CryoSAMU: Enhancing 3D Cryo-EM Density Maps of Protein Structures at Intermediate Resolution with Structure-Aware Multimodal U-Nets

Chenwei Zhang [1] **Khanh Dao Duc** [2]

Abstract

Enhancing cryogenic electron microscopy (cryo-EM) 3D density maps at intermediate resolution (4-8 Å) is crucial in protein structure determination. Recent advances in deep learning have led to the development of automated approaches for enhancing experimental cryo-EM density maps. Yet, these methods are not optimized for intermediate-resolution maps and rely on map density features alone. To address this, we propose CryoSAMU, a novel method designed to enhance 3D cryo-EM density maps of protein structures using structure-aware multimodal U-Nets and trained on curated intermediate-resolution density maps. We comprehensively evaluate CryoSAMU across various metrics and demonstrate its competitive performance compared to state-of-the-art methods. Notably, CryoSAMU achieves significantly faster processing speed, showing promise for future practical applications. Our code is available at https://github.com/chenwei-zhang/CryoSAMU.

1. Introduction

Cryogenic electron microscopy (Cryo-EM) has become one of the most prevalent techniques in structural biology for determining protein structures, thereby accelerating structure-based drug discovery (Renaud et al., 2018; Merk et al., 2016). Cryo-EM projects a series of 2D images, which are then reconstructed into 3D electron density maps, providing voxelized representations of proteins. While cryo-EM 3D maps serve as the basis for molecular structure determination, using raw maps is usually not possible as they often lack contrast due to various factors, including molecular motion and heterogeneity, imaging artifacts, and incoherent averaging of image data (Rosenthal, 2019). To address these limitations, various approaches have been developed to enhance map quality by sharpening or modifying map densities (Terwilliger et al., 2018b; Kimanius et al., 2016; Ramírez-Aportela et al., 2020; Sanchez-Garcia et al., 2021; Dai et al., 2023; Maddhuri Venkata Subramaniya et al., 2023; He et al., 2023). Traditional methods rely on B-factor correction, which can be applied globally (Terwilliger et al., 2018b; Kimanius et al., 2016) and locally (Kaur et al., 2021; Ramírez-Aportela et al., 2020). However, these methods struggle with maps exhibiting varying signal-to-noise ratios and lacking prior knowledge (e.g. local resolution) (He et al., 2023).

With recent advancements in deep learning (DL), fully data-driven methods have been developed to automatically enhance raw cryo-EM maps for protein structure modeling. Leveraging neural networks such as convolutional neural networks (CNNs) (LeCun & Bengio, 1995), generative adversarial networks (GANs) (Creswell et al., 2018), and Transformers (Vaswani et al., 2017), these methods achieved promising results in map enhancement. Yet, they are not optimized for intermediate-resolution maps (i.e., 4-8 Å (He et al., 2023)) and rely solely on a single modality—the density map itself—during neural network training, overlooking other relevant modalities such as structural information. This limitation restricts their ability to generalize across diverse protein structures and prevents them from fully leveraging complementary biological information. To address these shortcomings, we thus introduce **CryoSAMU**, a novel approach that combines 3D map features with structural embeddings derived from the pretrained protein language model ESM-IF1 (Hsu et al., 2022) to enhance 3D **Cryo**-EM density maps with **S**tructure-**A**ware **M**ultimodal **U**-Nets. Our main contributions are:

- We propose the first multimodal network that integrates structural information into a 3D U-Net model using cross-attention mechanisms for cryo-EM map enhancement.
- We develop a self-attention-based post-processing procedure for ESM-IF1's structural embeddings, effectively preserving both chain and residue relationships

[*]Equal contribution [1]Department of Computer Science, University of British Columbia, Vancouver, Canada [2]Department of Mathematics, University of British Columbia, Vancouver, Canada. Correspondence to: Chenwei Zhang <cwzhang@cs.ubc.ca>, Khanh Dao Duc <kdd@math.ubc.ca>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

while maintaining structural integrity.

- We train CryoSAMU on a curated dataset of joint density maps at intermediate resolution and associated protein structures, optimizing it for map enhancement.
- We benchmark CryoSAMU against state-of-the-art approaches across various evaluation metrics over diverse tested samples. We achieve competitive level of performance but with significantly faster processing speeds (approximately 4.2 to 16.7 times), making our method well-suited for large-scale and practical applications.
- Our ablation study demonstrates significant improvement brought from integrating structural information.

2. Related Work

2.1. Existing map enhancement methods

Conventional map enhancement (sharpening) approaches, including Phenix Autosharpen (Terwilliger et al., 2018b) and RELION postprocessing (Kimanius et al., 2016), are based on global B-factor correction. This technique enhances the amplitude of high-frequency Fourier components in raw cryo-EM maps. However, global B-factor-based methods encounter difficulties with maps exhibiting heterogeneous local resolutions, often leading to over- or under-sharpening in specific regions. Despite local B-factor-based sharpening algorithms (Kaur et al., 2021; Ramírez-Aportela et al., 2020) have been developed to alleviate this limitation, these methods still suffer from poor accuracy in estimating the local resolution of maps, which is crucial for precise local B-factor sharpening.

DeepEMhancer (Sanchez-Garcia et al., 2021) is a pioneering DL-based fully automatic method that leverages a 3D U-Net model to mimic local sharpening effects and enhance map features. Subsequently, CryoFEM (Dai et al., 2023) that employs convolutional neural networks (CNNs) and EM-GAN (Maddhuri Venkata Subramaniya et al., 2023) that utilizes generative adversarial networks (GANs) have been introduced to further enhance cryo-EM maps. Most recently, with the emergence of vision transformers, EM-Ready (He et al., 2023), which adopts a Swin transformer architecture (Liu et al., 2021), has shown superior performance in enhancing map quality for accurate protein structure modeling.

2.2. Protein large language models

The advancement of protein large language models (pLLMs) has enabled unprecedented insights into protein structure, function, and evolution (Rives et al., 2021; Lin et al., 2023; Hayes et al., 2025; Hsu et al., 2022; Brandes et al., 2022; Heinzinger et al., 2023; Dauparas et al., 2022). In analogy to human texts, protein sequences are treated as "biological texts" and input into pLLMs to capture contextual information inherent in the sequences. Notable examples of such models include the ESM family (Rives et al., 2021; Lin et al., 2023; Hayes et al., 2025), which are pretrained on vast datasets of protein sequences using the masked language modeling strategy, allowing them to develop rich representations that encapsulate evolutionary information.

Addressing the inverse problem of predicting protein sequences from given structures, ESM-IF1 (Hsu et al., 2022) has been developed. Trained on 12 million protein structures derived from AlphaFold2 (Jumper et al., 2021), ESM-IF1 predicts protein sequences from backbone atom coordinates. It is specifically designed to encode both sequence and structural information, including backbone geometry, side chain orientations, and secondary structure elements. These traits make ESM-IF1 a compelling choice for generating structure-aware embeddings that complement the map-only modality.

3. Method

3.1. Dataset of protein structures and density maps

Our dataset was built with a set of cryo-EM density maps at resolutions from 4.0 Å to 7.9 Å from the EMDB databank (Lawson et al., 2016) and their associated protein structures from the PDB databank (Berman et al., 2002). To ensure that density maps are properly aligned with their corresponding PDB structures, we excluded maps and PDBs from the dataset if: (i) maps contain extensive regions without or misaligned corresponding PDB structures; (ii) maps contain other macromolecules except proteins; (iii) PDB structures contain backbone atoms only and/or unknown residues. Furthermore, to enhance training efficiency, we measured the correlation between map-PDB pairs using ChimeraX (Pettersen et al., 2004), and removed pairs with correlation score lower than 0.65 to filter out incomplete mappings. To avoid data redundancy, we measured the sequence identity between PDB structures, and retained only one if identity is greater than 30 %. As a result, a total of 384 pairs of cryo-EM maps and associated PDB structures remained. Among these data, 247 (∼65 %), 62 (∼15 %), and 75 (∼20 %) map-PDB pairs were selected as training, validation, and test sets, respectively. Details are listed in Supplementary Tables S1 to S3.

