Assessing ConvNeXt V1 and V2 for Protein Classification on the SHREC 2020 Cryo-Electron Tomography Dataset

Faris H. Rizk

April 4, 2025

Abstract

Cryo-electron tomography (cryo-ET) enables three-dimensional visualization of macromolecular complexes within their native cellular environments. However, its inherently low signal-to-noise ratio—caused by the strong interaction of electrons with biological specimens, which limits the allowable electron dose—and its restricted resolution pose significant challenges for accurate protein particle classification. This study evaluates the performance of two state-of-the-art convolutional neural network (ConvNet) architectures—ConvNeXt V1 and ConvNeXt V2—pretrained on the ImageNet dataset for classifying protein particles in the simulated SHREC 2020 cryo-ET dataset. The report introduces a comprehensive preprocessing pipeline that extracts orthogonal triplet views from 48^3 voxel subtomograms centered on each particle's coordinates and fine-tunes the ConvNeXt models using a gradual unfreezing strategy in conjunction with class imbalance penalization. The ConvNeXt V1 and V2 models achieved F1 scores of 0.9887 and 0.9857, respectively, along with strong precision and recall metrics, demonstrating state-of-the-art performance on several protein classes compared to SHREC 2020 challenge leaders. These findings highlight the discriminative power of ConvNeXt-based feature representations for cryo-ET particle classification. The full implementation and source code are publicly available at: `https://github.com/faris-hamdi/cryoet-convnext-shrec2020`.

1 Introduction

Cryo-electron tomography (cryo-ET) is a cutting-edge imaging technique that enables the three-dimensional visualization of macromolecular complexes within their near-native cellular context. Despite its tremendous potential to reveal the structural details of biological molecules, cryo-ET inherently suffers from low signal-to-noise ratios and limited resolution due to the constraints imposed by electron dose and imaging conditions Gubins et al. (2020). These challenges necessitate robust computational methods for the localization and classification of particles, which are crucial for subsequent subtomogram averaging and high-resolution structure determination.

Historically, conventional techniques such as template matching have been widely used in cryo-ET for particle detection and classification. However, these approaches often struggle with high noise levels and variability in particle orientation and positioning Gubins et al. (2020). The advent of deep learning has marked a significant shift in this domain, as learning-based methods—especially those leveraging convolutional neural networks (ConvNets)—have demonstrated superior performance in overcoming these limitations. Recent contest benchmarks, such as those presented in the SHREC 2020 cryo-ET track, have highlighted the efficacy of such methods in classifying particles within simulated tomograms Gubins et al. (2020).

In parallel with advancements in cryo-ET, the field of visual recognition has experienced a renaissance with the introduction of modern ConvNet architectures. ConvNeXt V1 Liu et al. (2022) and its subsequent evolution, ConvNeXt V2 Woo et al. (2023), embody a new generation of ConvNets that integrate design principles from vision Transformers while retaining the simplicity and efficiency of traditional convolutional architectures. These models have been pre-trained on large-scale datasets such as ImageNet and have achieved state-of-the-art performance on various tasks, including image classification, object detection, and semantic segmentation Liu et al. (2022); Woo et al. (2023).

In this work, the application of pre-trained ConvNeXt V1 and V2 models has been investigated to classify protein particles in a simulated cryo-ET volume. By leveraging these architectures' superior feature representation and scalability, our report aims to benchmark their performance on a task representative of the challenges faced in cryo-ET data analysis. The following sections detail the methodology employed, the experimental setup derived from the SHREC 2020 contest guidelines Gubins et al. (2020), and the comparative analysis of the two ConvNeXt variants on this challenging dataset.

