An Active Learning Framework for ML-Assisted Labeling of Cryo-EM Micrographs

Robert Kiewisz
Simons Machine Learning Center
New York Structural Biology Center
Convent Ave. 89, NY 10027
rkiewisz@nysbc.org

Tristan Bepler
Simons Machine Learning Center
New York Structural Biology Center
Convent Ave. 89, NY 10027
tbepler@n.org

Abstract

Single-particle cryo-electron microscopy (cryo-EM) has grown significantly as a tool for discerning biological macromolecule structures. A fundamental step in this technique is the accurate identification of individual protein particles from micrographs laden with noise. Machine learning models, specifically convolutional neural networks like ResNet, have shown promise by reducing dependence on manual methods and adapting to the intricate features within the micrographs. However, challenges persist due to low signal-to-noise ratios, resulting in false positives or missed detections. Analogous challenges in computer vision have found respite in active learning, a method that combines automated systems with human intervention for refined outcomes. This paper presents a novel approach for cryo-EM particle picking based on active learning and logistic regression. Our method employs the pre-trained convolutional-based model from the Topaz particle picking software. This model is used for the initial feature extraction and subsequently refines particle predictions through a logistic regression with a human feedback loop. Complementing this, we introduce the Napari plugin, enhancing user interaction with the micrograph and facilitating intuitive model training. This approach allowed us to achieve $\sim 10\%$ average precision improvement over the Topaz pre-trained model with only 100 labeled particles. [1]

1 Introduction

Single-particle cryo-electron microscopy (cryo-EM) has emerged as a revolutionary technique for determining the structures of biological macromolecules at near-atomic resolution. In the cryo-EM pipeline, a critical step is the identification and selection of individual protein particles from noisy micrographs [1]. This task can be achieved by the template matching approach. Template matching in cryo-EM involves using a representative image (template) to identify and select particles from raw micrograph images through a cross-correlation. Or, by manually labeling in cryo-EM micrograph every particle visually observed by an experienced user. While template methods are quick and effective when particles match the templates, manual picking, considered a golden standard, is meticulous, labor-intensive, and can have human bias. Recently, machine learning (ML) has been harnessed for particle picking, offering automated particle identification from micrographs [2, 3, 4, 5, 6]. These models benefit from data-driven learning, reducing reliance on pre-set templates or human intervention. Several techniques have been adopted, for example, convolutional neural networks (CNNs) [6, 5, 7], which have been pivotal due to their ability to automatically and adaptively learn spatial hierarchies of features from the micrographs. Among the various architectures, ResNet (Residual Networks) stands out [6, 8]. ResNet deep architecture, characterized by its unique skip

[1] The source code will be available with camera-ready submission.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

connections, enables the training of a large number of layers without the hindrance of vanishing gradients. This depth allows ResNet to generate intricate feature maps. When applied to cryo-EM micrographs, these feature maps can be processed to select maxima peaks, pinpointing the centers of particles, thus providing a robust method for accurate particle identification. However, while this method has achieved impressive results, there is a notable challenge due to the low signal-to-noise ratio (SNR) in cryo-EM data [6]. CNN methods can often generate false-positive picks or even miss certain particles altogether (Figure 1A). Addressing these discrepancies is essential to ensure the robustness of the automated particle-picking process, as well as generate more reliable data that can be used to further improve the particle-picking process. False positives and missed detections in cryo-EM particle picking, mirror issues seen in broader computer vision domains. These issues within the computer vision domain are often tackled with an active learning approach. Active learning is a strategy aimed at training machine learning models more efficiently by selectively acquiring the data that is most beneficial for the model. In the context of CNN, and more specifically image recognition tasks, active learning can significantly reduce the amount of labeled data required to achieve good performance. Initially, a small portion of data is labeled and used to train a basic model. This model is then utilized to analyze the unlabeled data, identifying instances where it is uncertain about the predictions. These uncertain instances are then manually labeled by human annotators and added to the training dataset, incrementally improving the model accuracy. Over several iterations of this process, the CNN becomes progressively better at the task with a much smaller labeled dataset compared to a conventionally trained model.