3.2. Multimodal representations of protein structures

Generating 3D target maps from protein structures For input experimental maps (denoted as *ExpMaps*) in training and validation sets, we simulated the corresponding target maps (denoted as *TgtMaps*) from associated protein structures using the `StructureBlurrer` package in TEMPy2 (Cragnolini et al., 2021). The simulation was performed with a grid interval of 1 Å and a resolution cut-

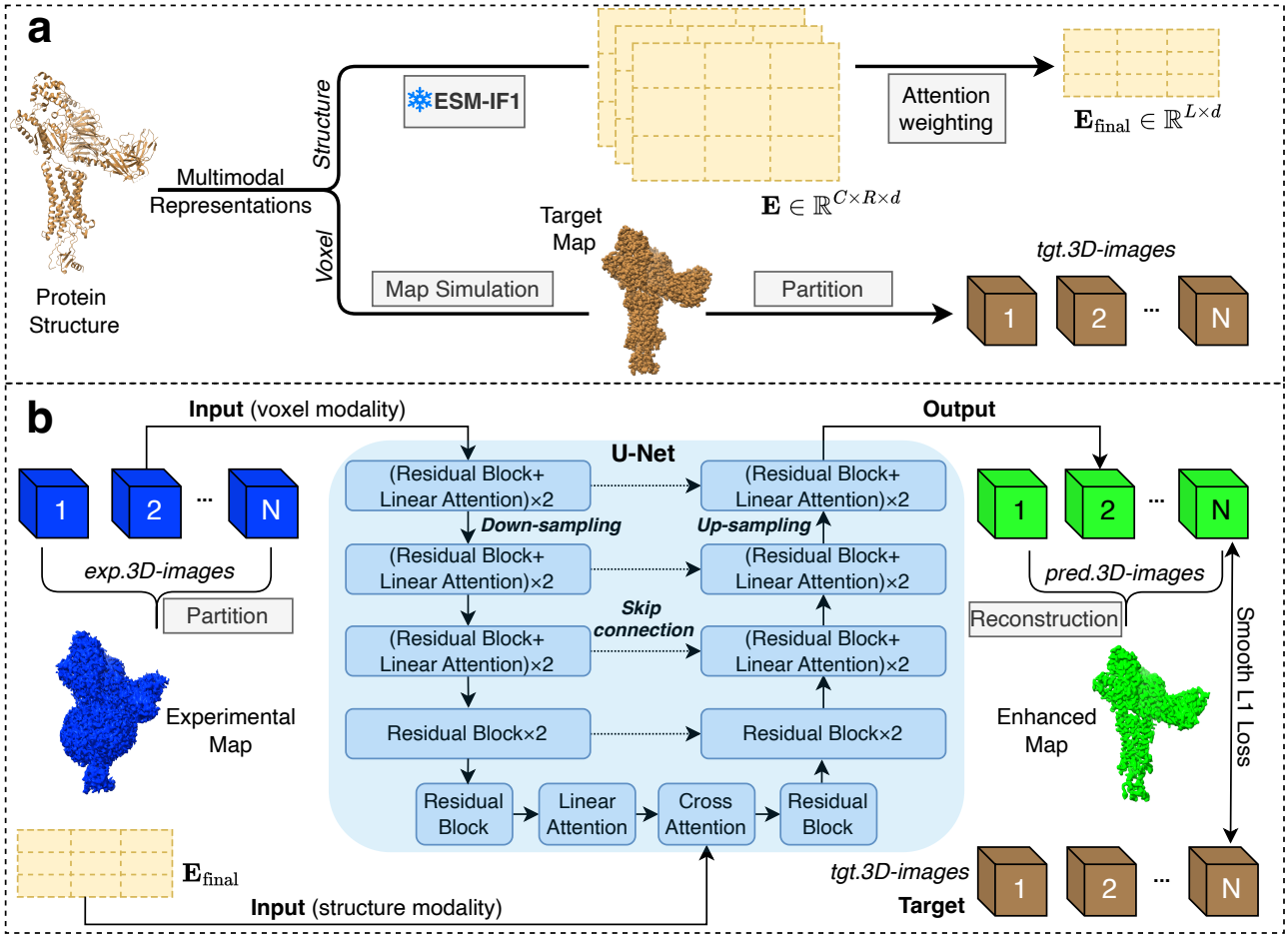

Figure 1. Overview of the CryoSAMU framework. **a** Generating protein multimodal representations: structure features are derived from a frozen pretrained ESM-IF1 model with self-attention weighting for a fixed-size representation; map voxel features are simulated via resolution-lowering point spread function and partitioned into smaller cubes. **b** The CryoSAMU architecture. The experimental map is partitioned into smaller cubes and processed by a U-Net with residual blocks and linear attention modules. Structural embeddings are integrated into the bottleneck layer with cross-attention mechanism. The output cubes are reconstructed into the full-size enhanced map. The illustrated protein complex is a CX3CL1-US28-G11iN18-scFv16 in TL-state (PDB-7RKF, EMDB-24496, reported resolution of 4.00 Å) (Tsutsumi et al., 2022).

off at 2 Å, based on the convolution of atom points with resolution-lowering point spread functions. Given a PDB structure with M atoms, the simulated density ρ at grid point x is calculated by:

$$\rho(\mathbf{x}) = \sum_{i=1}^{M} \theta Z_i e^{-k|\mathbf{x}-\mathbf{r}_i|^2}, \tag{1}$$

where Z_i and $\mathbf{r}_i$ refer to the atomic number and the position vector of the i-th heavy atom, respectively. Here, θ is a scaling factor and k is defined based on resolution (He et al., 2023; 2022).

Resampling 3D maps We first resampled both *ExpMaps* and *TgtMaps* to 1 Å/voxel since the cryo-EM maps vary in voxel size. Subsequently, we normalized the density values to a range of 0 to 1 using the 99.9th percentile density value of each map. Due to GPU memory constraints, we partitioned *ExpMaps* and *TgtMaps* into smaller 3D subvolume pairs (denoted as *exp.3D-images* and *tgt.3D-images*) with size of $64\times64\times64$, the largest feasible size that allows for a sufficient batch size (See Figure 1a.). To mitigate boundary artifacts during truncation, we applied zero-padding of 64 voxels on each side along all dimensions. As a result, a total of 29829 *exp-tgt* image pairs were yielded for network training and 4642 for validation.

Generating structural embeddings We employed ESM-IF1 (Hsu et al., 2022) to generate protein structural embeddings, which will serve as an additional modality for network training. Specifically, we derived the embeddings by first extracting backbone coordinates (N, Cα, and C atoms)

from a PDB file, ensuring that only standard residues with complete backbone information are included. We then fed these coordinates into ESM-IF1 to generate embeddings for each protein chain. Since the lengths of chains varied, we applied zero-padding to standardize the embeddings.

Fixed-size representation with attention weighting Following the generation of embeddings, we implemented self-attention weighting to create fixed-size representations while preserving the intrinsic relationships between chains and residues. To this end, we computed attention weights based on embedding similarity to identify the most informative regions. Specifically, given a PDB structure containing C chains and R residues per chain, its structural embedding derived from ESM-IF1 is denoted as $\mathbf{E} \in \mathbb{R}^{C\times R\times d}$, where $d = 512$ is the embedding dimension. We carried out the refinement process in several steps. First, we computed chain-level embeddings by averaging across residues:

$$\mathbf{E}_{\text{chain}} = \frac{1}{R}\sum_{j=1}^{R}\mathbf{E}_{:,j,:}, \quad \mathbf{E}_{\text{chain}} \in \mathbb{R}^{C\times d}. \tag{2}$$

Next, we computed the similarity matrix to determine the relative importance of each chain:

$$\mathbf{S} = \mathbf{E}_{\text{chain}} \cdot \mathbf{E}_{\text{chain}}^{T}, \quad \mathbf{S} \in \mathbb{R}^{C\times C}, \tag{3}$$

where each element $\mathbf{S}_{ij}$ represents the similarity between chain i and chain j: $\mathbf{S}_{ij} = \mathbf{E}_{\text{chain},i} \cdot \mathbf{E}_{\text{chain},j}^{T}$. To derive attention weights, we applied a column-wise softmax function to $\mathbf{S}$:

$$\mathbf{W}_{ij} = \frac{\exp(\mathbf{S}_{ij})}{\sum_{k=1}^{C}\exp(\mathbf{S}_{ik})}. \tag{4}$$

We then aggregated these weights across chains to assign a single importance weight per chain:

$$\mathbf{w}_i = \frac{1}{C}\sum_{j}^{C}\mathbf{W}_{ij}, \quad i = 1, 2, \ldots, C. \tag{5}$$

These weights $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, .., \mathbf{w}_C]$ reflect the relative importance of each chain, and we leveraged them to aggregate chain-level embeddings into a unified representation:

$$\mathbf{E}_{\text{pooled}} = \sum_{i=1}^{C} w_i \mathbf{E}_{i,:,:}, \quad \mathbf{E}_{\text{pooled}} \in \mathbb{R}^{R\times d}. \tag{6}$$

We further measured the importance of each residue in $\mathbf{E}_{\text{pooled}}$ using a residue-level similarity matrix. Following the same procedure as the chain-level weighting, we obtained a scalar weight α_j for each residue j, where $j = 1, 2, \ldots, R$. Finally, we applied min-max normalization and resampled the embedding $\mathbf{E}_{\text{pooled}}$ based on the attention weights to a fixed-size representation, $\mathbf{E}_{\text{final}} \in \mathbb{R}^{L\times d}$, where $L = 800$. When the input length $R > L$, we selected the top-L residues with the highest attention weights. Conversely, when $R < L$, we sorted the residues by their attention weights and repeated them $\lceil L/R \rceil$ times to reach the target length, ensuring each resulting embedding maintains rich representations and consistent dimensions. Compared to simple averaging or top-L residue selection, we note that this hierarchical attention strategy preserves the distinct contributions of each residue through explicit weighting, and can capture biologically informative patterns across diverse protein complexes.

