

Supplementary Information for CryoPROS: Correcting Misalignment Caused by Preferred Orientation Using AI-generated Auxiliary Particles

Hui Zhang^{1†}, Dihan Zheng^{2†}, Qiurong Wu^{3,4,5}, Nieng Yan^{3,4,5,6,7}, Han Peng^{8,9}, Qi Hu^{8,9}, Ying Peng¹⁰, Zhaofeng Yan¹⁰, Zuoqiang Shi^{2,11}, Chenglong Bao^{2,6,11*} and Mingxu Hu^{3,7*}

¹Qiuzhen College, Tsinghua University, Beijing, China.

²Yau Mathematical Sciences Center, Tsinghua University, Beijing, China.

³Beijing Frontier Research Center for Biological Structure (Tsinghua University), Beijing, China.

⁴Tsinghua-Peking Joint Center for Life Sciences, Tsinghua University, Beijing, China.

⁵School of Life Sciences, Tsinghua University, Beijing, China.

⁶State Key Laboratory of Membrane Biology, School of Life Sciences, Tsinghua University, Beijing, China.

⁷Institute of Bio-Architecture and Bio-Interactions, Shenzhen Medical Academy of Research and Translation, Shenzhen, China.

⁸School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China.

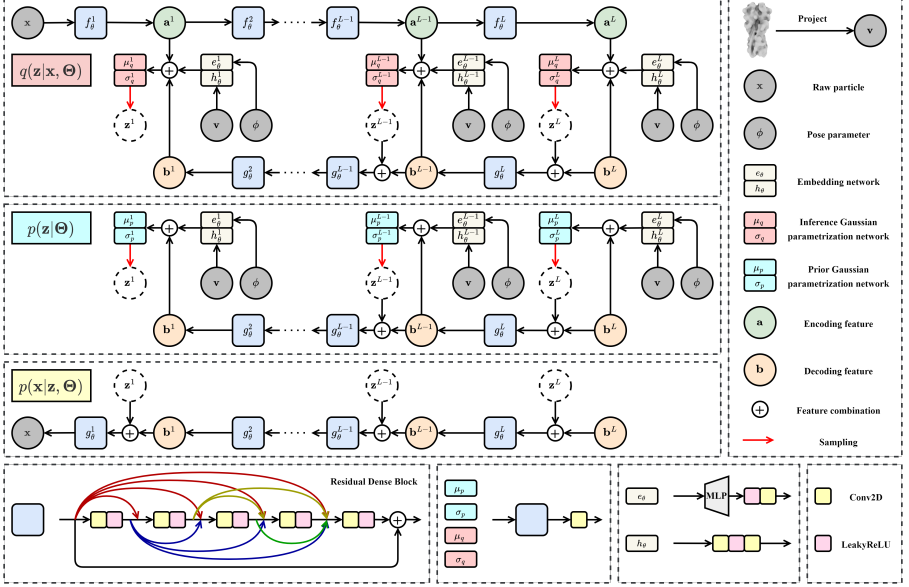
⁹Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China.

¹⁰School of Biomedical Sciences, Hunan University, Changsha, China.

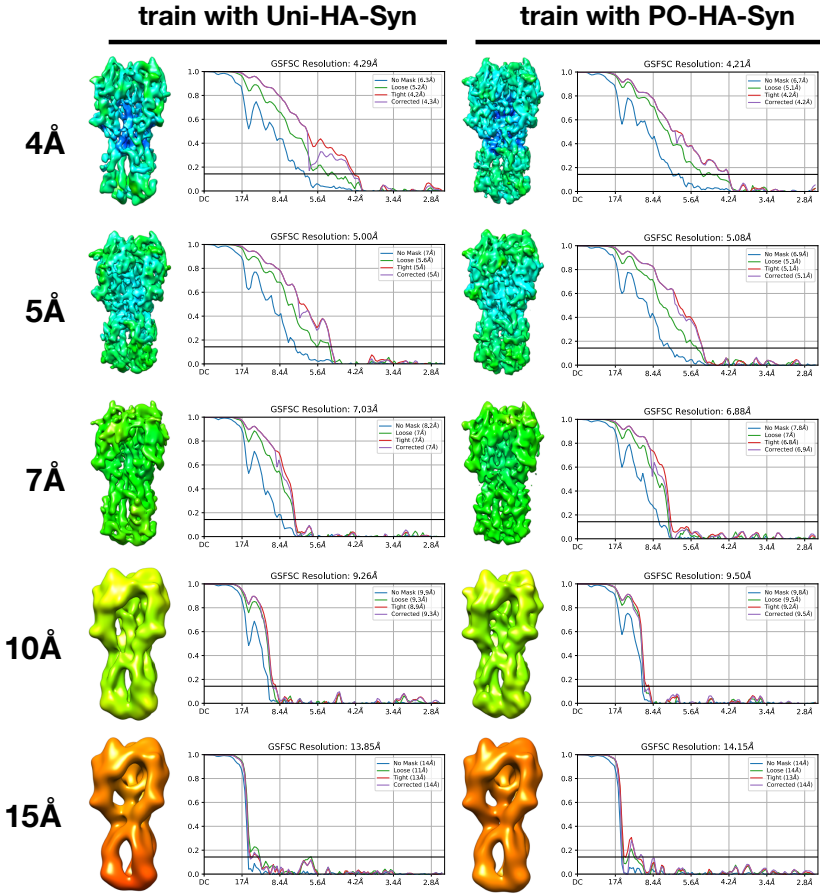
¹¹Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, China.

*Corresponding author(s). E-mail(s): clbao@mail.tsinghua.edu.cn;
humingxu@smart.org.cn;

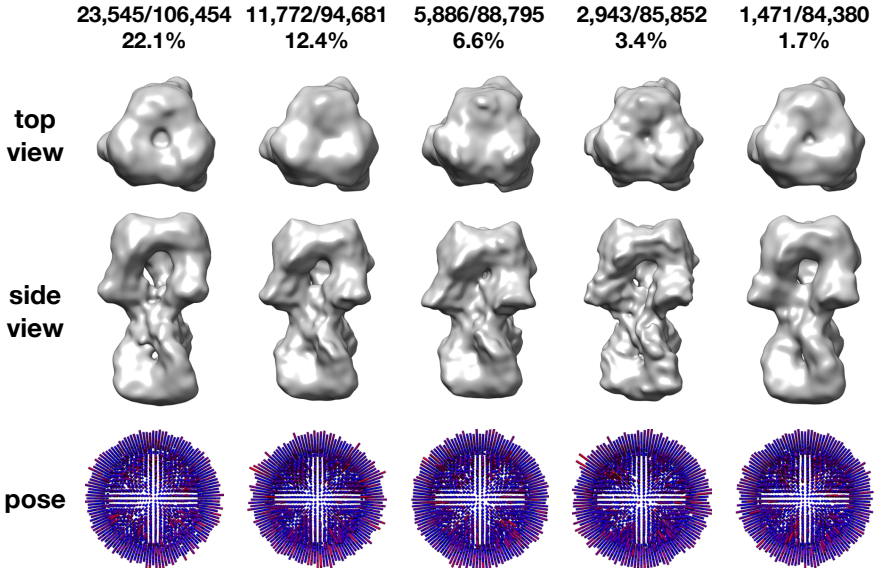
[†]These authors contributed equally to this work.



