CRYOLVM: SELF-SUPERVISED LEARNING FROM CRYO-EM DENSITY MAPS WITH LARGE VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cryo-electron microscopy (cryo-EM) has revolutionized structural biology by enabling near-atomic-level visualization of biomolecular assemblies. However, the exponential growth in cryo-EM data throughput and complexity, coupled with diverse downstream analytical tasks, necessitates unified computational frameworks that transcend current task-specific deep learning approaches with limited scalability and generalizability. We present CryoLVM, a foundation model that learns rich structural representations from experimental density maps with resolved structures by leveraging the Joint-Embedding Predictive Architecture (JEPA) integrated with SCUNet-based backbone, which can be rapidly adapted to various downstream tasks. We further introduce a novel histogram-based distribution alignment loss that accelerates convergence and enhances fine-tuning performance. We demonstrate CryoLVM's effectiveness across three critical cryo-EM tasks: density map sharpening, density map super-resolution, and missing wedge restoration. Our method consistently outperforms state-of-the-art baselines across multiple density map quality metrics, confirming its potential as a versatile model for a wide spectrum of cryo-EM applications.

1 INTRODUCTION

All biological functions are governed by events orchestrated at the molecular level by macromolecular assemblies (Liao et al., 2013; Gestaut et al., 2022). Therefore, elucidating their intricate structures not only deepens our understanding of underlying molecular mechanisms, but also lays the groundwork for advancements in fields like drug discovery (Congreve et al., 2020). Cryo-electron microscopy has emerged as a pivotal technique for resolving biomacromolecular structures (Kühlbrandt, 2014; Nogales & Mahamid, 2024), with the number of deposited density maps in the Electron Microscopy Data Bank (EMDB) (wwPDB Consortium, 2023) growing exponentially (Lawson et al., 2016). However, cryo-EM faces several intrinsic pitfalls that complicate accurate structure determination. Single particle analysis (SPA) cryo-EM suffer from poor signal-to-noise ratio, attenuated high-frequency information, and anisotropic resolution due to low-dose imaging requirements and sample heterogeneity (Rosenthal & Henderson, 2003; Liu et al., 2025). Cryo-electron tomography (cryo-ET) encounters additional limitations from radiation damage, missing wedge artifacts and substantial noise that obscures fine structural details (Liu et al., 2022a; Tao et al., 2018). These challenges hinder direct structural interpretation from raw cryo-EM maps, particularly at resolutions lower than 4 Å, necessitating extensive processing pipelines with substantial manual intervention and domain expertise (Terwilliger et al., 2018a; Terashi & Kihara, 2018).

Building on these observations, the cryo-EM community has increasingly embraced machine learning approaches across the entire structural determination pipeline. For map reconstruction, cryo-DRGN (Zhong et al., 2021) employs variational autoencoders to learn continuous latent representations of conformational heterogeneity, while 3DFlex (Punjani & Fleet, 2023) models non-rigid motions through coordinate-based neural networks. For postprocessing, DeepEMhancer (Sanchez-Garcia et al., 2021) adopts a 3D U-Net architecture trained on pairs of experimental maps and LocScale-sharpened targets for automated masking and local sharpening, while EMReady (He et al., 2023) implements a Swin-Conv-UNet framework combining residual convolution for local modeling with Swin transformer for non-local feature extraction. For atomic model building, early deep learning approaches include A2-Net (Xu et al., 2019), DeepTracer (Pfab et al., 2021), and ModelAngelo (Jamali et al., 2024). These methods employ deep neural networks to perform de novo recognition and tracing of amino acids from densities, thereby enabling automated model building at expert-level quality. Despite these advances, current deep learning methods in cryo-EM remain predominantly task-specific and rely on supervised training paradigms requiring labeled input data, constraining dataset scales and yielding models with limited generalizability.

Foundation models harness self-supervised pretraining on large-scale data to develop transferable representations with emergent capabilities, facilitating efficient adaptation to diverse downstream applications (Wei et al., 2022; Achiam et al., 2023; Kaplan et al., 2020). Notably, protein language models trained on extensive sequence corpora have become transformative tools for structure prediction (Lin et al., 2023), protein design (Verkuil et al., 2022), and function annotation (Hayes et al., 2025). However, foundation models remain largely unexplored in the cryo-EM density map domain. Zhou et al. (2024) introduced CryoFM, a flow-based foundation model for cryo-EM density maps that demonstrates versatility as a generative prior across multiple tasks. Nevertheless, CryoFM was trained and evaluated exclusively on curated high-quality density maps, with assessments conducted primarily on synthetic noise-corrupted maps. Consequently, its robustness and performance on genuinely noisy, low-resolution experimental maps from real-world cryo-EM workflows remain unvalidated.

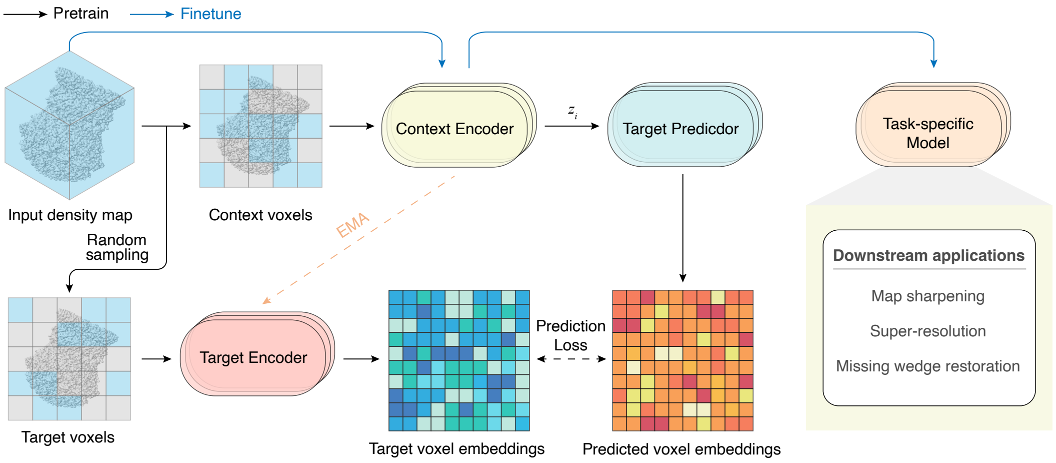


Figure 1: CryoLVM framework. During pretraining, CryoLVM leverages JEPA to learn rich structural representations from cryo-EM density maps. Input density maps are split into non-overlapping 3D patches and random sets of 3D patches are masked to produce context and target patches. The Target Predictor receives context embeddings from the Context Encoder along with positional information of masked target patches, and predicts the corresponding Target Encoder outputs. A regression loss is applied to masked tokens, encouraging alignment between predicted and target voxel embeddings. The weights of Target Encoder are updated via an exponential moving average (EMA) of the Context Encoder weights. Following pretraining, we evaluate the model through fine-tuning on three downstream cryo-EM tasks.

We present CryoLVM, a foundation model that pioneers the application of Joint-Embedding Predictive Architecture (JEPA) (Assran et al., 2023) in the cryo-EM density map domain, learning self-supervised structural representations that facilitate efficient adaptation across multiple downstream tasks. JEPA bridges the gap between generative and invariance-based methods by predicting in the abstract representation space rather than pixel space, eliminating dependencies on hand-crafted data augmentations while preserving high-level semantic information. It also achieves computational efficiency, converging significantly faster than reconstruction-based methods while maintaining transferability across diverse downstream tasks. During pretraining, we employ SCUNet-based encoder (Zhang et al., 2023) within the JEPA framework, training on experimental cryo-EM density maps with resolved structures collected from the EMDB. Through fine-tuning with task-specific models and comprehensive evaluation on three downstream tasks, CryoLVM attains state-of-the-art results across most metrics, showcasing robust performance and efficacy in processing noisy experimental maps and its promise as a foundation model for broader cryo-EM applications.

The main contributions of this work are summarized as follows:

- We propose CryoLVM, the first foundation model for cryo-EM that employs JEPA with SCUNet as backbone instead of standard vision transformer to learn semantically rich representations of cryo-EM density maps;
- We develop a novel histogram-based distribution alignment loss $\mathcal{L}_{\text{HistKL}}$ that enhances convergence speed and improves fine-tuning performance;
- We conduct comprehensive experiments across three downstream tasks, demonstrating that CryoLVM outperforms existing methods including DeepEMhancer, EMReady, EM-GAN (Subramaniya et al., 2021), and IsoNet (Liu et al., 2022a) on most evaluation metrics.

2 Related Work

In this section, we review the background and most related work on computational methods for cryo-EM density map processing, with particular emphasis on the three downstream applications addressed in this study.

Density map sharpening Cryo-EM density maps suffer from resolution-dependent amplitude falloff that attenuates high-frequency contrast, necessitating post-processing methods to restore interpretability. Traditional map sharpening approaches include global methods like phenix.auto_sharpen (Terwilliger et al., 2018b) and RELION (Scheres, 2015) that apply uniform B-factor correction, and local methods such as LocScale (Jakobi et al., 2017) that use spiral phase transformation for spatially adaptive enhancement. Recent deep learning approaches have advanced this field significantly: DeepEMhancer employs a 3D U-Net architecture to mimic LocScale's local sharpening effects (Sanchez-Garcia et al., 2021), while EMReady utilizes a 3D Swin-Conv-UNet framework with combined smooth L1 and structural similarity losses to optimize both local and non-local features of experimental cryo-EM density maps (He et al., 2023).

Density map super-resolution Model building from low-resolution cryo-EM maps remains a major impediment, as current methods like ModelAngelo exhibit marked performance degradation beyond 4 Å resolution (Jamali et al., 2024). For protein identification methods like CryoDomain (Dai et al., 2025), their accuracy also deteriorates as resolution exceeds 6-8 Å. This limitation motivates the development of super-resolution techniques, which aim to estimate high-resolution maps from their low-resolution counterparts and thereby facilitate structural determination. EM-GAN represents an early deep learning approach in this domain, employing 3D generative adversarial networks to enhance experimental cryo-EM maps in the 3-6 Å resolution range (Subramaniya et al., 2021).