3.3. The model architecture

We proposed a structure-aware multimodal 3D U-Net, as depicted in Figure 1b. The network contains an encoder, bottleneck, and decoder, interconnected by skip connections.

Encoder The input to the encoder is a 3D volume with a single channel. The encoder comprises four hierarchical layers. The first three layers each consist of two residual blocks, with each block incorporating a group normalization, a SiLU activation (Elfwing et al., 2018), and a dropout (p=0.2), followed immediately by a linear self-attention module with 4 heads (Katharopoulos et al., 2020) to capture long-range (global) dependencies across voxels. The channel depth progressively increases as features are abstracted. In the fourth layer, only residual blocks are employed, producing a higher-level feature representation without the addition of attention modules.

Bottleneck At the bottleneck layer, the feature representation is first refined by a residue block and then by a linear self-attention module. Subsequently, a cross-attention block is introduced to fuse and align the volumetric features with structural embeddings using multi-head attention with 4 heads, where queries are derived from the volume features and keys/values from structural embeddings. This process enables structural conditioning while preserving spatial relationships. A second residual block is then applied to further fuse the combined features from both modalities.

Decoder The decoder follows a symmetric architecture to the encoder. Feature maps are progressively upsampled using nearest-neighbor interpolation combined with 3D convolutions, and skip connections incorporate corresponding features from the encoder. Finally, a group normalization, a SiLU activation, and a concluding 3D convolution project the processed features to a single output channel.

3.4. Network training and inference

Protein structural embeddings provide an additional modality containing structure information, serving as key-value pairs in the attention mechanism when training. However, since these embeddings are unavailable during validation and inference, we implemented a specialized mode in which

the network bypasses the cross-attention operation. In this mode, the network relies exclusively on feedforward transformations with residual connections. This design aligns with the well-established principle of conditional training, where auxiliary modalities, despite not available at inference, are used during training to improve the quality and generalization of learned representations from the primary input (Shen et al., 2023). Moreover, this design also maintains consistency between training and validation/inference phases while preserving the feature representations.

Training During training, CryoSAMU accepts an *exp.3D-image* and its corresponding structural embedding as input and generates an enhanced 3D image (denoted as *pred.3D-image*). Previous studies have shown that L1 loss performs well in similar tasks (Sanchez-Garcia et al., 2021; Dai et al., 2023). However, to improve training stability in the presence of noisy data and outliers that are common in cryo-EM maps, we employed the smooth L1 loss to encourage the generator to minimize the difference between the output *pred.3D-image*, X, and the target *tgt.3D-image*, Y:

$$\text{SmoothL1Loss}(X, Y) = \begin{cases} 0.5(X-Y)^2, & \text{if } |X-Y|<1, \\ |X-Y| - 0.5, & \text{otherwise.} \end{cases} \tag{7}$$

Moreover, to enhance the network's robustness, we employed *TorchIO* (Pérez-García et al., 2021) for data augmentation, including random Gaussian noise, anisotropy, and blurring.

Inference During inference, the input experimental map was first zero-padded and divided into smaller cubes ($64 \times 64 \times 64$), following the same strategy used for training data. Each cube was then individually processed by the trained neural network to generate the enhanced cube. These enhanced cubes were subsequently reassembled to reconstruct the map as its original dimensions. To prevent loss of spatial information and ensure smooth transition between cubes, only the central $50 \times 50 \times 50$ voxels from each enhanced cube were used in the final reconstruction, following the method proposed by Si et al. (Si et al., 2020).

Implementation The network was implemented in PyTorch 2.6.0 with CUDA 12.4, running under Python 3.12.8. Training was conducted using a distributed data parallel (DDP) strategy across two computational nodes connected via NVLink, with each node equipped with four NVIDIA A100 GPUs of 40 GB VRAM. This setup supported a maximum batch size of 18 per GPU. The network was trained over 95 epochs, requiring approximately 63 computational hours. The AdamW (Loshchilov & Hutter, 2017) optimizer was used with an initial learning rate of 0.0001, along with a cosine annealing learning rate scheduler. To improve training performance while maintaining accuracy, automatic mixed precision training was applied. Additionally, gradient clipping (set to 0.5) was applied to prevent gradient explosion.

4. Experiments and Results

We conducted a comprehensive study to assess the performance of CryoSAMU using a test set of 75 intermediate-resolution cryo-EM density maps and associated PDB structures across a wide range of evaluation metrics.

4.1. Visualization and quantification of map enhancement

We first visualized a CryoSAMU-enhanced map alongside its associated deposited map using UCSF ChimeraX (Pettersen et al., 2004). For a fair comparison, both sets of maps were illustrated with the same volume, which requires contour level adjustments owing to differences in their volume ranges. Specifically, we first presented the deposited map at its recommended contour level and volume, then adjusted the contour level of the corresponding CryoSAMU-enhanced map to match the recommended volume. In addition, we also visualized both maps at a higher contour level with the same volume.

As displayed in Figure 2b, CryoSAMU significantly suppressed noise in the lip nanodisc regions (highlighted by dashed boxes in Figure 2a) of the deposited map for EMDB-24496 (PDB-7RKF). Moreover, the deposited map at a smaller volumes missed certain structural regions corresponding to the protein structures, as highlighted by black boxes in Figure 2a. In contrast, the CryoSAMU-enhanced maps exhibited better alignment with the corresponding protein structures, revealing more structural details, as demonstrated by black boxes in Figure 2b. Similar visual results were observed for another protein structure (see Supplementary Figure S1). Furthermore, residue-level real-space correlation coefficient (RSCC) measurements (Afonine et al., 2018) in Figure 2c suggested significant improvements. Specifically, Chains A, B, and C exhibit RSCC increases compared to the deposited map, with correlations rising from 0.835 to 0.860, 0.848 to 0.865, and 0.801 to 0.836, respectively. In addition, 84.9%, 73.5%, and 90.6% of residues in Chains A, B, and C, respectively, showcased higher RSCC scores. Consistent RSCC improvements were also observed in other samples (see Supplementary Figure S2).

4.2. Benchmark I: improvement of real and Fourier space correlations

We then benchmarked CryoSAMU against other state-of-the-art methods, including Autosharpen (Terwilliger et al., 2018b), DeepEMhancer (Sanchez-Garcia et al., 2021), and EMReady (He et al., 2023), in terms of both real-space and reciprocal-space (i.e., Fourier-space) correla-