Supplementary Figure 1: Network architecture. The hierarchical conditional VAE framework comprises three principal components: the conditional encoding model $q(\mathbf{z}|\mathbf{x}, \Theta)$, the conditional prior model $p(\mathbf{z}|\Theta)$, and the decoding model $p(\mathbf{x}|\mathbf{z}, \Theta)$. The raw particle $\mathbf{x}$ undergoes a hierarchical encoding process to obtain the latent variable $\mathbf{z}$. Within each layer, a residual dense block serves as the feature extraction network. The Gaussian parameterization network is composed of the residual dense block and a convolutional layer. The pose embedding network e_θ is constructed with an MLP and a convolutional layer, while the projection embedding network h_θ is formulated from an assembly of convolutional layers.

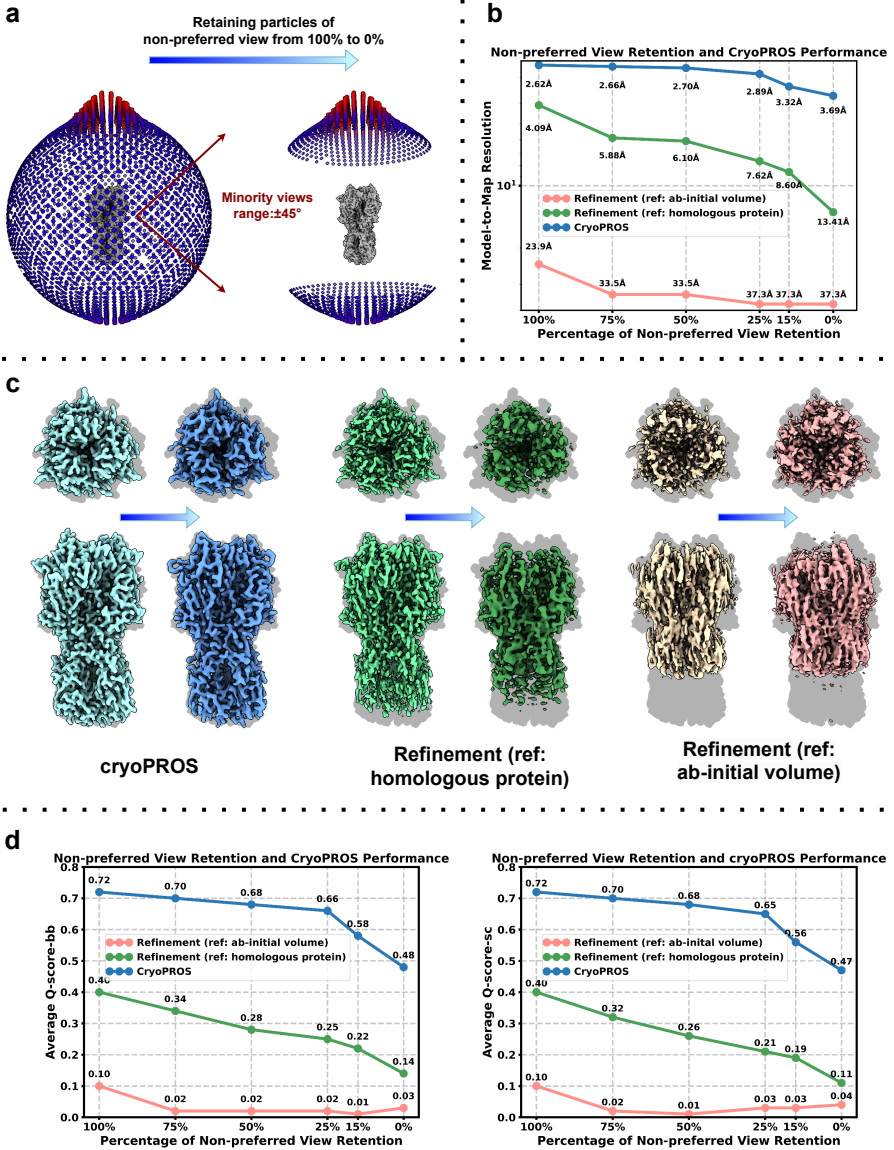


Supplementary Figure 2: The generative module can produce auxiliary particles that result in a density map that meets expectations. The generative module underwent training with varied low-pass filtered (4Å, 5Å, 7Å, 10Å, and 15Å) maps of the ground-truth atomic model (PDB ID: 6WXB), represented across five rows. It was also trained with two datasets: Uni-HA-Syn and PO-HA-Syn, represented in two columns to indicate whether the dataset's pose was uniform or preferentially oriented. Following the training phase, the generative module produced 10,000 auxiliary particles. Subsequently, non-uniform refinement procedures by cryoSPARC were applied to these 10,000 auxiliary particles, after their poses had been discarded. The density maps, featuring surfaces colored to represent local resolution (ranging from blue for 4Å to red for 15Å), along with the FSCs, were both illustrated.