Missing wedge restoration in cryo-ET The missing wedge problem in cryo-ET stems from physical limitations during tilt-series acquisition, where specimen geometry and mechanical constraints restrict image collection to an angular range of typically ±60° (Lučić et al., 2005). Incomplete angular sampling creates characteristic wedge-shaped gaps in Fourier space, leading to anisotropic resolution with severe artifacts along the beam direction, manifesting as structural elongation and distortion that compromise reconstruction fidelity (Wiedemann & Heckel, 2024). While classical approaches have implemented iterative reconstruction algorithms including SIRT (Gilbert, 1972) and ART (Gordon et al., 1970; Yan et al., 2019), alongside constrained optimization techniques like ICON (Deng et al., 2016) that impose binary assumptions such as density positivity, these methods offer limited recovery of missing information. Deep learning has revolutionized missing wedge restoration through learning complex priors directly from tomographic data. IsoNet, built upon a 3D U-Net architecture, iteratively recovers missing information by training on paired datasets—generated by rotating subtomograms to 20 orientations and imposing additional missing wedge artifacts—to map degraded inputs to less-degraded targets (Liu et al., 2022a).

3 Methodology

In this section, we describe the implementation of CryoLVM for pretraining on unlabeled cryo-EM density maps and its application to downstream tasks. First, we elaborate on the CryoLVM model architecture, encompassing the pretraining stage, the fine-tuning and inference phases for downstream tasks. Then, we present our proposed loss function tailored for the three targeted downstream tasks.

3.1 Model Architecture

CryoLVM combines Joint-Embedding Predictive Architecture (JEPA) with SCUNet backbone for efficient self-supervised learning on cryo-EM density maps. Our design addresses the unique challenges of volumetric biological data through strategic architectural modifications and domain-specific optimizations.

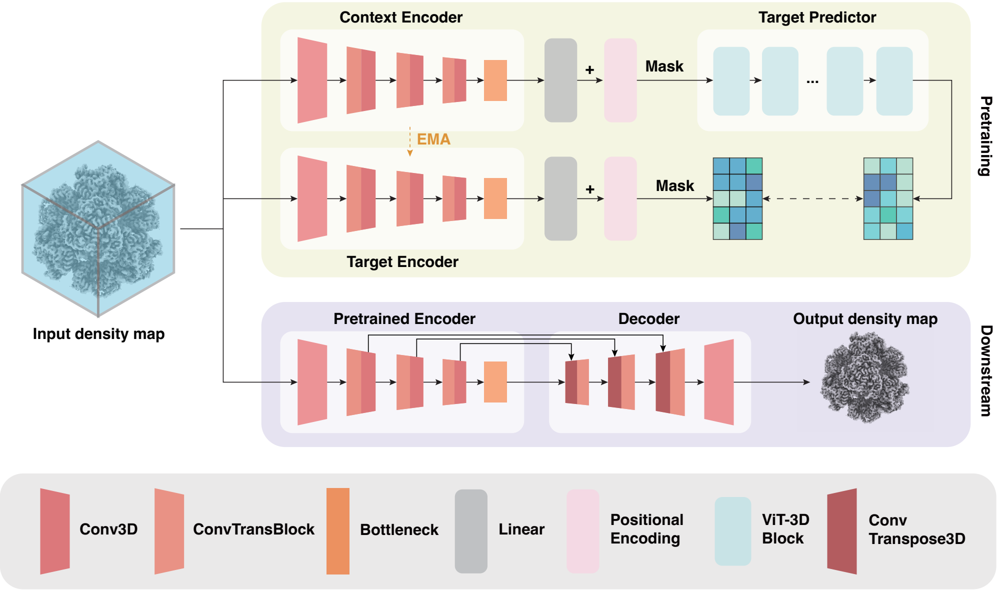


Figure 2: CryoLVM architecture. The Context Encoder and Target Encoder use hierarchical swin-conv (SC) blocks for multi-scale feature extraction, with outputs converted to patch embeddings and combined with 3D sinusoidal positional encodings. The Target Predictor employs transformer blocks to predict target representations. For downstream tasks, task-specific decoder with upsampling SC blocks are jointly fine-tuned with the pretrained encoder.

We adopt JEPA over traditional masked autoencoders for its representation-space prediction paradigm. Unlike voxel-level reconstruction that amplifies noise in low-SNR cryo-EM densities, JEPA learns semantic features by predicting masked region representation from visible context. This approach naturally filters noise while preserving structural information critical for downstream tasks. Our pretraining architecture comprises three components: Context Encoder f_{θ_c}, Target Encoder $f_{\bar{\theta}_t}$ (stop-gradient), and Target Predictor g_ϕ. The training objective minimizes prediction loss:

$$\mathcal{L}_\text{p} = \mathbb{E}_{x,M}[\Sigma_{i\in M}\mathrm{SmoothL1}_\beta(g_\phi(f_{\theta_\text{c}}(x_\text{context}), z_i) - f_{\bar{\theta}_\text{t}}(x_i))], \tag{1}$$

where M denotes masked patches, z_i denotes spatial position information, and $\mathrm{SmoothL1}_\beta$ provides robust optimization with reduced sensitivity to outliers compared to L2 loss.

Different from previous studies, we select SCUNet over conventional 3D Vision Transformer as the backbone for its hybrid architecture that adeptly combines local and global modeling. SCUNet's swin-conv (SC) blocks partition features into dual paths: Swin Transformers capture long-range dependencies while residual convolutions preserve local details. This architectural choice is particularly well-suited for cryo-EM applications, where local convolutions extract atomic-level structural features and global attention mechanisms model cross-regional spatial relationships across multiple scales. Our Context Encoder and Target Encoder consist of three downsampling SC blocks, three 3D convolution blocks and a bottleneck SC block. Within each SC block, input features are bifurcated into parallel pathways: the convolution branch employs residual $3 \times 3 \times 3$ convolutions with Filter Response Normalization, while the transformer branch utilizes 3D windowed multi-head self-attention with window size $4 \times 4 \times 4$. Feature fusion is achieved through $1 \times 1 \times 1$ convolutions, enabling simultaneous local feature preservation and global context aggregation. During pretraining, encoder outputs are converted to patch embeddings via linear transformation and incorporated with 3D sinusoidal positional encodings to preserve spatial relationships within the feature grid. Masks are then applied to these patch embeddings to generate context and target embeddings. The Target Predictor follows the JEPA framework, comprising standard transformer blocks. Final predictions are mapped back to encoder embedding dimension through a linear projection, ensuring compatibility with target outputs for loss computation. For downstream applications, task-specific decoders utilize upsampling SC blocks and 3D transposed convolution blocks, jointly fine-tuned with the pretrained encoder on labeled datasets for each target task.

3.2 Loss For Downstream Tasks

Across all three downstream tasks, we employ a unified composite loss function that combines standard reconstruction error $\mathcal{L}_{\text{MSE}}$ with distributional alignment loss $\mathcal{L}_{\text{HistKL}}$.

To enforce statistical consistency between predicted and target densities, we design a novel histogram-based distribution alignment loss, denoted as $\mathcal{L}_{\text{HistKL}}$. The idea is to align the predicted density distribution with the ground-truth density distribution via a differentiable histogram and a divergence measure. Given predicted density X and target density $X^\star$, we first construct their soft histograms using Gaussian kernel weighting:

$$h(x)_j = \frac{1}{N} \sum_{i}^{N} \exp\left(-\frac{1}{2} \left(\frac{x_i - c_j}{\sigma} \right)^2 \right), \tag{2}$$

where c_j denotes the center of the j-th bin, σ controls the smoothness, and N is the number of total voxels.

Then, we quantify the distributional divergence between the two histograms $p = h(X)$ and $q = h(X^\star)$ using Jensen–Shannon (JS) divergence, which is based on Kullback-Leibler (KL) divergence:

$$D_{\text{JS}}(p \parallel q) = \tfrac{1}{2} \sum_k p_k \log \frac{p_k}{m_k} + \tfrac{1}{2} \sum_k q_k \log \frac{q_k}{m_k}, \tag{3}$$

with $m = \frac{1}{2}(p + q)$.

The proposed histogram KL loss is thus defined as:

$$\mathcal{L}_{\text{HistKL}}(X, X^\star) = D_{\text{JS}}\big(H(X) \,\big\|\, H(X^\star)\big) \,. \tag{4}$$

The final objective combines both loss components:

$$\mathcal{L}_{\text{total}} = \alpha \, \mathcal{L}_{\text{MSE}} + (1 - \alpha) \, \mathcal{L}_{\text{HistKL}}, \tag{5}$$

where hyperparameter $\alpha \in [0, 1]$ balances reconstruction accuracy and distributional alignment.

4 EXPERIMENTS

In this section, we introduce the pretraining dataset of CryoLVM and evaluate the effectiveness of CryoLVM across three critical downstream applications. We conducted comprehensive experiments to assess model performance on density map sharpening, super-resolution, and missing wedge restoration tasks, demonstrating CryoLVM's versatility as a foundation model for diverse cryo-EM computational challenges.

4.1 Pretraining Dataset

To enable effective learning of structural semantics embedded within cryo-EM density maps, we assembled our pretraining dataset by leveraging the training subset of Cryo2StructData (Giri et al., 2024), currently the most comprehensive publicly accessible repository in this domain. This curated collection comprises 7,392 high-resolution experimental density maps representing proteins and macromolecular complexes with resolutions spanning 1-4 Å. All density maps underwent standardized preprocessing procedures: 1) voxel sizes were uniformly resampled to 1 Å spacing, 2) density values were clipped to the 0.01-0.99 percentile range to mitigate outlier effects and subsequently normalized to $[0, 1]$. To ensure rigorous evaluation and prevent data leakage, we systematically excluded density maps present in the test sets of baseline methods used for downstream task comparisons, yielding a final pretraining corpus of 7,302 density maps. For CryoLVM input generation, we applied random spatial cropping to extract volumes of size 48^3.

4.2 Density Map Sharpening

Cryo-EM density maps suffer from resolution-dependent amplitude falloff that attenuates high-frequency contrast, making direct structural interpretation challenging. Density map sharpening aims to computationally recover high-frequency information while preserving structural integrity, effectively reversing the B-factor decay that obscures fine molecular details.