2 Methodology

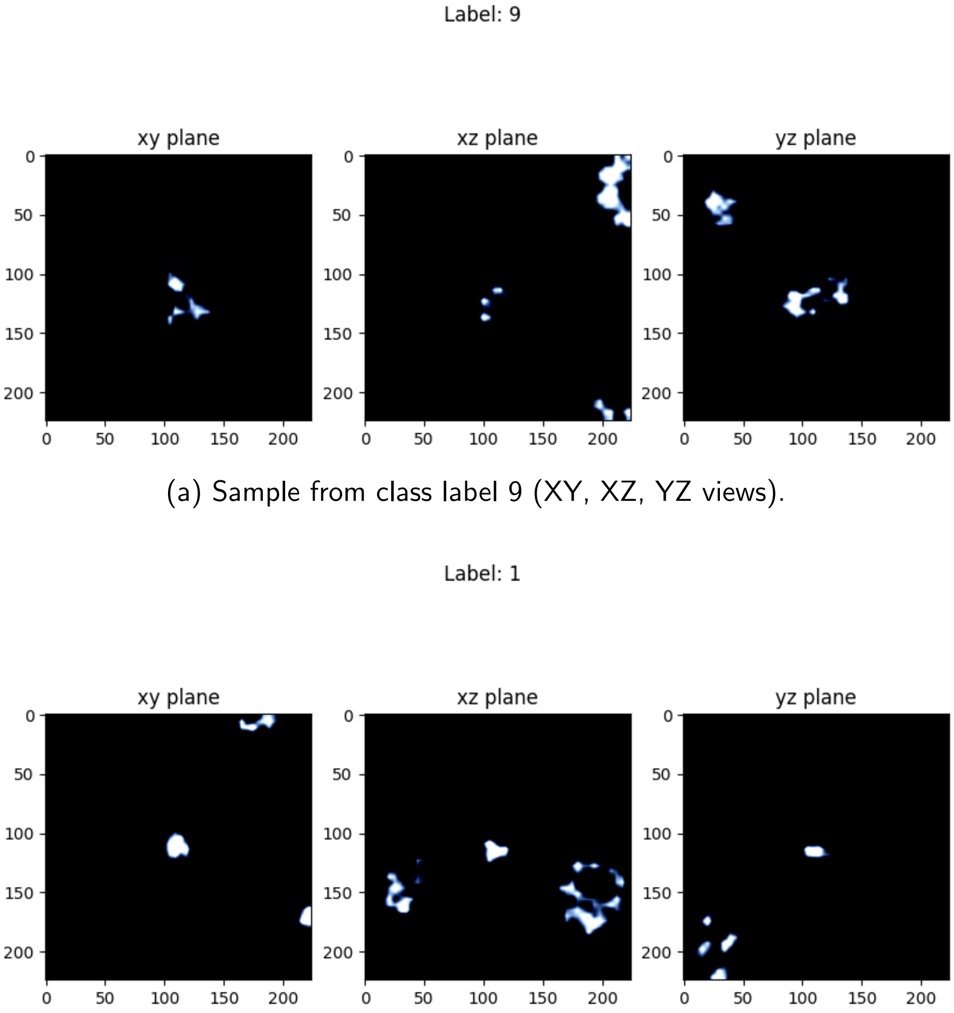
This study leverages pre-trained ConvNeXt V1 and V2 models to classify protein particles within simulated cryo-electron tomograms. The methodology encompasses four main components: data acquisition and preprocessing, model adaptation and training, evaluation, and inference. Each of these components is detailed below.

2.1 Dataset and Preprocessing

The evaluation uses simulated tomograms provided by the SHREC 2020 challenge Gubins et al. (2020). Each tomogram comprises, on average, 2500 particles distributed across 12 distinct macromolecular classes. To support the development and benchmarking of learning-based methods, nine tomograms (indexed 0 to 8) are used for training and validation. In contrast, the tenth tomogram (index 9) is reserved exclusively as a test set to allow fair comparison with SHREC 2020 submissions Gubins et al. (2020). The original dataset is publicly accessible via the SHREC 2020 official repository[1], and both the raw and processed data are also mirrored in a Google Drive folder[2] to facilitate reproducibility. Figure 1 presents representative examples of extracted particles from class labels 1 and 9, illustrating the diversity and structure of the data across the orthogonal planes.