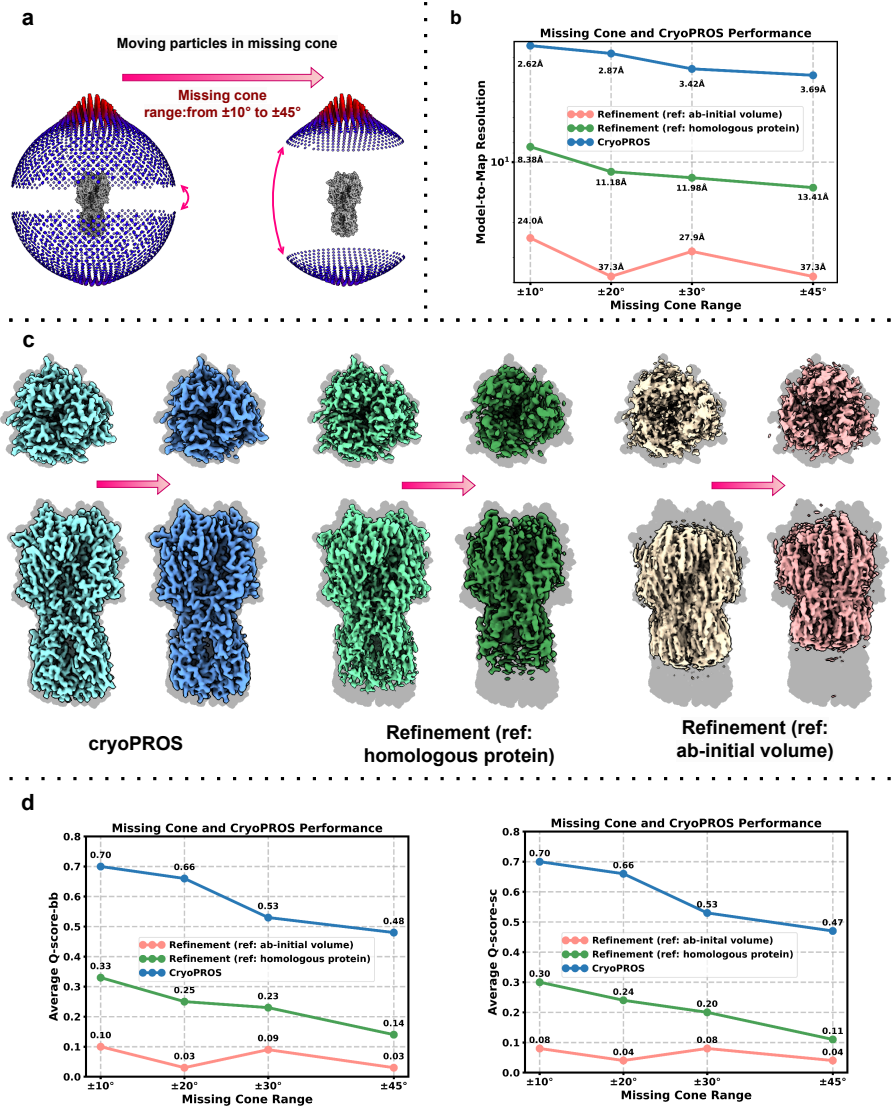


Supplementary Figure 3: Generative model adept at handling datasets with highly preferentially oriented particles during training.

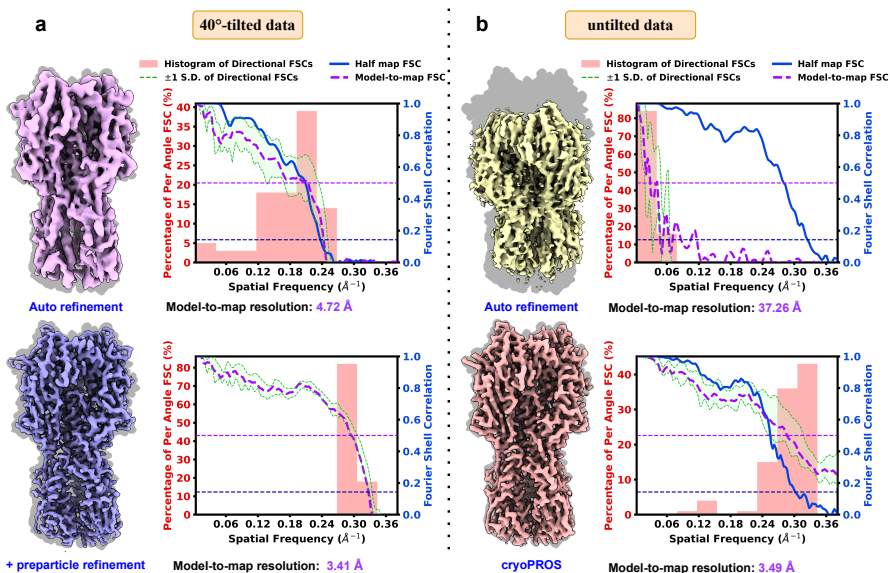
Successive halvings of non-preferred views were performed on the PO-HA-Syn dataset, reducing their proportion from 22.1% (after the first halving) to 1.7% (after the fifth halving). The number of particles from non-preferred views, the total number of particles, and the ratio of non-preferred views are displayed at the top. The generative module was trained using 10Å low-pass maps of the ground-truth atomic model (PDB ID: 6WXB), incorporating ground-truth poses from particles in preferred views as well as from particles in non-preferred views after multiple rounds of halving. Following training, 10,000 uniformly distributed particles were generated from the generative module. Non-uniform refinement was then applied to these particles after discarding their poses. The top and side views of the density (with a contour level of 0.15) and the pose distribution of the generated particles after refinement are depicted.



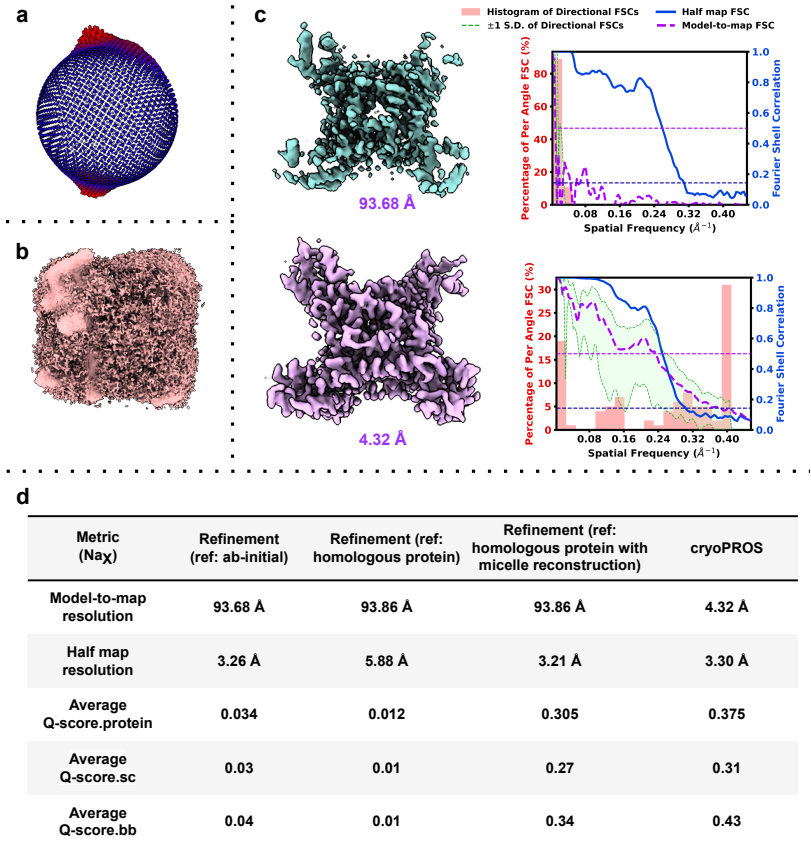
Supplementary Figure 4: Impact of non-preferred view reduction on cryoPROS performance. **a**, Pose distribution after removing non-preferred views. Visualized for PO-HA-Syn-Re100 to PO-HA-Syn-Re00, showcasing the progressive reduction of non-preferred views. **c**, Comparison of density maps determined by cryoPROS, refinement (ref: homologous protein), and refinement (ref: ab-initial) methods for the data spanning from PO-HA-Syn-Re00 to PO-HA-Syn-Re100. **b**, **d**, Performance metrics across all datasets, along with results of conventional refinement methods using initial volumes of ab-initial and homologous protein respectively. Metrics include model-to-map resolution, average Q-score-bb, and average Q-score-sc.