Table 1: Evalutation metrics of different methods on the density map sharpening test set.

	phenix.map_model_cc			Chimera.MapQ
	$\mathrm{CC}_{\mathrm{box}}$ ↑	$\mathrm{CC}_{\mathrm{mask}}$ ↑	$\mathrm{CC}_{\mathrm{peaks}}$ ↑	Q-score ↑
Deposited	0.744	0.788	0.659	0.338
DeepEMhancer	0.695	0.679	0.659	0.323
EMReady	0.878	0.802	0.791	0.424
CryoLVM	**0.894**	**0.821**	**0.806**	**0.444**

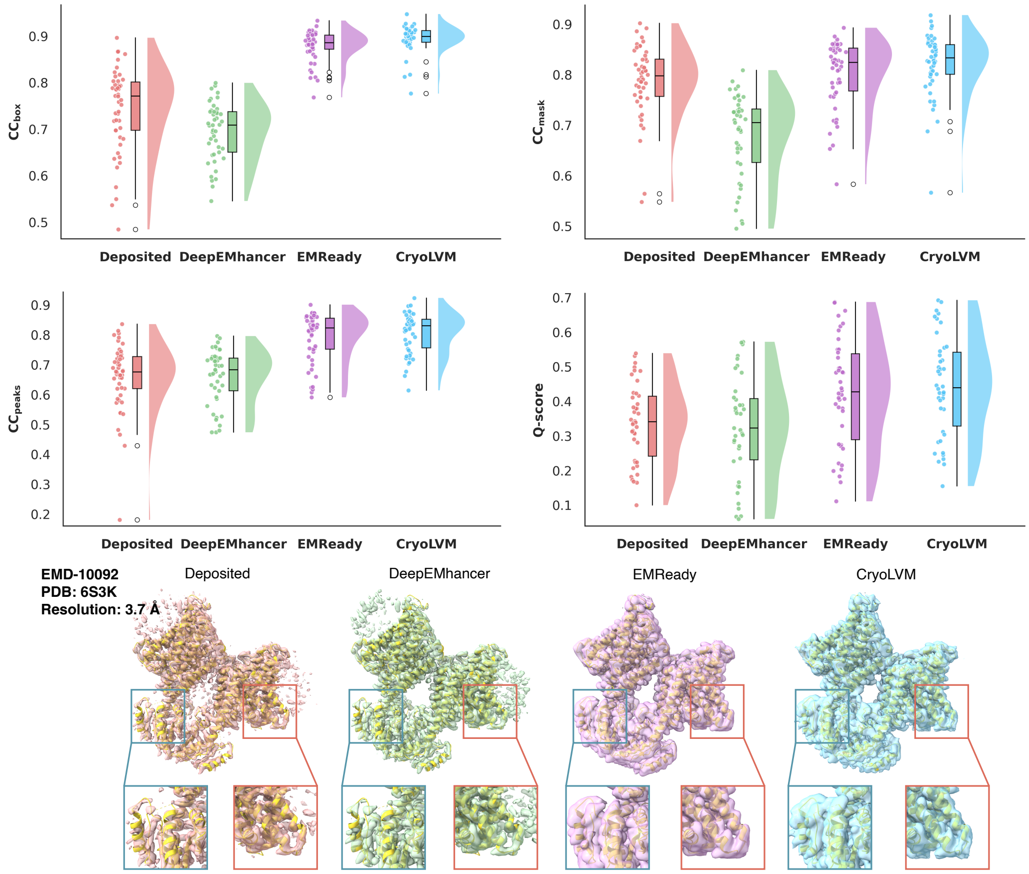


Figure 3: Comparative evaluation of density map sharpening performance across baseline and proposed methods. Cross-correlation metrics ($\mathrm{CC}_{\mathrm{box}}$, $\mathrm{CC}_{\mathrm{mask}}$, $\mathrm{CC}_{\mathrm{peaks}}$) were calculated via phenix.map_model_cc (Afonine et al., 2018), which quantify the agreement between density maps and their associated atomic models over different spatial regions. Q-score was computed using Chimera.MapQ (Pintilie et al., 2020; 2025), providing an independent assessment of map quality based on local atom-to-density correlation and atomic resolvability.

Following the established training paradigm of EMReady, we adopted a supervised learning approach for density map sharpening using paired experimental and target maps. Target density maps were simulated from their corresponding structures using Chimera.molmap (Pettersen et al., 2004), with the resolution parameter matched to that of the paired experimental maps.

We evaluated CryoLVM on density map sharpening using a comprehensive set of quality metrics computed through established cryo-EM analysis tools. As shown in Tab. 1, CryoLVM demonstrates superior performance across all metrics, achieving the highest scores in $\mathrm{CC_{box}}$ (0.894), $\mathrm{CC_{mask}}$ (0.821), $\mathrm{CC_{peaks}}$ (0.806), and Q-score (0.444). These improvements indicate enhanced overall quality between sharpened maps, suggesting better preservation of structural features and more accurate high frequency signal recovery. The violin plots in Fig. 3 also illustrate that CryoLVM consistently achieves tighter distributions with higher median values, indicating robust performance across diverse map types and resolution ranges. Visualization for EMD-10092 in Fig. 3 showcases CryoLVM's ability to enhance fine structural details such as improved definition of secondary structure elements while maintaining overall molecular topology.

4.3 Density Map Super-resolution

Cryo-electron microscopy (cryo-EM) frequently produces density maps at intermediate resolutions that preclude direct atomic model building. Even modest resolution improvements can greatly benefit downstream structural analysis. Fine-grained density features are indispensible for accurately determining backbone geometry and side-chain conformations.

Table 2: Evalutation metrics of different methods on the density map super-resolution test set

	phenix.mtriage			CryoRes
	$\mathrm{d_{model}}$(Å) ↓	FSC-0.143(Å) ↓	FSC-0.5(Å) ↓	Resolution(Å) ↓
Deposited	3.66	3.39	4.64	3.81
DeepEMhancer	3.27	2.72	4.86	3.47
EMGAN	2.49	2.70	5.46	4.18
CryoLVM	**2.33**	**2.58**	**4.58**	**3.39**

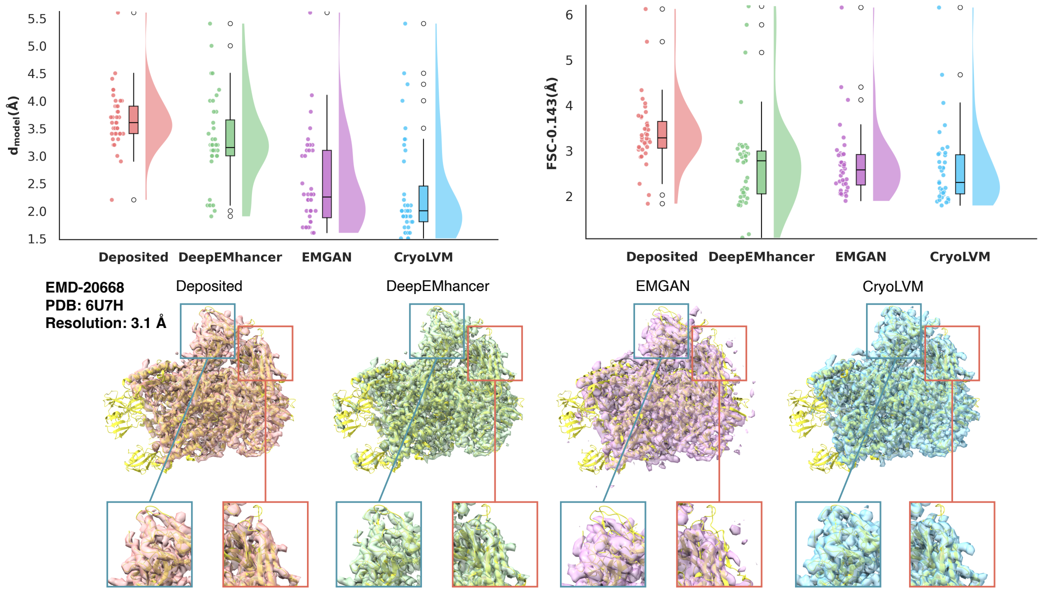


Figure 4: Comparison of map super-resolution performance between different methods. FSC-based metrics computed using phenix.mtriage (Afonine et al., 2018). Local resolution estimates obtained via CryoRes (Dai et al., 2023); global resolution represents the average of voxel-wise predictions.

We developed a supervised training protocol for CryoLVM based on paired density maps. Training data consists of experimental maps at 4-6 Å paired with target maps simulated from their corresponding PDB structures. Target maps were generated using Chimera.molmap with the resolution parameter set to 1.8 Å. This paired supervision constrains the model to learn a robust mapping from medium-resolution experimental densities to near-atomic structural detail, enabling effective super-resolution of cryo-EM maps.

We benchmarked CryoLVM for density map super-resolution against DeepEMhancer and EMGAN. Under the phenix.mtriage (Afonine et al., 2018) metrics (Fig. 4), CryoLVM achieves the best FSC-based resolutions ($\mathrm{d_{model}}$=2.33Å, FSC-0.143 = 2.58, FSC-0.5 = 4.58), significantly outperforming both DeepEMhancer and EMGAN. Similarly, when assessed with CryoRes (Dai et al., 2023), CryoLVM attains 3.39Å versus 3.47Å and 4.18Å for the baselines. Visualizations for EMD-20668 (PDB: 6U7H, 3.1 A) are shown in Fig. 4 . In the two magnified insets, CryoLVM improves backbone connectivity, and aligns much more closely with ground truth. CryoLVM not only enhances better visual interpretability but also improves quantitative resolution metrics, facilitating more accurate downstream structure determination from intermediate-resolution cryo-EM maps.

4.4 Missing Wedge Restoration

Cryo-electron tomography (cryo-ET) faces critical technical constraints: radiation sensitivity limits total electron dose, resulting in low signal-to-noise ratios, while specimen geometry restricts tilt angles to approximately ±60° rather than the ideal ±90°. As shown in Fig. 5, this limited angular sampling creates characteristic wedge-shaped gaps in Fourier space. Missing wedge produces anisotropic resolution with degraded Z-axis information and structural distortions. We simulated missing wedge artifacts by applying a Fourier transform to the map and then masking the frequency domain with a wedge-shaped filter.

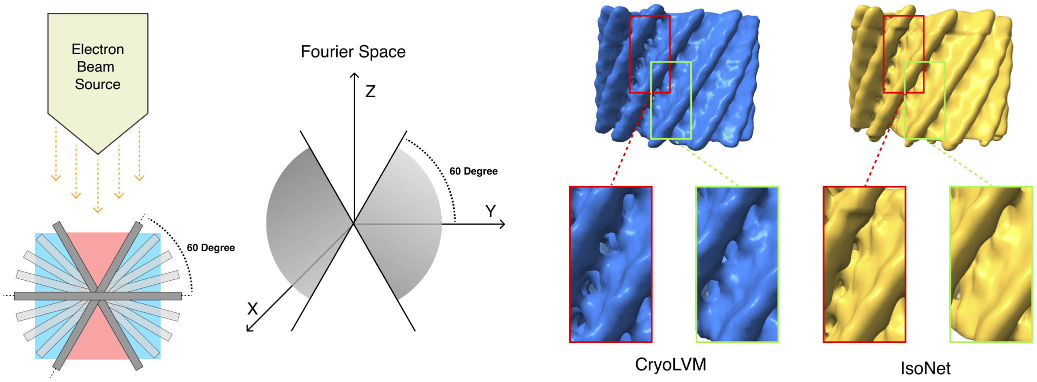


Figure 5: Visualization and schematic analysis of the missing wedge problem and its restoration. Left: schematic illustration of the missing wedge problem in cryo-electron tomography. Right: case result for missing wedge task. The original density map is EMD-5331 from EMDB and is added wedge-shaped mask for model processing.