[1] https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Y2ZMRH

[2] https://drive.google.com/drive/folders/1ROgxmjFOAZoFKB19cl94RgeFqtbcCfjz?usp=sharing



(a) Sample from class label 9 (XY, XZ, YZ views).

(b) Sample from class label 1 (XY, XZ, YZ views).

Figure 1: Example subtomogram slices (XY, XZ, YZ views) from two classes in the dataset: label 9 and label 1. Each sample illustrates the three orthogonal planes extracted from the 48^3 voxel subvolume centered on the particle.

The preprocessing pipeline begins by extracting cubic subtomograms of size 48^3 voxels, centered at the ground truth coordinates of each protein particle. From each subtomogram, three orthogonal 2D slices (XY, XZ, YZ planes) are extracted to construct a triplet-view representation. These views are then subjected to the same transformations used in the ImageNet training pipeline—namely, resizing and normalization—to ensure compatibility with the ConvNeXt model inputs. Subsequently, each slice undergoes slice-wise normalization, and the resulting triplets, along with their associated class labels, are stored in `.pt` format to accelerate training and validation. The training and validation sets are formed by an 80/20 split of the processed data from tomograms 0 to 8, while tomogram 9 is the test set. A class-weighting scheme was employed during training to account for class imbalance observed in the training data (illustrated in Figure 2). The penalization weights, ordered by class labels from 0 to 11, are defined as: `weights` = `[0.0773, 0.1030, 0.0835, 0.0671, 0.0695, 0.0828, 0.0841, 0.0928, 0.0923, 0.0816, 0.0752, 0.0908]`.

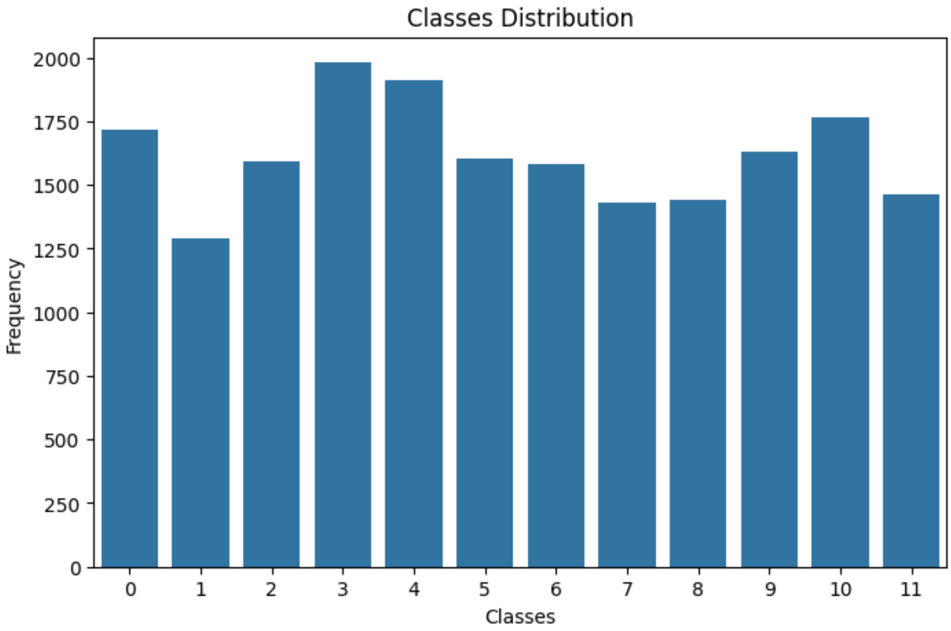


Figure 2: Distribution of protein particle classes in the training set. The observed imbalance motivated the use of class weighting during training.

