CryoMAE: Few-Shot Cryo-EM Particle Picking with Masked Autoencoders

Chentianye Xu*, Xueying Zhan*†, Min Xu†
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
chentiax@cs.cmu.edu, xueyingz@andrew.cmu.edu, mxu1@cs.cmu.edu

Abstract

Cryo-electron microscopy (cryo-EM) emerges as a pivotal technology for determining the architecture of cells, viruses, and protein assemblies at near-atomic resolution. Traditional particle picking, a key step in cryo-EM, struggles with manual effort and automated methods' sensitivity to low signal-to-noise ratio (SNR) and varied particle orientations. Furthermore, existing neural network (NN)-based approaches often require extensive labeled datasets, limiting their practicality. To overcome these obstacles, we introduce ***cryoMAE****, a novel approach based on few-shot learning that harnesses the capabilities of Masked Autoencoders (MAE) to enable efficient selection of single particles in cryo-EM images. Contrary to conventional NN-based techniques, cryoMAE requires only a minimal set of positive particle images for training yet demonstrates high performance in particle detection. Furthermore, the implementation of a self-cross similarity loss ensures distinct features for particle and background regions, thereby enhancing the discrimination capability of cryoMAE. Experiments on large-scale cryo-EM datasets show that cryoMAE outperforms existing state-of-the-art (SOTA) methods, improving 3D reconstruction resolution by up to* 22.4%*. Our code is available at:* https://github.com/xulabs/aitom.

1. Introduction

Cryo-EM is vital for obtaining high-resolution images of biological entities, such as cells and proteins, at cryogenic temperatures, significantly minimizing radiation damage [9, 12]. It revolutionizes structural biology, especially through single-particle analysis (SPA), allowing for detailed examinations of molecular structures in their near-native state [7, 8, 20]. The process starts with sample preparation, where specimens are vitrified in a thin ice layer to maintain their native state. Researchers then use a transmission electron microscope to gather multiple 2D projection images from different angles. Image processing includes denoising and identifying particles for 3D reconstruction. Fig. 1 shows a simplified workflow of SPA using cryo-EM [35].

Particle picking is a pivotal step in cryo-EM for isolating individual protein particles from micrographs for further analysis. The quality of particle picking significantly influences the accuracy and resolution of the reconstructed particle structure in the following steps. Challenges in particle picking include the low SNR and varied particle orientations in cryo-EM micrographs, necessitating a large sample size for accurate 3D reconstructions [1]. Moreover, manual picking is inefficient, time-consuming, labor-intensive, error-prone, and introduces dataset inconsistencies [5]. Mis-identifications, or false positives, further compromise reconstruction quality. These issues highlight the need for improved particle selection techniques to enhance both the efficiency of particle identification and the overall quality of cryo-EM reconstructions, emphasizing the reduction of false positives and the increase of true positives [19].

Various semi-automated and automated cryo-EM particle picking methods have been developed in response to this need. Traditional methods are categorized into template-free [21] and template-based methods [22, 24, 25, 27] Template-free methods like Difference of Gaussians (DoG) [31] are noise-sensitive and less effective for irregular particles. Template-based approaches struggle with particle variability and are ill-suited for novel structures, limiting their efficacy in complex cryo-EM analysis. With the advent of deep learning, NN-based particle picking methods [1,32,33,36] have been proposed, marking a significant evolution in the field. Among them, crYOLO [32] and Topaz [1] are notable for their widespread application. While crYOLO is recognized for its efficiency in particle detection, it occasionally misses real particles. Topaz, though capable of identifying particles with limited labeled data, is susceptible to false positives and duplicates. Recently, CryoTransformer [6] has been proposed to utilize Transformer [2, 29] for particle picking, which also faces the issue of outputting

* These authors contributed equally to this work.
† Corresponding authors.

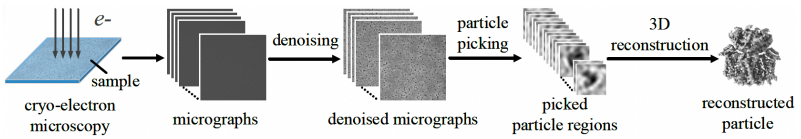


Figure 1. In cryo-EM with SPA, electron beams capture numerous 2D images of proteins within a cryogenically preserved sample. These images are subsequently denoised and subjected to particle picking, facilitating the reconstruction of 3D structure of protein.

numerous false positives. Despite claims of minimal data requirements, these methods still require large-scale labeled datasets for improved performance. They also exhibit limited generalization to unseen data, restricting their applicability in diverse cryo-EM research settings.

In this study, we present **cryoMAE**, a cutting-edge cryo-EM particle picking approach, drawing inspiration from MAE [14]. Leveraging the few-shot learning paradigm, cryoMAE is meticulously designed to first learn representative particle features from a limited set of cryo-EM particle regions efficiently, cryoMAE then detects and extracts particles from query micrographs by comparing the latent features generated for exemplars against those from regions within the query micrographs. The operation of cryoMAE unfolds in two stages. Initially, it trains on a curated set of particle regions and a broader selection of unlabeled regions from a reference micrograph, utilizing a self-supervised approach. We introduce a unique self-cross similarity loss, ensuring the cryoMAE encoder generates distinct latent features for particle and non-particle areas. Subsequently, the trained encoder analyzes query micrographs, extracting and comparing latent features to exemplar features to ascertain particle locations through similarity scoring.

The performance of cryoMAE was evaluated using the CryoPPP particle picking dataset [5], showcasing significant enhancements. Particles selected by our model from this dataset exhibit up to 22.4% (average 11.1%) improvement in resolution compared to those picked using current SOTA models. Remarkably, these results were achieved using just a few labeled exemplars (*e.g.*, 15) per protein type, highlighting cryoMAE's efficient use of limited data.

Our contributions are summarized as follows:

1. We introduce cryoMAE, an innovative two-stage few-shot learning method specifically designed for SPA cryo-EM particle picking task, which markedly diminishes the reliance on extensive, labeled datasets.

2. We propose a novel formulation of self-cross similarity loss, aiming to promote the model capacity to differentiate between particle objects and background regions.