Fourier-domain wedge masking. Given a 3D volume V, let its Fourier transform be $\widehat{V}(\mathbf{k})$ at frequency $\mathbf{k} \in \mathbb{R}^3$. To simulate the missing wedge, we introduce an indicator mask $M(\mathbf{k})$ that preserves only frequencies within the permitted tilt cone:

$$M(\mathbf{k}) = \begin{cases} 1, & |\arctan(k_z/\sqrt{k_x^2 + k_y^2})| \leq \theta_{\max}, \\ 0, & \text{otherwise.} \end{cases}$$

The corrupted spectrum is then

$$\widetilde{V}(\mathbf{k}) = M(\mathbf{k})\,\widehat{V}(\mathbf{k}),$$

and the simulated tomogram with missing wedge effect is obtained by inverse Fourier transform,

$$V_{\mathrm{mw}}(\mathbf{x}) = \mathcal{F}^{-1}\{\widetilde{V}(\mathbf{k})\}.$$

Here, $\theta_{\max}$ denotes the experimental tilt limit (commonly around 60°).

We evaluated restoration performance via phenix.mtriage (Afonine et al., 2018) and report the map-model FSC-based resolution metrics $d_{\min}$ (Å) at two cutoff thresholds, $\text{FSC} = 0.143$ and $\text{FSC} = 0.5$. Lower $d_{\min}$ values indicate higher resolution and thus better restoration quality. Our results demonstrate that CryoLVM consistently outperforms IsoNet baseline across all evaluation metrics(Tab. 3): at $\text{FSC} = 0.143$, CryoLVM reduces $d_{\min}$ from 10.448 Å to 10.094 Å ($\sim 3.39\%$); at $\text{FSC} = 0.5$, CryoLVM improves from 12.361 Å to 11.447 Å ($\sim 7.39\%$). These quantitative improvements in $d_{\min}$ reflect CryoLVM's greater ability to recover high-frequency structural information that is typically lost due to missing wedge artifacts.

Fig. 5 depicts a missing-wedge case, where CryoLVM is able to reconstruct fine-scale, high-frequency features that IsoNet fails to maintain. The magnified insets highlight a critical difference: while CryoLVM correctly reconstructs continuous pore-like channels, IsoNet produces fragmented or occluded segments that result in topologically incorrect representations of these hollow structures. The FSC curves comparing reconstructed maps to reference models show that CryoLVM maintains a higher correlation across mid-to-high spatial frequencies and reaches established cutoff thresholds at higher spatial frequencies than IsoNet (Fig. 6), indicating better resolution restoration.

Table 3: Performance of different methods in missing wedge restoration task. The score is computed between two half maps (predicted map and ground truth map). Additional results are in Appendix G.6.

	phenix.mtriage	
Method	FSC-0.143 ↓	FSC-0.5 ↓
IsoNet	10.448	12.361
CryoLVM	**10.094**	**11.447**

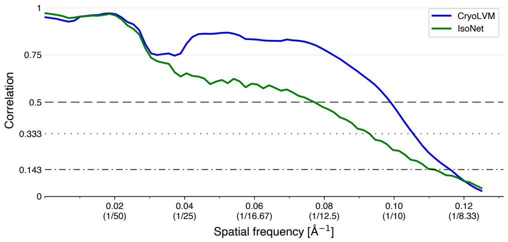


Figure 6: FSC score versus spatial frequency for predicted maps of IsoNet and CryoLVM (EMD-5331).

4.5 Ablation & Discussion

To validate our design choices and understand the contribution of individual components, we conducted ablation studies examining key architectural decisions and hyperparameter configurations. In this section, we briefly present the key findings.

SCUNet-based models achieve better performance than their ViT counterparts across all evaluated downstream tasks. To demonstrate that adopting SCUNet as the backbone over ViT yields consistently better performance across all downstream tasks, we compared models with different backbones trained using identical hyperparameters; detailed results are presented in Appendix G.1.

Composite loss of MSE and HistKL leads to accelerated convergence and enhanced downstream performance. To validate the contribution of the proposed histogram KL loss, we compared fine-tuning CryoLVM on the density map super-resolution task with the composite loss against with MSE alone; extended results are provided in Appendix G.2.

5 Conclusion

In this paper, we present CryoLVM, the first foundation model for cryo-EM density maps that employs Joint-Embedding Predictive Architecture with SCUNet-based backbone to learn rich structural representations. We introduce a novel histogram-based distribution alignment loss that accelerates convergence and enhances fine-tuning performance. CryoLVM consistently outperforms established baselines across three critical downstream tasks, showing robust performance on genuinely noisy, low-resolution experimental maps encountered in real-world cryo-EM workflows. We envision that this foundation model approach will facilitate broader adoption of AI-driven methods in cryo-EM and contribute to structural biology discoveries.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Pavel V Afonine, Bruno P Klaholz, Nigel W Moriarty, Billy K Poon, Oleg V Sobolev, Thomas C Terwilliger, Paul D Adams, and Alexandre Urzhumtsev. New tools for the analysis and validation of cryo-em maps and atomic models. *Biological Crystallography*, 74(9):814–840, 2018.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Miles Congreve, Chris De Graaf, Nigel A Swain, and Christopher G Tate. Impact of gpcr structures on drug discovery. *Cell*, 181(1):81–91, 2020.

Muzhi Dai, Zhuoer Dong, Kui Xu, and Qiangfeng Cliff Zhang. Cryores: local resolution estimation of cryo-em density maps by deep learning. *Journal of Molecular Biology*, 435(9):168059, 2023.

Muzhi Dai, Zhuoer Dong, Weining Fu, Kui Xu, and Qiangfeng Cliff Zhang. Cryodomain: Sequence-free protein domain identification from low-resolution cryo-em density maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 119–127, 2025.

Yuchen Deng, Yu Chen, Yan Zhang, Shengliu Wang, Fa Zhang, and Fei Sun. Icon: 3d reconstruction with 'missing-information'restoration in biological electron tomography. *Journal of structural biology*, 195(1):100–112, 2016.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024.

Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Sunet: Swin transformer unet for image denoising. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2333–2337. IEEE, 2022.

Daniel Gestaut, Yanyan Zhao, Junsun Park, Boxue Ma, Alexander Leitner, Miranda Collier, Grigore Pintilie, Soung-Hun Roh, Wah Chiu, and Judith Frydman. Structural visualization of the tubulin folding pathway directed by human chaperonin tric/cct. *Cell*, 185(25):4770–4787, 2022.

Peter Gilbert. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of theoretical biology*, 36(1):105–117, 1972.

Nabin Giri, Liguo Wang, and Jianlin Cheng. Cryo2structdata: A large labeled cryo-em density map dataset for ai-based modeling of protein structures. *Scientific Data*, 11(1):458, 2024.

Richard Gordon, Robert Bender, and Gabor T Herman. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of theoretical Biology*, 29(3):471–481, 1970.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Jiahua He, Tao Li, and Sheng-You Huang. Improvement of cryo-em maps by simultaneous local and non-local deep learning. *Nature communications*, 14(1):3217, 2023.

Arjen J Jakobi, Matthias Wilmanns, and Carsten Sachse. Model-based local density sharpening of cryo-em maps. *Elife*, 6:e27131, 2017.

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, and Sjors HW Scheres. Automated model building and protein identification in cryo-em maps. *Nature*, 628(8007):450–457, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.

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.

Maofu Liao, Erhu Cao, David Julius, and Yifan Cheng. Structure of the trpv1 ion channel determined by electron cryo-microscopy. *Nature*, 504(7478):107–112, 2013.

Chuchao Lin, Changjun Zou, and Hangbin Xu. Scnet: A dual-branch network for strong noisy image denoising based on swin transformer and convnext. *Computer Animation and Virtual Worlds*, 36 (3):e70030, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yun-Tao Liu, Heng Zhang, Hui Wang, Chang-Lu Tao, Guo-Qiang Bi, and Z Hong Zhou. Isotropic reconstruction for electron tomography with deep learning. *Nature communications*, 13(1):6482, 2022a.

Yun-Tao Liu, Hongcheng Fan, Jason J Hu, and Z Hong Zhou. Overcoming the preferred-orientation problem in cryo-em with self-supervised deep learning. *Nature methods*, 22(1):113–123, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022b.

Vladan Lučić, Friedrich Förster, and Wolfgang Baumeister. Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.*, 74(1):833–865, 2005.

Eva Nogales and Julia Mahamid. Bridging structural and cell biology with cryo-electron microscopy. *Nature*, 628(8006):47–56, 2024.

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

Jonas Pfab, Nhut Minh Phan, and Dong Si. Deeptracer for fast de novo cryo-em protein structure modeling and special studies on cov-related complexes. *Proceedings of the National Academy of Sciences*, 118(2):e2017525118, 2021.

Grigore Pintilie, Jingcheng Zhang, Thomas D Goddard, Wah Chiu, and David C Gossard. Measurement of atom resolvability in cryo-em maps with q-scores. *Nature Methods*, 17(3):328–334, 2020.

Grigore Pintilie, Chenghua Shao, Zhe Wang, Brian P Hudson, Justin W Flatt, Michael F Schmid, Kyle Morris, Stephen K Burley, and Wah Chiu. Q-score as a reliability measure for protein, nucleic acid, and small molecule atomic coordinate models derived from 3dem density maps. *bioRxiv*, 2025.

Ali Punjani and David J Fleet. 3dflex: determining structure and motion of flexible proteins from cryo-em. *Nature Methods*, 20(6):860–870, 2023.

Peter B Rosenthal and Richard Henderson. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology*, 333(4):721–745, 2003.

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.

Sjors HW Scheres. Semi-automated selection of cryo-em particles in relion-1.3. *Journal of structural biology*, 189(2):114–122, 2015.

Yingjun Shen, Haizhao Dai, Qihe Chen, Yan Zeng, Jiakai Zhang, Yuan Pei, and Jingyi Yu. Draco: A denoising-reconstruction autoencoder for cryo-em. *arXiv preprint arXiv:2410.11373*, 2024.