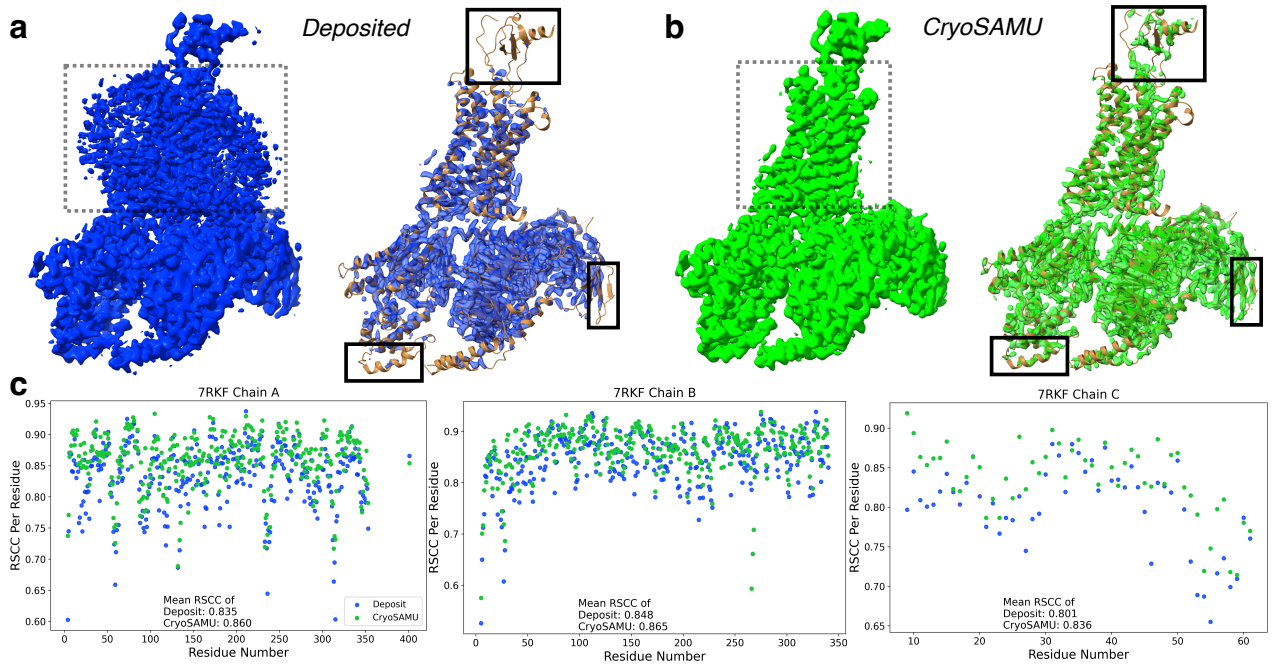

Figure 2. Visual and quantitative comparison of deposited (blue) and CryoSAMU-enhanced (green) maps, with superimposed corresponding PDB structures (brown). **a, b**: Maps are shown at two contour levels. Left: recommended contour level (volume = 85.74e3). Right: higher contour level (volume = 22.57e3). **c**: RSCC comparisons between deposited and CryoSAMU-enhanced maps. The example protein is a CX3CL1-US28-G11iN18-scFv16 in TL-state (PDB-7RKF, EMDB-24496, reported resolution of 4.00 Å) (Tsutsumi et al., 2022).

Metric	Deposit	Autosharpen	DeepEMhancer	EMReady	**CryoSAMU**(ours)	CryoSAMU (w/o struct.)
CC_box ↑	0.731	0.679	0.618	0.862	0.834	0.751
CC_peaks ↑	0.750	0.722	0.611	0.774	0.753	0.698
CC_volume ↑	0.594	0.542	0.534	0.729	0.691	0.571
FSC05 ↓	6.124	6.147	5.283	4.668	5.108	6.434

Table 1. Comparison of different methods across various metrics. See Section 4.2.

tions, across a test set of 75 primary maps. For real-space correlation, we computed three correlation metrics using `phenix.map_model_cc` (Afonine et al., 2018) for each map-model pair (where the model refers to a protein structure): CC_box, CC_volume, and CC_peaks. These metrics differ based on the choice of map regions used in the calculations. CC_box considers the entire map. CC_volume and CC_peaks focus on regions with the highest density values. However, CC_volume selects grid points only around atomic centers, while CC_peaks selects points located anywhere within the volume. For all three metrics, higher values indicate better map performance. For Fourier-space correlation, we computed Fourier shell correlation (FSC) using `phenix.mtriage` (Afonine et al., 2018), and reported the unmasked map-model FSC05 values. FSC values are typically represented as a function of the inverse map resolution, where lower value indicates better map resolution.

The average real-space CC and FSC values are listed in Table 1. According to the violin plots shown in Figure 3, CryoSAMU-enhanced maps demonstrated significant improvements over the deposited maps in terms of CC_box and CC_volume, with average values increasing from 0.731 to 0.834 and from 0.594 to 0.691, respectively. The average CC_peaks score showed a slightly increase from 0.750 to 0.753. These results indicate that CryoSAMU effectively enhances deposited maps in both the entire region and the highest-density regions. In contrast, maps processed by Autosharpen and DeepEMhancer exhibited lower scores across all three metrics. EMReady showed slightly better improvements than CryoSAMU across all three metrics. For FSC05 scores, CryoSAMU outperformed the deposited map, Autosharpen, and DeepEMhancer, achieving an average value of 5.108 Å. However, it slightly underperformed compared to EMReady, which achieved an average value of 4.668 Å.

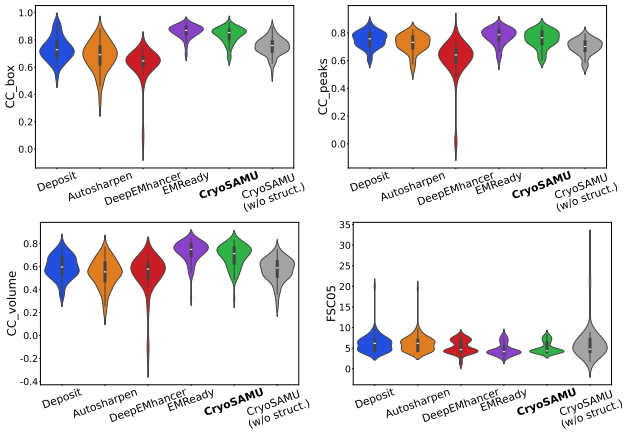

Figure 3. The violin plots for comparison of different methods across four evaluation metrics (see Section 4.2) over 75 test samples.

These results demonstrate that both CryoSAMU and EMReady consistently enhance the deposited maps in terms of correlations in both real and Fourier spaces.

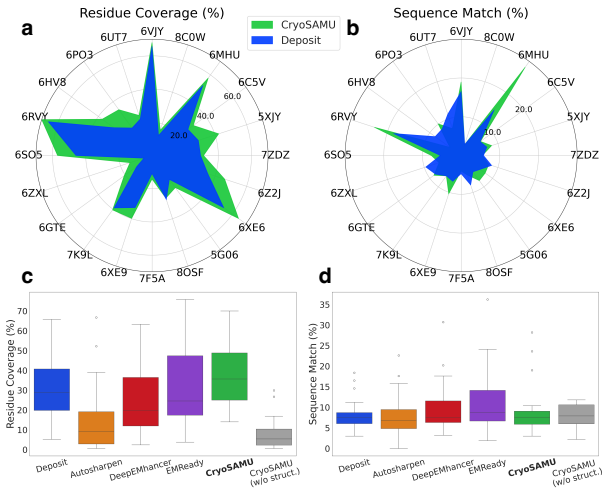

Figure 4. **a-b**: The polar plots for comparison of protein structures constructed from deposited (blue) and CryoSAMU-enhanced (green) maps, using metrics of (**a**) residue coverage and (**b**) sequence match. **c-d**: The box-whisker plots for comparison of different methods across two evaluation metrics over 20 test samples. See Section 4.3.

4.3. Benchmark II: improvement of protein structure modeling

As the goal of enhancing cryo-EM density maps is to improve the performance of protein structure modeling from density maps (i.e., map interpretability), we benchmarked protein structures constructed from CryoSAMU-enhanced maps against those processed by other methods. Specifically, we used a standard structure modeling tool, known as `phenix.map_to_model` (Terwilliger et al., 2018a), to construct protein structures from 20 maps enhanced by the different tested methods. These maps were randomly selected from the test dataset to ensure that they were not exposed during training, as listed in Supplementary Table S4. To evaluate these structures, we used `phenix.chain_comparison` (Terwilliger et al., 2018a) to compare the constructed structures against their corresponding ground-truth PDB protein structures. We reported two metrics: residue coverage and sequence match. The residue coverage indicates the fraction of residues in the query structure that match the corresponding residues in the target structure within 3.0 Å, regardless of residue type. The sequence match indicates the percentage of matched residues that share identical residue types between the query and target structures.

Method	Residue Coverage(%) ↑	Sequence Match(%) ↑
Deposit	31.71	8.42
Autosharpen	16.00	8.13
DeepEMhancer	24.31	10.0
EMReady	31.61	11.38
CryoSAMU(ours)	38.03	9.33
CryoSAMU(w/o struct.)	8.08	8.13

Table 2. Comparison of average residue coverage and sequence match across different methods.

