3D Conditional Diffusion Models for Synthetic Cryo-ET Particle Subtomograms

Gal Passi, Tomer Cohen

GitHub

Introduction

Over the past decade, cryo-electron tomography (cryoET) has emerged as a central technique in in-situ structural biology, allowing cells to be visualised as three-dimensional (3D) volumes at molecular resolution[1,2] . However CryoET tomograms are affected by strong reconstruction artefacts and variable resolution across the tomogram resulting from limited tilt range (otherwise known as missing wedge), and overlapping structures in densely populated regions[3,4]. Together these make automatic identification of protein or protein-complexes a challenging 3D detection problem [5,6].

Limited training data is another major bottleneck with segmentation pipelines still requiring substantial amounts of carefully curated labels to reach high performance [7,8]. With manual annotation of tomograms being time-consuming and variable [9] automated approaches to augment and generate training data are essential. Recent works generate synthetic tomograms from atomic models or density maps [10], providing large training sets that support benchmarks like the CZII CryoET Object Identification dataset. However, even these simulations cannot fully capture experimental noise, imaging artefacts and specimen heterogeneity , leaving a domain gap that limits the generalisation of models trained purely on synthetic data[10,11].

Learning-based generative models have therefore been explored as a means to bridge the gap between idealised simulations and experimental subtomograms. CryoETGAN introduced a Wasserstein GAN architecture that translates simulated density maps into more realistic cryoET appearances that better match the texture and contrast of experimental tomograms [12]. More recently, denoising diffusion probabilistic models (DDPMs) achieved state-of-the-art image synthesis quality. Eschweiler *et al.* [13] use a DDPM to synthesize annotated 2D and 3D electron microscopy (EM) images from masked inputs, enabling segmentation models trained purely on synthetic data. Lu *et al.* introduce **EMDiffuse**[14], a diffusion model for EM denoising and super-resolution that outperforms regression and CycleGAN baselines while better preserving ultrastructure. However these have not been implemented on CroET data.

Existing cryoET data synthesis methods typically operate either in 2D, at the level of tilt-series or projected slices, or as unconditioned generators that cannot guarantee the presence of a specific particle type at a desired location. We propose an end-to-end approach, training diffusion models on three-dimensional, cryoET tomograms. Furthermore, our architecture

supports conditional generation, able to produce specific, protein or protein-complexes at a desired location. To this end, large-scale benchmarks such as the CZII CryoET Object Identification competition have made curated tomograms and particle annotations publicly available, creating an opportunity to directly learn the distribution of 3D particle-centred subtomograms from real data.

In this work, we investigate whether 3D conditional DDPM, implemented as a U-Net–based architecture following Ho *et al.* (2020)[15], can be used to generate realistic, labelled cryo-ET particle volumes. We focus on the CZII/Kaggle CryoET Object Identification dataset of denoised tomograms[16], and train a diffusion model that is conditional on key particle class: apo-ferritin, β-amylase, β-galactosidase, 80S ribosome, thyroglobulin, and virus-like particles. Our overarching goal is to characterise the visual realism and diversity of these synthetic volumes which can be used in the future to augment training data for CryoET object detection tasks.

Methods

Dataset Curation

We trained and evaluated our model on the CZII CryoET Object Identification dataset[15], which comprises denoised cryo-electron tomograms and curated particle annotations across six macromolecular complexes (**Table 1**). We restricted our experiments to the experimental training subset of the CZII CryoET Object Identification phantom dataset, which comprises 7 annotated tomograms with ground-truth particle centres (N=1,269 particle centers).

We partition the annotated particle-centers across all tomograms into a train and validation sets corresponding to an approximate 90/10 split, ensuring no overlapping samples. For each annotated particle, we extracted a 3D cubic subtomogram of size $64 \times 64 \times 64$ from the (rounded) particle-center value. If a cube would extend beyond the tomogram boundaries, we extended it using zero-padding. Each particle center is paired with the particle class using a one-hot encoding vector. Intensity normalization was applied per subtomogram to account for inter and intra scan variations, using z-score normalization:

$$(1)\ \bar{x} = \frac{x - \mu_{patch}}{\sigma_{patch} + \epsilon}$$

Where μ_{patch}, σ_{patch} are the mean and standard-deviation over all voxels in the subtomogram and $\epsilon = 10^{-8}$ added for numerical stability.

No geometric data augmentation (rotations, flips or elastic deformations) was applied. The model is trained directly on denoised subtomograms.

Particle name	Approximate Dimensions (pixels)	Number of instances in dataset	Difficulty
apo-ferritin	(12,12,12)	375	Easy
beta-amylase	(13,13,13)	87	Impossible
beta-galactosidase	(18,18,18)	112	Hard
ribosome	(30,30,30)	331	Easy
thyroglobulin	(26,26,26)	251	Hard
virus-like-particle	(27,27,27)	113	Easy

Table 1. Dataset Summary. Difficulty is determined by the CZII challenge.

Conditional Diffusion Model Architecture

We implemented a 3D conditional Denoising Diffusion Probabilistic Model (DDPM) built on a U-Net backbone, following the original DDPM formulation of Ho *et al.* (2020)[15]. The model takes as input: a noisy subtomogram $x_t \in R^{1\times64\times64\times64}$ and three conditioning variables: diffusion time step t, particle class and normalised 3D coordinates and predicts the corresponding clean subtomogram x_0.

3D U-Net backbone: consists of a 3-level 3D encoder–decoder U-Net that operates on a single-channel patch $x_t \in R^{1\times64\times64\times64}$. The first conditional residual block maps the input to 32-feature channels at full resolution, followed by a 1-strided 3D convolution with a $3 \times 3 \times 3$ kernel and 2-stride convolution for down-sampelling. Each subsequent encoder applies the same procedure, doubling the number of feature-channels while downsampling spatial resolution. The encoder outputs a 128-channel $8 \times 8 \times 8$ matrix. Encoder blocks are preceded by group normalization and a SiLU nonlinearity[17] , with a residual skip path ($1 \times 1 \times 1$ convolution) . The diffusion time step, particle class and normalised 3D coordinates are first embedded into a shared conditioning vector, which is linearly projected and added as a channel-wise bias to the intermediate activations inside each residual block [18].
Skip connections are used between each encoder block and its corresponding decoder block. The encoder blocks are followed by a bottleneck $3 \times 3 \times 3$ convolution.

The decoder architecture mirrors the encoder applying three upsampling steps. Each decoder block consists of a $2 \times 2 \times 2$ convolution using 2-stride upsampling the spatial dimension back to $64 \times 64 \times 64$. channels are concatented to the corresponding encoder block via the skip

connection and to the conditional bias derived from the time, class and coordinate embeddings. Channel downsampling using 3D residual convolution block, preceded by group normalization and SiLU activations is then applied. In the last block the number of channels is reduced from 32 to 1, yielding the predicted clean subtomogram $x_0 \in R^{1\times64\times64\times64}$.

Conditional encoding: Conditioning is encoded as a single vector c built from the concatenation of a sinusoidal time embedding[19], a linear projection of the one-hot class label, and an MLP over the normalised 3D coordinates. The conditioning vector is added as a per-channel bias at every encoder and decoder level.

Diffusion Process and Noise Schedule

We follow the DDPM process described by Ho et al. (2020).

Forward diffusion: For each uniformly sampled timestep $t \in Uniform(0, T-1)$, a noise volume x_t is sampled using the standard DDPM forward process with fixed variance schedule:

$$(2)\ x_t = \sqrt{\alpha_t x_0} + \sqrt{1-\bar{\alpha}_t} \bullet \epsilon,\ \epsilon \in N(0,1)$$

We use a linear diffusion schedule with $T = 1,000$ steps over $\beta_1 = 10^{-4}$ to $\beta_T = 2 \times 10^{-2}$ where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

Backward diffusion (sampling): Given a target class and 3D coordinates the conditional DDPM generator initiates a uniform Gaussian noise subtomogram $x_T = N(0,1) \in R^{1\times64\times64\times64}$. We iteratively denoise the noisy image using the standard closed-form DDPM posterior over $t = T, T-1, ..., 1$, this is:

$$(3)\ p(x_{t-1}|x_t) = N(\mu_t(x_t, \widehat{x_0}),\ \sigma_t^2 I)$$

Training Procedure

Loss function: We train the model by minimizing the mean squared error (MSE) between the predicted $\widehat{x_0}$ and ground-truth (clean) patch x_0 over randomly sampled timestamps and gaussian noises which is the standard DDPM training objective :

$$(4)\ MSE = E_{x_0,t,\epsilon}[||\widehat{x_0}(x_t, t, class) - x_0||^2]$$

Hyperparameters: the model is trained using the Adam optimizer[20], with a constant learning rate of $1\times\ 10^{-4}$ and otherwise default parameters. We use a batch size of 16 and train for 100 epochs, training samples are shuffled between epochs.

Data and code availability: all code used for training and data curation was uploaded to Cryo-ET-Diff github repository. The training set used is publicly available through the CZII Kaggle dataset. We uploaded the trained model weights to zenedo.org (17726084).

Results

Training and Validation Loss

The training process demonstrates stable convergence across 100 epochs. The loss curves indicate rapid decrease in the first 10–15 epochs and gradual stabilization around an MSE of **0.38-0.42.** The validation loss closely tracks training loss, indicating **no overfitting** (**Figure 1**). Together, these trends suggest that the model successfully learns a consistent mapping from noisy to clean subtomograms across the different particle classes.

Figure 1. Train and validation loss. Loss computed using MSE (Methods) over 100 epochs.

Qualitative Evaluation of Generated Samples

For each of the six particle classes, we compare real subtomograms to five randomly generated samples from our model, conditioned on the corresponding particle class.
All figures are provided in the Figure section below.

Key qualitative observations

1. **Apo-ferritin.**
   Surprisingly, the generated structures do not show the typical ring-like or hollow-core patterns observed in experimental subtomograms. This is unexpected given the relatively large number of apo-ferritin examples in the dataset (375). We hypothesize that the model struggles with this class due to its very small physical size-only about (12, 12, 12) voxels-making its features extremely subtle.

2. **β-amylase**
   The generated volumes display elongated and streak-like densities that match the overall contrast and directional texture seen in real subtomograms. This is a particularly encouraging result given the difficulty of this particle (often considered “impossible”) and the limited number of training examples (87).

3. **β-galactosidase**
   Generated samples reproduce the multi-lobe patterns and granular contrast characteristic of the real subtomograms for this class.

4. **Ribosome**
   Complex, high-density internal patterns are generally reproduced. The generated patches capture realistic texture and density variations.

5. **Thyroglobulin**
   Generated samples contain speckled, high-contrast regions similar to those observed in the real data.

6. **Virus-like particle**
   Spherical morphologies are partially reproduced. The generated samples tend to deviate slightly from the near-perfect circularity present in some real virus-like particles.

Realism and variability

Across all particle classes, the generated volumes exhibit class-specific texture and morphology. Importantly, the samples display diversity, indicating that the model does not collapse to a single mode.Noise patterns and contrast distributions also closely resemble those found in experimental subtomograms. Overall, the model successfully learns a multimodal, class-conditional generative distribution.

Discussion

Our results show that a 3D conditional DDPM can effectively learn the distribution of cryo-ET particle-centered subtomograms directly from real experimental data.

Unlike slice-based 2D methods, our approach performs full 3D generation and incorporates explicit conditioning on particle type. We achieve high similarity between generated and real subtomograms in both texture and intensity for several particle types, including some considered highly difficult (or nearly impossible) to model due to limited data or weak structural signatures.

Despite these promising results, several limitations remain.First, some generated volumes appear slightly smoother than their real counterparts, consistent with DDPMs trained using L2 reconstruction loss. Second, the model still struggles to reproduce the perfectly round morphology of certain particles-most notably virus-like particles and apo-ferritin.

