Delta G$ Prediction

PPIformer (Bushuiev et al., 2023) is an SE(3)-equivariant model designed to generalize across diverse protein–binder variants. It captures geometric relationships within protein complexes while incorporating a thermodynamically inspired loss formulation to improve the accuracy of prediction. **ProBASS** (Gurusinghe et al., 2025) leverages both sequence and structural information through pre-trained protein language models ESM2. It generates embeddings for protein–protein interaction mutants and fine-tunes the model on experimentally measured $\Delta\Delta G$ datasets, demonstrating the effectiveness of combining sequence- and structure-based representations for mutation effect prediction. **DGCddG** (Jiang et al., 2023) is a graph convolution-based approach for predicting mutation-induced changes in binding affinity. It applies multi-layer graph convolutions to learn contextualized residue representations from protein complex structures, and uses these features to model the impact of mutations. **MFFN** (Zhang et al., 2025a) is a multi-scale feature fusion network for $\Delta\Delta G$ prediction that reduces reliance on handcrafted biological features. It models protein complexes at multiple scales and incorporates attention mechanisms along with feature extraction modules to capture diverse structural patterns relevant to binding affinity changes.

D.3.2 Pretrained Models

ESM-1b (Rives et al., 2021) is a Transformer-based protein language model trained on large-scale amino acid sequences using the masked language modeling objective. By learning to recover masked residues, it captures evolutionary constraints and contextual dependencies within protein sequences, providing informative sequence embeddings for downstream tasks. **ESM2** (Lin et al., 2022) employs a deeper and more scalable Transformer architecture, trained on millions of protein sequences to effectively capture long-range dependencies and rich evolutionary patterns. In our experiments, we use ESM2-150M as the primary baseline due to its strong performance and moderate computational cost. In addition, we include smaller variants, ESM2-8M and ESM2-35M, in the ablation study to investigate the impact of model scale on downstream performance. **ProtBert** (Elnaggar et al., 2021) is a bidirectional Transformer-based protein language model trained on large-scale sequence datasets using language modeling objectives. It is further enhanced with auxiliary tasks such as GO annotation prediction, allowing it to capture both sequence-level patterns and functional semantics, which improves its applicability to various downstream tasks. **OmegaFold** (Wu et al., 2022) is a protein structure-aware language model that learns informative structural representations. It employs a memory-efficient self-attention architecture to capture spatial dependencies and model three-dimensional protein structures without relying on MSA, providing useful structure-level features for downstream tasks.

D.4 Evaluation Metrics

We evaluate CryoProt on four downstream tasks, including regression tasks (e.g., protein flexibility prediction, binding affinity prediction, and $\Delta\Delta G$ prediction) and a binary classification task (active site identification).

Regression Tasks. We adopt Root Mean Square Error (RMSE), Pearson correlation coefficient, and Spearman rank correlation coefficient. Given ground-truth values $\{y_i\}_{i=1}^{S}$ and predictions $\{\hat{y}_i\}_{i=1}^{S}$,

RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(y_i - \hat{y}_i)^2}, \tag{19}$$

where S denotes the total number of samples. The Pearson correlation coefficient is computed as

$$\text{Pearson} = \frac{\sum_{i=1}^{S}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{S}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{S}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{20}$$

the Spearman rank correlation coefficient is defined as the Pearson correlation between ranked variables

$$\text{Spearman} = \frac{\sum_{i=1}^{S}(r_i - \bar{r})(\hat{r}_i - \bar{\hat{r}})}{\sqrt{\sum_{i=1}^{S}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{S}(\hat{r}_i - \bar{\hat{r}})^2}}, \tag{21}$$

where r_i and $\hat{r}_i$ denote the ranks of y_i and $\hat{y}_i$, respectively. For residue-level tasks, S corresponds to the total number of residues across all proteins, whereas for protein-level tasks, S denotes the number of protein samples.