The average metric scores from all methods are listed in Table 2. Figure 4a and b provide a detailed comparison for each individual test examples in terms of residue coverage and sequence match, respectively. The polar plots clearly showcase that after CryoSAMU enhancement, 19 out of 20 samples exhibited an improvement in residue coverage on deposited maps, with the average score increasing from 31.71% to 38.03%; and 55% of samples exhibited an improvement in sequence match, with the average score increasing from 8.42% to 9.33%. Furthermore, we benchmarked CryoSAMU against other methods, as shown in Figure 4c and d. CryoSAMU achieved the highest residue coverage score among all methods, although its correlation scores were slightly lower than those of EMReady. The sequence match score of CryoSAMU was slightly lower than EMReady and DeepEMHancer, while still better than the deposited maps. We hypothesize that the performance differences observed across metrics reflect a fundamental trade-off between local voxel-level agreement and global model-building capability. Correlation metrics primarily capture local agreement between predicted and reference densities at individual voxels, while residue coverage assesses the method's ability to generate interpretable structures suitable for atomic model reconstruction. The integration of structural embeddings in CryoSAMU appears to enhance the continuity and interpretability of generated density maps,

thereby improving atomic model reconstruction capabilities. This improvement in global structural quality may come at the cost of slightly reduced point-to-point correlations, suggesting that CryoSAMU prioritizes biologically meaningful structural features over perfect local density matching to achieve better performance in protein structure modeling.

4.4. Benchmark III: processing time

To evaluate the scalability of CryoSAMU in practice, we recorded the time required to generate each enhanced map of all 75 test samples and compared it against the processing time of other methods. Figure 5 shows the wall-clock time plotted against the volume size of input experimental maps, ranging from the order of 10^6 to 10^8 Å^3. For a fair comparison, all methods were run on the same workstation equipped with an AMD Ryzen Threadripper 2950X Processor of 32 CPUs and an NVIDIA GeForce RTX 2080 Ti of 12 GB VRAM. Each method was executed with the maximum batch size that our GPU can accommodate: approximately 12 for DeepEMhancer, 64 for EMReady, and 24 for CryoSAMU. CryoSAMU (shown in green) displayed the

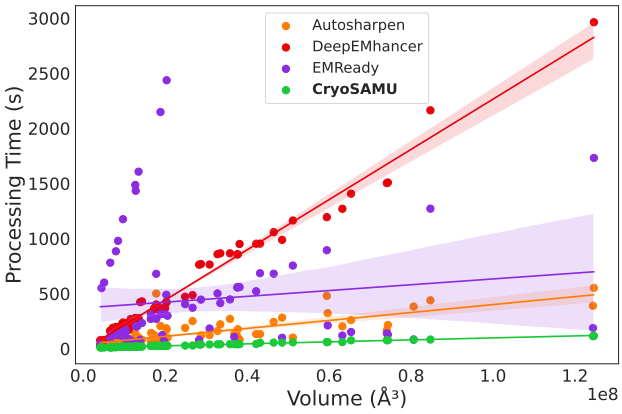

Figure 5. The scatter plot of map processing time against map volume. Each dot represents the processing time for an individual map based on its volume. The shaded area around the regression line denotes the confidence interval of the regression estimate.

minimum processing time across maps of varying volumes. Its weak linear dependency on map volume and tight confidence interval around its fit line indicate that CryoSAMU has both optimal scalability and consistent performance. In contrast, DeepEMhancer exhibited a strong linear correlation between processing time and map volume, indicating poor scalability as volume size increases. EMReady showed a wider confidence interval in its linear fit, reflecting high variability in processing time. Notably, several outliers at lower volumes showed significantly longer processing time compared to other methods. For a significantly large map with a volume size of 1.25×10^8 Å^3, CryoSAMU took only 116.48 seconds for generating an enhanced map, while

Autosharpen	DeepEMhancer	EMReady	**CryoSAMU**
138±118	544±517	441±500	32±23

Table 3. Average processing time in seconds of different methods.

Autosharpen, DeepEMHancer, and EMReady took 552.19, 2963.10, and 1731.719 seconds, respectively. Table 3 lists the average processing time for each method. CryoSAMU achieved an average processing time of 32.49 seconds, approximately 13.6 times faster than EMReady, while generating comparably enhanced maps. These results suggest that CryoSAMU scales efficiently with increasing map volume, making it a promising tool for practical applications.

4.5. Ablation study

We finally conducted an ablation study to evaluate the impact of integrating structural modality. We compared CryoSAMU with (w/) and without (w/o) structural embeddings using 75 test samples for correlation evaluation and 20 test samples for protein structure modeling assessment.

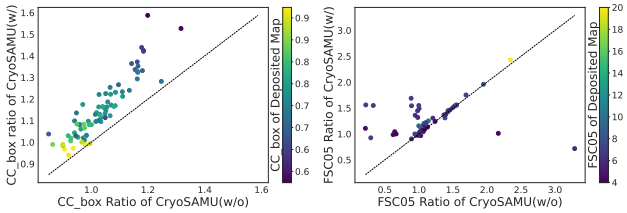

Figure 6. Pairwise comparison of enhanced/deposited ratios for CryoSAMU (w/) and (w/o). Each point represents a single map. Point colors encode the deposited map quality.

Figure 6 shows that CryoSAMU (w/) outperforms CryoSAMU (w/o) in both CC_box and FSC05, with improvements of 98.7% and 77.3%, respectively, as indicated by scatter points above the diagonal line. Moreover, the most significant gains (points far from 1.0) were observed in poorer-quality deposited maps (colored in purple), which tend to have lower deposited CC_box or higher FSC05 values. Table 1 lists the average real- and Fourier-space metrics, indicating that incorporating structural embeddings derived from ESM-IF1 led to a significant improvement of map enhancement, as also reported in Figure 3. In terms of protein structure modeling, residue coverage significantly raised from 8.08% to 38.03%, while sequence match raised from 8.13% to 9.33% with the integration of structural embeddings. This suggests that structural information helps complement map regions with poor resolutions, artifacts, or noise, thereby increasing the completeness (higher residue coverage) and improving accuracy (higher sequence match) during structure modeling. These findings underscore the importance of integrating structural modality to enable the

network to develop structural awareness beyond learning solely from 3D density maps.

5. Conclusion and Discussion

In this work, we introduce CryoSAMU, the first structure-aware multimodal network for enhancing cryo-EM density maps at intermediate resolution of protein structures. Our approach combines 3D map features with corresponding structural features through cross-attention mechanisms. Notably, structural embeddings are used solely as auxiliary supervision during training to guide the model toward learning biologically-grounded and structure-aware representations from the density maps, which aligns well with the principle of conditional training (Shen et al., 2023). In addition, we develop a self-attention weighting algorithm to produce fixed-size representations of structural embeddings derived from the pretrained ESM-IF1 model, preserving inter-chain and residue relationships while maintaining structural integrity. Our benchmark results demonstrate that CryoSAMU preforms competitively with existing cutting-edge methods, closely approaching the performance of EMReady, the current leading tool for cryo-EM density map enhancement. Notably, CryoSAMU achieves the fastest processing speed among all tested methods, positioning it as a promising solution for large-scale and practical applications. Furthermore, our ablation study reveals that incorporating an additional structural modality significantly boosts CryoSAMU's performance across all evaluation metrics, suggesting a new avenue for future cryo-EM research to explore the effective integration of multimodal data during network training.

Despite CryoSAMU demonstrating superior performance in enhancing cryo-EM maps, its current architecture—based on residual convolutions within a U-Net framework—is primarily designed to capture local information. In practice, capturing global context and long-range dependencies across map voxels could further improve performance. This could be addressed by adopting more hierarchical architectures, such as the Swin Transformer (Liu et al., 2021), which facilitates feature extraction over larger receptive fields. Moreover, incorporating supplementary loss terms, such as the Structural Similarity Index Measure (SSIM) loss, could mitigate overfitting and enhance training efficiency (He et al., 2023). These will be explored in our future work. Furthermore, we plan to explore other cutting-edge pLLMs for structural embeddings, such as ESM3—a multimodal generative language model that offers significantly improved structural representations (Hayes et al., 2025). Unlike structure-based models, ESM3 accepts amino acid sequences as input, which are more readily available. Additionally, we aim to expand our dataset by including high-resolution maps, which could increase the robustness of the model and further elevate its performance.

References

Afonine, P. V., Klaholz, B. P., Moriarty, N. W., Poon, B. K., Sobolev, O. V., Terwilliger, T. C., Adams, P. D., and Urzhumtsev, A. New tools for the analysis and validation of cryo-em maps and atomic models. *Acta Crystallographica Section D: Structural Biology*, 74(9):814–840, 2018.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907, 2002.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

Chen, H., Liu, Y., and Li, X. Structure of human dispatched-1 provides insights into hedgehog ligand biogenesis. *Life science alliance*, 3(8), 2020.