To overcome these limitations we suggest the following for future work: incorporating geometric augmentations for the subtomograms (non-trivial due to the missing wedge), scaling to larger patch sizes (e.g., 96 or 128) which will require larger GPU memory, conditioning on orientation, local neighborhood context, or multi-scale representations. In addition, we also leave for future work evaluating whether training particle detection models on a mixture of synthetic and real subtomograms improves downstream performance.

Conclusion

We implemented a fully 3D, class-conditional diffusion model for generating synthetic cryo-ET subtomograms. Using curated experimental data from the CZII CryoET Object Identification dataset, our method can synthesize realistic, diverse volumes for six particle classes.

The generated subtomograms match the texture, intensity, and morphological characteristics of real data, suggesting that diffusion-based generative modeling is a promising direction for data augmentation in 3D particle detection and bridging the domain gap between simulations and experimental tomograms.

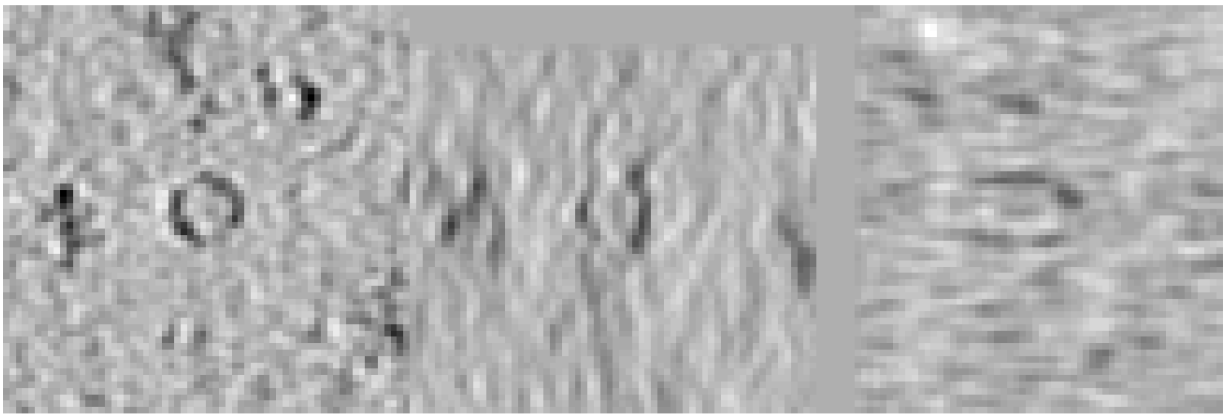
Figures:

Output example for each particle class:

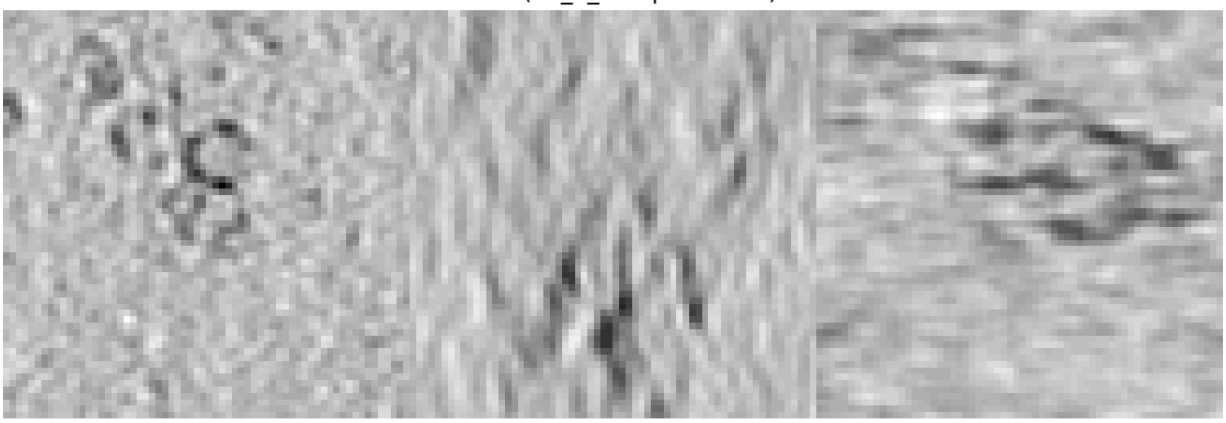
Apo-Ferritin:

Real examples from the dataset:

Clean (TS_73_6 - apo-ferritin)

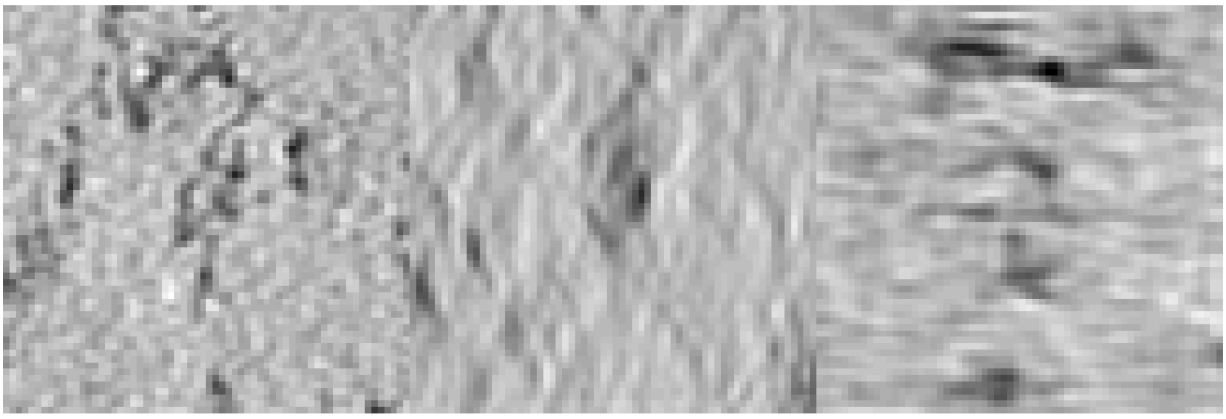


Clean (TS_5_4 - apo-ferritin)



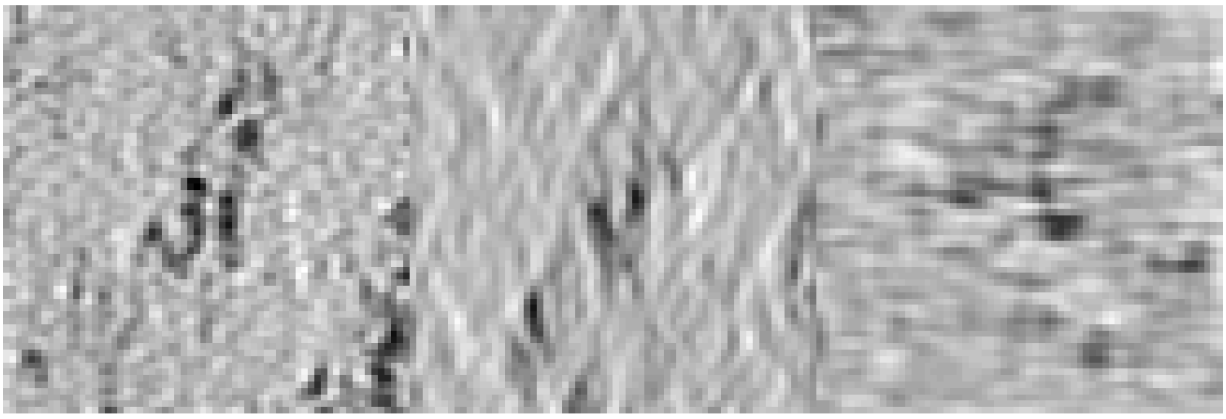
Samples from our trained model:

apo-ferritin - Sample 0



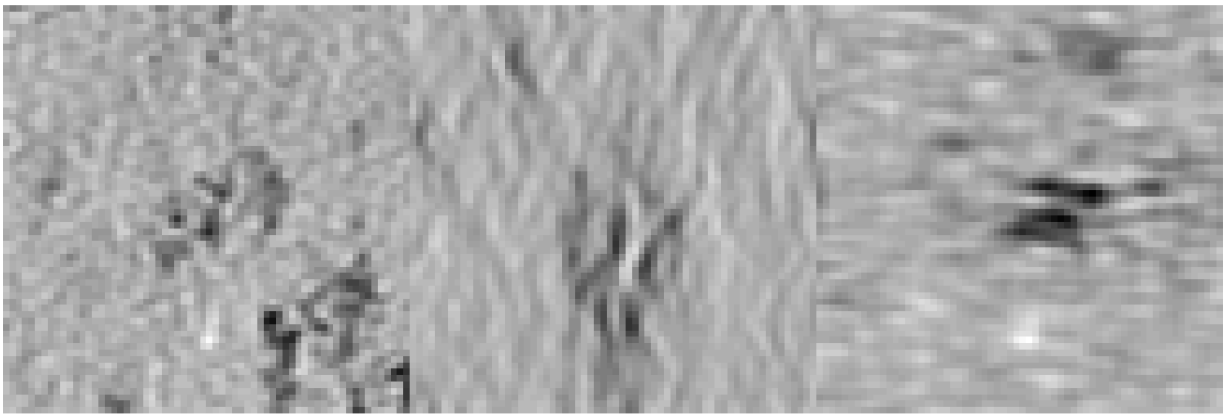
XY XZ YZ

apo-ferritin - Sample 1



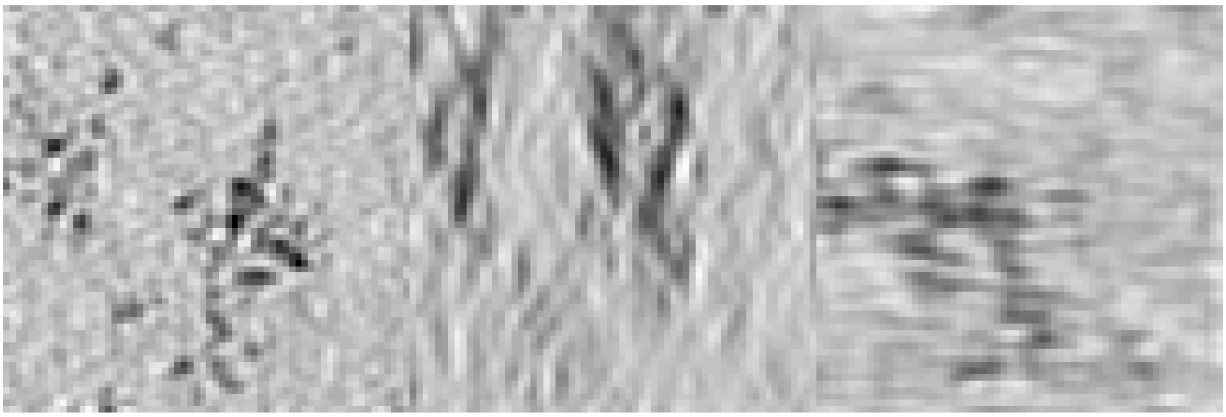
XY XZ YZ

apo-ferritin - Sample 2



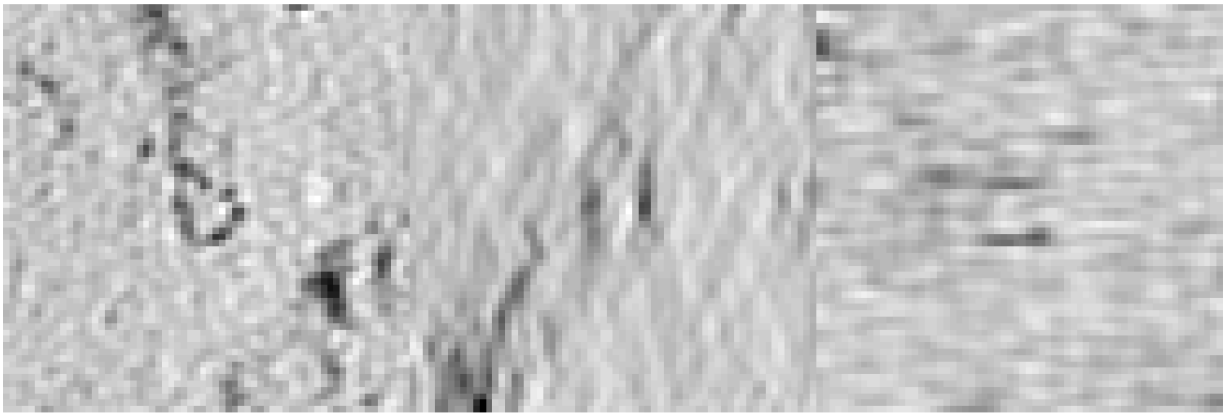
XY XZ YZ

apo-ferritin - Sample 3



XY XZ YZ

apo-ferritin - Sample 4

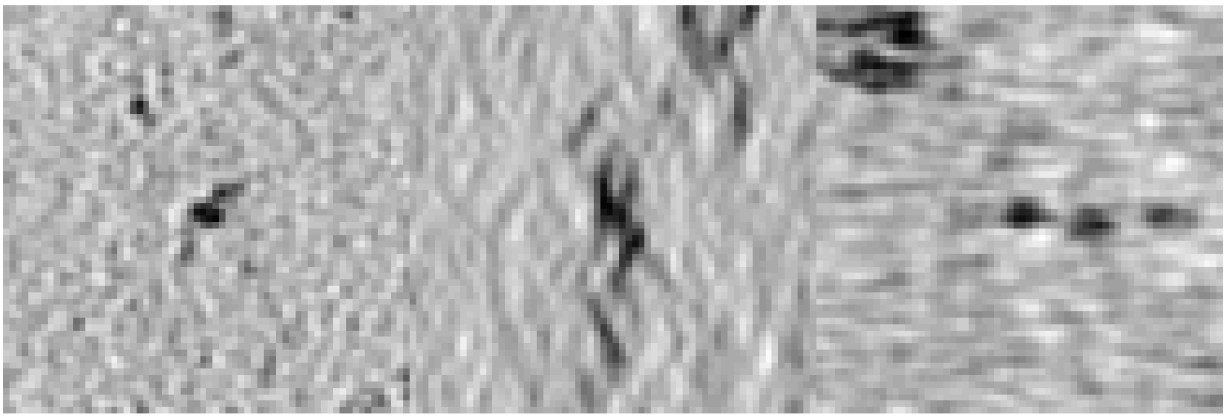


XY XZ YZ

Beta-Amylase:

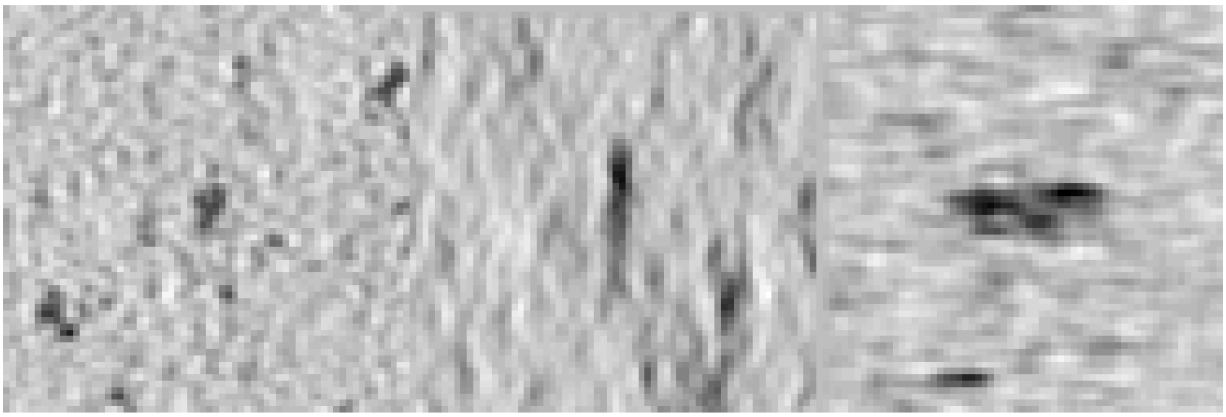
Real examples from the dataset:

Clean (TS_5_4 - beta-amylase)



XY XZ YZ

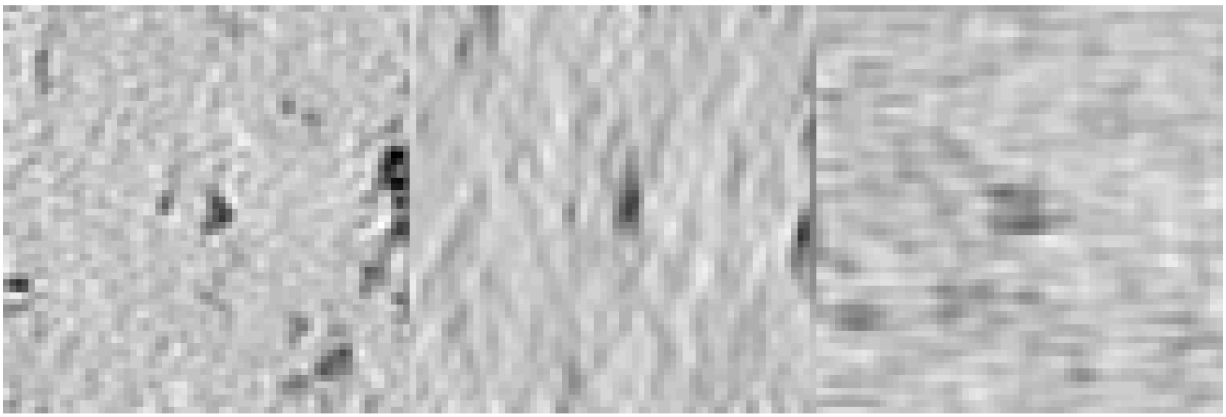
Clean (TS_6_6 - beta-amylase)



XY XZ YZ

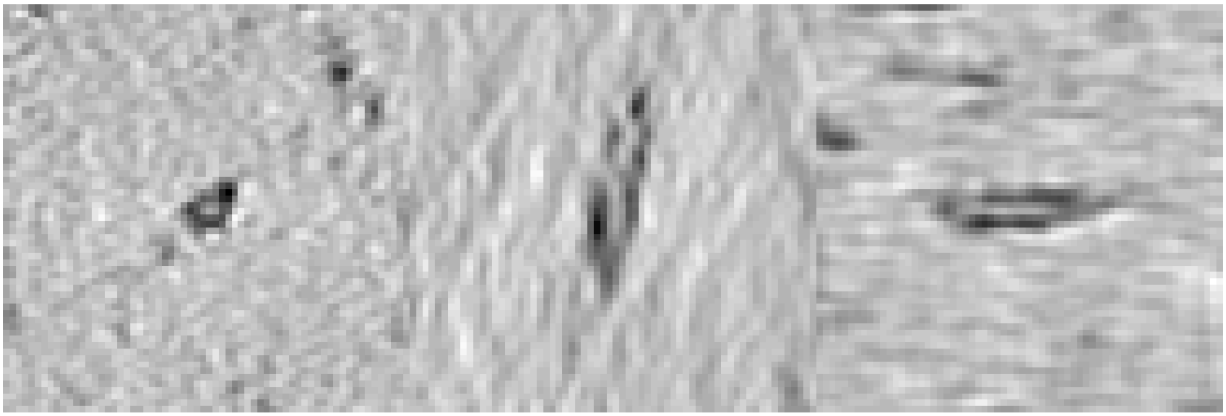
Samples from our trained model:

beta-amylase - Sample 0



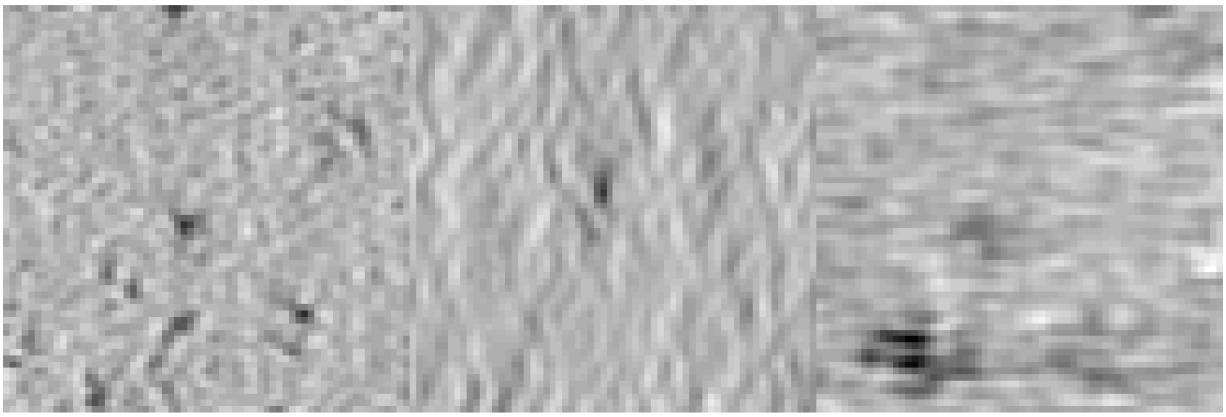
XY XZ YZ

beta-amylase - Sample 1



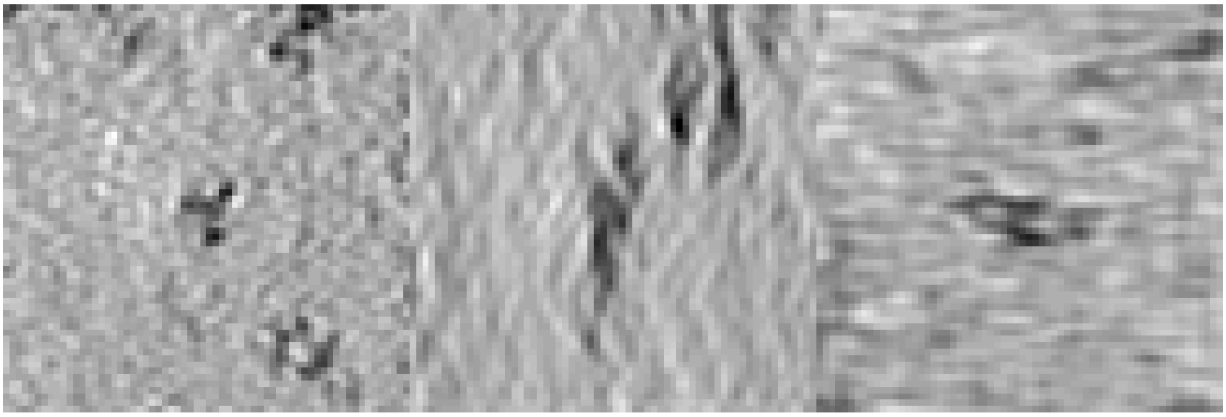
XY XZ YZ

beta-amylase - Sample 2



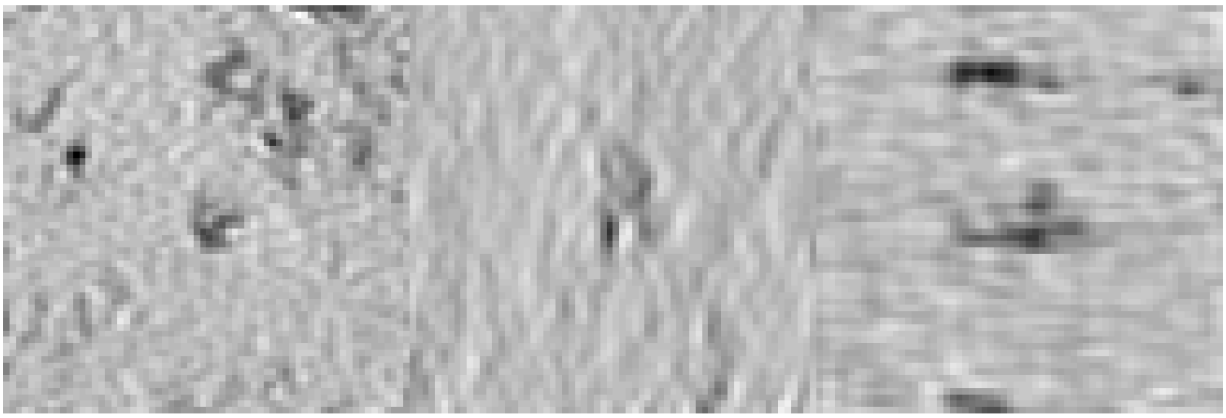
XY XZ YZ

beta-amylase - Sample 3



XY XZ YZ

beta-amylase - Sample 4

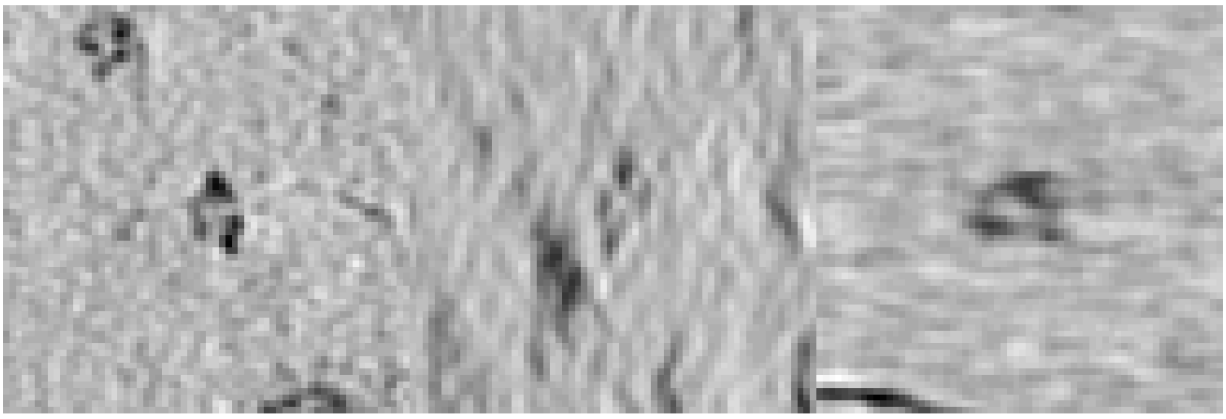


XY XZ YZ

Beta-Galactosidase:

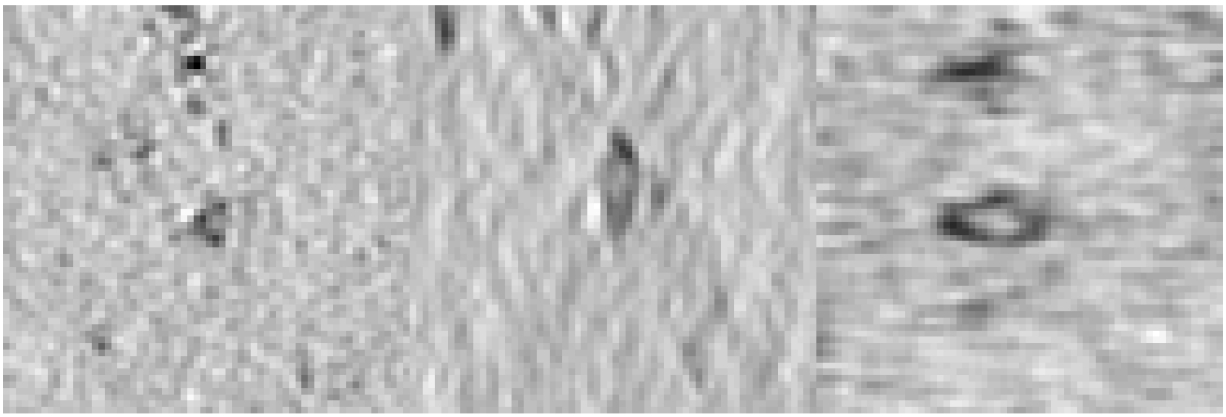
Real examples from the dataset:

Clean (TS_99_9 - beta-galactosidase)



XY XZ YZ

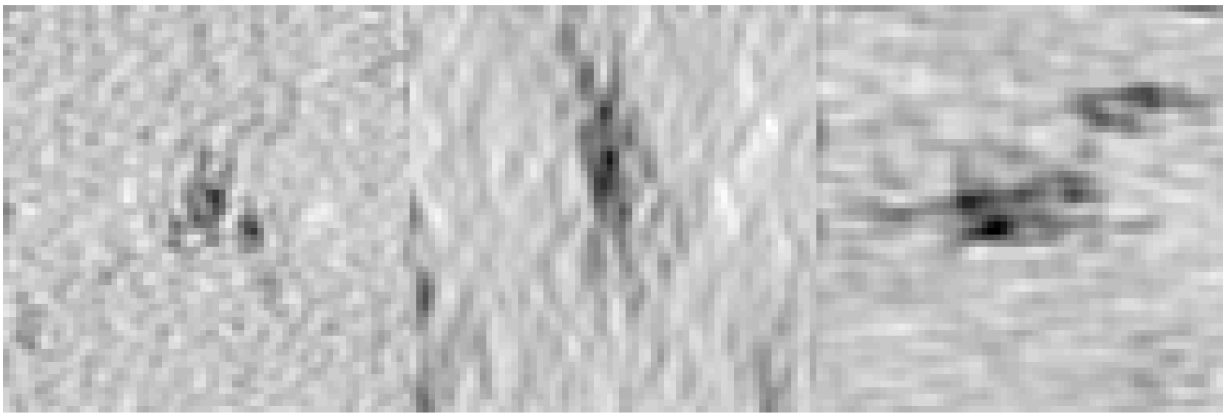
Clean (TS_73_6 - beta-galactosidase)



XY XZ YZ

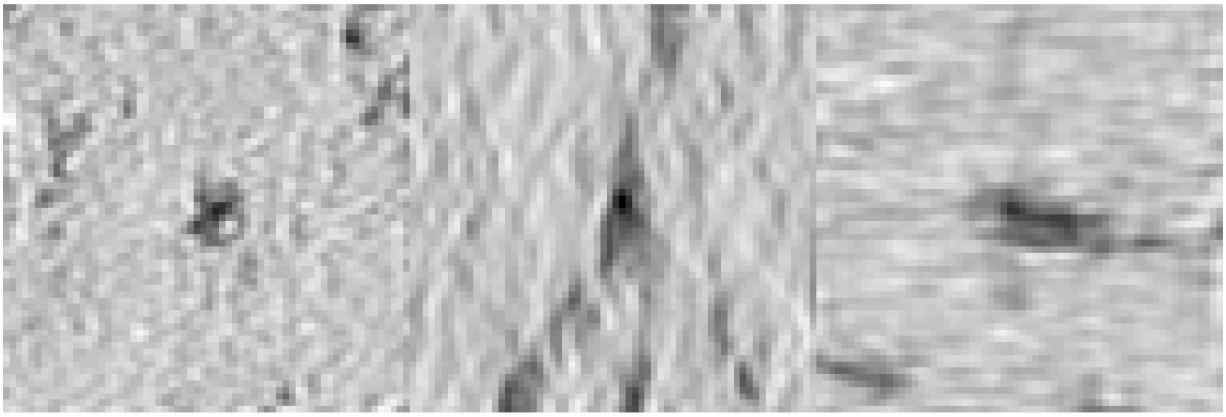
Samples from our trained model:

beta-galactosidase - Sample 0



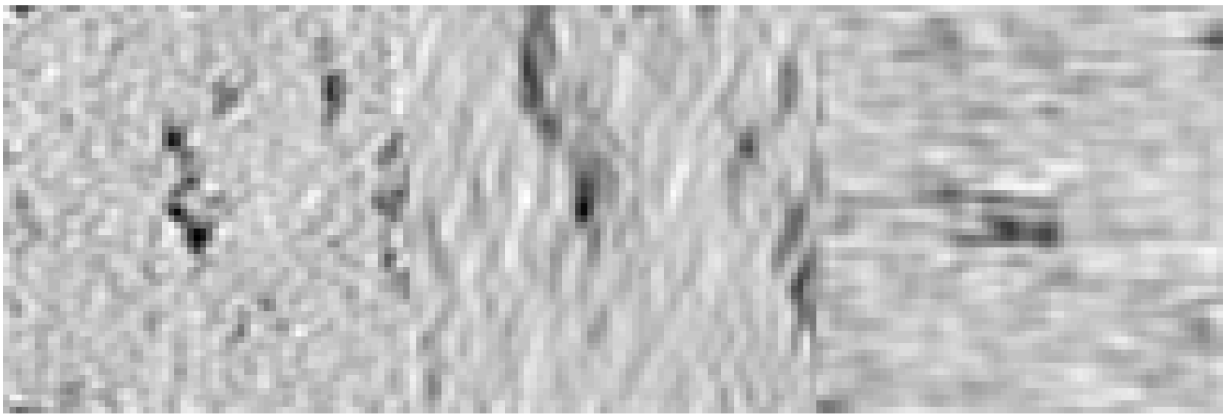
XY XZ YZ

beta-galactosidase - Sample 1



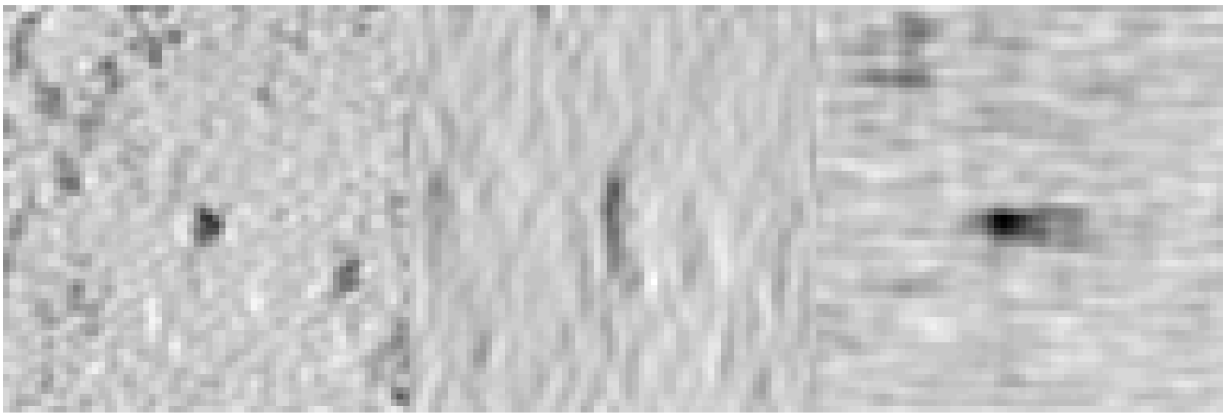
XY XZ YZ

beta-galactosidase - Sample 2



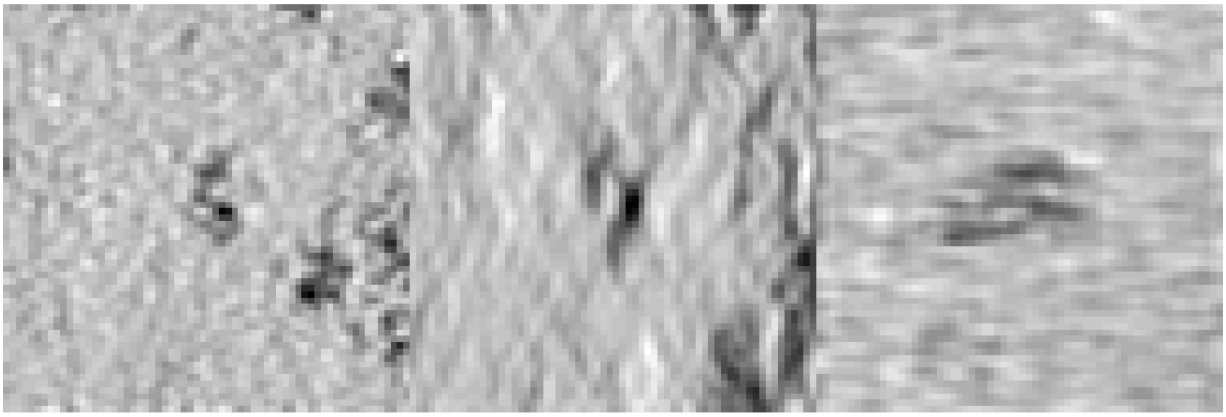
XY XZ YZ

beta-galactosidase - Sample 3



XY XZ YZ

beta-galactosidase - Sample 4

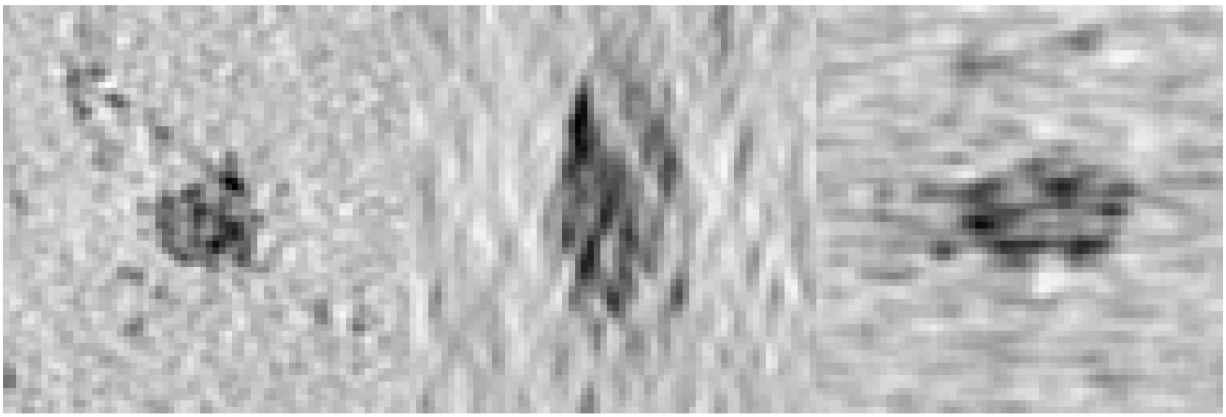


XY XZ YZ

Ribosome:

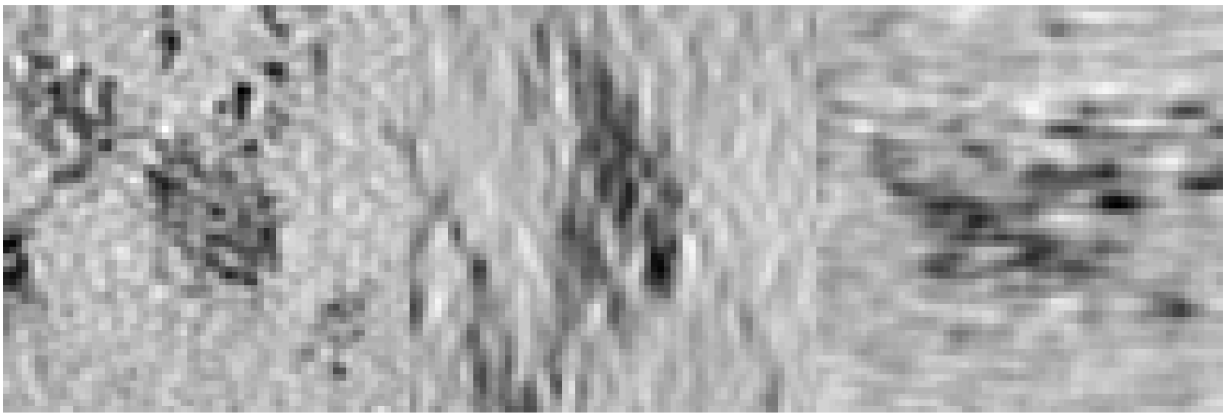
Real examples from the dataset:

Clean (TS_6_6 - ribosome)



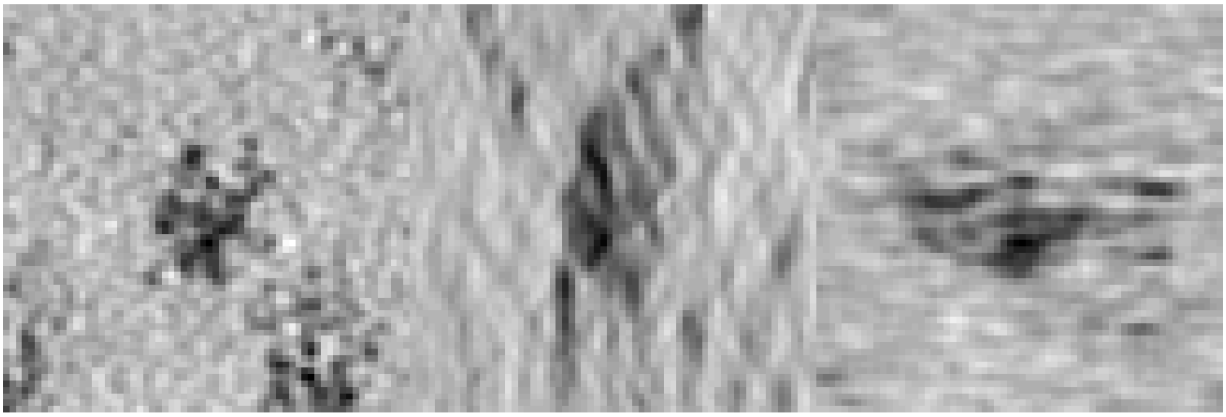
XY XZ YZ

Clean (TS_73_6 - ribosome)



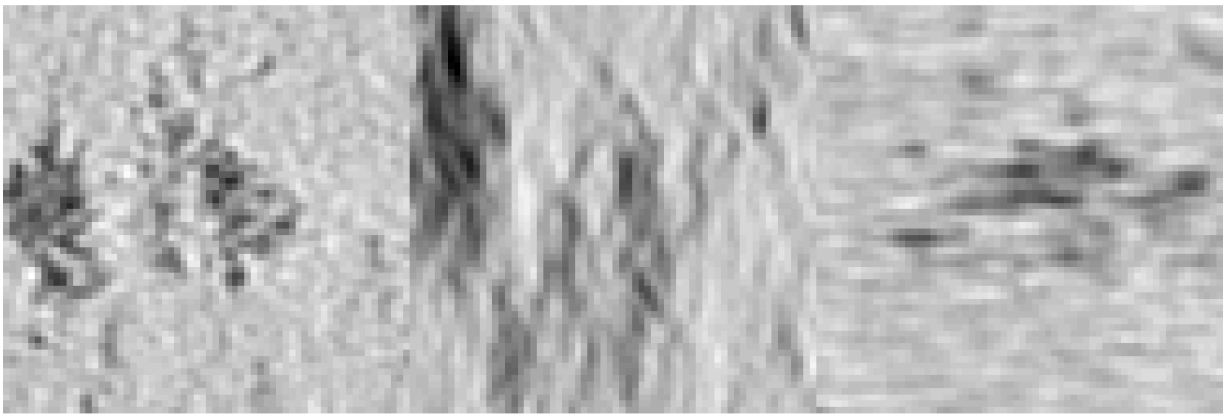
XY XZ YZ

Clean (TS_73_6 - ribosome)



XY XZ YZ

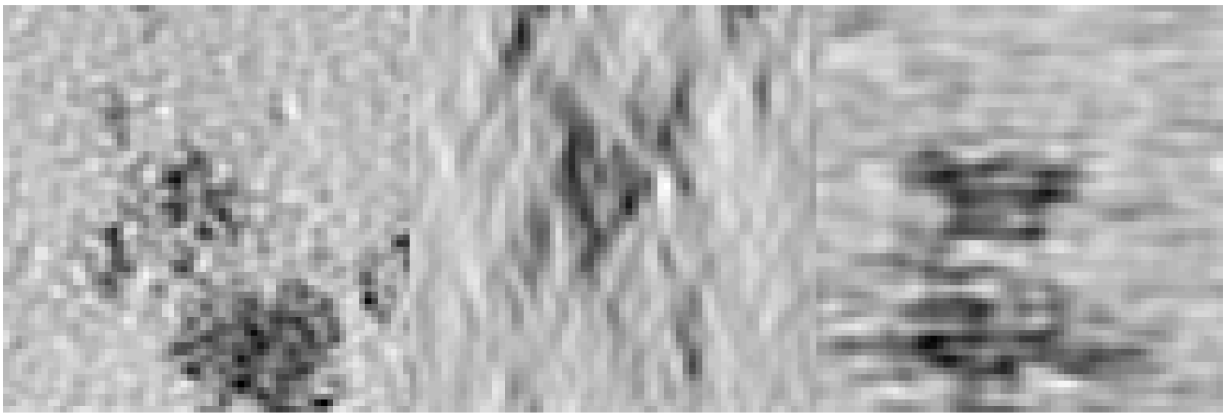
Clean (TS_69_2 - ribosome)



XY XZ YZ

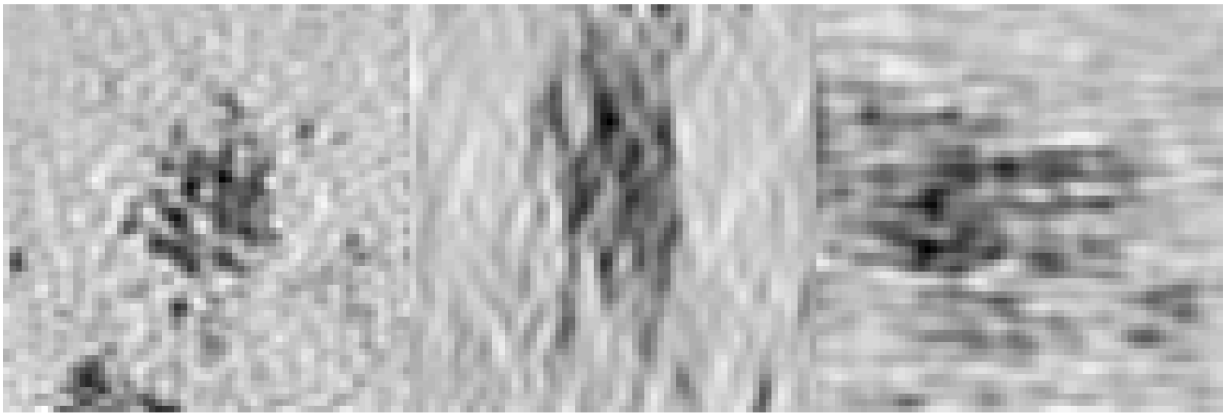
Samples from our trained model:

ribosome - Sample 0



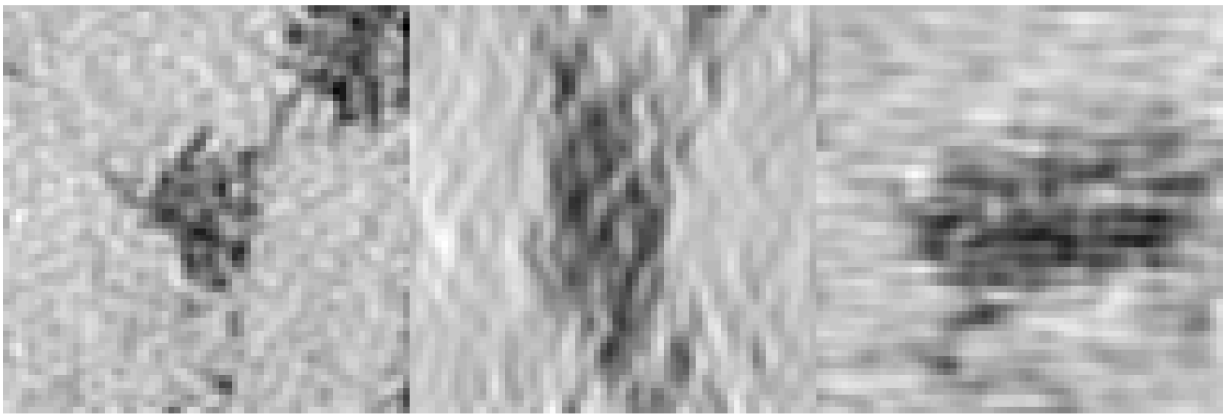
XY XZ YZ

ribosome - Sample 1



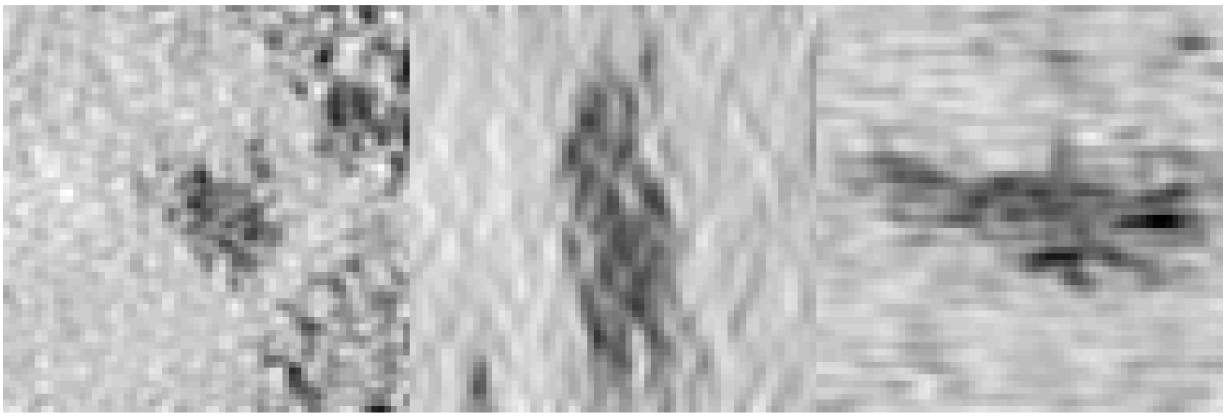
XY XZ YZ

ribosome - Sample 2



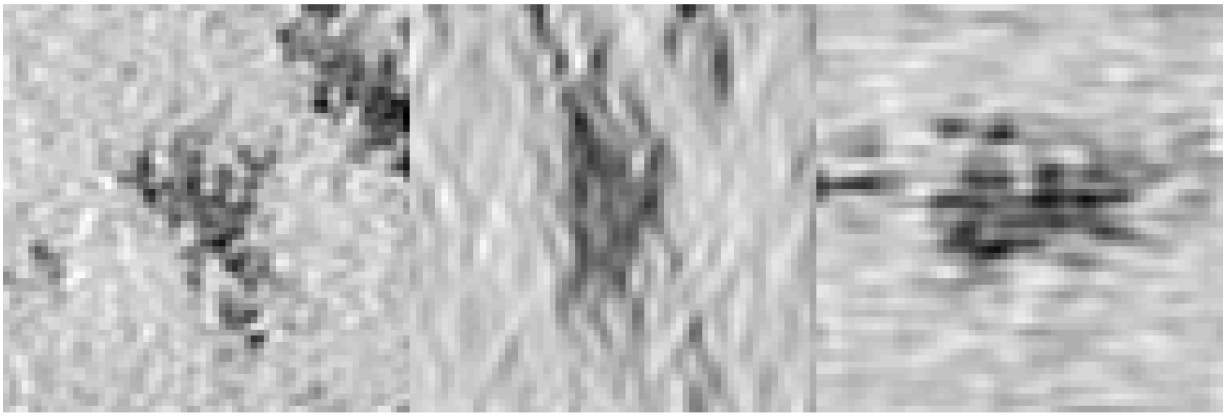
XY XZ YZ

ribosome - Sample 3



XY XZ YZ

ribosome - Sample 4

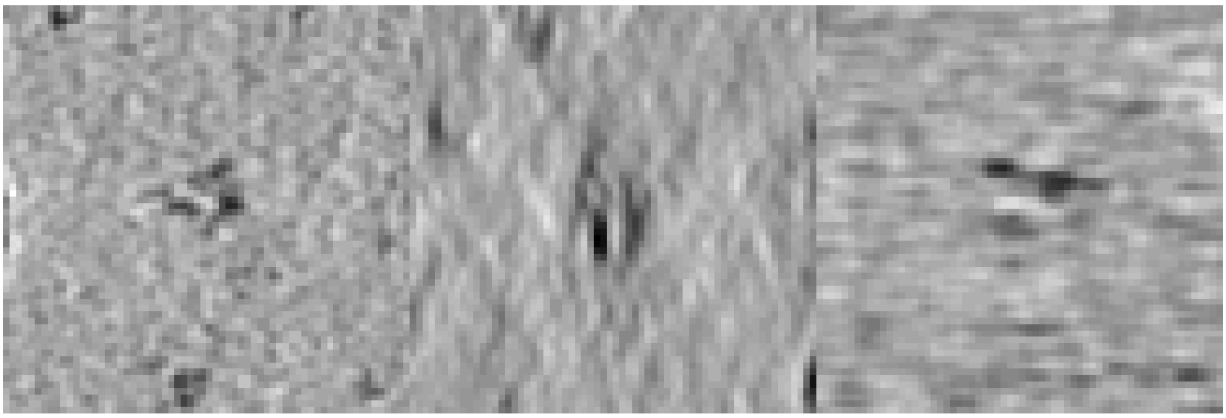


XY XZ YZ

Thyroglobulin:

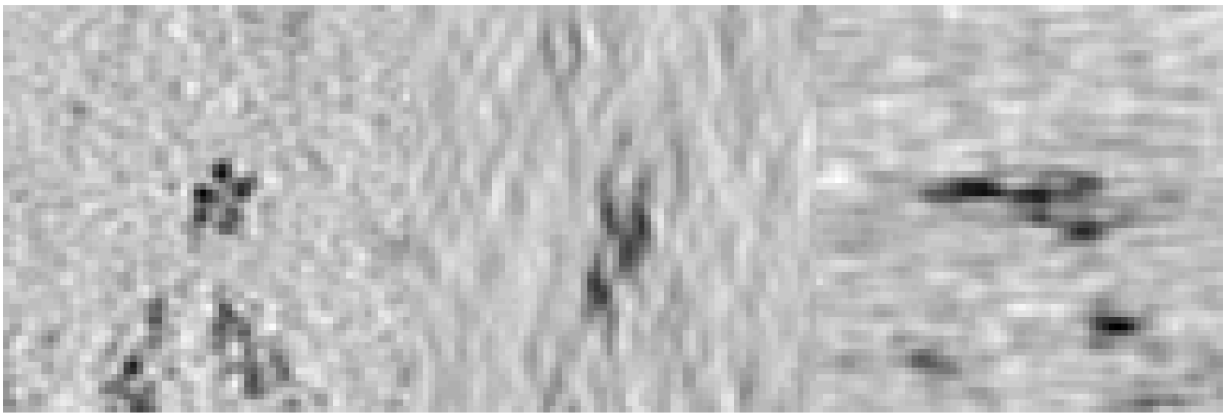
Real examples from the dataset:

Clean (TS_86_3 - thyroglobulin)



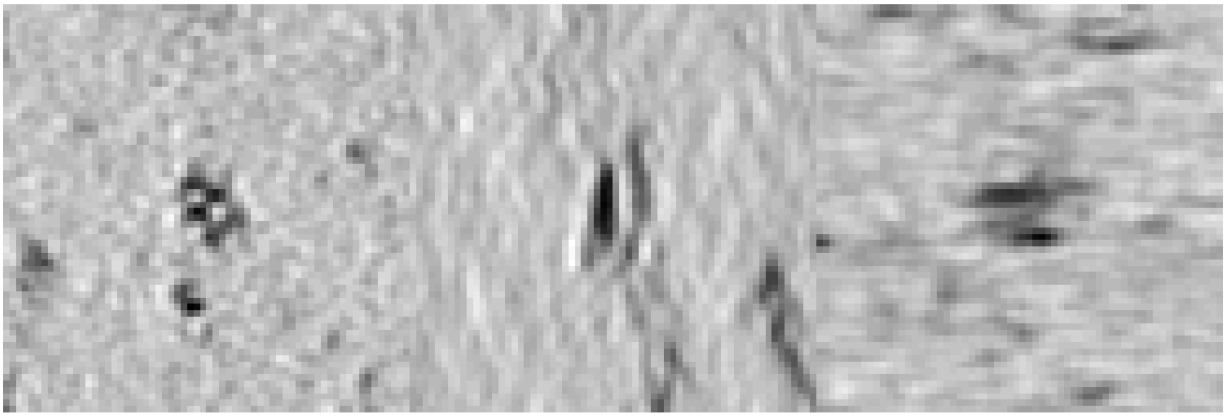
XY XZ YZ

Clean (TS_99_9 - thyroglobulin)



XY XZ YZ

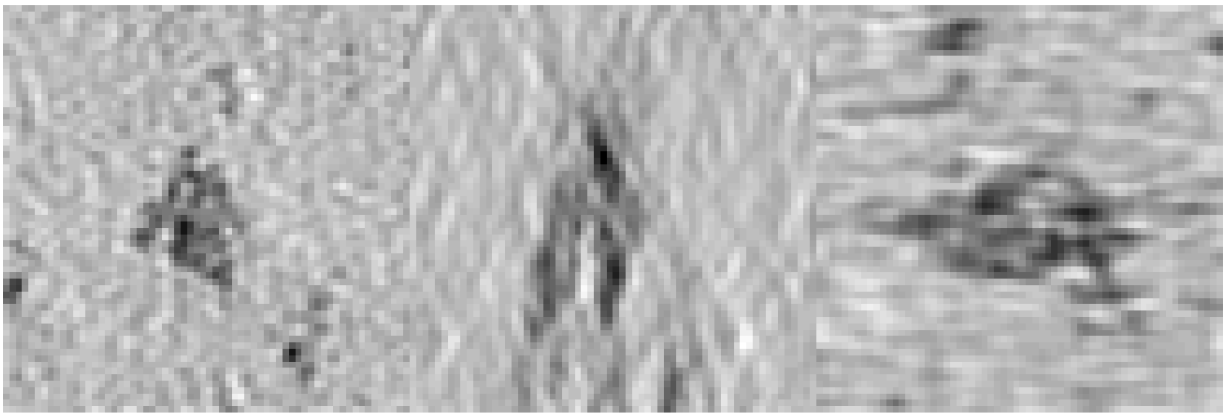
Clean (TS_86_3 - thyroglobulin)