Sai Raghavendra Maddhuri Venkata Subramaniya, Genki Terashi, and Daisuke Kihara. Super resolution cryo-em maps with 3d deep generative networks. *Biophysical Journal*, 120(3):283a, 2021.

Chang-Lu Tao, Yun-Tao Liu, Rong Sun, Bin Zhang, Lei Qi, Sakar Shivakoti, Chong-Li Tian, Peijun Zhang, Pak-Ming Lau, Z Hong Zhou, et al. Differentiation and characterization of excitatory and inhibitory synapses by cryo-electron tomography and correlative microscopy. *Journal of Neuroscience*, 38(6):1493–1510, 2018.

Genki Terashi and Daisuke Kihara. De novo main-chain modeling for em maps using mainmast. *Nature communications*, 9(1):1618, 2018.

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

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

Robert Verkuil, Ori Kabeli, Yilun Du, Basile IM Wicky, Lukas F Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *BioRxiv*, pp. 2022–12, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Simon Wiedemann and Reinhard Heckel. A deep learning method for simultaneous denoising and missing wedge reconstruction in cryogenic electron tomography. *Nature Communications*, 15(1): 8255, 2024.

The wwPDB Consortium. EMDB—the Electron Microscopy Data Bank. *Nucleic Acids Research*, 52(D1):D456–D465, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1019. URL `https://doi.org/10.1093/nar/gkad1019`.

Kui Xu, Zhe Wang, Jianping Shi, Hongsheng Li, and Qiangfeng Cliff Zhang. A2-net: Molecular structure estimation from cryo-em density volumes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1230–1237, 2019.

Rui Yan, Singanallur V Venkatakrishnan, Jun Liu, Charles A Bouman, and Wen Jiang. Mbir: A cryo-et 3d reconstruction method that effectively minimizes missing wedge artifacts and restores missing information. *Journal of structural biology*, 206(2):183–192, 2019.

Yang Yan, Shiqi Fan, Fajie Yuan, and Huaizong Shen. A comprehensive foundation model for cryo-em image processing. *bioRxiv*, pp. 2024–11, 2024.

Jiakai Zhang, Shouchen Zhou, Haizhao Dai, Xinhang Liu, Peihao Wang, Zhiwen Fan, Yuan Pei, and Jingyi Yu. Cryofastar: Fast cryo-em ab initio reconstruction made easy. *arXiv preprint arXiv:2506.05864*, 2025.

Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023.

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.

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.

A Statement of The Use of Large Language Models

Large language models were used solely to polish the academic writing of this paper. They were not involved in research ideation, experimental design, data analysis, or any other substantive aspects of the work. The authors take full responsibility for the content of the manucript.

B Additional related work

B.1 Foundation models for Cryo-EM

The emergence of foundation models has begun transforming the cryo-EM data processing landscape, with recent works addressing distinct stages of the cryo-EM structural determination pipeline. DRACO (Shen et al., 2024) introduced a denoising-reconstruction autoencoder pretrained over 270,000 cryo-EM movies or micrographs, demonstrating strong generalization capabilities across micrograph-level tasks including denoising, micrograph curation, and particle picking. CryoFastAR (Zhang et al., 2025) pioneered geometric foundation modeling for cryo-EM by directly predicting particle poses for unordered, noisy 2D projection images, facilitating acceleration in ab initio reconstruction compared to traditional iterative optimization approaches. Cryo-IEF (Yan et al., 2024) presented a comprehensive foundation model pretrained on approximately 65 million particle images and showed excellent performance in tasks such as classifying particles from different structures, clustering particles by pose, and assessing image quality. These works operate on 2D cryo-EM images from the data acquisition and reconstruction stages and complement CryoLVM to address complete cryo-EM workflow from particle image processing through density map post-processing and analysis.

B.2 JEPA applications in biological context

Joint-Embedding Predictive Architecture (JEPA), first introduced by Assran et al. (2023). with I-JEPA, has gained significant traction in computational biology due to its ability to learn semantic representations through prediction in abstract embedding space rather than raw pixel-level reconstruction. Brain-JEPA (Dong et al., 2024) provides the most direct precedent for CryoLVM by successfully applying JEPA to 3D spatiotemporal biological data using functional Magnetic Resonance Imaging (fMRI). This pioneering brain dynamics foundation model incorporates two innovative techniques tailored to volumetric biological data: Brain Gradient Positioning (BGP), which establishes functional coordinate system for brain functional parcellation and enhances the positional encoding of different Regions of Interest (ROI), and Spatiotemporal Masking, which is designed for the unique characteristics of fMRI data to tackle with heterogeneous time-series patches. Brain-JEPA achieves state-of-the-art performance in dempgraphic prediction, disease diagnosis, and trait prediction, demonstrating JEPA's effectiveness for complex 3D biological volumes with temporal dependencies.

B.3 SCUNet applications in image denoising

The SCUNet framework represents a breakthrough in hybrid architecture for image denoising, effectively combining global modeling capabilities of Swin Transformers (Liu et al., 2021) with local feature extraction advantages of convolutional neural networks. Zhang et al. (2023) introduced SCUNet in their seminal work, proposing Swin-Conv (blocks) as the core building components of a UNet backbone architecture, where each SC block processes input through a 1×1 convolution followed by evenly splitting feature maps into two groups that are respectively fed into Swin Transformer blocks and residual convolutional layers. Subsequent developments further validated SCUNet's effectiveness across diverse denoising applications: SUNet (Fan et al., 2022) applies Swin Transformer layers as basic building blocks within UNet architecture for image denoising, achieving significant performance improvements on fixed-size input denoising tasks, while SCNet (Lin et al., 2025) proposes a dual-branch fusion network that combines Swin Transformer branches with ConvNext (Liu et al., 2022b) branches, employing Feature Fusion Blocks with joint spatial and channel attention for adaptive output merging, demonstrating robustness under severe noise conditions and proving effective in real-world applications like mural image denoising.

C Additional Details of CryoLVM

C.1 Implementation Details

Fig. 7 illustrates the detailed architecture of CryoLVM's Context Encoder and Target Encoder. Both encoders share an identical hierarchical design, with Target Encoder parameters updated via exponential moving average (EMA) of Context Encoder during pretraining. The encoder architecture begins with an initial $3 \times 3 \times 3$ convolution applied to the input density, followed by three consecutive down-sampling stages. Each stage consists of a Swin-Conv Block and a $2 \times 2 \times 2$ convolution. Within each Swin-Conv Block, feature map is first passed through a $1 \times 1 \times 1$ convolution, and then split evenly into two feature map groups, each of which is then fed into a swin transformer (SwinT) block and a residual $3 \times 3 \times 3$ convolution (RConv) block separately; after that, feature maps are concatenated and passed through a final $1 \times 1 \times 1$ convolution to integrate local and global modeling.

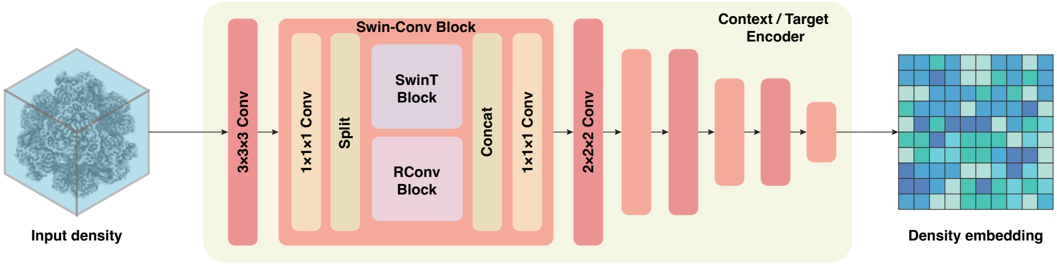


Figure 7: Detailed architecture of CryoLVM's Context Encoder and Target Encoder.

Fig. 8 illustrates the complete architecture of CryoLVM including a pretrained encoder and a task-specific decoder for downstream applications. The modular design enables effective transfer learning, where the encoder captures generalizable density map representations in the pretraining stage, while the decoder specializes these features for particular cryo-EM tasks. The task-specific decoder employs a symmetric up-sampling strategy to progressively reconstruct high-quality density maps from the encoded density embeddings. It starts with three consecutive up-sampling stages, each comprising a $2 \times 2 \times 2$ transposed convolution followed by a Swin-Conv Block. The transposed convolution performs spatial up-sampling while reducing feature dimensionality, and the Swin-Conv Block refines the upsampled features. A critical component is the skip connections between corresponding encoder and decoder layers. Specifically, skip connections link the convolution layers of encoder with the transposed convolution layers of decoder. These connections preserve multi-scale structural information that might otherwise be lost. After that, feature map is passed through a final $3 \times 3 \times 3$ convolution to produce output density.

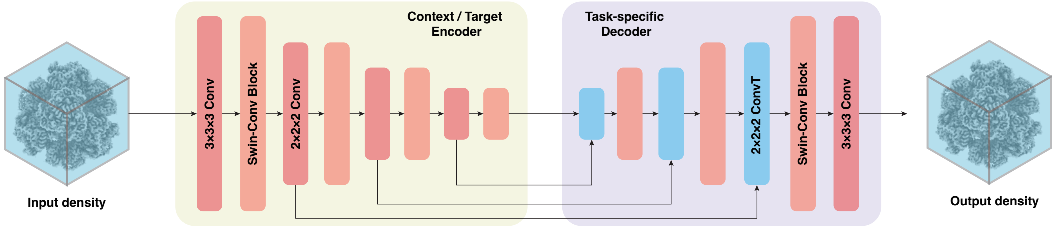


Figure 8: Detailed implementation of pretrained encoder and task-specific decoder for downstream application.

D Data Curation and Processing

D.1 PreProcessing for Training Acceleration and Memory Reduction

Pretraining To construct a high-quality pretraining dataset of cryo-EM density maps with high-quality structural information for CryoLVM, we focused on relatively high-resolution experimental density maps with corresponding resolved PDB(Berman et al., 2000) structures deposited in EMDB (wwPDB Consortium, 2023). We utilized Cryo2StructData, currently the most comprehensive publicly available depository in this domain, containing 7,600 cryo-EM density maps with associated structural annotations. We assembled our pretraining dataset from the training subset of Cryo2StructData, which comprises 7,392 density maps representing proteins and macromolecular complexes with resolutions spanning 1-4 Å. We first excluded density maps present in the test sets of baseline methods used for downstream task comparisons to guarantee meticulous evaluation and prevent data leakage, resulting in a final pretraining corpus of 7,302 density maps. Next, we preprocessed the pretraining density maps as follows: 1) voxel sizes were uniformly resampled to 1 Å spacing, 2) density values were clipped to the 0.01-0.99 percentile range to mitigate outlier effects and subsequently normalized to [0,1], 3) density maps are converted to Pytorch tensor format with bFloat16 precision, providing computational speedup and memory reduction.