Cragnolini, T., Sahota, H., Joseph, A. P., Sweeney, A., Malhotra, S., Vasishtan, D., and Topf, M. Tempy2: a python library with improved 3d electron microscopy density-fitting and validation workflows. *Acta Crystallographica Section D: Structural Biology*, 77(1):41–47, 2021.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Dai, X., Wu, L., Yoo, S., and Liu, Q. Integrating alphafold and deep learning for atomistic interpretation of cryo-em maps. *Briefings in Bioinformatics*, 24(6):bbad405, 2023.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.

He, J., Lin, P., Chen, J., Cao, H., and Huang, S.-Y. Model building of protein complexes from intermediate-resolution cryo-em maps with deep learning-guided automatic assembly. *Nature Communications*, 13(1):4066, 2022.

He, J., Li, T., and Huang, S.-Y. Improvement of cryo-em maps by simultaneous local and non-local deep learning. *Nature Communications*, 14(1):3217, 2023.

Heinzinger, M., Weissenow, K., Sanchez, J. G., Henkel, A., Mirdita, M., Steinegger, M., and Rost, B. Bilingual language model for protein sequence and structure. *bioRxiv*, pp. 2023–07, 2023.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

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

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Kaur, S., Gomez-Blanco, J., Khalifa, A. A., Adinarayanan, S., Sanchez-Garcia, R., Wrapp, D., McLellan, J. S., Bui, K. H., and Vargas, J. Local computational methods to improve the interpretability and analysis of cryo-em maps. *Nature communications*, 12(1):1240, 2021.

Kimanius, D., Forsberg, B. O., Scheres, S. H., and Lindahl, E. Accelerated cryo-em structure determination with parallelisation using gpus in relion-2. *elife*, 5:e18722, 2016.

Lawson, C. L., Patwardhan, A., Baker, M. L., Hryc, C., Garcia, E. S., Hudson, B. P., Lagerstedt, I., Ludtke, S. J., Pintilie, G., Sala, R., et al. Emdatabank unified data resource for 3dem. *Nucleic acids research*, 44(D1):D396–D403, 2016.

LeCun, Y. and Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

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

Liu, J.-J., Niu, C.-Y., Wu, Y., Tan, D., Wang, Y., Ye, M.-D., Liu, Y., Zhao, W., Zhou, K., Liu, Q.-S., et al. Cryoem structure of yeast cytoplasmic exosome complex. *Cell research*, 26(7):822–837, 2016.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Maddhuri Venkata Subramaniya, S. R., Terashi, G., and Kihara, D. Enhancing cryo-em maps with 3d deep generative networks for assisting protein structure modeling. *Bioinformatics*, 39(8):btad494, 2023.

Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M. I., Pragani, R., Boxer, M. B., Earl, L. A., Milne, J. L., et al. Breaking cryo-em resolution barriers to facilitate drug discovery. *Cell*, 165(7):1698–1707, 2016.

Pérez-García, F., Sparks, R., and Ourselin, S. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, 2021.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13): 1605–1612, 2004.

Ramírez-Aportela, E., Vilas, J. L., Glukhova, A., Melero, R., Conesa, P., Martínez, M., Maluenda, D., Mota, J., Jiménez, A., Vargas, J., et al. Automatic local resolution-based sharpening of cryo-em maps. *Bioinformatics*, 36 (3):765–772, 2020.

Renaud, J.-P., Chari, A., Ciferri, C., Liu, W.-t., Rémigy, H.-W., Stark, H., and Wiesmann, C. Cryo-em in drug discovery: achievements, limitations and prospects. *Nature reviews Drug discovery*, 17(7):471–492, 2018.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., 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.

Rosenthal, P. B. Interpreting the cryo-em map. *IUCrJ*, 6(1): 3–4, 2019.

Sanchez-Garcia, R., Gomez-Blanco, J., Cuervo, A., Carazo, J. M., Sorzano, C. O. S., and Vargas, J. Deepemhancer: a deep learning solution for cryo-em volume post-processing. *Communications biology*, 4(1):874, 2021.

Shen, Y., Wang, X., Gao, P., and Lin, M. Auxiliary modality learning with generalized curriculum distillation. In *International Conference on Machine Learning*, pp. 31057–31076. PMLR, 2023.

Si, D., Moritz, S. A., Pfab, J., Hou, J., Cao, R., Wang, L., Wu, T., and Cheng, J. Deep learning to predict protein backbone structure from high-resolution cryo-em density maps. *Scientific reports*, 10(1):4282, 2020.

Terwilliger, T. C., Adams, P. D., Afonine, P. V., and Sobolev, O. V. A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. *Nature methods*, 15(11):905–908, 2018a.

Terwilliger, T. C., Sobolev, O. V., Afonine, P. V., and Adams, P. D. Automated map sharpening by maximization of detail and connectivity. *Biological Crystallography*, 74 (6):545–559, 2018b.

Tsutsumi, N., Maeda, S., Qu, Q., Vögele, M., Jude, K. M., Suomivuori, C.-M., Panova, O., Waghray, D., Kato, H. E., Velasco, A., et al. Atypical structural snapshots of human cytomegalovirus gpcr interactions with host g proteins. *Science advances*, 8(3):eabl5442, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xu, C., Lu, P., Gamal El-Din, T. M., Pei, X. Y., Johnson, M. C., Uyeda, A., Bick, M. J., Xu, Q., Jiang, D., Bai, H., et al. Computational design of transmembrane pores. *Nature*, 585(7823):129–134, 2020.

A. Supplementary Material

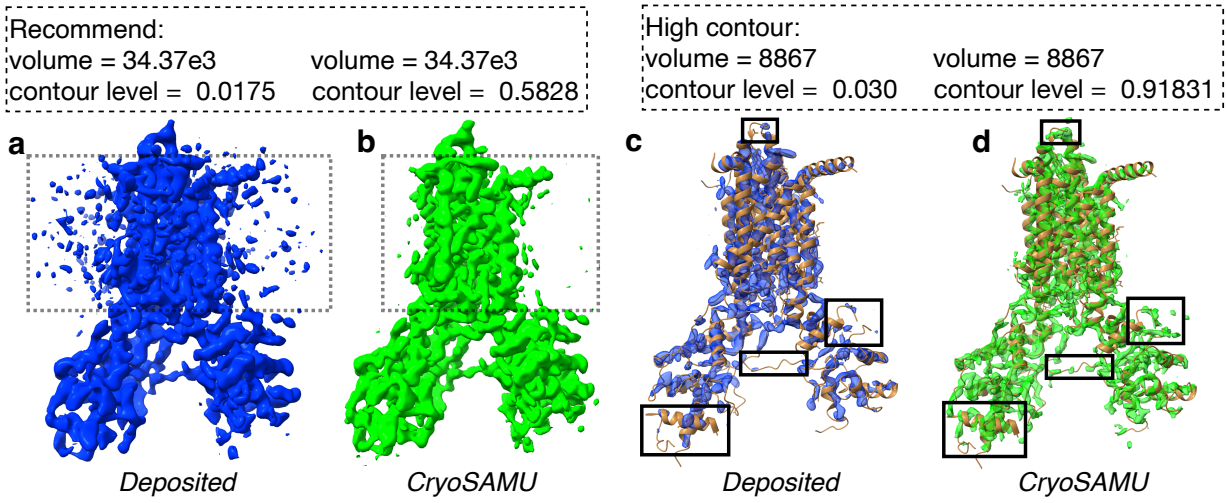

Figure S1. Visualizations of deposited (blue) and CryoSAMU-enhanced (green) maps. The corresponding PDB structures (brown) are superimposed on the maps. **a**: Human Dispatched-1 (PDB-6XE6, EMDB-22144, reported resolution of 4.53 Å) (Chen et al., 2020). **a-b**: Maps displayed at the recommended contour level. **c-d**: Maps displayed at a higher contour level. Visualizations were produced by UCSF ChimeraX (Pettersen et al., 2004). The protein structure modeling completeness and accuracy improved after CryoSAMU enhancement. For instance, residue coverage increased from 53.5% to 64.5%, as well as sequence match increased from 6.1% to 7.7%.