This paper aims to improve upon existing ML methods for particle picking in cryo-EM by introducing a human-in-the-loop active learning framework. By doing so, we aim to minimize false positive particle prediction. Using the Topaz [6] pre-trained model for feature extraction from cryo-EM micrographs, we train a logistic regression model to iteratively refine its predictions based on user feedback on automatically drawn true and false positive examples. To further simplify this ask we also developed the Napari [9] plugin. The Napari is an open-source, multi-dimensional image viewer for Python, which is designed to make it easy to visualize and annotate large datasets. We utilized it in an active learning framework by enabling users to interactively label or annotate data, where the model predictions are uncertain, thereby facilitating a more efficient training process. Moreover, we demonstrated that our approach allows us to improve the prediction of particles from micrographs in only a few iterations.

The main contributions of this paper are:

1. Introduction of active learning logistic regression framework tailored for cryo-EM particle picking which allows rapid model improvement with automatically labeled data.

2. Empirical evaluations indicate significant advancements in particle-picking accuracy over the existing methods.

3. Development of the interactive framework that harmonizes machine learning with human expertise through iterative feedback loops.

2 Method

We begin our process by utilizing a pre-trained ResNet16 model from Topaz to serve as our initialization point. This pre-trained model provides a feature map with a depth of 128 hidden layers. With this feature map at hand, we apply a non-maximum suppression algorithm [6], enabling us to capture preliminary particle predictions. Based on the model confidence, we subsequently derive regions on the feature map about which the model was least certain. A human operator is then asked to review and refine these initial predictions. Once this initialization phase is concluded, the model proceeds to generate a fresh set of particle center predictions. From these predictions, we select the next set of coordinates that exhibit the highest degree of uncertainty and solicit the human operator to label these selections. This iterative loop, involves model prediction and human correction, till the operator labels N particles. At the culmination of these iterations, the model produces a set of particle coordinates with a confidence score.