XY XZ YZ

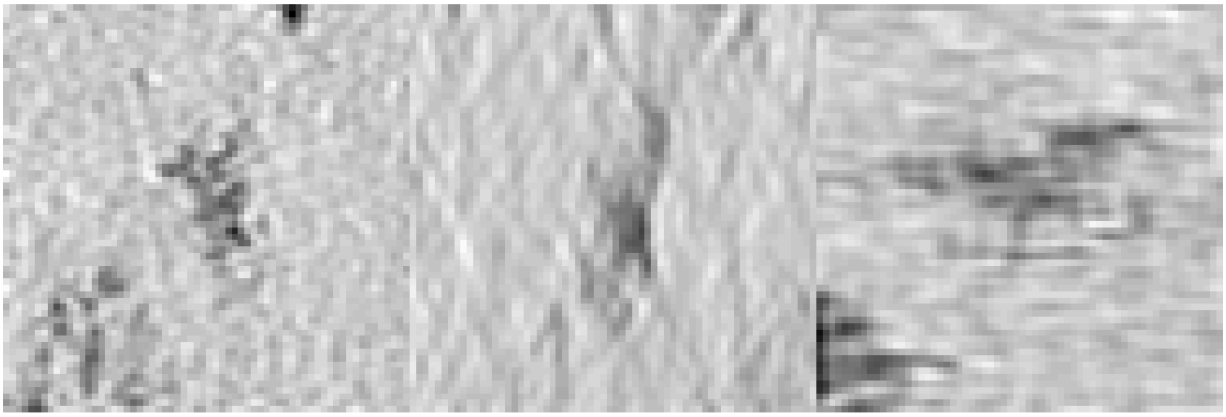
Samples from our trained model:

thyroglobulin - Sample 0



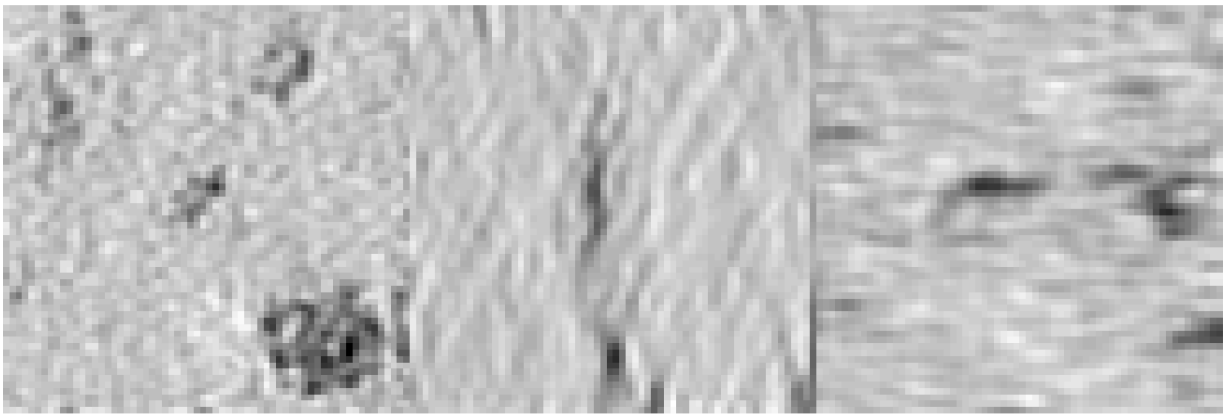
XY XZ YZ

thyroglobulin - Sample 1



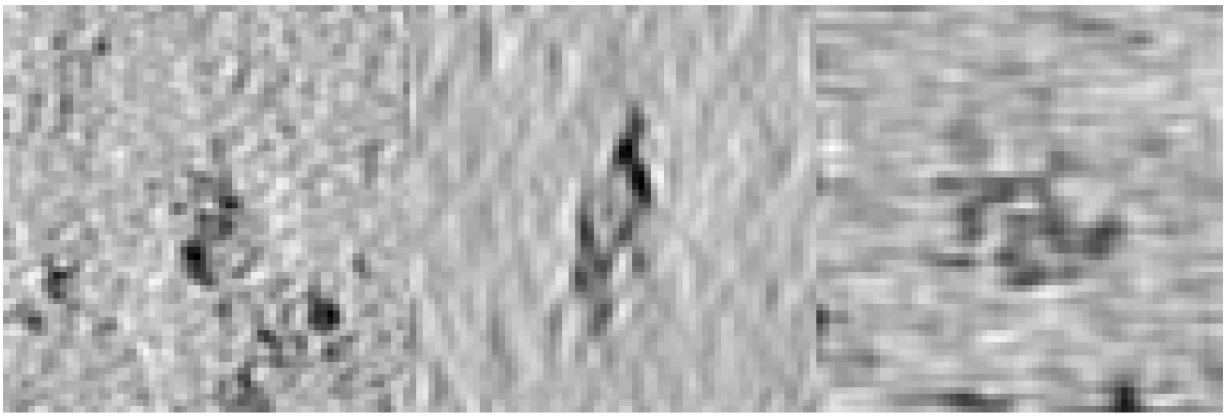
XY XZ YZ

thyroglobulin - Sample 2



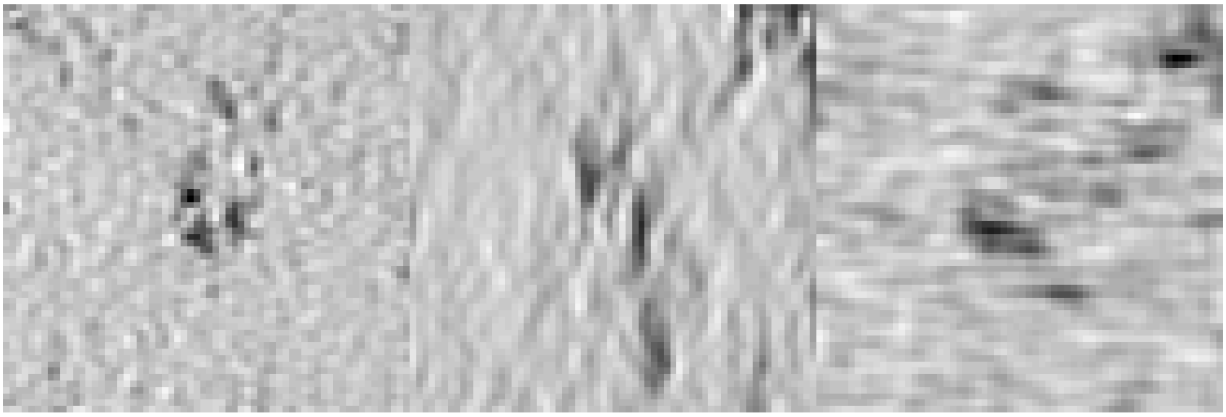
XY XZ YZ

thyroglobulin - Sample 3



XY XZ YZ

thyroglobulin - Sample 4

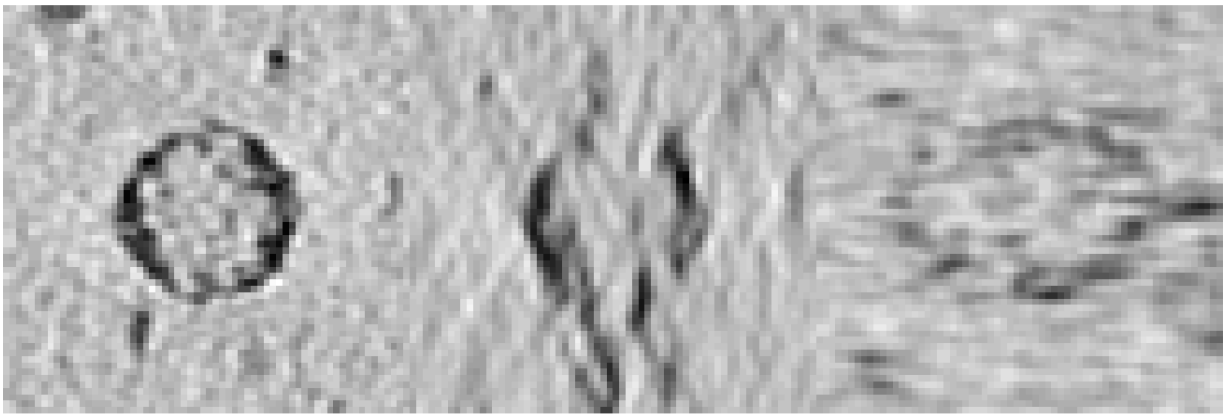


XY XZ YZ

Virus-Like-Particle:

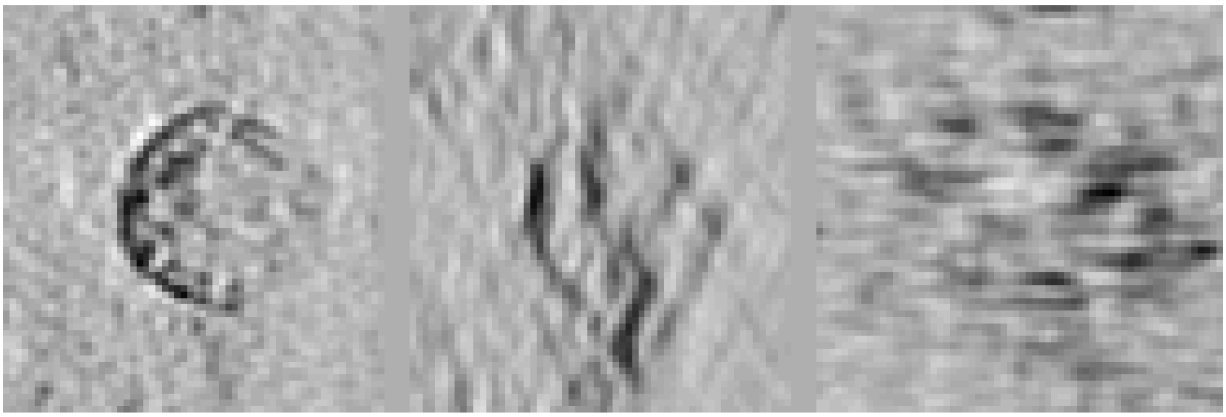
Real examples from the dataset:

Clean (TS_69_2 - virus-like-particle)



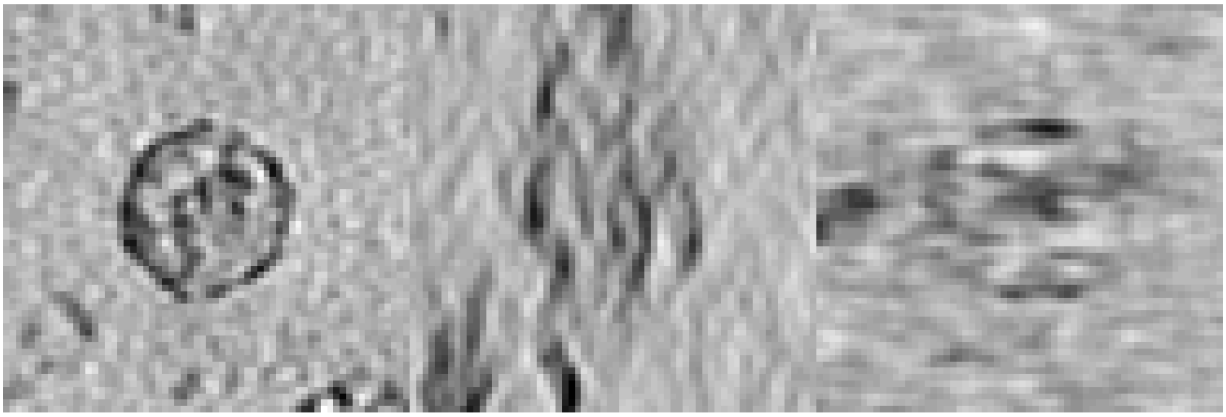
XY XZ YZ

Clean (TS_86_3 - virus-like-particle)



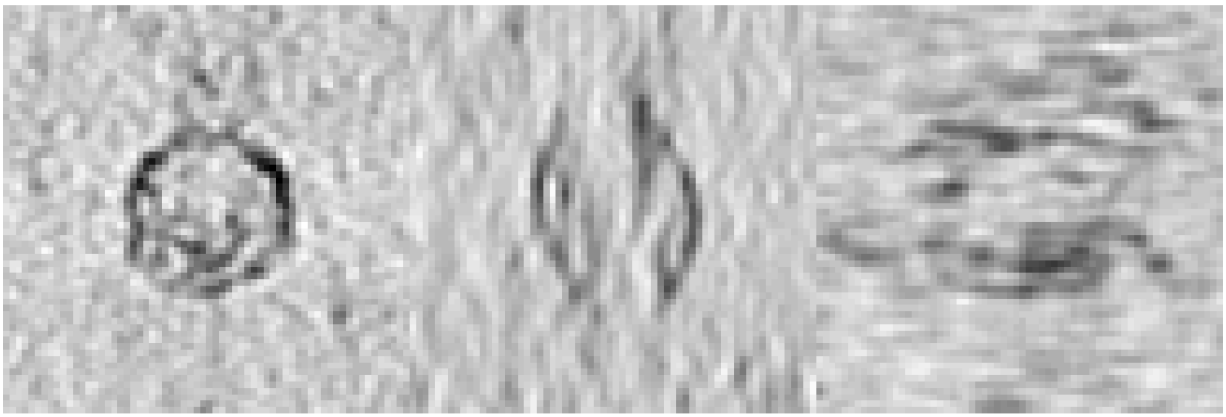
XY XZ YZ

Clean (TS_6_6 - virus-like-particle)