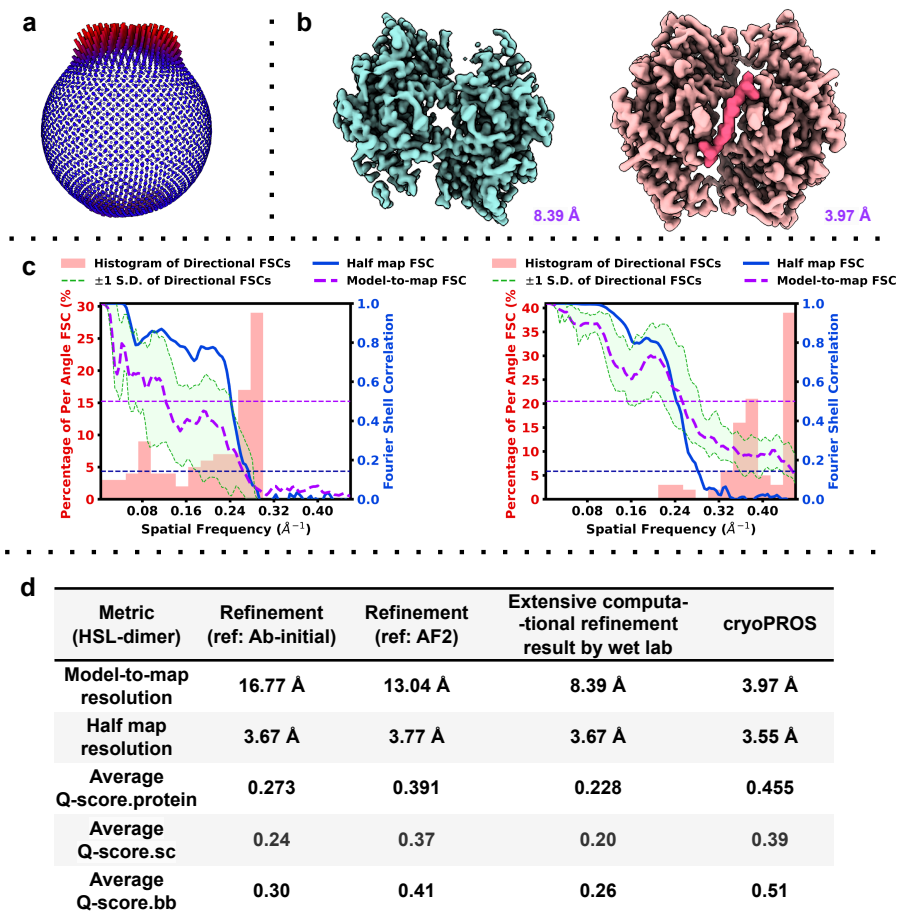
Supplementary Figure 5: Impact of simulated missing cone variation on cryoPROS performance. **a**, Orientation distributions varying the range of the missing cone along the preferred orientation axis. Visualized for PO-HA-Syn-MW10 to PO-HA-Syn-MW45, showcasing the progressive reduction in available views. **c**, Comparison of density maps determined by cryoPROS, refinement (ref: homologous protein), and refinement (ref: ab-initial) methods for the data spanning from PO-HA-Syn-MW10 to PO-HA-Syn-MW45. **b,d**, Performance metrics across all datasets, along with results of cryoPROS, conventional refinement methods using initial volumes of ab-initial and homologous protein respectively. Metrics include model-to-map resolution, Average Q-score-bb, and Average Q-score-sc.



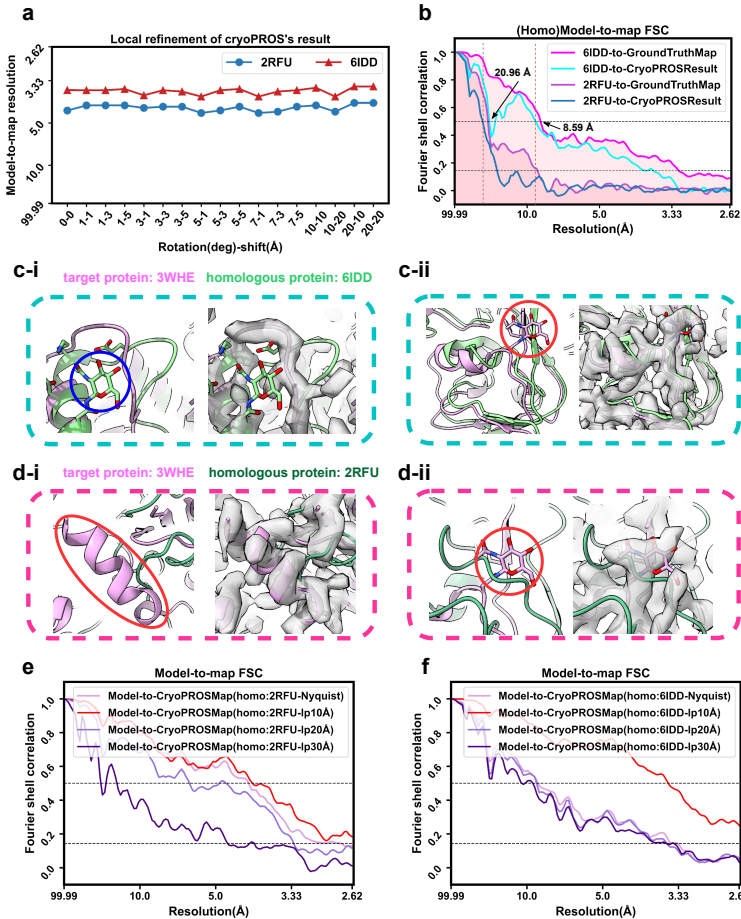
Supplementary Figure 6: Comparison of density plots and 3D FSC results for 40° tilted and untilted data. In panel **a**, the top row displays density plots calculated by autorefinement for 40° tilted data, along with their corresponding map-to-model 3D FSCs. The bottom row presents results obtained by autorefinement with per-particle refinement for the same tilted data. Panel **b** shows autorefinement results for untilted data in the top row, and cryoPROS results for untilted data in the bottom row, both accompanied by their respective 3D FSC results.