3. Our experimental findings indicate that cryoMAE achieves up to 22.4% improvement in the resolution of 3D particle reconstructions compared to SOTA NN-based particle picking methods.

2. Related Work

2.1. Particle Picking

A variety of approaches exist for cryo-EM particle picking, including automated and semi-automated techniques. Template matching, use predefined reference images and cross-correlation for particle identification [30, 34], performing best with known particle structures but limited by template quality and diversity. In contrast, template-free methods bypass the need for templates, employing computer vision techniques to distinguish particles. For instance, the DoG method emphasizes particles by contrasting two differently blurred image versions, improving visibility but risking noise amplification in low-SNR scenarios.

NN-based particle picking methods [1,32,33,36] provide significant advances in cryo-EM, offering more accurate, efficient, and accessible solutions. These methods can learn from a wide range of particle shapes, sizes, and orientations directly from the training data, making them more adaptable to different datasets without the need for specific templates. CrYOLO [32] and Topaz [1] are distinguished for their advanced particle picking capabilities in cryo-EM. CrYOLO leverages the You Only Look Once framework [23] for particle detection, and Topaz employs convolutional neural networks (CNNs) with positive-unlabeled (PU) learning. Recently, CryoTransformer [6] has been introduced to employ Transformer [2, 29] for particle picking. Despite their strengths, crYOLO may overlook true particles, while Topaz and CryoTransformer are prone to recognizing numerous false positives and duplicates [6, 10]. They require extensive labeled datasets, demanding significant time and resources. Our cryoMAE, utilizing few-shot learning, offers high efficiency using a minimal number of exemplars. It effectively reduces false negatives and positives, and minimizes reliance on large labeled datasets, representing a significant leap in cryo-EM particle picking technology.

2.2. Masked Autoencoders

MAEs were initially introduced by He et al. [14], drawing inspiration from the BERT model [4], a transformative approach in natural language processing. MAEs bring the innovative concept of masking into the realm of computer vision, a technique where random sections of an image are obscured (masked) before being processed by an encoder. Subsequently, a decoder attempts to reconstruct these masked sections. [14] demonstrated that masking a

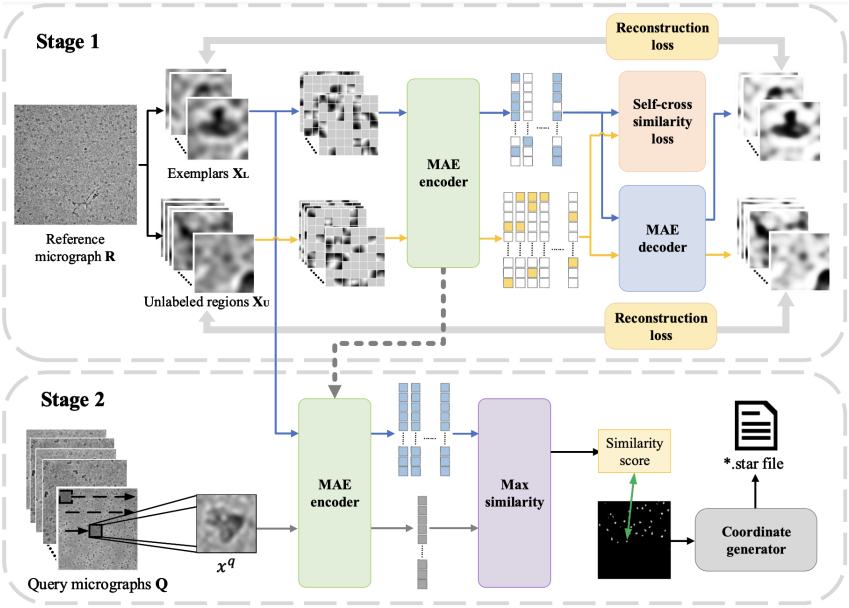


Figure 2. Overview of the two-stage cryoMAE framework: Stage 1 illustrates the training phase with a mix of labeled particle and unlabeled regions, employing reconstruction loss and self-cross similarity loss. Stage 2 depicts the particle picking process, where the trained MAE encoder assesses query micrographs, leveraging latent feature comparisons to identify particle positions accurately.

substantial portion of the image (up to 75%) compels the model to learn deeper and more comprehensive representations of the data. In our study, we harness the exceptional feature extraction capabilities of MAEs to discern unique features of particles, thereby enhancing the efficiency and accuracy of particle picking in cryo-EM.

2.3. Contrastive Learning

Contrastive Learning concentrates on increasing the similarity between representations of positive pairs while simultaneously differentiating those of negative pairs. The first concept of contrastive loss was introduced for dimensionality reduction and embedding learning to preserve semantic similarity [13]. Further advances was made with the development of SimCLR [3], utilizing data augmentation techniques to enhance the robustness of visual representations. He et al. [15] introduced Momentum Contrast, building dynamic dictionaries in contrastive learning, which ensures the consistency of the representations for negative samples across learning process. Prior works have explored the integration of contrastive learning in cryo-EM/ET applications. Tang et al. [28] introduced SimCryoCluster, leveraging contrastive learning to effectively classify high-noise, unlabeled cryo-EM single-particle images, enhancing its utility in cryo-EM protein structure determination. Huang et al. [16] developed a 3D particle detection framework that incorporates a debiased voxel-level contrastive learning module, which utilizes information from both annotated and unannotated data. In our research, we leverage the principles of contrastive learning to develop a unique contrastive loss mechanism called self-cross similarity loss. This innovation enables our model to effectively discriminate between regions with and without particles.

3. Methodology

In this section, we detail cryoMAE, starting with defining the few-shot cryo-EM particle picking problem, followed by our two-stage framework.