Classification Task. For active site identification, we adopt Area Under the Precision-Recall Curve (AUPRC), Matthews Correlation Coefficient (MCC), and False Positive Rate (FPR). Given binary labels and predictions, we compute true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), where TP denotes correctly predicted positive samples, TN denotes correctly predicted negative samples, FP denotes incorrectly predicted positive samples, and FN denotes incorrectly predicted negative samples. Based on these quantities, precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}. \tag{22}$$

AUPRC is computed as the area under the precision–recall curve obtained by varying the classification threshold. MCC and FPR are defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{23}$$

$$\text{FPR} = \frac{FP}{FP + TN}. \tag{24}$$

E Additional Ablation Studies

The effects of cross-box interactions and cryo-EM maps on protein flexibility prediction are analyzed in Section 4.4. Here, we provide the full ablation results for this task in Table 7, with additional analyses of local resolution, synthesized maps, pretraining tasks, and sequence encoder variants.

Local resolution. We calculated voxel-wise local resolution using ResMap[3] and added the resolution value of each voxel as an additional feature channel to the original box representation. Incorporating explicit local resolution information brings only marginal changes, suggesting that CryoProt can already capture local density quality variations from the original density representations.

Synthesized maps. We examine whether synthesized density maps can replace experimental maps or expand the pretraining dataset. For synthetic replacement, we replace all experimental maps with ChimeraX-synthesized maps. For synthetic augmentation, we add 8,927 synthesized maps for proteins previously excluded because of their lower map resolution (4–8 Å), expanding the pretraining set to 29,457 proteins. Replacement and augmentation yield RMSE values of 0.521 and 0.503, respectively, compared with 0.484 for CryoProt. Neither setting improves performance over pretraining with experimental maps alone.

Sequence encoder variants. We replace the default ESM2-150M sequence encoder with ESM-1b, ESM2-8M, ESM2-35M (Lin et al., 2022), ProtBert, ProtAlbert, or ProtXLNet (Elnaggar et al., 2021) to assess sensitivity to encoder choice. ESM2-150M achieves the lowest RMSE and highest

[3] https://resmap.sourceforge.net/

correlations among the tested encoders. Within the ESM2 family, RMSE decreases from 0.509 to 0.499 and 0.484 as model size increases from 8M to 35M and 150M parameters, respectively. These results support the choice of ESM2-150M as the sequence encoder for CryoProt.

Ablation results for active site identification, binding affinity prediction, and $\Delta\Delta G$ prediction are reported in Tables 8, 9, and 10, respectively.

Table 7: Ablation study on protein flexibility prediction. The best results are in bold.

Ablation	Model	RMSE ↓	Pearson ↑	Spearman ↑
Box interactions	w/o MLA	0.546 ± 0.019	0.271 ± 0.020	0.278 ± 0.025
	w/ local-window attention	0.519 ± 0.018	0.291 ± 0.019	0.296 ± 0.023
	w/ self-attention	0.497 ± 0.017	0.302 ± 0.018	0.307 ± 0.021
Cryo-EM maps	w/ synthesized maps	0.521 ± 0.019	0.286 ± 0.021	0.293 ± 0.023
	w/ PDB structure	0.536 ± 0.020	0.276 ± 0.020	0.284 ± 0.022
	w/ sequence	0.551 ± 0.023	0.263 ± 0.024	0.270 ± 0.026
Local resolution	w/ local resolution	0.489 ± 0.020	0.307 ± 0.038	0.312 ± 0.043
Synthesized maps	w/ synthesized maps	0.521 ± 0.019	0.286 ± 0.021	0.293 ± 0.023
	Synthetic augmentation	0.503 ± 0.017	0.299 ± 0.018	0.303 ± 0.022
Pretraining tasks	w/o pretraining task2	0.498 ± 0.015	0.308 ± 0.016	0.309 ± 0.021
	w/o pretraining task3	0.508 ± 0.016	0.302 ± 0.017	0.306 ± 0.022
	w/o pretraining task2 + 3	0.532 ± 0.018	0.284 ± 0.019	0.291 ± 0.024
Sequence encoder	ESM-1b	0.518 ± 0.017	0.295 ± 0.018	0.298 ± 0.023
	ESM2-8M	0.509 ± 0.016	0.301 ± 0.017	0.304 ± 0.022
	ESM2-35M	0.499 ± 0.015	0.307 ± 0.016	0.309 ± 0.021
	ProtBert	0.503 ± 0.015	0.304 ± 0.016	0.307 ± 0.021
	ProtAlbert	0.506 ± 0.016	0.300 ± 0.017	0.303 ± 0.022
	ProtXLNet	0.504 ± 0.015	0.302 ± 0.016	0.305 ± 0.021
	CryoProt	**0.484 ± 0.021**	**0.310 ± 0.041**	**0.315 ± 0.046**