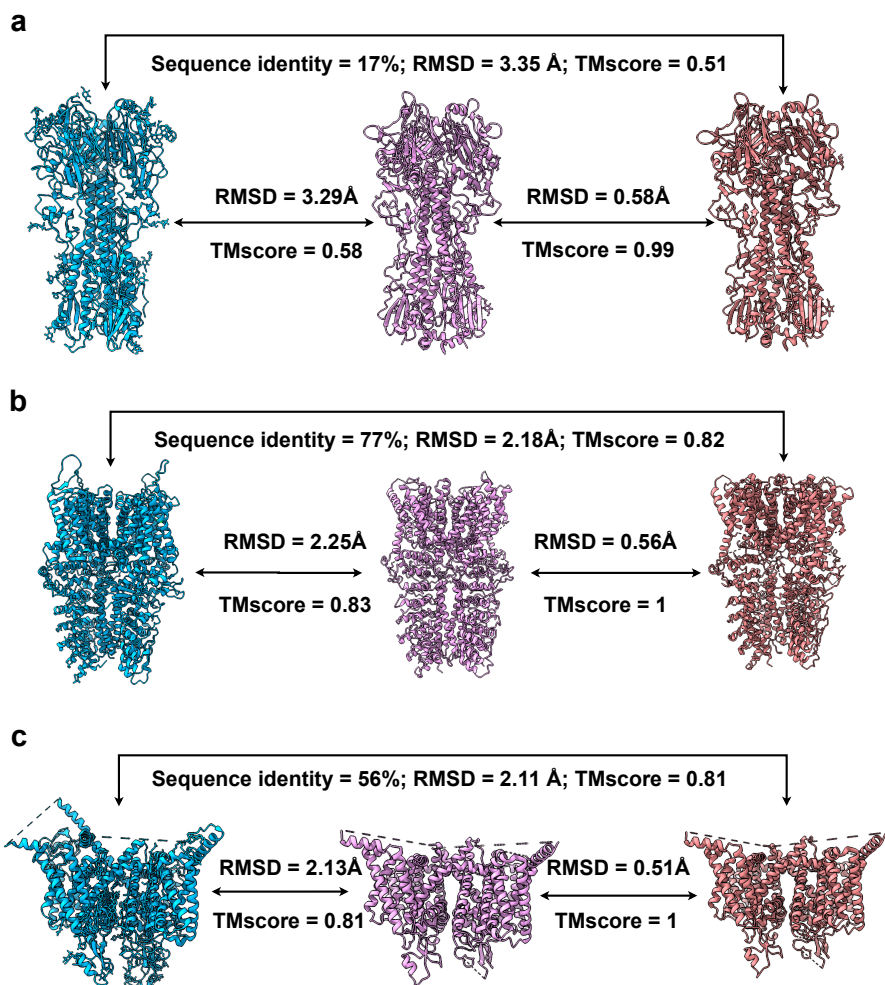
Supplementary Figure 7: Enhanced Na_X structure using cryoPROS and ReconDisMic. **a**, Pose distribution of Na_X after mitigating orientation estimation bias with cryoPROS. **b**, Reconstruction result of micelles obtained using the ReconDisMic algorithm. **c**, Reconstruction density maps from conventional refinement (cyan) and cryoPROS (pink), accompanied by map-to-model 3D FSCs on the right. Light coral depicts the histogram of directional FSCs, green surrounds the ± 1 standard deviation of directional FSCs, the blue curve represents the half-maps FSC, and the purple dashed curve signifies the model-to-map FSC. **d**, Quantitative comparison of density maps in **c** using multiple metrics: model-to-map resolution, half-maps resolution, and average Q-scores for protein, backbone (bb), and side chains (sc).



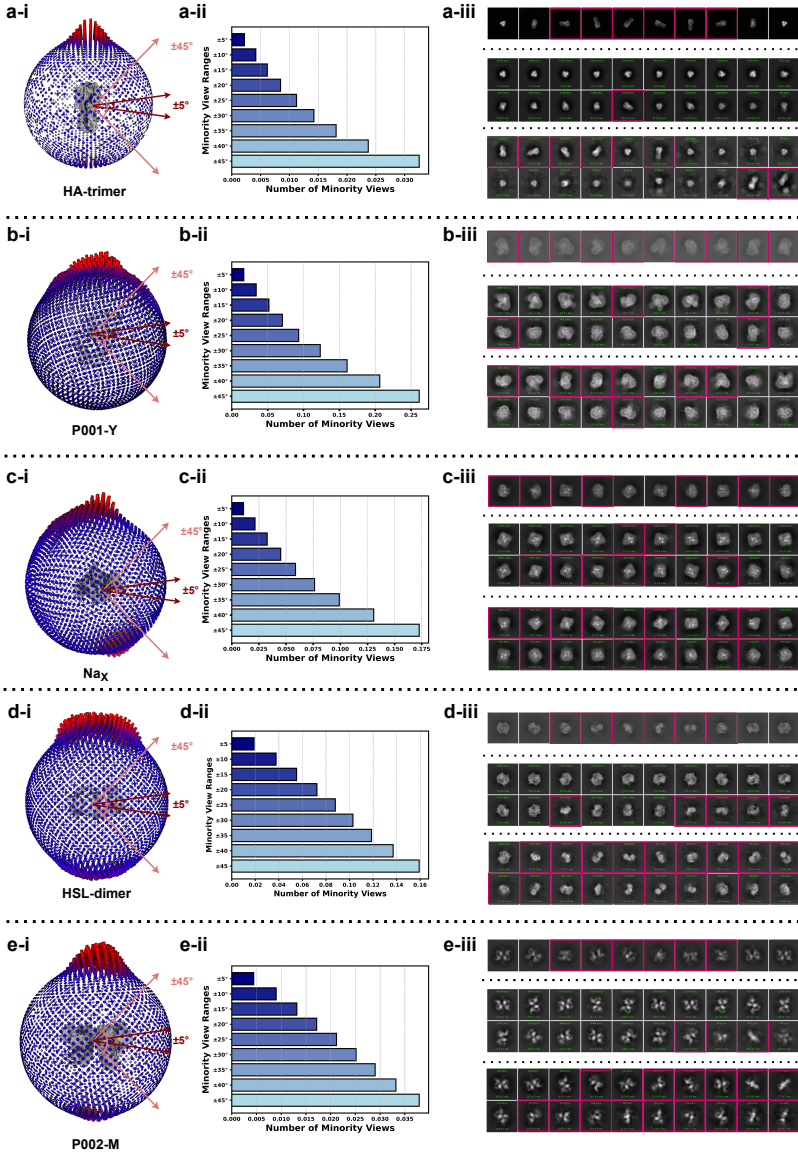
Supplementary Figure 8: Enhanced HSL-dimer structure using cryoPROS and ReconDisMic. **a**, Pose distribution of HSL-dimer after mitigating orientation estimation bias with cryoPROS. **b-c**, Following the same convention as in Supplementary Fig. 7c, except the density map obtained via cryoPROS is shown in magenta, and the newly observed density map, indicative of a new conformation, is highlighted in red. **d**, Following the same convention as in Supplementary Fig. 7d.