3.1. Overview

3.1.1 Problem setup

Given a reference micrograph $\mathbf{R}$, containing the target particles for analysis, we first randomly select a reference micrograph $\mathbf{R}$ and manually label m (m is a small number, *e.g.* 15) particle regions $\mathbf{x}_i^l$ as exemplars ($\mathbf{X_L} = \{\mathbf{x}_i^l\}_{i=1}^m$), and randomly crop additional n regions $\mathbf{x}_j^u$ from the same cryo-EM micrograph as unlabeled regions ($\mathbf{X_U} = \{\mathbf{x}_j^u\}_{j=1}^n$). The remaining micrographs containing the same particle are query micrograph set $\mathbf{Q}$. Our goal is to leverage the limited set of exemplars $\mathbf{X_L}$ and unlabeled regions extracted from $\mathbf{R}$ to detect the particles within $\mathbf{R}$ and $\mathbf{Q}$.

As depicted in Fig. 2, our framework unfolds in two stages. In stage 1, cryoMAE is trained using a mixture of labeled exemplars $\mathbf{X_L}$ and unlabeled regions $\mathbf{X_U}$ from $\mathbf{R}$. This training process is guided by both mean squared error reconstruction loss and a novel self-cross similarity loss, which helps the model distinguish between regions with and without particles. In stage 2, trained MAE encoder scans query micrographs to identify particles, comparing latent features of regions against those of exemplars to determine similarity scores. Regions with higher similarity scores are

identified as more likely to contain particles.

3.2. Stage 1: Training on One Reference Micrograph

3.2.1 Model training

For each protein type represented by multiple micrographs, we select a reference micrograph $\mathbf{R}$ with manually annotated regions $\mathbf{X_L}$ as exemplars and crop random unlabeled regions $\mathbf{X_U}$ from the remaining parts of $\mathbf{R}$. As discussed in [1], particle regions are sparse within micrographs, making most unlabeled regions likely non-particle areas. These images are resized to 224×224 and further processed into 16×16 patches during training, which are then subjected to random masking at a rate of 75%. This process transforms exemplar and unlabeled regions into $\hat{\mathbf{x}}_i^l$ for labeled exemplars and $\hat{\mathbf{x}}_j^u$ for unlabeled regions, respectively. The cryoMAE encoder then generates latent features for these regions, denoted as $\mathbf{E}(\hat{\mathbf{x}}_i^l)$ and $\mathbf{E}(\hat{\mathbf{x}}_j^u)$ respectively. Subsequently, the MAE decoder utilizes the generated latent features to reconstruct the original input images. This reconstruction is achieved through a self-supervised process, with the original images serving as the supervisory signal. This masking encourages the model to focus on global features of cryo-EM images, enhancing understanding of particle structures and generalizing across conditions. Such a focus is crucial for overcoming the limited training data challenge in the cryo-EM field, improving the model's performance in particle detection and generalization.

Training cryoMAE incorporates both particle and unlabeled regions to bolster model robustness. Exclusive training on particle images could lead MAE to converge towards a homogeneous latent feature space for any given input, escalating the false positive rate by assigning high similarity scores indiscriminately. By including unlabeled regions, cryoMAE learns to recognize features of non-particle spaces, avoiding overfitting to a solely particle-focused feature space. This broader training approach refines the model's ability to distinguish between particle and background regions, markedly lowering false positive rates by assigning more accurate similarity scores to non-particle areas. However, adding unlabeled regions faces some challenges: 1) the diverse background noise in cryo-EM, ranging from crystalline ice contamination and malformed particles to grayscale background regions, which demands a nuanced approach for accurate differentiation; 2) merely incorporating unlabeled data might not prompt the model to learn features unique to particles against complex backgrounds. To optimize the training efficiency of cryoMAE few-shot particle datasets and reduce overfitting risks, while also accounting for a wide range of background noise, we introduced a pre-training phase. Pre-training cryoMAE on a broader set of unlabeled regions better represents background variability. Further, introducing a self-cross similarity loss specifically addresses these noise issues, enhancing the model's ability to discern particles from backgrounds.

3.2.2 Self-cross similarity

Inspired by [26], we develop a self-cross similarity loss to foster distinct latent features for particles and background within cryo-EM images, enhancing the model's ability to differentiate between these regions. This approach aims to increase the disparity in the feature space, thereby improving the precision of particle identification. As illustrated in Fig. 2, the MAE encoder's latent features are utilized not only for image reconstruction by the decoder but also are evaluated using the self-cross similarity loss, further detailed in Fig. 3. The cosine similarity between two feature vectors $\mathbf{a}$ and $\mathbf{b}$ is calculated as $S_{cos}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

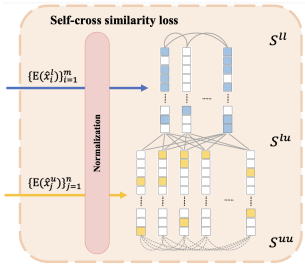


Figure 3. Self-cross similarity loss.

The self-similarity S_{self} is calculated as the mean cosine similarity among the features of positive regions:

$$S_{self} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} S_{cos}(\mathrm{E}(\hat{\mathbf{x}}_i^l), \mathrm{E}(\hat{\mathbf{x}}_j^l)). \quad (1)$$

Similarly, the cross similarity S_{cross} is the mean cosine similarity between features of positive and unlabeled regions:

$$S_{cross} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} S_{cos}(\mathrm{E}(\hat{\mathbf{x}}_i^l), \mathrm{E}(\hat{\mathbf{x}}_j^u)). \quad (2)$$

S_{self} measures the similarity among latent features of exemplars, reflecting the internal consistency of particle features. This is crucial for the model to identify and enhance particle-specific patterns, facilitating better distinction from background noise. The goal is for S_{self} to increase, indicating stronger similarity within particle groups. Conversely, S_{cross} assesses the similarity between exemplar features and those of negative regions, capturing the distinctiveness between particles and background. The objective is for S_{cross} to decrease, reducing feature similarity between particles and background. Self-cross similarity loss L_{SCS} is designed to optimize these dynamics, improving model's ability to differentiate between particles and backgrounds:

$$L_{SCS}(S_{cross}, S_{self}) = \max\left(\tau, 1 + \alpha \cdot S_{cross} - (1-\alpha) \cdot S_{self}\right). \quad (3)$$

α balances self and cross-similarity contributions, and τ sets a minimum difference threshold between them, limiting further distinction efforts beyond it.