Table 8: Ablation study on active site identification. The best results are in bold.

Ablation	Model	AUPRC ↑	MCC ↑	FPR ↓
Box interactions	w/o MLA	0.640 ± 0.021	0.575 ± 0.019	0.026 ± 0.015
	w/ local-window attention	0.728 ± 0.019	0.641 ± 0.020	0.025 ± 0.002
	w/ self-attention	0.738 ± 0.018	0.651 ± 0.019	0.023 ± 0.002
Cryo-EM maps	w/ synthesized maps	0.724 ± 0.020	0.636 ± 0.021	0.025 ± 0.002
	w/ PDB structure	0.711 ± 0.022	0.621 ± 0.023	0.027 ± 0.002
	w/ sequence	0.694 ± 0.024	0.603 ± 0.025	0.029 ± 0.002
Local resolution	w/ local resolution	0.740 ± 0.021	0.661 ± 0.020	0.022 ± 0.002
Synthesized maps	Synthetic replacement	0.724 ± 0.020	0.636 ± 0.021	0.025 ± 0.002
	Synthetic augmentation	0.737 ± 0.019	0.655 ± 0.020	0.023 ± 0.002
Pretraining tasks	w/o pretraining task2	0.713 ± 0.019	0.638 ± 0.017	0.023 ± 0.013
	w/o pretraining task3	0.701 ± 0.020	0.624 ± 0.018	0.023 ± 0.014
	w/o pretraining task2 + 3	0.662 ± 0.021	0.588 ± 0.019	0.025 ± 0.015
Sequence encoder	ESM-1b	0.695 ± 0.020	0.618 ± 0.018	0.023 ± 0.014
	ESM2-8M	0.688 ± 0.019	0.611 ± 0.017	0.024 ± 0.014
	ESM2-35M	0.706 ± 0.018	0.629 ± 0.016	0.023 ± 0.013
	ProtBert	0.718 ± 0.017	0.641 ± 0.015	0.022 ± 0.012
	ProtAlbert	0.711 ± 0.018	0.633 ± 0.016	0.022 ± 0.013
	ProtXLNet	0.714 ± 0.018	0.636 ± 0.016	0.023 ± 0.013
	CryoProt	**0.746 ± 0.023**	**0.669 ± 0.019**	**0.021 ± 0.001**

Table 9: Ablation study on binding affinity prediction. The best results are in bold.