2.2 Model Adaptation and Training Pipeline

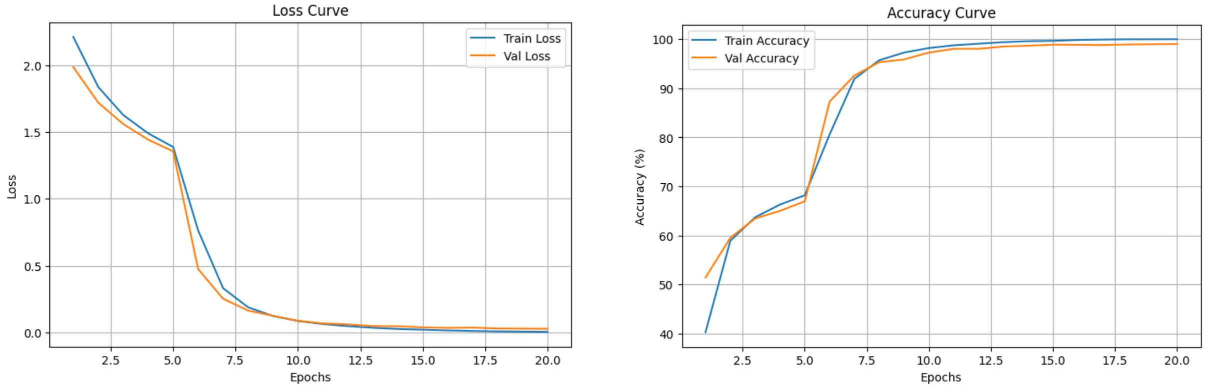
ConvNeXt V1 and ConvNeXt V2 architectures, pre-trained on the ImageNet dataset and accessed via the Hugging Face model hub, were adopted as the backbone for classification Liu et al. (2022); Woo et al. (2023). For this task, the original prediction head of each model was replaced and fine-tuned to output predictions across the 12 protein classes defined in the SHREC 2020 dataset.

Each training sample comprises a triplet of images corresponding to the three orthogonal planes (XY, XZ, YZ) extracted from the 48^3 subtomogram centered on the particle. Each view is represented as a three-channel image to maintain compatibility with the pre-trained models. Both ConvNeXt variants were fine-tuned for up to 20 epochs using a batch size of 32, and training was accelerated using mixed-precision computations. Optimization was performed using the AdamW optimizer Loshchilov and Hutter (2019), with a learning rate of 1×10^{-4} for the prediction head and a lower rate of 1×10^{-5} for the unfrozen backbone parameters. An early stopping strategy was employed to prevent overfitting: training was halted if validation accuracy did not improve for three consecutive epochs.

To further support model adaptation, a gradual unfreezing strategy was implemented. Only the prediction head was trainable during the initial epochs, and stage 3 of the ConvNeXt backbone was unfrozen after the fifth epoch. This approach enabled progressive refinement of deeper layers as training progressed. The classification task was supervised using a cross-entropy loss function, enhanced by class weighting to address the imbalance in class distribution observed in the training set. All training was conducted on a single NVIDIA Tesla T4 GPU.

2.3 Evaluation and Inference

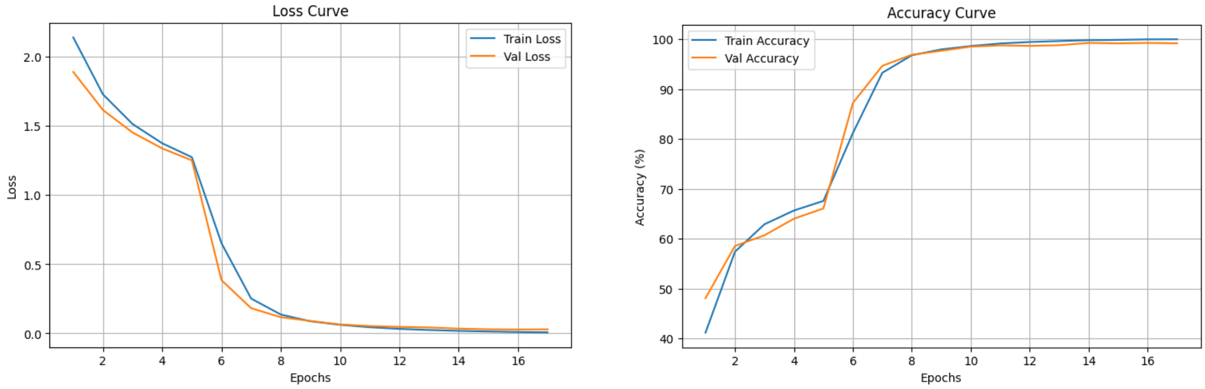
The best-performing models were selected based on their validation accuracy throughout training. ConvNeXt V1 achieved a final validation accuracy of 99.02% at epoch 20. At the same time, ConvNeXt V2 reached a peak accuracy of 99.23% at epoch 14, at which point training was terminated due to the early stopping criterion Liu et al. (2022); Woo et al. (2023). Training progress was monitored through accuracy and loss metrics, as illustrated in Figures 3 and 4. Both models demonstrated stable convergence and effective generalization, with no evidence of overfitting.