3.2.3 PU learning

Inspired by [1], we identify a limitation in our previous loss function design, which treats unlabeled data as negative. Randomly cropped training regions may unintentionally include particles, potentially confusing the model's distinction between labeled particles and background noise. This overlap complicates training, as the model could wrongly link particle features with the background, undermining our strategy to reduce background similarity scores and challenging the model's ability to learn discriminatively. To enhance the loss formulation, we accommodate the potential inclusion of particles in unlabeled regions. Acknowledging that a certain proportion ($\hat{\pi}$) of these samples may harbor particles, we modify the representation of features for these unlabeled samples.

We adjust feature representation by implementing a weighting scheme grounded in the estimated probability $\hat{\pi}$ that an unlabeled region harbors a particle, alongside a complementary weight $1-\hat{\pi}$ for regions likely devoid of particles. This approach enhances the model's capacity to differentiate between particle-laden regions and background, optimizing the use of unlabeled data in training and improving particle identification accuracy. The presence of particles in unlabeled regions necessitates a recalibration of similarity calculations, introducing a deeper analysis of self-similarity among potential positives and their cross-similarity with potential negatives within the unlabeled data:

$$\hat{S}_{self} = \frac{1}{(m+\hat{\pi}n)^2}\left[S^{ll} + 2\hat{\pi}S^{lu} + \hat{\pi}^2 S^{uu}\right], \tag{4}$$

$$\hat{S}_{cross} = \frac{1}{(m+\hat{\pi}n)\times(1-\hat{\pi})n}\left[(1-\hat{\pi})S^{lu} + \hat{\pi}(1-\hat{\pi})S^{uu}\right]. \tag{5}$$

S^{ll}, S^{lu}, and S^{uu} measure the sums of cosine similarities among exemplars, between exemplars and unlabeled regions, and among unlabeled regions, respectively. Detailed derivation is shown in Appendix A. The refined self-cross similarity loss, $\hat{L}_{SCS}(\hat{S}_{cross}, \hat{S}_{self})$, adeptly captures the complexity of similarity within data subsets. By refining these calculations, we refine these metrics to account for the intricate characteristics of unlabeled data, facilitating a more discerning and efficacious training regimen.

The total loss of cryoMAE, taking into account the reconstruction loss:

$$L_{total} = L_{MSE} + \beta \cdot \hat{L}_{SCS}. \tag{6}$$

Here β adjusts the weight of the self-cross similarity loss in the overall loss function, balancing reconstruction accuracy with discriminative learning.

3.3. Stage 2: Particle Picking on Query Micrographs

In stage 2, our model undertakes particle picking by utilizing the MAE encoder to scan query micrographs and extract features from each sliding window region, as detailed in **Stage 2** of Fig. 2. This stage does not employ masking for the input regions. The extracted latent features are then matched against those of exemplars through cosine similarity, assigning similarity scores to each region based on the highest similarity. Following the completion of the sliding process on a micrograph, these similarity scores are ranked. It is crucial to recognize the variability in the imaging states of different micrographs, where a single threshold does not work well. Therefore, we adopt a density-based method to determine the most suitable cutoff threshold for each micrograph automatically. This process involves calculating the average distance of each score to its k nearest neighbors, and finding the score where the rate of change in these average distances is maximized as the cutoff threshold. Refer to Algorithm 1 in Appendix B for pseudo code. Coordinates of all regions with similarity scores exceeding this threshold are recorded in a `.star` file. The `.star` format is widely used in cryo-EM to document particle coordinates, aiding in subsequent steps like 3D reconstruction using CryoSPARC.

4. Experiments

This section evaluates cryoMAE through ablation studies, sensitive analysis, and qualitative visualizations.

4.1. Experimental Setup

4.1.1 Datasets

We evaluated cryoMAE using five datasets from the CryoPPP database [5], the only expert-curated cryo-EM dataset with particle coordinate labels. To ensure fair comparisons with other methods, we excluded particles used in those methods' pre-training (details in Appendix C). Thus, our evaluation set included only the official CryoPPP test sets: EMPIAR IDs 10081, 10093, 10345, 10532, and 11056.

To further test our approach, we gathered five more unlabeled micrograph datasets from EMPIAR database [17] (details in Appendix C), which provides raw, high-resolution cryo-EM images. As these lack coordinate labels, we used 3D reconstruction resolution as the evaluation metric.

4.1.2 Baselines

We utilized crYOLO [32], Topaz [1], and CryoTransformer [6] as our baselines. For detailed information on these models, please refer to Appendix D.

4.1.3 Evaluation metrics

Our evaluation metrics include precision, recall, F1 scores, and 3D reconstruction resolution. A true positive occurs when a picked particle region overlaps with a groundtruth, achieving an intersection over union (IoU) of 0.5 or higher. False positives include picked regions that either have an IoU less than 0.5 with any groundtruth region or represent multiple detections for a single groundtruth. False negatives are groundtruths that remain undetected.

4.1.4 3D reconstruction

CryoSPARC [22] is used to perform 3D reconstructions. The detailed workflow can be found in Appendix E.

4.1.5 Implementation Details

Data pre-processing We applied identical denoising techniques to cryo-EM images as in [11]. For each protein type, we selected 15 particle exemplars. Given the diverse orientations of particles, we augmented the dataset by rotating exemplars by $90°$, $180°$, and $270°$. We included randomly cropped regions as unlabeled samples, maintaining a $2:1$ ratio of unlabeled to positive samples for training.

Network architecture The MAE model with 12 encoder blocks (embedding dimension 768 and 12 attention heads) and 8 decoder blocks (embedding dimension 512 and 16 attention heads) is used.

Training and inference details The pre-training phase used unlabeled regions from a variety of cryo-EM images including EMPIAR-10075, 10077, 10096, 10291, 11051, 11057, and 11183, each with 300 micrographs. For each micrograph, we extracted 10 random regions. The pre-training learning rate is $1e-5$, optimizer is Adam [18], and epochs are 100. Then cryoMAE underwent fine-tuning on specific particle datasets. To retain essential features, we froze the first six encoder blocks. Fine-tuning extended for 5000 epochs on the few-shot training set. The parameters of $\hat{L}_{SCS}$ were: $\alpha = 0.7$, $\tau = 0.02$, $\beta = 5$, and $\hat{\pi} = 0.018$. The sliding stride is set to 28 during particle picking.