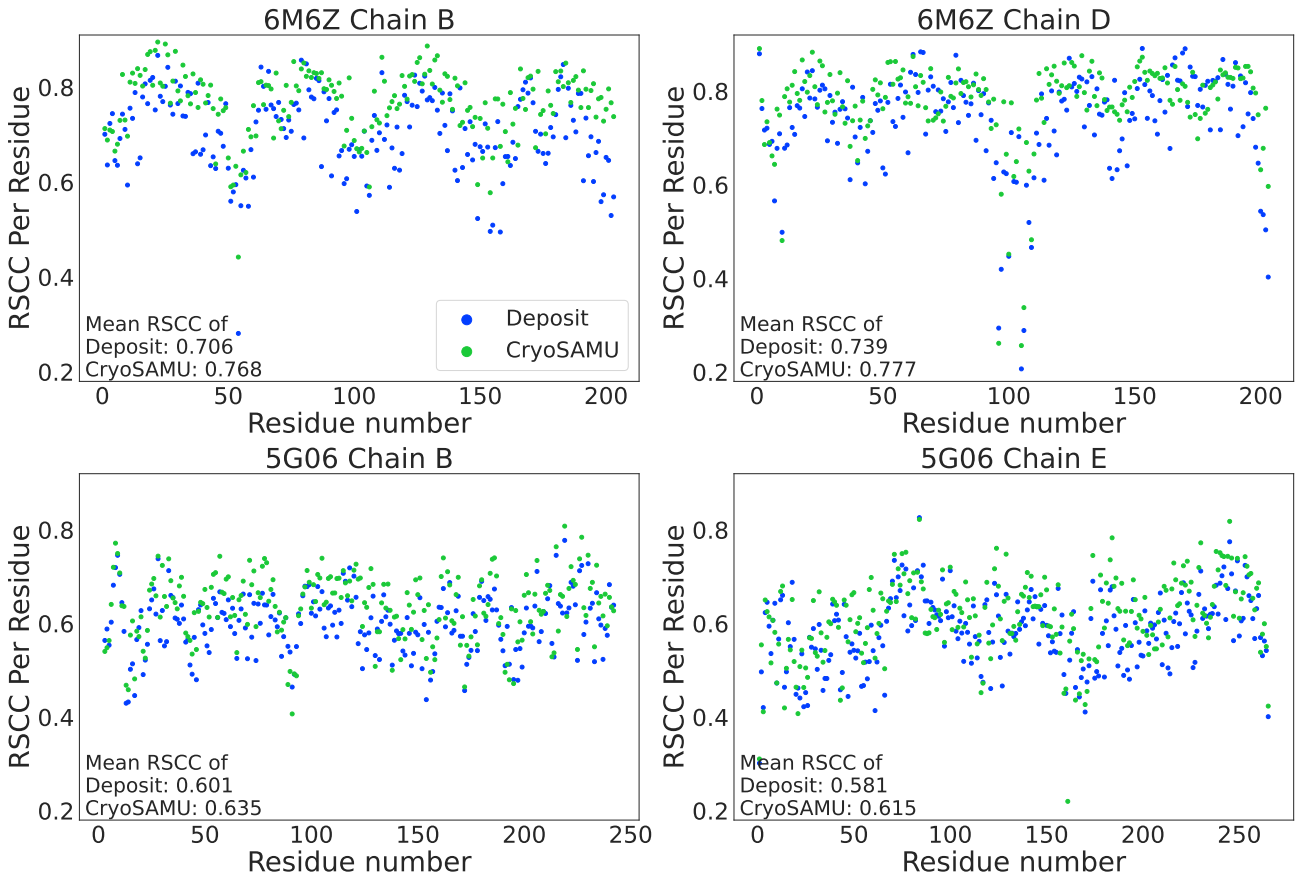

Figure S2. The real-space correlation coefficient (RSCC) comparison between deposited and CryoSAMU-enhanced maps. Top: a transmembrane nanopore TMH4C4 (PDB-6M6Z, EMDB-30126, reported resolution of 5.9 Å) (Xu et al., 2020). Bottom: a yeast cytoplasmic exosome (PDB-5G06, EMDB-3366, reported resolution of 4.2 Å) (Liu et al., 2016). In the first example of PDB-6M6Z, both Chain B and Chain D exhibited significant RSCC improvements compared to the deposited maps, increasing from 0.706 to 0.768 and from 0.739 to 0.777, respectively. In addition, 85.2% of residues in Chain B and 70.9% of residues in Chain D showed an increase in their RSCC scores. In the second example of PDB-5G06, the average RSCC scores increased from 0.601 to 0.635 for Chain B, with 82.9% of 240 residues showing improvement; and increased from 0.581 to 0.615 for Chain E, with 76.2% of 265 residues perform better.

PDB ID	EMDB ID	Res. (Å)	PDB ID	EMDB ID	Res. (Å)	PDB ID	EMDB ID	Res. (Å)
6VU8	21388	4.14	6BO4	7118	4	6RWA	10036	4
6ROW	4975	4.5	6T6V	10387	4.5	3J1P	5410	6.5
6U9F	20696	4.35	6AJ2	9631	4	5JZT	8187	7.4
5WSN	6685	4.3	7KXY	23067	4.4	6D3R	7793	4.3
6VFI	21173	4.5	6WYK	21967	4	6LXE	30006	4.2
6NY1	8994	4.2	8I4T	35183	5.2	6ZPO	11342	4
8BTZ	16244	5.39	7JGG	22326	4.9	7BG4	12170	4.2
7KSR	23024	4.1	6ZN2	11309	4.3	6VFK	21185	4.3
7YR6	34047	4.8	6WCZ	21618	4	6C14	7328	4.5
6M66	30114	4.1	6JI1	9832	4.1	2Y9J	1874	6.4
6IXH	9747	4	6V8P	21108	4.1	7CG3	30349	5.1
7Q3Y	13797	4.34	7R7T	24304	4.5	6GZV	0103	4
6EL1	3885	6.1	7PTS	13636	5.71	6NYB	0541	4.1
7Y1Q	33570	5.03	7O24	12698	4.8	5UZ7	8623	4.1
6Y5K	10700	4.2	5KEL	8240	4.3	6VEJ	21363	4.3
3J2W	5577	5	5NG5	3636	6.5	7PTX	13642	4.03
6BOA	7122	4.2	7ELE	31182	4.9	5LVC	4112	4.2
5ZQZ	6940	4.2	6R3B	4717	4.5	6ALF	8585	4.05
6PW9	20501	4	6N1Q	9317	5.2	6BVF	7294	4
6SIH	10210	4.7	6F2D	4173	4.2	6SCT	0126	4.69
6S5T	10100	4.15	7O42	12716	4.1	7OZ3	13119	4.46
7BGJ	12181	6.9	6ZYY	11581	4.4	6OMA	20122	7.2
6MZC	9298	4.5	5LY6	4118	4.5	7KAL	22774	4
6P6F	20261	4.5	6UZ2	20950	4.2	6NT8	0505	6.5
7MO7	23919	4.8	6W1C	21509	5.3	7ND2	12273	4
8EKI	28204	4.5	7L30	23147	4.4	7YMX	33946	4.44
5FLC	3213	5.9	6TEB	10479	4.1	6U9E	20695	4.2
7BE9	12154	4.2	6OFJ	20047	4.5	5FN2	3237	4.2
6WBI	21590	4.4	7NTF	12588	5.32	7FIF	31595	6.5
6UCV	20729	4.1	7K1V	22630	4.6	6XPE	22286	4.1
6JSH	9881	5.1	6ZPI	11340	4.5	6V85	21095	4.4
5ZAL	6905	4.7	6BP7	7125	4.9	7AL3	11815	4.8
5OF4	3802	4.4	6M1D	30041	4.5	6OLM	20117	4.4
3J9P	6267	4.2	5H64	6668	4.4	8BAH	15948	4.13
7A1D	11606	4.19	5LC5	4032	4.35	6H3L	0135	4.2
6K9K	9949	7.8	6RGL	4876	5.4	2YEW	1886	6
6PPB	20432	4.3	6G2D	4342	5.4	7B6E	12053	4.5
8GP3	34188	4.8	6OKR	20102	4.2	4UQQ	2685	7.6
6Z0S	11022	5.7	6CE7	7461	7.4	6D83	7453	4.3
6UC2	20725	4.5	7KS3	23015	5.8	6D7L	7823	4
7Q55	13826	5.7	5OAF	3773	4.1	6S2E	10088	4.2
6O2M	0608	6.3	6JT0	9883	4	6LBA	0868	4.1
6ZQJ	11366	4.2	6XYE	10649	4.3	6OCE	20017	4.9
7K2V	22647	6.6	7KEU	22233	5	6HZ5	0311	4.2
7QJ0	14005	5.32	7B6D	12052	4.3	6POF	20414	4.3
6SSM	10298	4.3	6VEF	21156	4.08	5XMK	6734	4.2
5Y3R	6803	6.6	6X0V	21985	4.5	6MRW	9214	4.3
6EGX	3866	4.1	7ZJ4	14740	4.43	8J5Z	35996	4.75
7JW1	22513	4.2	4BTG	2364	4.4	5FKX	3204	6.1
7CA3	30323	4.5	6R5K	4728	4.8	6S8F	10120	4
6BBM	7076	4.1	6E15	8954	5.1	5N8Y	3602	4.7
5W1R	8751	4.4	6W1S	21514	4.02	5VFR	8665	4.9
7CTF	30463	4.8	7B6H	12054	5.4	7MOB	23923	5
6KNB	0723	6.9	6KLE	0709	4.5	8S91	40234	4.3
5LJO	4061	4.9	6CFZ	7469	4.5	5GRS	9537	5.4
6C21	7332	5.2	5TQW	8436	5.6	6UWM	20924	5.9
7ZC6	14622	4.27	5KGF	8246	4.54	6OGD	20053	4.4
7KTT	23029	4.17	5Z1F	6875	4.8	6YTV	10924	4.4
6U8Y	20692	4	8FTK	29439	4.56	7PTQ	13633	4.08
8HMF	34898	4.6	8POG	17791	4.15	7R0J	14223	4.23
8I9J	35274	6.39	5OYG	3861	4.1	5OWX	3856	5.2
4PT2	5917	4.6	6B5B	7055	5.2	7DN5	30781	4.11
6TQE	10549	4.3	6V69	21060	4.2	6DVW	8919	4.3
8I6Q	35203	4.23	6LT4	0967	4.5	7MDI	23773	4.3
8C1C	16378	4.1	6D6V	7821	4.8	7YL9	33215	4.7
6I2T	4400	5.7	8FNW	29328	6.73	7VH1	31983	4.2
8CA1	16515	4.3	5Y5Y	6811	4.7	7SQT	25394	4
6ZGD	11202	4.1	5N9Y	3605	4.2	5YYS	6859	4.2
5VHW	8685	7.8	6SGY	10188	4.6	5YZ0	6862	4.7
6ZLU	11274	4.2	8ECI	28016	4	6PWP	20510	4.1
8A5Y	15199	4.9	5FVM	3329	6.1	6V9I	21121	5.2
6U1S	20613	7.6	6ZVT	11470	7	6VXH	21437	4
6V9H	21120	4.1	8HEU	34692	4.6	7R9E	24324	4
6C05	7322	5.15	6TGB	10497	5.5	6EZ8	3984	4
7Q5S	13846	4.47	3J94	6204	4.2	7DXK	30912	4.1
6ZPG	11338	4.6	7NBN	12260	7	5TCP	8398	4.3
7D7R	30610	4	6SZA	10351	6	7DL2	30708	4.4
6LQI	0946	4.5	5WC0	8794	4.4	8GAA	29849	4.24
6PYH	20524	4.3	6W4P	21536	6.6	6A69	6987	4.1
6OJ3	20086	4.5	6SHS	10204	4.4	6CA0	7439	5.75
6VOA	21259	4	6HCG	0193	4.3	6OUA	20200	4.2
6ZY4	11549	4.1	3IZI	5245	6.7	6PSN	20459	4.6
5J8V	8073	4.9						