Ablation	Model	RMSE ↓	Pearson ↑	Spearman ↑
Box interactions	w/o MLA	0.575 ± 0.041	0.445 ± 0.076	0.448 ± 0.079
	w/ local-window attention	0.528 ± 0.040	0.478 ± 0.059	0.486 ± 0.060
	w/ self-attention	0.507 ± 0.038	0.503 ± 0.055	0.509 ± 0.057
Cryo-EM maps	w/ synthesized maps	0.531 ± 0.041	0.471 ± 0.061	0.478 ± 0.064
	w/ PDB structure	0.548 ± 0.038	0.451 ± 0.057	0.459 ± 0.061
	w/ sequence	0.572 ± 0.043	0.417 ± 0.060	0.425 ± 0.063
Local resolution	w/ local resolution	0.495 ± 0.039	0.516 ± 0.080	0.520 ± 0.084
Synthesized maps	Synthetic replacement	0.531 ± 0.041	0.471 ± 0.061	0.478 ± 0.064
	Synthetic augmentation	0.501 ± 0.039	0.511 ± 0.079	0.515 ± 0.083
Pretraining tasks	w/o pretraining task2	0.509 ± 0.037	0.507 ± 0.080	0.511 ± 0.083
	w/o pretraining task3	0.528 ± 0.038	0.492 ± 0.078	0.498 ± 0.081
	w/o pretraining task2 + 3	0.563 ± 0.041	0.461 ± 0.076	0.465 ± 0.079
Sequence encoder	ESM-1b	0.541 ± 0.040	0.472 ± 0.075	0.476 ± 0.078
	ESM2-8M	0.532 ± 0.039	0.483 ± 0.077	0.487 ± 0.080
	ESM2-35M	0.518 ± 0.038	0.497 ± 0.079	0.501 ± 0.082
	ProtBert	0.523 ± 0.038	0.492 ± 0.078	0.496 ± 0.081
	ProtAlbert	0.527 ± 0.039	0.488 ± 0.077	0.492 ± 0.080
	ProtXLNet	0.525 ± 0.038	0.490 ± 0.078	0.494 ± 0.081
	CryoProt	**0.486 ± 0.039**	**0.524 ± 0.082**	**0.527 ± 0.087**

Table 10: Ablation study on $\Delta\Delta G$ prediction. The best results are in bold.

Ablation	Model	RMSE ↓	Pearson ↑	Spearman ↑
Box interactions	w/o MLA	1.295 ± 0.042	0.548 ± 0.024	0.438 ± 0.010
	w/ local-window attention	1.151 ± 0.035	0.586 ± 0.029	0.454 ± 0.022
	w/ self-attention	1.096 ± 0.034	0.602 ± 0.027	0.476 ± 0.020
Cryo-EM maps	w/ synthesized maps	1.164 ± 0.036	0.579 ± 0.030	0.448 ± 0.021
	w/ PDB structure	1.213 ± 0.039	0.556 ± 0.032	0.421 ± 0.024
	w/ sequence	1.286 ± 0.042	0.523 ± 0.034	0.392 ± 0.026
Local resolution	w/ local resolution	1.061 ± 0.034	0.605 ± 0.025	0.482 ± 0.008
Synthesized maps	Synthetic replacement	1.164 ± 0.036	0.579 ± 0.030	0.448 ± 0.021
	Synthetic augmentation	1.075 ± 0.035	0.601 ± 0.026	0.479 ± 0.009
Pretraining tasks	w/o pretraining task2	1.108 ± 0.036	0.596 ± 0.025	0.476 ± 0.008
	w/o pretraining task3	1.162 ± 0.039	0.585 ± 0.024	0.468 ± 0.009
	w/o pretraining task2 + 3	1.248 ± 0.042	0.561 ± 0.024	0.451 ± 0.010
Sequence encoder	ESM-1b	1.205 ± 0.041	0.568 ± 0.024	0.456 ± 0.010
	ESM2-8M	1.168 ± 0.040	0.579 ± 0.025	0.463 ± 0.009
	ESM2-35M	1.120 ± 0.038	0.592 ± 0.025	0.472 ± 0.008
	ProtBert	1.135 ± 0.039	0.586 ± 0.025	0.468 ± 0.009
	ProtAlbert	1.148 ± 0.040	0.582 ± 0.024	0.465 ± 0.009
	ProtXLNet	1.142 ± 0.039	0.584 ± 0.025	0.467 ± 0.009
	CryoProt	**1.036 ± 0.033**	**0.611 ± 0.025**	**0.488 ± 0.007**