4.2. Overall Performance

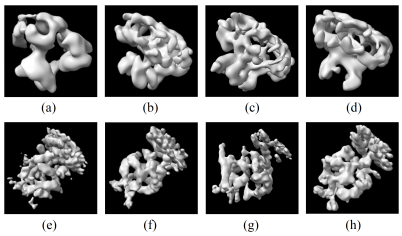


Figure 4. 3D reconstructions from crYOLO, Topaz, CryoTransformer, and cryoMAE: (a)-(d) for 10081, (e)-(h) for 10093.

The performance comparison of crYOLO, Topaz, CryoTransformer, and cryoMAE on the CryoPPP dataset is presented in Table 1, with the corresponding 3D reconstructions shown in Fig. 4. The performance comparison on the additional datasets is provided in Table 2. The number of groundtruth (GT) particles in Table 1 represents the number of provided labeled particles in the CryoPPP dataset. The total number of picked particles (including false positives) used for generating the 3D reconstruction of a particle can be calculated using the formula:

$$\text{Total Particles} = \text{TP} + \text{FP} = \frac{\text{\# of GT particles} \times \text{Recall}}{\text{Precision}} \quad (7)$$

Table 1 reveals high precision of crYOLO but its tendency to miss many particles. This is particularly evident with 10081, where crYOLO shows strong performance due to pre-training on datasets that included 10081 particles. This pre-trained crYOLO model raises questions about its generalization to new particle types, where performance significantly drops, highlighting a generalization issue. Topaz and CryoTransformer score well in the recall, albeit with a high false positive rate. CryoMAE excels in both precision and recall, outperforming Topaz in all evaluated metrics for five particles and showing better recall than crYOLO, aside from 10081. It surpasses crYOLO in precision and F1 score, excluding 10081. Notably, precision and recall focus on whether the selected particles are correct, whereas *ab-initio* reconstruction resolution emphasizes whether the diversity of orientations among the selected particles is sufficient. CryoMAE leads to the highest 3D reconstruction resolution on all particles except one AAA-ATPase dataset, which indicates that it can select particles with a more diverse range of orientations.

4.3. Ablation Studies

This section validates the contributions of key cryoMAE components.

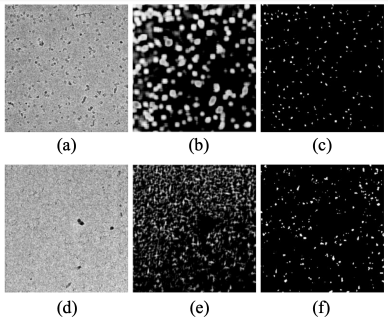


Figure 5. Similarity maps by cryoMAE w/ and w/o adjusted SCS loss. (a), (d) micrographs; (b), (e) w/o the loss; (c), (f) w/ the loss.

4.3.1 W/ and w/o self-cross similarity loss

We assessed the performance of cryoMAE across different configurations of self-cross similarity loss (with self-cross similarity loss, with unadjusted self-cross similarity loss, and with adjusted self-cross similarity loss) in Table 3, revealing optimal performance with adjusted loss. This finding highlights the crucial impact of self-cross similarity loss in enhancing feature extraction, making cryoMAE more discerning in particle selection and greatly lowering the chance of incorrect region identification. CryoMAE without the loss incorrectly scores many non-particle regions highly, evident from widespread white areas in Fig.5(b)(e). With this loss, cryoMAE specificity improves, accurately identifying particle regions, as shown in Fig.5(c)(f), reducing false scores for background areas.

Table 1. Performance comparison of crYOLO, Topaz, CryoTransformer, and cryoMAE on CryoPPP. More results are in Appendix F & G.

EMPIAR ID	Data information					Precision				Recall				F1 score				Resolution (Å)			
	Protein type	# of micrographs	# of GT particles	Image size	Particle diameter (px)	crYOLO	Topaz	Cryo-Transformer	Ours	crYOLO	Topaz	Cryo-Transformer	Ours	crYOLO	Topaz	Cryo-Transformer	Ours	crYOLO	Topaz	Cryo-Transformer	Ours
10081	Transport protein	300	39,352	(3710, 3838)	154	**0.705**	0.412	0.453	0.645	0.867	0.855	0.578	**0.939**	**0.777**	0.556	0.508	0.765	12.25	12.72	14.13	**11.32**
10093	Membrane protein	300	56,394	(3838, 3710)	172	0.380	0.328	0.252	**0.383**	0.355	0.209	**0.529**	0.497	0.367	0.255	0.342	**0.433**	11.64	11.62	13.42	**9.02**
10345	Signaling protein	300	15,894	(3838, 3710)	149	0.441	0.195	0.166	**0.473**	0.561	0.732	**0.912**	0.733	0.494	0.308	0.280	**0.575**	11.63	10.39	10.82	**10.27**
10532	Viral protein	300	87,933	(4096, 4096)	174	0.501	0.387	0.380	**0.503**	0.231	0.311	**0.504**	0.497	0.316	0.345	0.433	**0.500**	12.86	10.85	11.72	**9.92**
11056	Transport Protein	361	125,908	(5760, 4092)	164	0.690	0.453	0.576	**0.694**	0.465	0.578	0.555	**0.671**	0.556	0.507	0.566	**0.682**	-	-	-	-
Average	-	-	-	-	-	**0.543**	0.355	0.365	0.540	0.496	0.537	0.616	**0.667**	0.502	0.394	0.426	**0.591**	12.10	11.40	12.52	**10.13**

* The micrographs for EMPIAR-11056 were not available in the *.mrc* format, precluding us from performing a 3D reconstruction for this particle.

Table 2. Performance comparison of crYOLO, Topaz, CryoTransformer, and cryoMAE on additional datasets.