Density map sharpening We first constructed our test set by selecting 50 cryo-EM density maps from the established EMReady's benchmark, ensuring direct comparison with prior methods. To build our training and validation sets, we collected additional single-particle cryo-EM maps at 3-6 Å resolutions and their corresponding atomic structures from EMDB (wwPDB Consortium, 2023) and PDB (Berman et al., 2000). We applied stringent quality control criteria to filter map-model pairs: 1) do not contain backbone atoms only, 2) do not include unknown residues, 3) do not include missing chain, 4) have orthogonal map axis, 5) resolution is given by the FSC-0.143 threshold, 6) any chain in the model does not share $>30\%$ sequence identity with any chain in the test set models. This yielded a total of 400 pairs of map and model as the training set, and 70 pairs of map and model as the validation set. Target density maps were simulated from their corresponding PDB structures using Chimera.molmap (Pettersen et al., 2004) with the resolution parameter matched to that of the paired experimental maps. We preprocessed the paired maps as follows: 1) voxel sizes were uniformly resampled to 1 Å spacing, 2) density values were clipped to the 0.01-0.99 percentile range to mitigate outlier effects and subsequently normalized to [0,1], 3) density maps are converted to Pytorch tensor format with bFloat16 precision, providing computational speedup and memory reduction. During training, we employed a sliding window sampling strategy with stride 24 to generate 48^3 input volumes and augmented training data through random flips and random rotations.

Density map super-resolution We constructed a test set consisting of 40 single-particle cryo-EM density maps, a training set with 400 maps and a validation set with 50 maps. All density maps exhibit resolutions ranging from 2.3-6 Å and satisfy the quality control criteria described above for our density map sharpening experiment. Target density maps were simulated from their corresponding PDB structures using Chimera.molmap with the resolution parameter uniformly set to 1.8 Å. We preprocessed the paired maps following the same procedure from the density map sharpening experiment.

Missing wedge restoration In the procedure for restoring a missing wedge, input density maps come from Cryo-ET, which typically exhibit resolutions worse than 8 Å, even after subtomogram averaging. Thus, we chose maps with resolutions lower than 8 Å and applied angular masks to mimic the missing wedge effect. Given constraints from the Nyquist sampling theorem and computational efficiency, a voxel size of 4 Å was used instead of 1 Å, which was used for the density map super-resolution and sharpening tasks. In total, we obtained 400 training samples, 50 validation samples, and 40 test samples. The resolution distributions of the training and test sets are shown in Fig. 9.

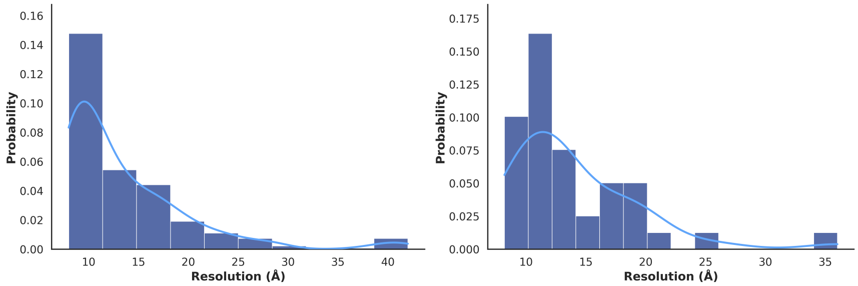


Figure 9: Resolution Distribution of Missing Wedge Restoration Task. (Left) Training Set. (Right) Test Set.

Algorithm 1: Gaussian-weighted fusion of patch predictions into a full density map

```
Input: Overlapping patch predictions $\{\hat{\mathbf{y}}_k\}_{k=1}^{K}$, their voxel coordinates $\{\mathbf{p}_k\}_{k=1}^{K}$, box size $B$,
       stride $s$, small $\varepsilon > 0$
Output: Reconstructed volume $\mathbf{M}$
Require: 3D Gaussian weight kernel $\mathbf{W}_{\text{ker}} \in \mathbb{R}^{B\times B\times B}$ with entries
  $w_{i,j,\ell} = \exp\big(-\frac{\|(i,j,\ell)-\mathbf{c}\|_2^2}{2\sigma^2}\big)$, normalized to $[1,3]$ (center larger, border smaller);
  zero-initialized accumulators $\mathbf{V},\ \mathbf{S}$ sized to the (padded) volume. (Implementation mirrors the
  sliding-window accumulation with per-voxel weights and final normalization.)
for $k = 1, 2, \ldots, K$ do
    Extract the region-to-update $\mathcal{R}_k \subset \mathbf{V}$ starting at $\mathbf{p}_k$ with size $B \times B \times B$
    // Weighted accumulation of prediction and weights
    $\mathbf{V}[\mathcal{R}_k] \leftarrow \mathbf{V}[\mathcal{R}_k] + \hat{\mathbf{y}}_k \odot \mathbf{W}_{\text{ker}}$
    $\mathbf{S}[\mathcal{R}_k] \leftarrow \mathbf{S}[\mathcal{R}_k] + \mathbf{W}_{\text{ker}}$
end
// Element-wise normalization to resolve overlaps
$\tilde{\mathbf{M}} \leftarrow \mathbf{V} \odot \max(\mathbf{S}, \varepsilon)^{-1}$
Crop $\tilde{\mathbf{M}}$ to the original (unpadded) spatial extent to obtain $\mathbf{M}$
return $\mathbf{M}$
```

D.2 PostProcessing After Prediction

We performed post-processing on the predicted density maps to ensure consistency, reduce artifacts, and improve interpretability of reconstructed volumes. Given patch-based predictions, overlapping regions were fused utilizing a Gaussian-weighted scheme (See in Algo. 1), which balances contributions between central and boundary voxels to promote continuity. This blending step reduces discontinuities at patch boundaries while maintaining local structural details. After weighted fusion, we applied normalization to maintain consistent voxel intensity distributions across the reconstructed volume. This step corrects for potential intensity variations arising from the overlapping patch-based prediction strategy, where edge regions of patches may exhibit different contrast characteristics than central regions due to reduced spatial context during inference. The final step involved cropping out padded regions introduced during the patch-wise sliding-window prediction step. Together, these steps generate a clean, continuous and interpretable density map suitable for model building, resolution evaluation, and quantitative validation. The Tab. 12 shows the post processing ablation study for missing wedge restoration task.

E Experimental Settings

For downstream training, we used the AdamW optimizer with an initial learning rate of 3×10^{-5}, weight decay of 1×10^{-3}, and a ReduceLROnPlateau (factor 0.9, patience 5). All experiments were run for 500 epochs using a batch size of 32, mixed precision (bfloat16), and DDP (Distributed Data Parallel). Input density maps were cropped into 48^3-voxel volumes, with a stride 24 during training and 48 during validation. To boost model robustness, we applied random flips along all three axes and random rotations as data augmentation. The backbone architecture was based on SCUNet, which takes a single input channel, uses a base dimension of 32 with head dimension 16, and consists of seven stages configured as [2,2,2,2,2,2,2]. Swin-Conv Blocks with a local attention window of 3^3 and zero drop-path rate were used; the output was one-channel density map. This design strikes a balance between local convolutional features and global attention, enabling efficient training on large multi-GPU clusters while preserving structural detail. The specific hyperparameters for downstream task training are listed in Tab. 4 and further architectural details are availble in Tab. 5.

Table 4: Hyperparameters for downstream task training

Parameter	**Value**	**Description**
optimizer	AdamW	Optimizer type
learning rate	3×10^{-5}	Initial LR for AdamW
weight_decay	1×10^{-3}	Weight decay coefficient
epochs	500	Number of training epochs
batch_size	35	Batch size per process
loss	$\alpha\, \mathcal{L}_{\text{MSE}} + (1-\alpha)\, \mathcal{L}_{\text{HistKL}}, \alpha \in [0,1]$	Reconstruction loss (ℓ_2)
lr_scheduler	ReduceLROnPlateau	Factor 0.9, patience 5 (min mode)
mixed precision	bfloat16	`autocast` enabled
distributed	DDP (NCCL)	Multi-GPU data parallel training
input size	48^3	Input/output volume size
train stride	24	Sliding-window stride (training)
val stride	48	Sliding-window stride (validation)
augmentation	RandomFlip/RandomRotate	3D flips and 90° rotations
num_workers	10	DataLoader workers per process
pin_memory	True	Page-locked host memory

Table 5: Model architecture of SCUNet

Component	**Setting**	**Description**
backbone	SCUNet	Swin-Conv U-Net for volumetric data
input channels	1	`in_nc = 1`
stages / blocks	[2, 2, 2, 2, 2, 2, 2]	Blocks per stage (`config`)
base dim	32	Channel width at first stage (`dim`)
head_dim	16	Attention head dimension
window_size	3	Local attention window (3×3×3)
input_size	48	Input size for model
drop_path_rate	0	Stochastic depth disabled
activation	(inherited by SCUNet)	Follows SCUNet default
output	1 channel	Reconstructed high-frequency density

F Evaluation Metrics

Fourier Shell Correlation (FSC). The Fourier Shell Correlation (FSC) is a standard metric for measuring the resolution of reconstructed density maps in cryo-EM. It is computed between two independently reconstructed half-maps $F_1(\mathbf{k})$ and $F_2(\mathbf{k})$, as the normalized cross-correlation of their Fourier coefficients within a frequency shell S:

$$\mathrm{FSC}(S) = \frac{\sum_{\mathbf{k}\in S} F_1(\mathbf{k})\, F_2^*(\mathbf{k})}{\sqrt{\sum_{\mathbf{k}\in S} |F_1(\mathbf{k})|^2 \, \sum_{\mathbf{k}\in S} |F_2(\mathbf{k})|^2}}\,, \tag{6}$$

where $F_i(\mathbf{k})$ denotes the Fourier transform of the i-th half-map, * indicates complex conjugation, and the summation is performed over all Fourier coefficients $\mathbf{k}$ within the shell S. The FSC curve typically decays with increasing frequency, and the resolution is conventionally defined at the spatial frequency where FSC drops below a certain threshold (e.g., 0.143 or 0.5).

d$_{\text{model}}$. The d$_{\text{model}}$ metric serves as an effective resolution indicator that quantifies the level of structural detail present in cryo-EM maps. It measures the effective resolution at which the atomic model agrees with the experimental density, providing a practical evaluation of density map's interpretability.