F Additional Parameter Sensitivity Analysis

We further investigate the sensitivity of CryoProt to the number of attention heads and the mask ratio on protein flexibility prediction. As shown in Fig. 6, increasing the number of attention heads from 2 to 8 consistently improves performance, indicating that a moderate number of heads helps the model capture diverse interaction patterns among spatial boxes. However, performance slightly decreases when the number of heads is further increased to 16, suggesting that excessive attention heads may introduce redundant or overly fragmented interaction patterns without providing addi-

tional useful information. For the mask ratio, the best overall performance is achieved at 0.15. A smaller mask ratio may provide insufficient reconstruction difficulty and therefore limit the learning of informative contextual representations, whereas a larger ratio removes too much local information and makes the reconstruction task unnecessarily challenging. These results indicate that a moderate mask ratio provides a better balance between representation learning difficulty and available contextual information, leading to more effective pretraining.

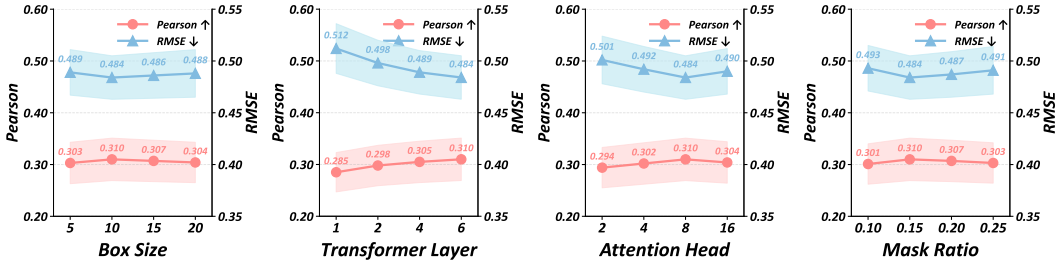


Figure 6: Parameter sensitivity analysis.

G Case on Top-L Coverage

Definition of Top-L Coverage. Given residue embeddings $\{\mathbf{E}_i\}_{i=1}^{L}$, where L is the protein sequence length, we compute the pairwise cosine similarity:

$$S_{ij} = \frac{\mathbf{E}_i \cdot \mathbf{E}_j}{\|\mathbf{E}_i\|_2 \|\mathbf{E}_j\|_2}. \tag{25}$$

We select the top-L residue pairs with the highest similarity scores from the strict upper triangle ($1 \le i < j \le L$), excluding self-pairs and duplicate pairs, and denote this set as $\mathcal{P}_{\text{top}}$. From the native structure, we compute the residue distance matrix $\mathbf{D}$, where D_{ij} is the C_α distance between residues i and j. Top-L coverage is defined as

$$\text{Top-}L = \frac{1}{L} \sum_{(i,j)\in\mathcal{P}_{\text{top}}} \mathbb{I}(D_{ij} < 10\,\text{Å}), \tag{26}$$

where $\mathbb{I}(\cdot)$ is the indicator function. A higher Top-L coverage indicates that a larger fraction of the selected residue pairs are spatially proximal.

Resolution setting. The resolution ranges used for pretraining and the Top-L analysis serve different purposes. During pretraining, we restrict cryo-EM maps to the 2–4 Å range to ensure sufficiently reliable local density information for box construction and representation learning. In contrast, the Top-L analysis is designed as an additional generalization study to examine whether the learned representations can preserve residue-level spatial relationships for proteins associated with a broader range of cryo-EM resolutions. We therefore evaluate proteins from 2–4 Å, 4–6 Å, and 6–8 Å intervals. Importantly, proteins from the lower-resolution intervals are used only for evaluation and are not introduced as additional pretraining data.

Interpretation of Top-L Coverage. Top-L coverage is used as an unsupervised probe of whether highly similar residue embeddings correspond to spatially proximal residue pairs, rather than as a direct measure of structure reconstruction. Since only a small fraction of all residue pairs are within 10 Å, relatively low absolute coverage values are expected without explicit distance or contact supervision. Nevertheless, CryoProt consistently outperforms sequence-only ESM2 representations, achieving a 16.2% relative improvement over the ESM2. This indicates that cryo-EM-aware pretraining enables the learned representations to capture richer residue-level spatial relationships.