Dataset	Data information			# of particles picked by				Resolution (Å)			
	Image size	Particle diameter (px)	# of micrographs	crYOLO	Topaz	Cryo-Transformer	Ours	crYOLO	Topaz	Cryo-Transformer	Ours
TcdA1	(4096, 4096)	352	80	9,145	9,467	10,543	10,319	10.26	10.07	9.41	**9.34**
80S ribosome	(7420, 7676)	420	300	20,930	22,482	22,709	22,417	9.57	8.02	8.07	**7.84**
AAA-ATPase	(7420, 7676)	560	300	17,802	19,315	19,023	18,880	9.04	**8.93**	8.95	9.00
PVY coat (truncated)	(3838, 3710)	210	300	33,089	36,074	33,361	35,140	8.73	8.21	8.66	**8.14**
Chicken CALHM1	(3710, 3838)	176	300	21,223	36,385	42,367	51,307	13.56	12.50	12.47	**11.99**
Average	-	-	-	-	-	-	-	10.23	9.55	9.51	**9.26**

Table 3. Comparison of cryoMAE supervised w/o SCS loss, w/ unadjusted SCS loss L_{SCS}, and w/ adjusted SCS loss $\hat{L}_{SCS}$.

EMPIAR ID	Precision			Recall			F1 Score		
	w/o	w/ L_{SCS}	w/ $\hat{L}_{SCS}$	w/o	w/ L_{SCS}	w/ $\hat{L}_{SCS}$	w/o	w/ L_{SCS}	w/ $\hat{L}_{SCS}$
10081	0.225	0.639	**0.645**	0.652	0.928	**0.939**	0.335	0.757	**0.765**
10093	0.143	0.376	**0.383**	0.216	0.493	**0.497**	0.172	0.427	**0.433**
10345	0.180	**0.474**	0.473	0.547	0.724	**0.733**	0.271	0.573	**0.575**
10532	0.177	0.495	**0.503**	0.269	0.478	**0.497**	0.213	0.486	**0.500**
11056	0.154	0.691	**0.694**	0.327	0.639	**0.671**	0.209	0.664	**0.682**
Average	0.176	0.535	**0.540**	0.402	0.652	**0.667**	0.240	0.581	**0.591**

We conduct 2D t-SNE visualizations to analyze the latent features of cryoMAE under varying conditions: trained on a dataset without unlabeled regions, on a dataset with unlabeled regions without the adjusted self-cross similarity loss, and on a dataset with unlabeled regions with the adjusted self-cross similarity loss. For each visualization, we randomly select a consistent set of 60 exemplars and 360 unlabeled regions from EMPIAR-10081 to ensure comparability across the three scenarios. The visualizations are in Fig. 6a, Fig. 6b and Fig. 6c, respectively.

As demonstrated in Fig. 6a, training exclusively on particle regions leads cryoMAE to generate homogeneous latent features for any input. This approach risks elevating the false positive rate by indiscriminately assigning high similarity scores, including to background regions. Fig. 6b illustrates that incorporating unlabeled regions enables cryoMAE to discern features of non-particle regions, thus mitigating over-fitting to a particle-exclusive feature space. Consequently, the model acquires a preliminary capability to differentiate between particle and background regions, although with limited clarity (as observed in the latent feature space 2D visualization, where the blue and yellow clusters are approximately but not distinctly separated). Further advancements are evident in Fig. 6c, where the introduction of adjusted self-cross similarity loss significantly enhances the model's ability to distinguish between background regions and particles. This improvement is illustrated by the distinct separation between the two clusters in the figure, despite the presence of some yellow points within the blue cluster. These exceptions, representing particle-containing regions within unlabeled areas, are considered reasonable.

4.3.2 W/ and w/o pre-training

Table 4 demonstrates how pre-training strategy markedly promotes model performance, with gains in precision and recall at 41.4%, and in F1 score at 36.8%. Without pre-training, cryoMAE shows reduced effectiveness, likely due to overfitting on the limited training data, hindering its generalization capabilities, especially in recognizing varied particle orientations and background noise variations.

4.3.3 Max and mean matching strategies

Table 5 presents a comparative study on two similarity score calculation methods for matching sliding regions against exemplar latent features: maximum vs. average cosine similarity. Table 5 reveals that maximum cosine similarity outperforms average cosine similarity, particularly in precision. This advantage is linked to the varied orientation distributions among particle exemplars. Maximum cosine similarity effectively matches regions to their closest exemplar across different orientations, ensuring optimal scores. Conversely, average cosine similarity dilutes scores for particles with diverse orientations, as it averages across all exemplars, including those with markedly different particle orientations from the target region. This dilution lowers similarity scores for such particles, reducing their distinctiveness from the background and making accurate particle identification more challenging amidst noise.

Table 4. W/ and w/o pre-training.

EMPIAR ID	Precision		Recall		F1 Score	
	w/	w/o	w/	w/o	w/	w/o
10081	**0.645**	0.352	**0.939**	0.919	**0.765**	0.509
10093	**0.383**	0.281	**0.497**	0.479	**0.433**	0.354
10345	**0.473**	0.209	**0.733**	0.581	**0.575**	0.307
10532	**0.503**	0.404	**0.497**	0.347	**0.500**	0.373
11056	**0.694**	0.664	**0.671**	0.579	**0.682**	0.619
Average	**0.540**	0.382	**0.667**	0.581	**0.591**	0.432

Table 5. Max and mean matching.

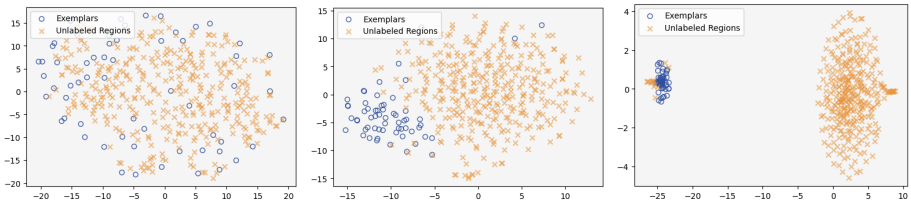
EMPIAR ID	Precision		Recall		F1 Score	
	Max	Mean	Max	Mean	Max	Mean
10081	**0.645**	0.595	0.939	**0.946**	**0.765**	0.731
10093	**0.383**	0.367	0.497	**0.548**	0.433	**0.440**
10345	**0.473**	0.396	**0.733**	0.718	**0.575**	0.510
10532	**0.503**	0.498	0.497	**0.502**	**0.500**	0.500
11056	**0.694**	0.606	**0.671**	0.650	**0.682**	0.628
Average	**0.540**	0.492	0.667	**0.673**	**0.591**	0.562

4.4. Sensitivity Analysis

In this section, we examined the impact of different numbers of exemplars and the sliding strides on performance.