Map–Model Correlation Coefficients (Phenix). The Phenix validation suite provides several correlation coefficients (CCs) to quantify the agreement between an atomic model and a cryo-EM density map (Afonine et al., 2018):

- CC_{box}: correlation computed over the entire map volume;
- $\mathrm{CC}_{\text{mask}}$: correlation restricted to voxels inside a molecular mask;
- $\mathrm{CC}_{\text{peaks}}$: correlation evaluated at the strongest density peaks;
- $\mathrm{CC}_{\text{volume}}$: correlation calculated for a user-specified molecular volume, typically representing the region occupied by the model.

These metrics probe complementary aspects of map-model agreement: global consistency, localized fit within the molecular envelope, and correspondence at prominent density features, and agreement within the molecular volume of interest respectively.

Q-score The Q-score quantifies atomic resolvability in cryo-EM maps (Pintilie et al., 2020). For each atom, the local density profile is extracted from the map and compared against a reference Gaussian profile, with the correlation the two serving as the Q-score. A higher Q-score indicates that the atomic density is well resolved and closely resembles the expected Gaussian profile, while lower values reflect poorer local resolvability. We report the average Q-score over all atoms in the model, as implemented in Chimera's MapQ plugin, to evaluate overall map quality.

CryoRes CryoRes (Dai et al., 2023) is a deep learning-based method for estimating local resolution from a single cryo-EM density map. Unlike traditional approaches, it does not require half-maps or multiple reconstructions, making it broadly applicable in scenarios where only a final map is available. CryoRes predicts voxel-wise resolution estimates that highlight spatial variability in map quality. These local resolution annotations provide complementary information to global metrics, offering a more fine-grained assessment of map reliability.

G Additional Experiment Results

G.1 Ablation Study of SCUNet-based Backbone vs ViT-based Backbone

The ablation results validate the effectiveness of SCUNet-based backbones relative to ViT backbones across density map sharpening, super-resolution, and missing wedge restoration. As shown in Tab. 6, 7, and 8, as well as in the radar plots, SCUNet consistently outperforms ViT on both correlation- and resolution-based metrics. For comparability, metrics where lower values indicate better performance (e.g., FSC-based metrics) were negated so that all scores follow a higher-is-better convention, consistent with $\mathrm{CC}_{\mathrm{box}}$ and related measures. The unified radar plot (Fig.10) summarizes these results, providing a holistic view of backbone performance across evaluation metrics.

Table 6: Evalutation metrics of SCUNet-based and ViT-based backbone models on density map sharpening task.

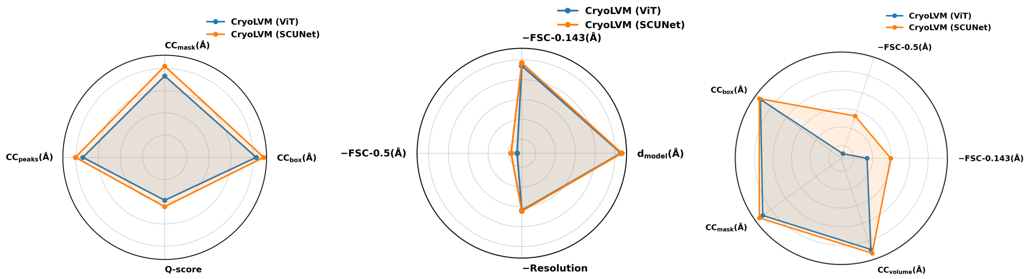
	phenix.map_model_cc			Chimera.MapQ
	$\mathrm{CC}_{\mathrm{box}} \uparrow$	$\mathrm{CC}_{\mathrm{mask}} \uparrow$	$\mathrm{CC}_{\mathrm{peaks}} \uparrow$	Q-score $\uparrow$
CryoLVM (ViT)	0.826	0.733	0.739	0.387
CryoLVM (SCUNet)	**0.894**	**0.821**	**0.806**	**0.444**

Table 7: Evalutation metrics of SCUNet-based and ViT-based backbone models on density map super-resolution task.

	phenix.mtriage			CryoRes
	d_{model}(Å) $\downarrow$	FSC-0.143(Å) $\downarrow$	FSC-0.5(Å) $\downarrow$	Resolution(Å) $\downarrow$
CryoLVM (ViT)	2.35	2.65	4.73	3.41
CryoLVM (SCUNet)	**2.33**	**2.58**	**4.58**	**3.39**

Table 8: Evalutation metrics of SCUNet-based and ViT-based backbone models on missing wedge restoration task.

	phenix.mtriage		phenix.map_model_cc		
	FSC-0.143(Å) $\downarrow$	FSC-0.5(Å) $\downarrow$	$\mathrm{CC}_{\mathrm{box}} \uparrow$	$\mathrm{CC}_{\mathrm{mask}} \uparrow$	$\mathrm{CC}_{\mathrm{volume}} \uparrow$
CryoLVM (ViT)	16.51	22.15	0.355	0.277	0.241
CryoLVM (SCUNet)	**10.09**	**11.45**	**0.391**	**0.391**	**0.348**



(a) Density map sharpening (b) Density map super-resolution (c) Missing wedge restoration

Figure 10: Radar plots comparing CryoLVM performance with SCUNet versus ViT backbones across three tasks.

G.2 Ablation Study of Composite Loss of MSE and HistKL

The experimental results summarized in Tab. 9 demonstrate that the proposed HistKL loss, when used in conjunction with the standard MSE loss, consistently outperforms the baseline MSE-only approach across all evaluation metrics in the density map super-resolution task. Beyond the improvement of density map quality, the incorporation of HistKL loss also accelerates the training convergence. Our analysis reveals that the composite loss configuration reaches optimal validation loss at epoch 107, whereas the MSE-only baseline requires 279 epochs to achieve its best validation loss—representing a 2.6× reduction in training time. These results highlight the effectiveness and efficiency of the HistKL term in guiding model optimization.

Table 9: Impact of different training losses on model performance.

	phenix.mtriage			CryoRes
	Resolution(Å) ↓	FSC-0.143(Å) ↓	FSC-0.5(Å) ↓	Resolution(Å) ↓
CryoLVM (MSE)	2.36	2.63	4.61	3.45
CryoLVM (Composite)	**2.33**	**2.58**	**4.58**	**3.39**

G.3 Ablation Study of Pretraining

To validate the effectiveness of our JEPA-based pretraining approach, we conducted two complementary ablation studies on the density map sharpening task. First, we compared our pretrained-then-finetuned model against a model trained from scratch using identical architectures and hyperparameters (Table 10). The pretrained model demonstrates consistent improvements across all metrics. These results confirm that self-supervised pretraining on high-quality cryo-EM density maps enables the model to learn transferable structural representations that enhance downstream task performance. Second, we evaluated JEPA against the widely-used Masked-Autoencoder (MAE) pretraining approach (Table 11). While both methods benefit from pretraining, JEPA consistently outperforms MAE across all metrics.

Table 10: Evalutation metrics of pretrained-then-finetuned and trained-from-scratch models on density map sharpening task.

	phenix.map_model_cc			Chimera.MapQ
	$\mathrm{CC_{box}}$ ↑	$\mathrm{CC_{mask}}$ ↑	$\mathrm{CC_{peaks}}$ ↑	Q-score ↑
CryoLVM (Scratch)	0.878	0.808	0.786	0.437
CryoLVM (Pretrain)	**0.894**	**0.821**	**0.806**	**0.444**

Table 11: Evalutation metrics of JEPA-pretrained and MAE-pretrained models on density map sharpening task.

	phenix.map_model_cc			Chimera.MapQ
	$\mathrm{CC_{box}}$ ↑	$\mathrm{CC_{mask}}$ ↑	$\mathrm{CC_{peaks}}$ ↑	Q-score ↑
CryoLVM (MAE)	0.881	0.813	0.791	0.441
CryoLVM (JEPA)	**0.894**	**0.821**	**0.806**	**0.444**

G.4 Ablation Study of PostProcessing Method

During inference, CryoLVM processes density maps using a sliding window approach with overlapping patches, necessitating an effective fusion strategy to merge predictions from overlapping regions into a coherent output volume. We conducted an ablation study comparing two postprocessing methods: mean-weighted fusion, which assigns equal weight to all overlapping predictions, and Gaussian-weighted fusion, which assigns higher weights to central regions of each patch and lower weights to boundary regions (as detailed in Alg. 1). To better differentiate between these two approaches and amplify their differences, we adopted a larger stride of 24 compared to the previously used value of 12.

As shown in Tab. 12, Gaussian-weighted fusion consistently outperforms mean-weighted fusion across all evaluation metrics for the missing wedge restoration task.

Table 12: Comparison of two postprocessing strategies

	phenix.mtriage		phenix.map_model_cc		
	FSC-0.143(Å) ↓	FSC-0.5(Å) ↓	$\mathrm{CC}_{\mathrm{box}}$ ↑	$\mathrm{CC}_{\mathrm{mask}}$ ↑	$\mathrm{CC}_{\mathrm{volume}}$ ↑
CryoLVM (Mean)	10.66	13.07	0.390	0.388	0.345
CryoLVM (Gaussian)	**9.92**	**11.47**	**0.391**	**0.391**	**0.348**

G.5 Additional Results of Super Resolution Tasks

This subsection presents extra results regarding the density map super-resolution task. As shown in Fig. 11, CryoLVM achieves state-of-the-art performance overall, outperforming all baselines in both the FSC-0.5 and the CryoRes metrics. Fig. 12 visualizes improvement in local resolution for the EMD-0023 example: the deposited map spans a local resolution range of 3.39 Å to 4.12 Å, whereas CryoLVM's predicted map compresses that range to a 2.31 Å to 3.45 Å. This illustrates that CryoLVM not only enhances global resolution measures but also recovers finer structural detail in regions where the original map is weaker.