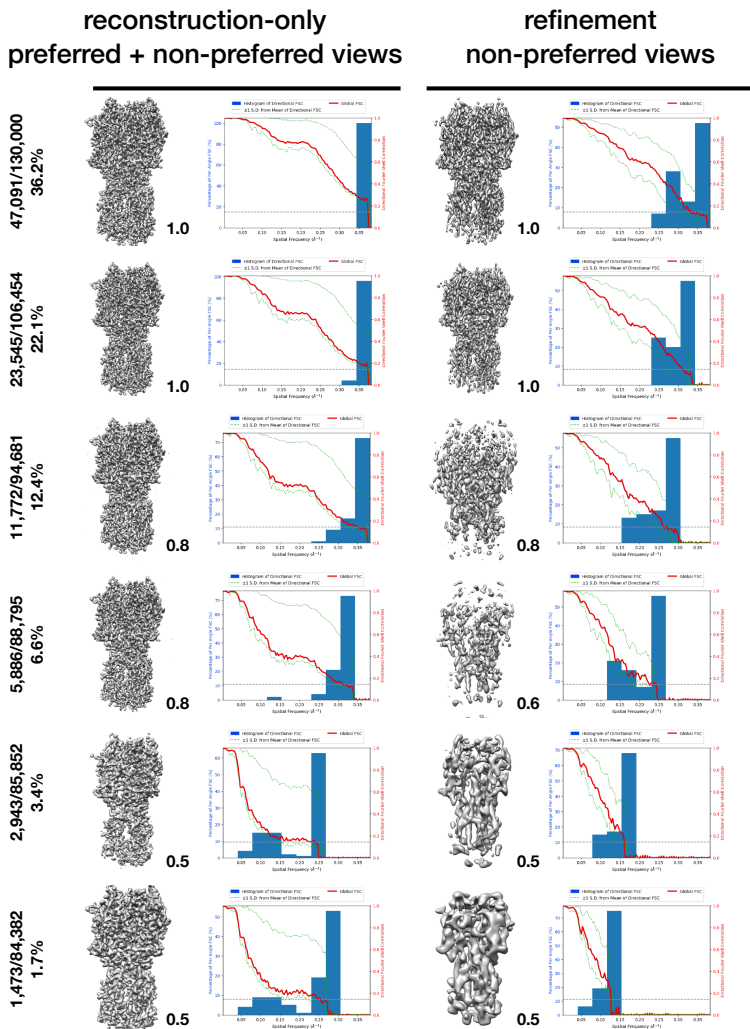
Supplementary Figure 9: Extended experimental results of model bias risk analysis of cryoPROS. **a**, Follow-up local refinement experiment on cryoPROS results initiated by two homologous proteins (6IDD and 2RFU). Variations in local search parameters (rotation and shift search range) show minimal changes in refinement outcomes. **b**, (Homologous Protein) Model-to-maps FSC curves. **c**, Two groups of detailed display of cryoPROS resulting map. The left side depicts a comparison between the homologous protein model and the target protein model. Structural segments specific to the homologous protein are marked by blue circles, while those unique to the target protein are marked by red circles. The right side displays both models embedded in the cryoPROS resulting density map, with the initial volume generated using the homologous protein 6IDD. **d**, Similar results to panel **c**, where cryoPROS employs the homologous protein 2RFU to generate the initial volume. **e**, (Target)Model-to-map FSC curves of the resulting map obtained by cryoPROS where the initial volume are lowpass filtered homologous protein (2RFU) map by various cutoff frequencies: Nyquist, 10Å, 20Å, and 30Å. **f**, Similar results to panel **e**, where cryoPROS employs the homologous protein 6IDD to generate the initial volume.



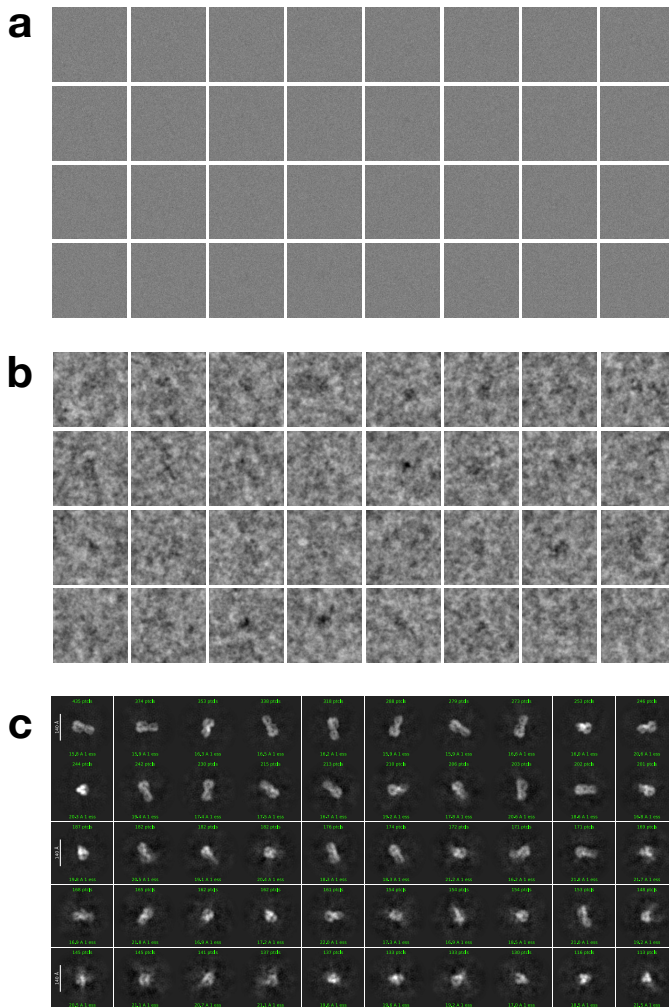
Supplementary Figure 10: Structural refinement results for three datasets: PO-HA-Syn, PO-TRAP1, and Na_X. Each panel (a, b, c) displays three structures: the target protein atomic model (center), cryoPROS structure refinement results (right), and the homologous protein atomic model used in cryoPROS initialize (left). Distances and similarities between the models are assessed using RMSD and TMscore metrics. The models employed in each dataset align with the corresponding experimental setup discussed in the main text.