4.4.1 Number of exemplars

Table 6 shows how the performance of cryoMAE varies with the number of exemplars. As expected, adding more



(a) trained on a dataset w/o unlabeled regions. (b) trained w/o adjusted self-cross similarity loss. (c) trained w/ adjusted self-cross similarity loss.

Figure 6. 2D t-SNE visualizations of cryoMAE latent feature space (on EMPIAR-10081).

Table 6. CryoMAE with various exemplar number settings $\{1, 5, 15, 25\}$.

EMPIAR ID	Precision				Recall				F1 Score			
	1	5	15	25	1	5	15	25	1	5	15	25
10081	0.167	0.401	0.645	**0.653**	0.774	0.896	0.939	**0.943**	0.275	0.554	0.765	**0.772**
10093	0.068	0.243	**0.383**	0.373	0.139	0.361	0.497	**0.508**	0.091	0.290	**0.433**	0.430
10345	0.102	0.183	**0.473**	0.460	0.585	0.706	0.733	**0.772**	0.174	0.291	0.575	**0.576**
10532	0.138	0.296	**0.503**	0.481	0.251	0.478	0.497	**0.507**	0.178	0.366	**0.500**	0.494
11056	0.287	0.289	0.694	**0.712**	0.353	0.485	0.671	**0.680**	0.317	0.362	0.682	**0.696**
Average	0.152	0.282	**0.540**	0.536	0.420	0.585	0.667	**0.682**	0.207	0.373	0.591	**0.594**

exemplars generally improves performance, owing to a more comprehensive representation of particle orientations in the similarity scoring process. This is particularly beneficial for particles with diverse orientations, as more exemplars increase the chance of capturing regions across different orientation states, improving recall. However, the improvement plateaus after a certain number of exemplars, with precision potentially decreasing. This is because particle orientations are limited, and once the diversity of these states is adequately covered, more exemplars offer little benefit and may even raise false positives by increasing the likelihood of background regions being mistakenly scored highly. Thus, considering the diminishing returns beyond 15 exemplars, we identify this count as the optimal number.

4.4.2 Sliding strides

Table 7. CryoMAE with various sliding strides $\{14, 28, 42, 56\}$.

EMPIAR ID	Precision				Recall				F1 Score			
	14	28	42	56	14	28	42	56	14	28	42	56
10081	0.584	0.645	0.665	**0.689**	**0.942**	0.939	0.856	0.772	0.721	**0.765**	0.748	0.728
10093	0.376	0.383	0.399	**0.411**	0.489	**0.497**	0.311	0.157	0.425	**0.433**	0.350	0.227
10345	0.446	0.473	0.478	**0.479**	**0.746**	0.733	0.551	0.502	0.558	**0.575**	0.512	0.490
10532	0.500	0.503	0.501	**0.503**	**0.604**	0.497	0.430	0.343	**0.547**	0.500	0.463	0.408
11056	0.689	0.694	0.692	**0.695**	**0.702**	0.671	0.606	0.522	**0.695**	0.682	0.646	0.596
Average	0.520	0.540	0.547	**0.555**	**0.697**	0.667	0.551	0.459	0.589	**0.591**	0.544	0.490
Average time (s)	356.4	87.2	38.0	**20.8**	-	-	-	-	-	-	-	-

Table 7 outlines the performance of cryoMAE across various sliding strides, noting that decreasing stride from 56 to 14 typically boosts recall but diminishes precision. This can be attributed to the fact that larger strides cause a certain particle to be present in fewer windows, minimizing duplicate detections and enhancing precision. However, this can result in lower similarity scores for many particles, as they're more likely to be close to window edges, which can reduce their likelihood of being selected and decrease recall. The F1 score, a precision-recall harmony measure, tends to improve with smaller strides. Yet, reducing stride size significantly lengthens processing time per query image. Considering the trade-off between time efficiency and accuracy, a 28-pixel stride is identified as the optimal.

4.4.3 Masking ratios

Table 8. CryoMAE w/ various masking ratios $\{25\%, 50\%, 75\%\}$.

EMPIAR ID	Precision			Recall			F1 Score		
	25%	50%	75%	25%	50%	75%	25%	50%	75%
10081	0.617	0.619	**0.645**	0.904	0.933	**0.939**	0.733	0.744	**0.765**
10093	0.379	0.392	**0.383**	0.489	0.496	**0.497**	0.427	0.438	**0.433**
10345	0.465	0.470	**0.473**	0.711	0.720	**0.733**	0.562	0.569	**0.575**
10532	0.489	0.497	**0.503**	0.479	0.492	**0.497**	0.484	0.494	**0.500**
11056	0.684	0.692	**0.694**	0.652	0.663	**0.671**	0.668	0.677	**0.682**
Average	0.527	0.534	**0.540**	0.647	0.661	**0.667**	0.575	0.585	**0.591**

Table 8 presents the performance metrics of cryoMAE when applying different masking ratios of 25%, 50%, and 75%. Previous studies [14] suggest that higher masking ratios can enable models to learn high-level, abstract features, although at the cost of potentially missing image-specific details. However, in cryo-EM imaging, low SNR and small particle sizes make it crucial to retain particle-specific information. Excessive masking could obscure essential features, diminishing model's capacity to accurately capture the underlying structures. In our framework, we address the challenges by adjusting exemplar sizes to ensure that each particle occupies a significant portion of input images.