Per-protein Top-L Coverage. To provide a more detailed case analysis, we report per-protein Top-L coverage results across different resolution ranges. Specifically, we randomly select 40 proteins from each resolution interval (2–4 Å, 4–6 Å, and 6–8 Å), resulting in a total of 120 proteins. For each protein, we report its PDB ID together with the corresponding Top-L coverage obtained by CryoProt, as summarized in Table 11. The results show that CryoProt consistently captures

meaningful residue-level spatial relationships across proteins with different cryo-EM resolutions, including lower-resolution cases. This further demonstrates the robustness and generalizability of the learned representations under varying data quality conditions.

Table 11: Top-L Coverage across proteins from different resolution ranges.

Range	Proteins (PDB ID / Top-L Coverage)			
2–4 Å	9CZL / 0.265	7SK9 / 0.289	5V7V / 0.145	7DQA / 0.089
	9CB9 / 0.185	8J8H / 0.072	8JPD / 0.196	8XQI / 0.104
	8X63 / 0.274	7SNF / 0.041	8FY4 / 0.082	9KCP / 0.506
	9DMY / 0.416	9QVF / 0.008	8Y0Q / 0.106	8EHS / 0.000
	9U9B / 0.144	6Z8D / 0.017	9JA5 / 0.026	9HIX / 0.264
	7WL3 / 0.151	7NSL / 0.454	9PKV / 0.186	9JBO / 0.647
	9T3P / 0.000	8B3A / 0.000	9J97 / 0.057	7STF / 0.250
	7P1K / 0.072	9UT9 / 0.121	9MOQ / 0.462	8XCY / 0.298
	6PUZ / 0.057	8SFJ / 0.064	9NOU / 0.132	7XBX / 0.164
	3J89 / 0.000	8G4E / 0.068	9J12 / 0.058	6WUJ / 0.160
4–6 Å	8E4C / 0.113	5LVC / 0.107	7B6Y / 0.375	8TIN / 0.323
	8VID / 0.151	6XE6 / 0.200	8W5Q / 0.090	9VKS / 0.135
	8T53 / 0.067	7RD8 / 0.280	8Y32 / 0.071	9IMH / 0.000
	7YOZ / 0.005	7JLX / 0.002	7WRJ / 0.065	6EGX / 0.124
	9R5K / 0.057	6R1T / 0.048	8I8C / 0.025	3JBF / 0.087
	5H1C / 0.008	7OZR / 0.072	7NIQ / 0.250	6T6V / 0.345
	5A6G / 0.649	7TFS / 0.144	9HM6 / 0.057	8VE6 / 0.009
	8AY4 / 0.117	7OZ3 / 0.009	6OFJ / 0.095	8JTS / 0.306
	6V6C / 0.159	7TT7 / 0.152	9B3H / 0.096	7RO5 / 0.468
	7SSG / 0.175	6MI8 / 0.203	8PQ5 / 0.132	9FH4 / 0.588
6–8 Å	8QI0 / 0.276	5NL2 / 0.066	5ABB / 0.339	3J15 / 0.162
	7QE0 / 0.118	7PGW / 0.046	7PEZ / 0.044	6WJG / 0.058
	9NTT / 0.187	6RRT / 0.012	8QHV / 0.307	8C9C / 0.344
	3JBJ / 0.052	5FL2 / 0.085	9EAD / 0.119	6LM3 / 0.195
	9DVC / 0.090	3J5V / 0.177	6BU9 / 0.168	8I2H / 0.271
	8PHT / 0.000	6HWW / 0.006	7FIF / 0.217	8FG2 / 0.049
	8WJO / 0.058	8I21 / 0.054	8ETT / 0.111	4CG6 / 0.295
	7XXL / 0.076	9TQB / 0.032	6N89 / 0.099	8E3C / 0.137
	7YLM / 0.052	8Z9V / 0.141	8SL3 / 0.174	8FMB / 0.141
	3MFP / 0.283	2I68 / 0.013	9FTM / 0.000	8BGG / 0.072