(a) Training and validation loss.

(b) Training and validation accuracy.

Figure 3: Training progress of the ConvNeXt V1 model.



(a) Training and validation loss.

(b) Training and validation accuracy.

Figure 4: Training progress of the ConvNeXt V2 model.

For inference, the test tomogram (tomogram 9) was processed using the same pipeline applied during training and validation. Each particle's three orthogonal views (XY, XZ, YZ) were passed through the model, and the resulting classification probabilities were summed across views. The class with the highest cumulative probability was then selected as the predicted label.

Inference was performed using the same NVIDIA Tesla T4 GPU employed during training. The complete classification of the test tomogram required 36.61 seconds for ConvNeXt V1 and 45.26 seconds for ConvNeXt V2, indicating that both models are accurate and computationally practical.

Overall, the methodology integrates consistent preprocessing, principled model adaptation, and efficient inference, offering a robust framework for benchmarking classification performance in cryo-electron tomograms using ConvNeXt architectures Gubins et al. (2020).

3 Results

The performance of the ConvNeXt V1 and ConvNeXt V2 models was evaluated on a simulated test tomogram (tomogram 9), which contains 2782 particles evenly distributed across 12 protein classes. Evaluation metrics include Accuracy, Precision, Recall, and F1 Score Gubins et al. (2020), along with inference time as a measure of computational efficiency. ConvNeXt V1 completed inference in 36.61 seconds, while ConvNeXt V2 required 45.26 seconds. Table 1 summarizes the overall classification performance of both models.

Table 1: Overall performance metrics for ConvNeXt V1 and ConvNeXt V2 on the test tomogram.

Model	Accuracy	Precision	Recall	F1 Score
ConvNeXt V1	0.9894	0.9887	0.9887	0.9887
ConvNeXt V2	0.9868	0.9868	0.9856	0.9857

Table 2 provides a detailed per-class performance breakdown, reporting each protein class's Support, Precision, Recall, and F1 Score. Both models demonstrate consistently high scores, with several classes achieving near-perfect metrics. Notably, both ConvNeXt variants performed equally well in identifying classes such as `1bxn`, `3cf3`, `4cr2`, and `4d8q`, where Precision, Recall, and F1 Score reached or approached 1.000.

Table 2: Per-class performance comparison between ConvNeXt V1 and ConvNeXt V2.

Class	Support	ConvNeXt V1			ConvNeXt V2		
		Precision	Recall	F1	Precision	Recall	F1
1bxn	177	1.000	1.000	1.000	1.000	1.000	1.000
1qvr	213	0.995	0.995	0.995	0.991	1.000	0.995
1s3x	168	0.964	0.970	0.967	0.993	0.893	0.940
1u6g	178	0.983	0.989	0.986	0.973	0.994	0.983
2cg9	176	0.983	0.989	0.986	1.000	0.966	0.983
3cf3	202	0.995	0.995	0.995	1.000	1.000	1.000
3d2f	168	0.982	0.982	0.982	0.988	1.000	0.994
3gl1	178	1.000	0.978	0.989	0.994	0.994	0.994
3h84	194	1.000	0.995	0.997	1.000	0.995	0.997
3qm1	177	0.961	0.972	0.966	0.902	0.989	0.943
4cr2	231	1.000	1.000	1.000	1.000	0.996	0.998
4d8q	210	1.000	1.000	1.000	1.000	1.000	1.000