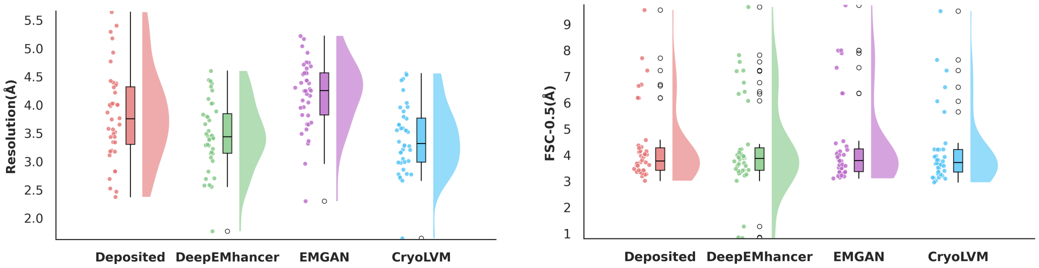


Figure 11: Comparison of CryoRes estimates and FSC-0.5 for density map super-resolution task.

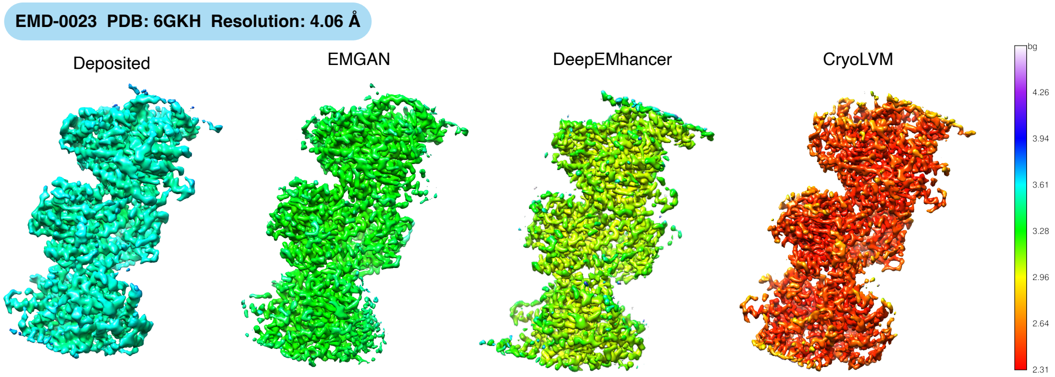


Figure 12: Comparison of local resolution maps between different methods on case EMD-0023. Local resolution maps are calculated via CryoRes and visualized through Chimera.

G.6 Additional Results of Missing Wedge Restoration

To provide thorough evaluation of missing wedge restoration performance, we present additional quantitative metrics and visualizations beyond the main results reported in Section 4.4. As shown in Tab. 13, CryoLVM achieves higher correlation coefficients in three complementary assesment criteria: 1) $\mathrm{CC}_{\mathrm{box}}$, which measures global structural coherence by evaluating correlation across the entire reconstruction volume; 2) $\mathrm{CC}_{\mathrm{mask}}$, which quantifies reconstruction accuracy within a maksed region; and 3) $\mathrm{CC}_{\mathrm{volume}}$, which assesses volumetric fidelity through voxel-wise correlation analysis of the complete density distribution. These improvements indicate that CryoLVM not only recovers missing angular information more accurately but also maintains superior structural consistency. Fig. 14 presents the FSC-based resolution metrics (FSC-0.143 and FSC-0.5), where CryoLVM exhibits significantly better resolution distributions, indicating more effective recovery of high-frequency structural information lost due to missing wedge effects. Fig. 15 provides an additional visualization case for EMD-5106.

Table 13: Addition CC results for missing wedge restoration task. Metrics computed using phenix.map_model_cc.

	phenix.map_model_cc		
	$\text{CC}_{\text{box}} \uparrow$	$\text{CC}_{\text{mask}} \uparrow$	$\text{CC}_{\text{volume}} \uparrow$
IsoNet	0.364	0.371	0.328
CryoLVM	**0.391**	**0.391**	**0.348**

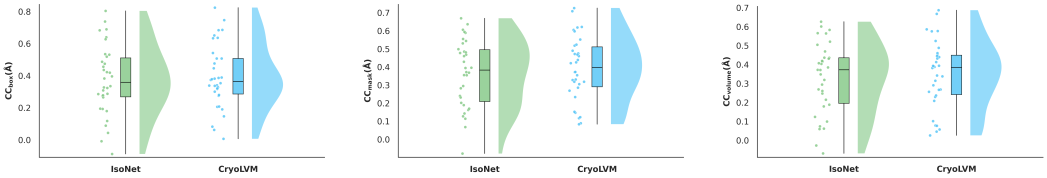


Figure 13: CC-based performance evaluation for missing wedge restoration task.

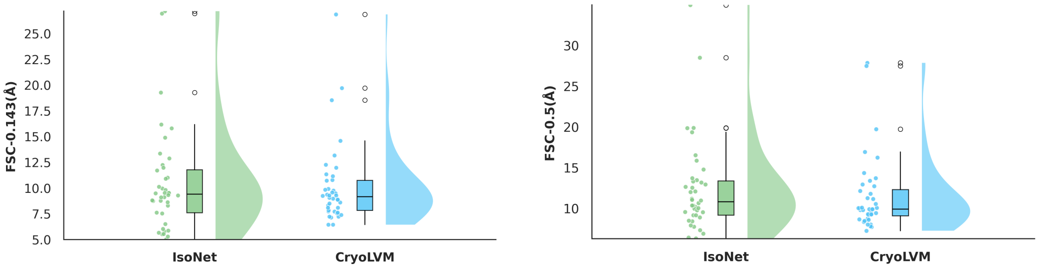


Figure 14: FSC-based performance evaluation for missing wedge restoration task.

G.7 Validation of Impact on Automated Model Building

To demonstrate the practical impact of CryoLVM processing on downstream structural biology workflows, we conducted a case study using ModelAngelo, a state-of-the-art automated model building tool, on EMD-6656 (PDB: 5H30, resolution: 3.5 Å). As shown in Fig. 16, we compared atomic structures built from the deposited experimental map versus the CryoLVM-sharpened map. The CryoLVM-processed map yielded quantifiable improvements in model quality: RMSD to the reference structure decreased from 0.68 Å to 0.58 Å, and sequence match increased from 92% to 94.3%. These gains show that enhanced map quality directly translates to more accurate and complete automated model building, reducing manual intervention requirements in structural determination workflows.

G.8 Protein Secondary Structure Classification

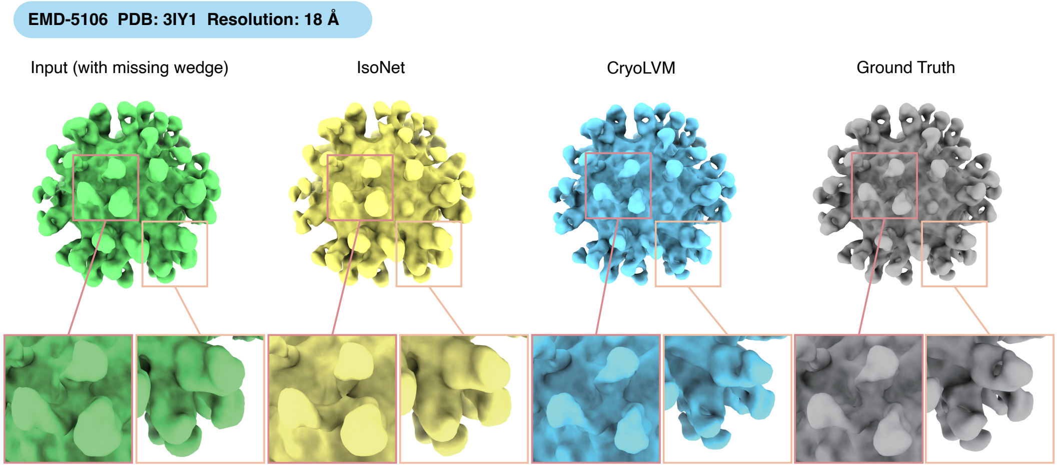

EMD-5106 PDB: 3IY1 Resolution: 18 Å
Input (with missing wedge)
IsoNet
CryoLVM
Ground Truth

Figure 15: Additional visualization case for missing wedge task. The original density map is EMD-5106 from EMDB and is added wedge-shaped mask for model processing.

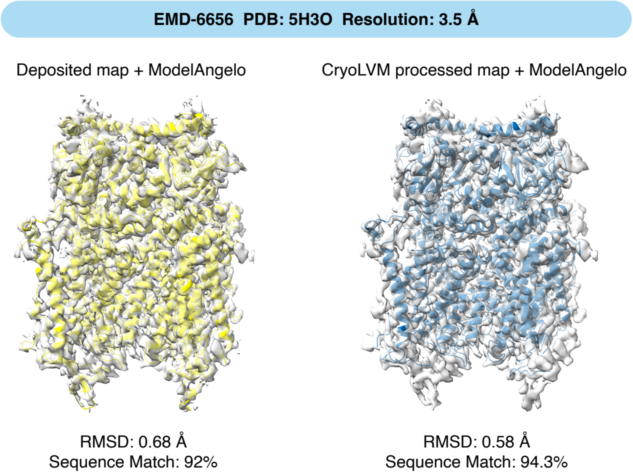

EMD-6656 PDB: 5H3O Resolution: 3.5 Å
Deposited map + ModelAngelo
CryoLVM processed map + ModelAngelo
RMSD: 0.68 Å
Sequence Match: 92%
RMSD: 0.58 Å
Sequence Match: 94.3%

Figure 16: Comparative model building quality assessment using ModelAngelo. Atomic structures were built automatically from (Left) the deposited experimental map EMD-6656 (resolution: 3.5 Å) and (Right) the same map after CryoLVM sharpening.

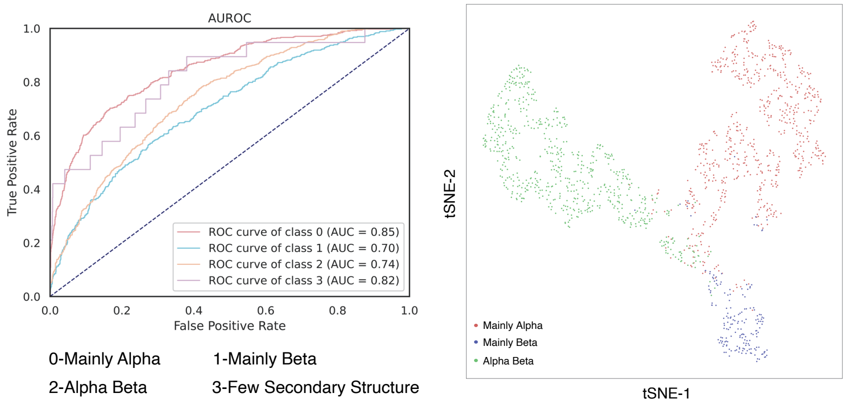


Figure 17: Protein secondary structure classification and representation visualization results.