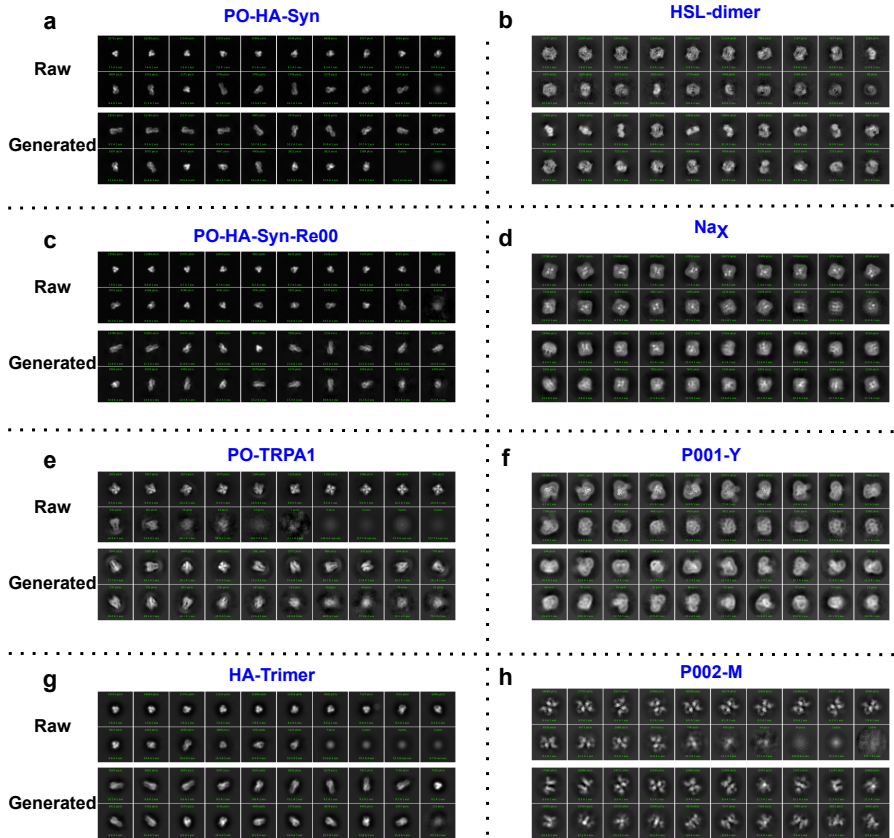
Supplementary Figure 11: Detection of particles from non-preferred views in experimental datasets with preferred orientations after repositioning with CryoPROS. a-e, Each row corresponds to one of five experimental datasets showing preferred orientations: HA-trimer, P001-Y, NaX, HSL-dimer, and P002-M. i, Pose distributions obtained via cryoRPOS are visualized as spherical plots. ii, Histograms indicate the proportion of particles from non-preferred views in specific angular ranges near the equator ($\pm 10^\circ$ to $\pm 45^\circ$). iii, 2D classification comparisons are presented in three sections: the top row shows random angle projections of the ground truth density map, the middle row shows 2D classifications of all particles, and the bottom row shows 2D classifications of particles from the non-preferred view subset (selected from cryoPROS-estimated poses within the $\pm 45^\circ$ range). Particles from non-preferred views are highlighted with red boxes.



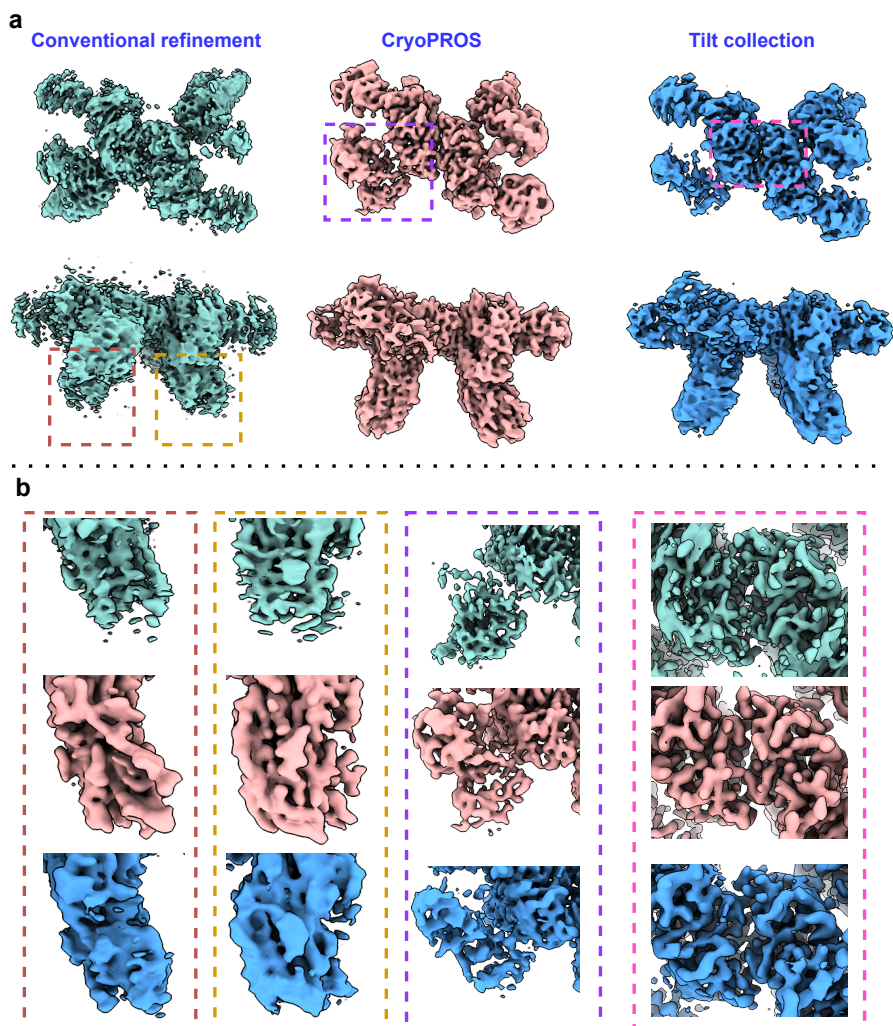
Supplementary Figure 12: A limited number of non-preferred views can sufficiently complement preferred views to generate a high-quality density map. Successive halvings of non-preferred views were conducted on the PO-HA-Syn dataset, reducing the proportion of non-preferred views from 36.2% to 1.7%. The number of particles from non-preferred views, along with the total number of particles, and the ratio of non-preferred views are labeled on the leftmost side. Reconstruction-only with ground-truth poses from particles in preferred views, along with particles from non-preferred views after several rounds of halving, is compared in parallel with refinement results using the remaining non-preferred views. The contour level is indicated in the bottom-right corner of the density map, while the model-to-map 3D-FSC curve is depicted on the right side of the density map.