To contextualize these results, Table 3 compares the per-class F1 scores of ConvNeXt V1 and V2 with the leading submissions from the SHREC 2020 challenge Gubins et al. (2020). Competing methods include 3D MS-D, DeepFinder, 3D ResNet, YOPO, Dn3DUnet, UMC, TM-T, and TM-F. Notably, these methods perform both localization and classification, while the proposed ConvNeXt models were evaluated strictly on classification using ground-truth particle positions. Despite this distinction, ConvNeXt V1 and V2 consistently match or exceed the performance of the top SHREC entries across many classes. Bold values highlight the best-performing method for each class.

Table 3: F1 score comparison across classes between SHREC 2020 methods and the ConvNeXt models. Bold indicates the best result in each column.

Method	1s3x	3qm1	3gl1	3h84	2cg9	3d2f	1u6g	3cf3	1bxn	1qvr	4cr2	4d8q
3D MS-D	0.192	0.408	0.437	0.416	0.368	0.461	0.492	0.719	0.948	0.851	0.942	0.964
DeepFinder	0.610	0.729	0.800	0.911	0.783	0.848	0.866	0.939	**1.000**	0.984	0.993	0.993
3D ResNet	0.193	0.185	0.405	0.407	0.334	0.445	0.491	0.628	0.906	0.719	0.868	0.817
YOPO	0.558	0.741	0.670	0.834	0.696	0.682	0.795	0.896	0.987	0.830	0.923	0.993
Dn3DUnet	0.529	0.577	0.569	0.674	0.332	0.523	0.462	0.676	0.925	0.684	0.907	0.974
UMC	0.661	0.827	0.839	0.947	0.855	0.873	0.899	0.981	0.997	0.980	**1.000**	0.997
TM-T	0.200	0.102	0.248	0.727	0.555	0.869	0.835	0.880	0.934	0.970	0.968	0.945
TM-F	0.319	0.219	0.207	0.660	0.589	0.808	0.815	0.945	0.939	0.966	0.968	0.945
ConvNeXt V1	**0.967**	**0.966**	0.989	**0.997**	**0.986**	0.982	**0.986**	0.995	**1.000**	**0.995**	**1.000**	**1.000**
ConvNeXt V2	0.940	0.943	**0.994**	**0.997**	0.983	**0.994**	0.983	**1.000**	**1.000**	**0.995**	0.998	**1.000**

4 Discussion

The experimental results confirm that ConvNeXt V1 and ConvNeXt V2 architectures deliver high classification performance on simulated cryo-electron tomograms. Overall accuracy and F1 scores exceeding 98% (Table 1) emphasize the effectiveness of transferring pre-trained ImageNet weights to this domain through task-specific fine-tuning on the SHREC 2020 dataset Gubins et al. (2020); Liu et al. (2022); Woo et al. (2023). Per-class evaluation (Table 2) reveals near-perfect scores for several protein categories, including `1bxn`, `3cf3`, `4cr2`, and `4d8q`, indicating that the models can distinguish well-defined structural features. Conversely, modest drops in performance for classes such as `1s3x` and `3qm1` suggest sensitivity to finer structural ambiguity or limited representation during training. These differences may stem from architectural distinctions or varying convergence dynamics during fine-tuning Liu et al. (2022); Woo et al. (2023).

A comparative analysis with leading methods from the SHREC 2020 challenge (Table 3) further validates the proposed approach. Unlike competing methods, which jointly perform localization and classification, the ConvNeXt models focus exclusively on classifying particles using ground-truth coordinates. Despite this scope limitation, ConvNeXt V1 and V2 outperform or match the state-of-the-art in per-class F1 scores across most categories. This result underscores the strength of the learned representations and the capacity of modern ConvNet architectures to generalize in structurally complex biomedical imaging domains.

