CryoGEN-II: Cryogenic Electron Tomography Reconstruction via Generative Network

Supplementary Material

6. Experiment Setup

6.1. Model Architecture

The proposed framework adopts a generative adversarial architecture built upon a customized 3D UNet++ backbone [31]. Specifically, both the encoder and decoder employs a four-level nested U-Net structure comprising residual convolutional blocks, each consisting of two sequential $3 \times 3 \times 3$ convolutions with InstanceNorm3d and PReLU activation, complemented by a learnable downsampling skip path (via a $3 \times 3 \times 3$ convolution) when input and output channel counts differ. Dense feature fusion is achieved through hierarchical upsampling (trilinear interpolation) and channel-wise concatenation across all coarser scales at each resolution level, culminating in a $1 \times 1 \times 1$ output convolution.

The discriminator reuses the same architecture as its feature extractor, but interprets its output as a realism score. Both modules perform input normalization: the generator applies per-sample percentile-based min–max scaling (defaulting to 4th–96th percentiles) before forward propagation and utilizes a wedge-aware masked reconstruction loss, inspired by [28].

6.2. Training Hyperparameters

To ensure a fair comparison with CryoGEN, CryoGEN-II adopts the same training hyperparameters as CryoGEN, with one deliberate modification to the batch size to promote stable and effective updates of the barycenter. Specifically, CryoGEN-II is optimized using the Adam optimizer. While CryoGEN employs a batch size of one for simulated shapes and real-world examples, we increase the batch size to 8 for simulated shapes and 10 for real-world examples—retaining a batch size of one only for protein subtomograms—to ensure sufficient gradient information for reliable barycenter estimation. The learning rate is fixed at 0.0004, with a linear warm-up phase applied over the first 10% of training steps, followed by a linear decay schedule for the remainder of training. In contrast to IsoNet, which progressively ramps up the noise scale during training, CryoGEN-II samples a noise level uniformly from the interval $(0, 1]$ at each step and scales it by a fixed maximum noise intensity. Additionally, the penalty coefficient λ is held constant during the first epoch and subsequently decayed linearly across the remaining epochs.

6.3. Computational Overhead

All experiments are conducted on a system equipped with two NVIDIA A100 GPUs, each with 40 GB of memory. Training on real-world examples is run for 20 epochs, converging in approximately one hour, whereas training on simulated shapes requires 10 epochs and completes in roughly 30 minutes under identical hardware conditions.

6.4. Data Processing

Inspired by [18], cropped subtomograms are treated as cube-shaped volumes with six faces, yielding 24 possible orientations for reorientation. However, we exclude the four rotations that preserve the original missing wedge along the X-Z direction, as these do not introduce new structural information. All 3D visualizations were rendered using UCSF ChimeraX.

7. Cryo-ET Dataset Details

7.1. Purified Ribosome

The ribosome tomograms are downloaded from the Electron Microscopy Pilot Image Archive (EMPIAR-10045). Following the same procedure as described by [18], the original volumes are binned six times, resulting in a final pixel size of 13.66 Å. CTF deconvolution is applied using IsoNet's Wiener filter. To exclude empty regions, a combined density and standard deviation mask is generated using IsoNet's default mask parameters. Subtomograms are then randomly extracted only from masked regions that contain biological signal. A total of 490 subtomograms of size $80 \times 80 \times 80$ voxels are extracted across the seven tomograms. Prior to downstream use, each subtomogram undergoes contrast inversion and percentile-based normalization such that 90% of voxel intensities are mapped to the range of $[0, 1]$ to avoid extreme values.

7.2. HIV Capsid

Raw tilt series for the immature HIV-1 capsid are downloaded from EMPIAR-10164. Following the same procedure outlined in [18] The movie stacks are drift-corrected using MotionCorr and reconstructed into tomograms via the WBP algorithm within IMOD, with defocus values determined by CTFFIND4. The resulting tomograms are binned eightfold to achieve a pixel

size of 10.8 Å. CTF deconvolution is then applied using IsoNet's Wiener-like filter, with the spectral signal-to-noise ratio fall-off parameter set to 0.7 and the deconvolution strength set to 1.0. A combined density and standard deviation mask is generated for each tomogram using IsoNet's default parameters. From the three processed tomograms (TS_01, TS_43, and TS_45), a total of 300 subtomograms of size $96 \times 96 \times 96$ voxels are randomly extracted within masked regions.