5. Conclusion

We introduce cryoMAE, a pioneering approach in few-shot learning tailored for the cryo-EM field, significantly reducing the dependence on extensive labeled datasets for particle picking. By harnessing the power of MAE and integrating a novel self-cross similarity loss, cryoMAE achieves superior performance in identifying particle regions amidst the challenges posed by low SNR and diverse particle orientations. Validations on the different datasets demonstrate cryoMAE's superiority over existing NN-based methods. This innovation not only streamlines the process of high-resolution protein structure determination but also makes it more accessible to a wider scientific audience, promising to accelerate discoveries in structural biology.

6. Acknowledgement

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629, DBI-2238093, IIS-2007595, IIS-2211597, and MCB-2205148. This work was supported in part by by UPMC Enterprises, Oracle Cloud credits and related resources provided by Oracle for Research, and the computational resources support from AMD HPC Fund.

References

[1] Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature methods*, 16(11):1153–1160, 2019. 1, 2, 4, 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[5] Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. A large expert-curated cryo-em image dataset for machine learning protein particle picking. *Scientific Data*, 10(1):392, 2023. 1, 2, 5

[6] Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. Cryotransformer: a transformer model for picking protein particles from cryo-em micrographs. *Bioinformatics*, 40(3):btae109, 2024. 1, 2, 5

[7] Ashwin Dhakal, Cole McKay, John J Tanner, and Jianlin Cheng. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Briefings in Bioinformatics*, 23(1):bbab476, 2022. 1

[8] Nabin Giri and Jianlin Cheng. Improving protein–ligand interaction modeling with cryo-em data, templates, and deep learning in 2021 ligand model challenge. *Biomolecules*, 13(1):132, 2023. 1

[9] Robert M Glaeser. Stroboscopic imaging of macromolecular complexes. *nature methods*, 10(6):475–476, 2013. 1

[10] Rajan Gyawali, Ashwin Dhakal, Liguo Wang, and Jianlin Cheng. Accurate cryo-em protein particle picking by integrating the foundational ai image segmentation model and specialized u-net. 2023. 2

[11] Rajan Gyawali, Ashwin Dhakal, Liguo Wang, and Jianlin Cheng. Accurate cryo-em protein particle picking by integrating the foundational ai image segmentation model and specialized u-net. *bioRxiv*, 2023. 6

[12] Rajan Gyawali, Ashwin Dhakal, Liguo Wang, and Jianlin Cheng. Cryovirusdb: a labeled cryo-em image dataset for ai-driven virus particle picking. *bioRxiv*, 2023. 1

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 3

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 8

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[16] Qinwen Huang, Ye Zhou, Hsuan-Fu Liu, and Alberto Bartesaghi. Accurate detection of proteins in cryo-electron tomograms from sparse labels. In *European Conference on Computer Vision*, pages 644–660. Springer, 2022. 3

[17] Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 2023. 5

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Hongjia Li, Ge Chen, Shan Gao, Jintao Li, Xiaohua Wan, and Fa Zhang. A transfer learning-based classification model for particle pruning in cryo-electron microscopy. *Journal of Computational Biology*, 29(10):1117–1131, 2022. 1

[20] Jacqueline LS Milne, Mario J Borgnia, Alberto Bartesaghi, Erin EH Tran, Lesley A Earl, David M Schauder, Jeffrey Lengyel, Jason Pierson, Ardan Patwardhan, and Sriram Subramaniam. Cryo-electron microscopy–a primer for the non-microscopist. *The FEBS journal*, 280(1):28–45, 2013. 1

[21] Long Pei, Min Xu, Zachary Frazier, and Frank Alber. Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. *BMC bioinformatics*, 17:1–13, 2016. 1

[22] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature methods*, 14(3):290–296, 2017. 1, 6

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[24] Alexander Rigort, David Günther, Reiner Hegerl, Daniel Baum, Britta Weber, Steffen Prohaska, Ohad Medalia, Wolfgang Baumeister, and Hans-Christian Hege. Automated segmentation of electron tomograms for a quantitative description of actin filament networks. *Journal of structural biology*, 177(1):135–144, 2012. 1

[25] Sjors HW Scheres. Semi-automated selection of cryo-em particles in relion-1.3. *Journal of structural biology*, 189(2):114–122, 2015. 1

[26] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9529–9538, 2022. 4

[27] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007. 1

[28] Huanrong Tang, Yaowu Wang, Jianquan Ouyang, and Jinlin Wang. Simcryocluster: a semantic similarity clustering method of cryo-em images by adopting contrastive learning. *BMC bioinformatics*, 25(1):77, 2024. 3

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[30] Marlene Vinzenz, Maria Nemethova, Florian Schur, Jan Mueller, Akihiro Narita, Edit Urban, Christoph Winkler, Christian Schmeiser, Stefan A Koestler, Klemens Rottner, et al. Actin branching in the initiation and maintenance of lamellipodia. *Journal of cell science*, 125(11):2775–2785, 2012. 2

[31] NR Voss, CK Yoshioka, M Radermacher, CS Potter, and B Carragher. Dog picker and tiltpicker: software tools to facilitate particle selection in single particle electron microscopy. *Journal of structural biology*, 166(2):205–213, 2009. 1

[32] Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, et al. Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. *Communications biology*, 2(1):218, 2019. 1, 2, 5

[33] Feng Wang, Huichao Gong, Gaochao Liu, Meijing Li, Chuangye Yan, Tian Xia, Xueming Li, and Jianyang Zeng. Deeppicker: A deep learning approach for fully automated particle picking in cryo-em. *Journal of structural biology*, 195(3):325–336, 2016. 1, 2

[34] Xiangrui Zeng, Anson Kahng, Liang Xue, Julia Mahamid, Yi-Wei Chang, and Min Xu. High-throughput cryo-et structural pattern mining by unsupervised deep iterative subtomogram clustering. *Proceedings of the National Academy of Sciences*, 120(15):e2213149120, 2023. 2

[35] Jiakai Zhang, Qihe Chen, Yan Zeng, Wenyuan Gao, Xuming He, Zhijie Liu, and Jingyi Yu. Genem: Physics-informed generative cryo-electron microscopy. *arXiv preprint arXiv:2312.02235*, 2023. 1

[36] Yanan Zhu, Qi Ouyang, and Youdong Mao. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC bioinformatics*, 18:1–10, 2017. 1, 2