XY XZ YZ

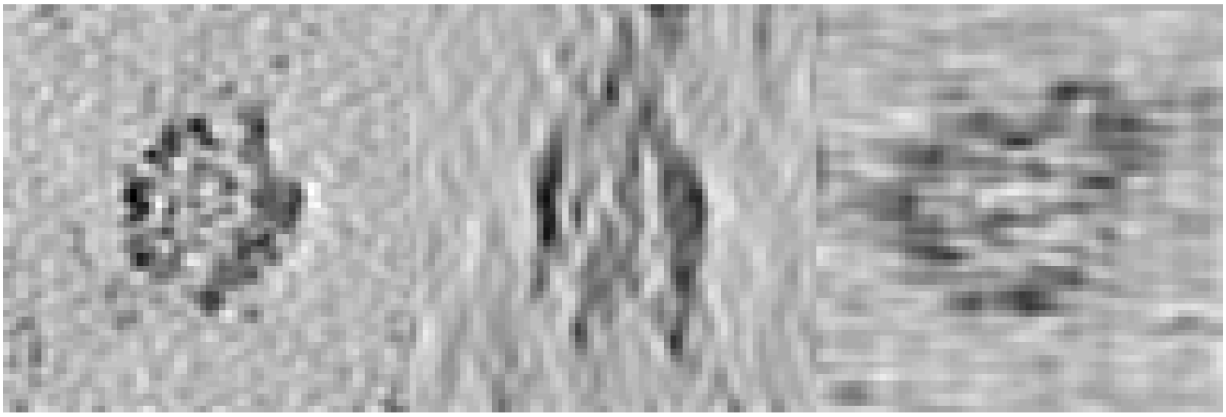
Clean (TS_6_6 - virus-like-particle)



XY XZ YZ

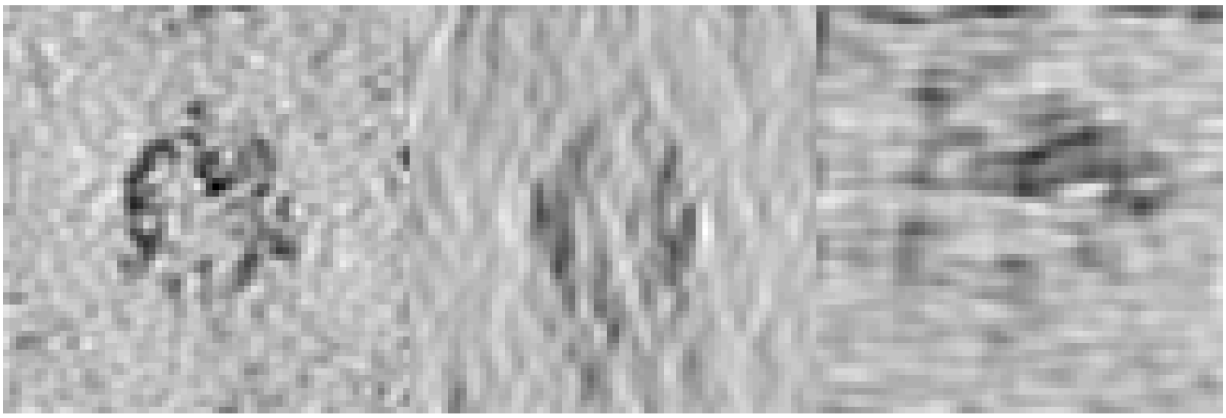
Samples from our trained model:

virus-like-particle - Sample 0



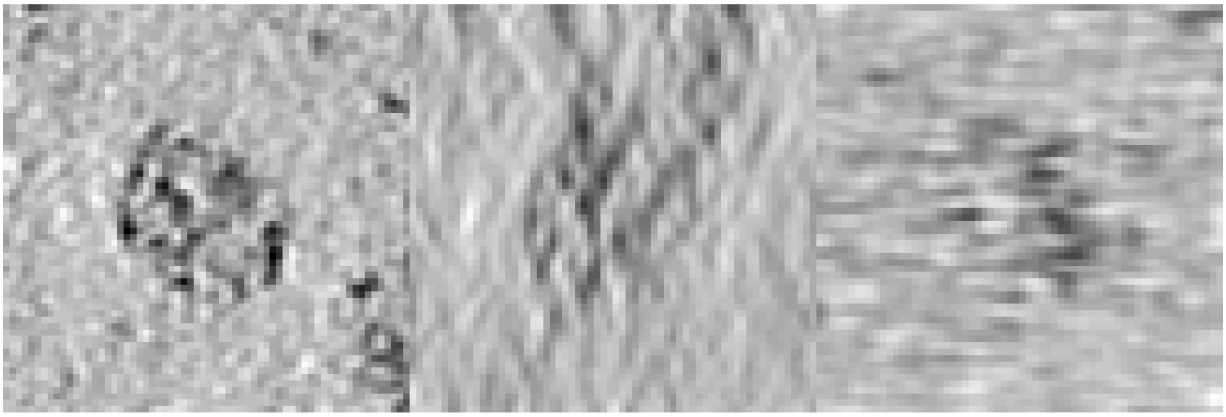
XY XZ YZ

virus-like-particle - Sample 1



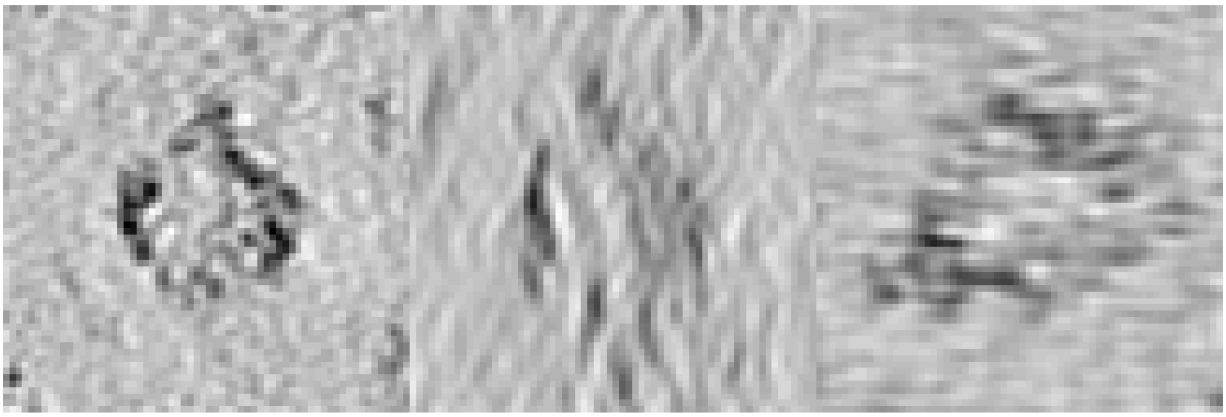
XY XZ YZ

virus-like-particle - Sample 2



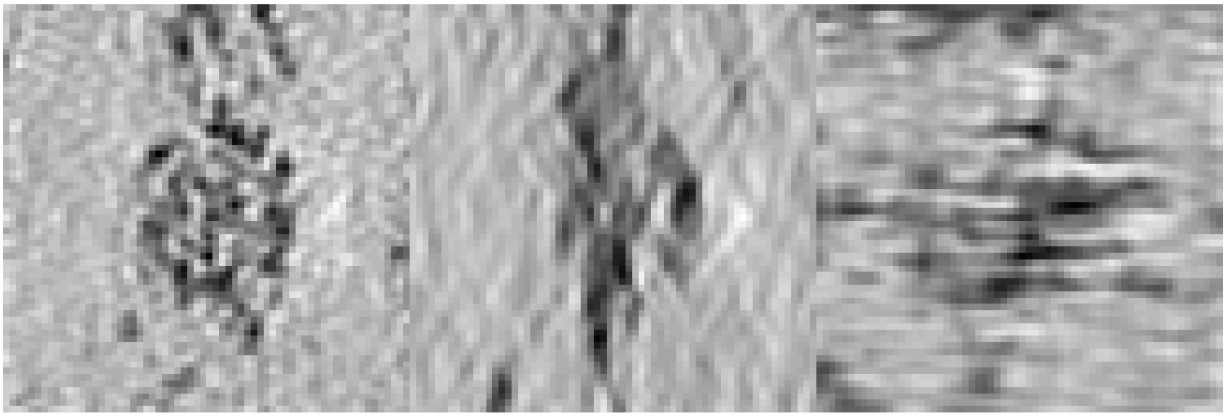
XY XZ YZ

virus-like-particle - Sample 3



XY XZ YZ

virus-like-particle - Sample 4



XY XZ YZ

References

1. Beck, M. & Baumeister, W. Cryo-Electron Tomography: Can it Reveal the Molecular Sociology of Cells in Atomic Detail? *Trends Cell Biol* **26**, 825–837 (2016).
2. Schiøtz, O. H., Klumpe, S., Plitzko, J. M. & Kaiser, C. J. O. Cryo-electron tomography: en route to the molecular anatomy of organisms and tissues. *Biochem Soc Trans* **52**, 2415–2425 (2024).

3. Yan, R., Venkatakrishnan, S. V., Liu, J., Bouman, C. A. & Jiang, W. MBIR: A Cryo-ET 3D Reconstruction Method that Effectively Minimizes Missing Wedge Artifacts and Restores Missing Information. *Journal of structural biology* **206**, 183 (2019).
4. Parkhurst, J. M. *et al.* Pillar data-acquisition strategies for cryo-electron tomography of beam-sensitive biological samples. *Acta Crystallogr D Struct Biol* **80**, 421–438 (2024).
5. Liu, G. *et al.* DeepETPicker: Fast and accurate 3D particle picking for cryo-electron tomography using weakly supervised deep learning. *Nat Commun* **15**, 2090 (2024).
6. Harastani, M., Patra, G., Kervrann, C. & Eltsov, M. Template Learning: Deep learning with domain randomization for particle picking in cryo-electron tomography. *Nat Commun* **16**, 8833 (2025).
7. Wagner, T. & Raunser, S. Cryo-electron tomography: Challenges and computational strategies for particle picking. *Curr Opin Struct Biol* **93**, 103113 (2025).
8. de Teresa-Trueba, I. *et al.* Convolutional networks for supervised mining of molecular patterns within cellular context. *Nat Methods* **20**, 284–294 (2023).
9. Hecksel, C. W. *et al.* Quantifying Variability of Manual Annotation in Cryo-Electron Tomograms. *Microsc Microanal* **22**, 487–496 (2016).
10. Purnell, C. *et al.* Rapid Synthesis of Cryo-ET Data for Training Deep Learning Models. *bioRxiv* (2023) doi:10.1101/2023.04.28.538636.
11. Peck, A. *et al.* A realistic phantom dataset for benchmarking cryo-ET data annotation. *Nat Methods* **22**, 1819–1823 (2025).
12. Wu, X. *et al.* CryoETGAN: Cryo-Electron Tomography Image Synthesis Unpaired Image Translation. *Front Physiol* **13**, 760404 (2022).
13. Eschweiler, D. *et al.* Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image datasets. *PLoS Comput Biol* **20**, e1011890 (2024).
14. Lu, C. *et al.* Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nat Commun* **15**, 4677 (2024).

15. Ho, J., Jain, A. & Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv [cs.LG]* (2020) doi:10.48550/ARXIV.2006.11239.
16. Website. CZII - CryoET Object Identification. https://kaggle.com/competitions/czii-cryo-et-object-identification, 2024. Kaggle.
17. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). (2016).
18. Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. (2017).
19. Continuous-Time Linear Positional Embedding for Irregular Time Series Forecasting. https://arxiv.org/html/2409.20092.
20. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. (2014).