The cumulative F1 scores visualized in Figure 5 demonstrate consistent performance across all 12 classes, including those with fewer training examples. Furthermore, the confusion matrices shown in Figure 6 indicate minimal misclassification, highlighting the models' robust ability to distinguish subtle class-level features.

From a computational standpoint, inference times of 36.61 seconds for ConvNeXt V1 and 45.26 seconds for ConvNeXt V2 confirm that the proposed method is accurate and

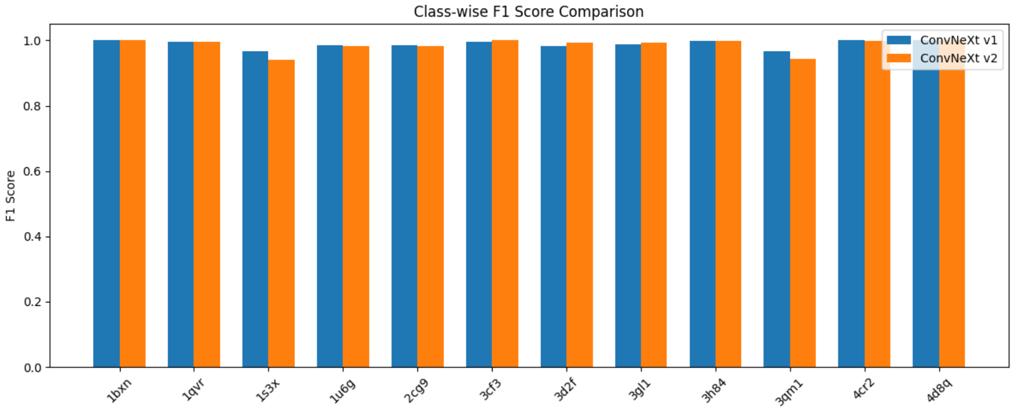
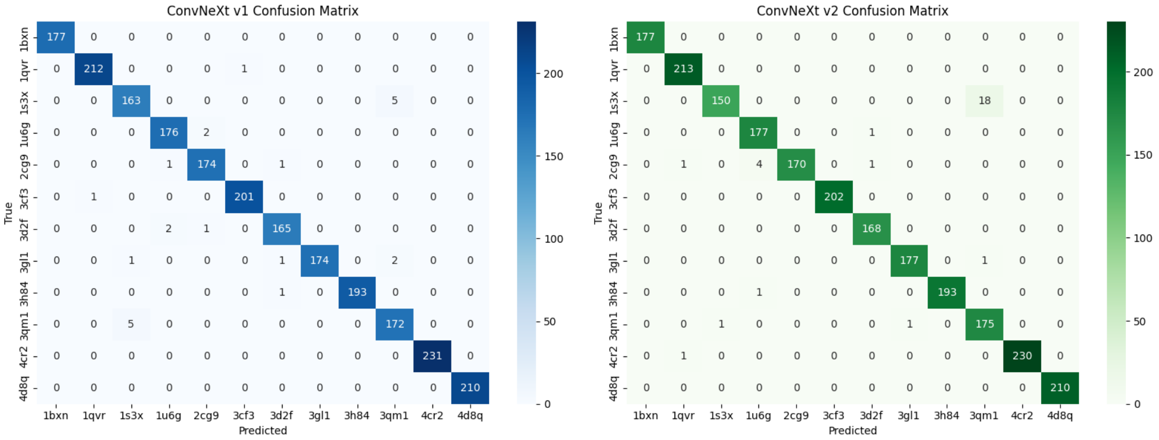


Figure 5: Cumulative F1 scores across the 12 protein classes for ConvNeXt V1 and ConvNeXt V2. The models consistently perform across all classes, including those with fewer examples.