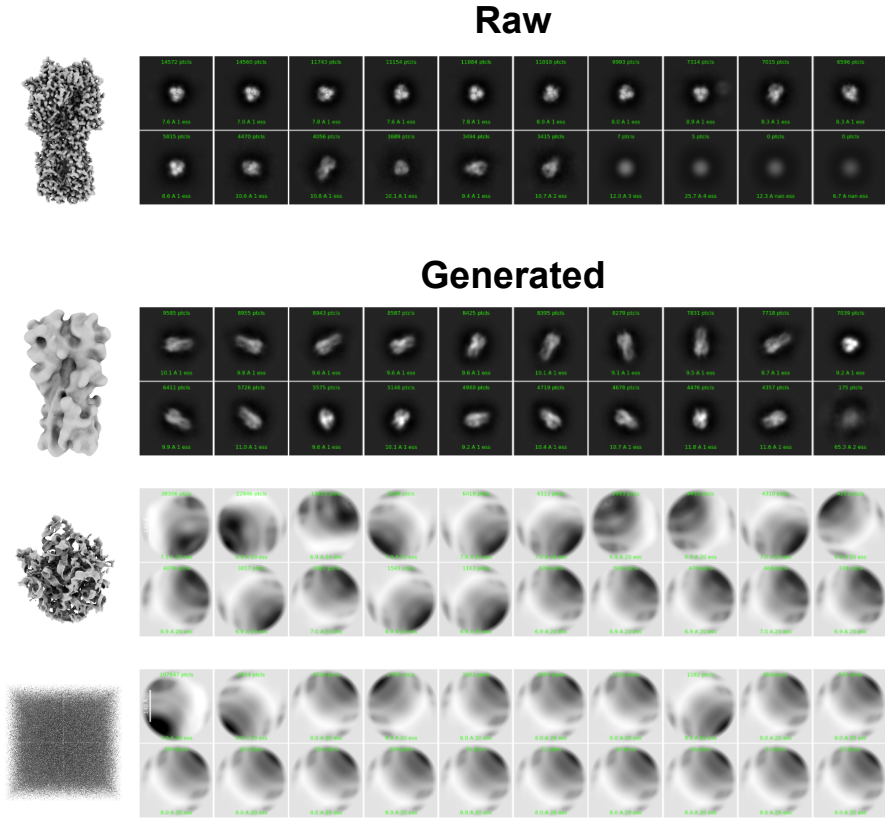
Supplementary Figure 13: Auxiliary particles produced by generative module of cryoPROS. The generative module was trained using the 10Å low-pass filtered map of the ground-truth atomic model (PDB ID: 6WXB) and the PO-HA-Syn dataset. After training, it generated 10,000 auxiliary particles. Raw auxiliary particles are depicted in **a**, while low-pass filtered auxiliary particles are shown in **b**. 2D classification of these 10,000 particles into 50 classes resulted in the findings shown in **c**.



Supplementary Figure 14: Comparative analysis of raw and generated particles. Averaged 2D classification comparisons showing the similar signal but distinct pose distributions between raw particle data (upper rows) and generated particles (lower rows) across datasets exhibiting preferred orientations. Panels **a** to **h** correspond to the following datasets: PO-HA-Syn, PO-HA-Syn-Re00, HSL-dimer, Na_X, PO-TRPA1, P001-Y, HA-trimer, and P002-M.



Supplementary Figure 15: Comparison of CryoPROS refinement with conventional and tilt collection strategies on dataset P002-M. **a**, Density map comparison between CryoPROS refinement (middle, in pink), conventional refinement (left, in green), and the tilt collection strategy (right, in blue). The results highlight the improved density map quality achieved by CryoPROS refinement compared to both conventional methods and the tilt collection strategy. **b**, Enlarged views of specific regions (highlighted with colored boxes) for detailed comparison across the three methods, providing an in-depth look at domain-level differences.



Supplementary Figure 16: Comparison of raw particles with CryoPROS-generated particles using varied initial volumes. The first row displays the target protein analyzed using HA-trimer data (left) and the original raw particles (right). The subsequent rows present the initial reference models used for training CryoPROS: a homologous protein 2RFU (low-pass filtered to 10Å), an unrelated ribosome structure (low-pass filtered to 10Å), and Gaussian noise, from top to bottom. On the right of each row are the corresponding CryoPROS-generated particles derived from these initial volumes.

Supplementary Table 1: Quantitative comparison on synthetic HA trimer dataset.

Metric (Syn HA-trimer)	Uni-HA-Syn- Reconstruction	PO-HA-Syn- Reconstruction	PO-HA-Syn- Refinement	PO-HA-Syn- cryoPROS
Model-to-map resolution	2.62Å	2.62Å	23.95Å	2.62Å
Half-maps resolution	2.62Å	2.62Å	2.62Å	2.66Å
Average Q-score.protein	0.794	0.783	0.094	0.72
Average Q-score.sc	0.80	0.79	0.10	0.72
Average Q-score.bb	0.78	0.78	0.10	0.72

Supplementary Table 2: Quantitative comparison on PO-TRPA1.

Metric (PO-subset of EM-10024)	PO- Reconstruction	PO- Refinement (ref: Ab-initial)	PO- Refinement (ref: AF2)	cryoPROS
Model-to-map resolution	7.44Å	91.17Å	11.40Å	8.10Å
Half-maps resolution	6.63Å	5.14Å	5.01Å	6.29Å
Average Q-score.protein	0.207	0.018	0.134	0.183
Average Q-score.sc	0.09	0.00	0.08	0.08
Average Q-score.bb	0.33	0.04	0.18	0.29

Supplementary Table 3: Quantitative comparison on HA trimer dataset. The non-tilted data is from EM-10096, and the 45° tilted data is from EM-10097.

Metric (HA trimer)	45° tilted data		non-tilted data		
	Refinement (ref: Ab-initial)	Per-particle refinement	Refinement (ref: Ab-initial)	Refinement (ref: Homo protein)	cryoPROS
Model-to-map resolution	4.72Å	3.41Å	37.26Å	22.36Å	3.49Å
Half-maps resolution	4.24Å	3.10Å	3.07Å	3.13Å	3.32Å
Average Q-score.protein	0.344	0.576	0.01	0.155	0.509
Average Q-score.sc	0.23	0.52	0.01	0.12	0.43
Average Q-score.bb	0.44	0.63	0.01	0.18	0.58

Supplementary Table 4: Comparison of model-to-map resolutions (Å) between density maps reconstructed using Gaussian noise substitution method and cryoPROS.

Method	HA trimer		Na _X	
	Target model (3WHE)	Homolog model (6IDD)	Target model (Na _X)	Homolog model (Nav1.6: 8FHD)
Gaussian noise substitution	4.14Å	9.58Å	40.23Å	31.28Å
cryoPROS	3.49Å	20.96Å	4.32Å	8.80Å

Supplementary Table 5: The performance of cryoPROS using different initial homologous proteins.

PDB ID (sequence similarity)	Half-maps resolution	Model-to-map resolution	Average Q-score
6IDD (47%)	3.32Å	3.49Å	0.509
2RFU (17%)	3.35Å	4.04Å	0.438