Table S1. List of all EMDB/PDB examples in training sets.

PDB ID	EMDB ID	Resolution (Å)	PDB ID	EMDB ID	Resolution (Å)
5MDX	3491	5.3	7JTH	22473	4
6RKW	4913	6.6	6QD0	4515	4.5
7OZ1	13118	4	6E0H	8948	4.1
3J17	5376	4.1	8C0V	16372	4.1
5ZFU	6927	6.7	8JNS	36450	4.2
6B40	7046	4.3	6VFJ	21174	5.35
5ZSU	6952	4.3	7CCS	30341	6.2
3J22	5465	6.3	6QVB	4646	4.3
5G4F	3436	7	5MKF	3524	4.2
6FSZ	4301	4.6	6V3G	21036	4
6QXM	4669	4.1	6R22	4707	5.5
6TSW	10567	4	6XJX	22216	4.6
6YRK	10890	4.1	6POD	20412	4.05
5FJ9	3179	4.6	6AYE	7018	4.1
6HS7	0264	4.6	3J7V	6034	4.6
5GQH	9535	4.5	4V1W	2788	4.7
5KBT	8230	6.4	7AHE	11784	4.1
6N52	0346	4	6WCJ	21611	6.3
6CES	7464	4	6M5V	30094	4.5
5O8O	3761	6.8	6BX3	7303	4.3
6R4O	4721	4.2	5G5L	3439	4.8
7Y59	33613	4.51	6Q0X	20555	4.2
7E8G	31018	4.5	3J6X	5942	6.1
3IYJ	5155	4.2	7RD8	24415	5.64
7ZZZ	15042	4.1	3JBC	5888	6.5
6M6A	30117	5	5FJ6	3186	7.9
7JPU	22423	5	6TY3	10615	6.3
6IBC	4447	4.4	6NI2	9375	4
6JXA	9892	4.3	6OR5	9032	4
6NT5	0502	4.1	7BST	30166	4.37
6DMW	7967	4.4	5ZBO	6746	4.1

Table S2. List of all EMDB/PDB examples in validation sets.

PDB ID	EMDB ID	Resolution (Å)	PDB ID	EMDB ID	Resolution (Å)
6FO0	4286	4.1	5OFO	3776	4.6
7ZDZ	14678	4.3	6YTK	10917	4.07
6MHU	9118	4	6V0B	20986	4.1
6ZXL	11524	4.2	6GDG	4390	4.1
6PO3	20408	4.28	5V5S	8636	6.5
6XE9	22145	4.3	6ZYX	11579	4.3
3J9T	6284	6.9	6CHS	7476	4.3
6WAZ	21582	4.1	5FWP	3340	7.2
6VJY	21220	4.3	6ND1	0440	4.1
6Z2J	11041	4	5TWV	8470	6.3
7U2B	26311	4.1	6GY6	0088	4
8OSF	17151	4	5I08	8069	4
8C0W	16373	4.7	8EDG	28034	4.64
6GTE	0063	4.07	7RYZ	24749	4.15
5G06	3366	4.2	6S6T	10105	4.1
6UT7	20868	4.3	6YTX	10925	6.23
7F5A	31463	6.4	7E2I	30958	4.07
6K4M	9915	4.5	6PEK	20327	4.2
7KDV	22830	4.59	6VK0	21222	4.1
6SO5	10266	4.2	5ONV	3835	4.1
6XE6	22144	4.5	3JC5	6535	4.7
6C5V	7344	4.8	7P16	13157	4.3
6RVY	10018	4.1	6PUR	20479	4.4
5XJY	6724	4.1	5IOU	8097	7
6HV8	0287	4.4	7KVC	23046	4.7
7K9L	22754	4.9	6R7Z	4748	5.14
6JPQ	9870	4.4	8POC	17788	4
6R7Y	4747	4.2	3JCL	6526	4
6L2T	0814	4.1	5ZBG	6911	4.4
6FE8	4241	4.1	8CLS	16718	4
6E10	8951	4.16	8IGG	35432	4.09
7K08	22589	4.7	6M6Z	30126	5.9
6EZO	4162	4.1	3J5L	5771	6.6
6R9T	4772	6.2	6POS	20418	4.12
8F6R	28889	4	6PQX	20455	4.6
6ZPH	11339	6.9	6HRB	0258	4
6U1N	20612	4	7D0G	30533	5
7RKF	24496	4			

Table S3. List of all EMDB/PDB examples in test sets.

PDB ID	EMDB ID	Resolution (Å)
7ZDZ	14678	4.3
5XJY	6724	4.1
6C5V	7344	4.8
6MHU	9118	4
8C0W	16373	4.7
6VJY	21220	4.3
6UT7	20868	4.3
6PO3	20408	4.28
6HV8	0287	4.4
6RVY	10018	4.1
6SO5	10266	4.2
6ZXL	11524	4.2
6GTE	0063	4.07
7K9L	22754	4.9
6XE9	22145	4.3
7F5A	31463	6.4
8OSF	17151	4
5G06	3366	4.2
6XE6	22144	4.5
6Z2J	11041	4

Table S4. List of all EMDB/PDB examples for protein structure modeling.