	1bxn	1qvr	1s3x	1u6g	2cg9	3cf3	3d2f	3gl1	3h84	3qm1	4cr2	4d8q
1bxn	177	0	0	0	0	0	0	0	0	0	0	0
1qvr	0	212	0	0	0	1	0	0	0	0	0	0
1s3x	0	0	163	0	0	0	0	0	0	5	0	0
1u6g	0	0	0	176	2	0	0	0	0	0	0	0
2cg9	0	0	0	1	174	0	1	0	0	0	0	0
3cf3	0	1	0	0	0	201	0	0	0	0	0	0
3d2f	0	0	0	2	1	0	165	0	0	0	0	0
3gl1	0	0	1	0	0	0	1	174	0	2	0	0
3h84	0	0	0	0	0	0	1	0	193	0	0	0
3qm1	0	0	5	0	0	0	0	0	0	172	0	0
4cr2	0	0	0	0	0	0	0	0	0	0	231	0
4d8q	0	0	0	0	0	0	0	0	0	0	0	210

	1bxn	1qvr	1s3x	1u6g	2cg9	3cf3	3d2f	3gl1	3h84	3qm1	4cr2	4d8q
1bxn	177	0	0	0	0	0	0	0	0	0	0	0
1qvr	0	213	0	0	0	0	0	0	0	0	0	0
1s3x	0	0	150	0	0	0	0	0	0	18	0	0
1u6g	0	0	0	177	0	0	1	0	0	0	0	0
2cg9	0	1	0	4	170	0	1	0	0	0	0	0
3cf3	0	0	0	0	0	202	0	0	0	0	0	0
3d2f	0	0	0	0	0	0	168	0	0	0	0	0
3gl1	0	0	0	0	0	0	0	177	0	1	0	0
3h84	0	0	0	1	0	0	0	0	193	0	0	0
3qm1	0	0	1	0	0	0	0	1	0	175	0	0
4cr2	0	1	0	0	0	0	0	0	0	0	230	0
4d8q	0	0	0	0	0	0	0	0	0	0	0	210

Figure 6: Confusion matrices for ConvNeXt V1 and ConvNeXt V2 on the test set. Both models demonstrate strong class discrimination with minimal confusion between classes.

practically efficient—this favorable trade-off between performance and speed positions ConvNeXt architectures as viable components in automated cryo-ET analysis workflows.

In summary, the study highlights the adaptability and discriminative strength of ConvNeXt-based models for particle classification in cryo-electron tomograms. It illustrates their broader applicability in high-resolution biomedical image analysis tasks.

5 Conclusion and Future Work

This report assessed the classification performance of two pre-trained ConvNeXt architectures—ConvNeXt V1 and ConvNeXt V2—for the task of protein particle identification in cryo-electron tomograms, following the SHREC 2020 challenge protocol. By employing a robust data preprocessing pipeline and an effective fine-tuning strategy—including gradual unfreezing and class imbalance penalization—both models achieved overall F1 scores exceeding 98%. Furthermore, they demonstrated strong per-class performance, often surpassing state-of-the-art approaches designed for joint localization and classification. These findings validate the effectiveness of modern convolutional architectures in capturing subtle structural distinctions within cryo-ET data, offering a streamlined alternative to more complex multi-task pipelines.

Future work could focus on integrating localization capabilities into the current classification framework to develop a unified end-to-end system for particle detection and classification. Further improvements may be achieved by leveraging transfer learning from larger, domain-specific biomedical datasets and adopting advanced data augmentation techniques to enhance robustness under varying imaging conditions. Lastly, validating the models on experimentally acquired cryo-ET datasets will be critical to ensure their practical applicability in real-world structural biology and biomedical research contexts.

References

Gubins, I., Chaillet, M. L., van der Schot, G., Veltkamp, R. C., Förster, F., Hao, Y., Wan, X., Cui, X., Zhang, F., Moebel, E., Wang, X., Kihara, D., Zeng, X., Xu, M., Nguyen, N. P., White, T., and Bunyak, F. (2020). SHREC 2020: Classification in cryo-electron tomograms. *Computers & Graphics*, 91:279–289.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. (2023). Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16133–16142.