H Consistency Between Inferred and Real Map Representations

To evaluate whether sequence-inferred map representations preserve information from real cryo-EM maps, we compare them with representations extracted directly by the Map Encoder. Specifically, we use 120 proteins from the Top-L coverage study, with 40 proteins in each resolution range. Representation consistency is evaluated using cosine similarity and Pearson correlation. As shown in Fig. 7, for proteins in the 2–4 Å range, the average cosine similarity and Pearson correlation reach 0.82 and 0.78, respectively, indicating strong agreement between the sequence-inferred and real map-derived representations. The consistency gradually decreases at lower resolutions, reaching 0.68 and 0.61 in the 6–8 Å range, likely due to the reduced structural detail in lower-resolution cryo-EM maps. Nevertheless, substantial agreement is retained across all resolution ranges. These results suggest that CryoProt can preserve map-related information in sequence-inferred representations without requiring cryo-EM maps during downstream inference.

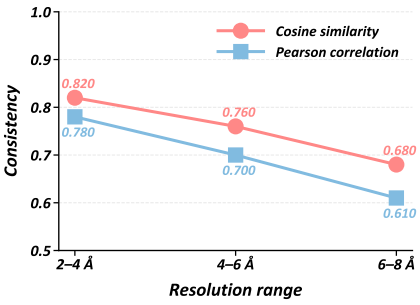


Figure 7: Consistency between inferred and real map representations.

I MORE VISUALIZATION RESULTS

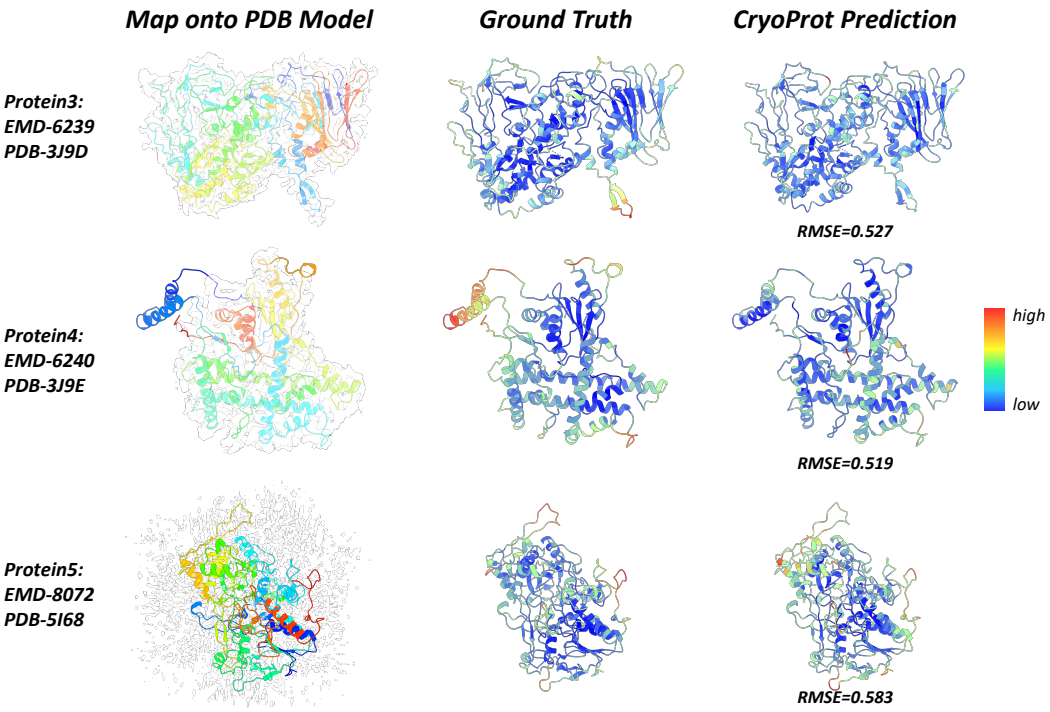


Figure 8: More visualization results.