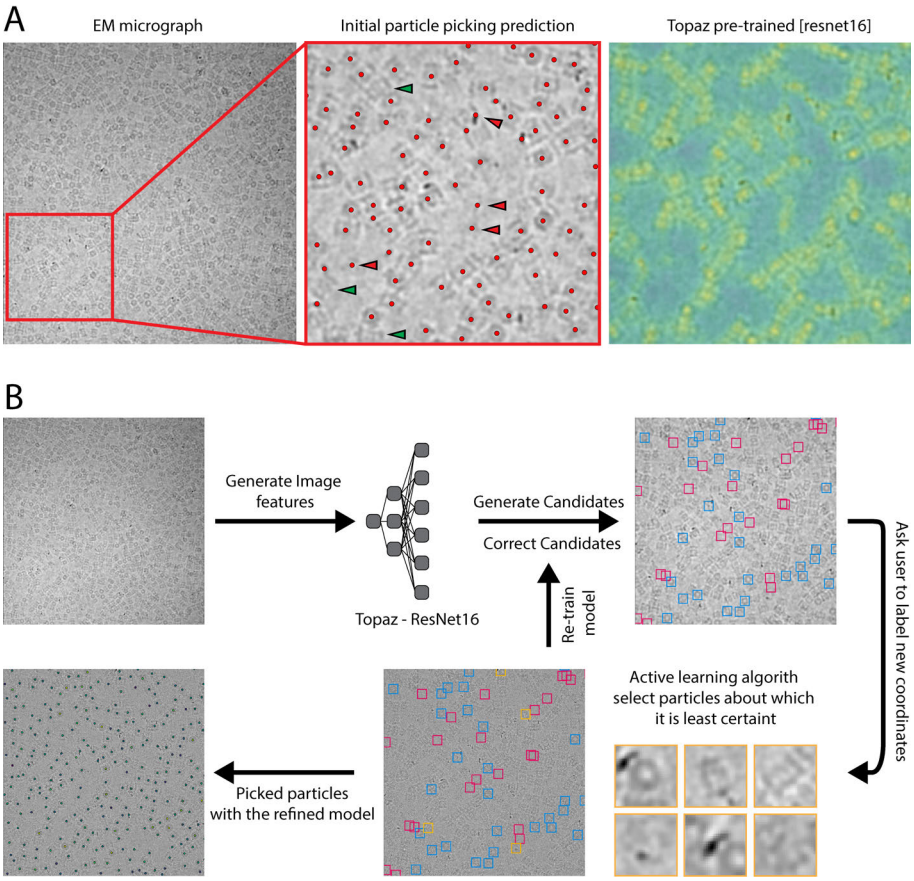

Figure 1: Particle picking and active learning cycle visualization: **A)** A denoised micrograph from the EMP-10025 dataset is depicted, showcasing particles identified by the Topaz ResNet16 model. Particle centers are marked with red circles. Instances of false positive and false negative picks are indicated using red and green arrows, respectively. The corresponding feature map generated by Topaz is displayed adjacent to it. **B)** The active learning process via the Napari plugin is illustrated. Initially, the micrograph is loaded, and the pre-trained ResNet16 model within Topaz is employed to derive feature maps. The model autonomously designates initial positive and negative particles, leaning on its inherent uncertainty measures. The user is subsequently prompted for label adjustments. This iterative feedback loop is typically executed around five times before finalizing the predicted particle centers.

2.1 Active learning strategy and loss function

To improve on the existing method for the particle picking problem from cryo-EM micrographs, we utilize a Binary Logistic Regression model [BLR]. Logistic regression is a simple yet powerful model that can effectively model binary classification tasks. Given its probabilistic interpretation and the ease with which it can be regularized, it forms a suitable choice for this problem. For training the model, we employed the Broyden–Fletcher–Goldfarb–Shanno (BFGS) [10, 11, 12, 13] optimization algorithm. BFGS is a quasi-Newton method that approximates the inverse of the Hessian matrix, which in turn is used to update the model parameters. Given the nature of our loss function and the size of our dataset, BFGS provides a suitable balance between convergence speed and computational efficiency. Our loss function integrates three components (Eq. 4): (1) a Binary Cross Entropy term (Eq. 1), which measures the difference between model predictions and true labels, (2) an L2 regularization (Eq. 2), to constrain model complexity and guard against over-fitting, (3) penalty on the expected π (Eq. 3), to ensure that the model does not over-fit to the labeled data and respects the prior knowledge about the data distribution, improving its generalization capability.

In our active learning approach, the core principle revolves around the entropy-based acquisition function (Eq. 5), which assists in determining the data points that should be labeled next. Entropy is a measure of uncertainty, and in the context of our model, locations with higher entropy represent areas where the model exhibits the greatest uncertainty in its predictions. To compute this, we utilize the function that first calculates the logarithmic probabilities of the positive and negative classes using the logarithmic sigmoid function on the model output logits. Subsequently, entropy is derived from these log probabilities. To ensure diverse sampling and avoid multiple selections of the same candidate locations, a peak finding algorithm is incorporated on top of the entropy scores. Once identified, these peaks, representing regions of high uncertainty, are sorted and used to ask a user to label as positive or negative peaks of higher uncertainty.

2.2 Napari plugin for particle picking.

To seamlessly integrate human expertise with the automated predictions of the Topaz software, we introduced a dedicated Napari plugin tailored to our active learning approach. This tool empowers users to make initial predictions with Topaz, harnessing its pre-trained ResNet16 model, and subsequently refine these predictions by guiding our BLR model (as illustrated in Figure 2). When a user loads a micrograph into the tool, it first undergoes automatic pre-processing described in the following section 3. Following this, the plugin showcases an initial selection of particles and accompanies the user with label generation. This allows for interactive correction of these selections and the generation of the final predictions.

3 Experiments

Dataset. A subset of aligned and summed micrographs and star files containing published particle sets were retrieved from the Electron Microscopy Public Image Archive (EMPIAR) for EMPIAR-10025 dataset [14]. The summary of the used dataset can be found in the table 3. For the ground truth, curated in-house particle sets were provided by the New York Structural Biology Center with an average of $\sim$ 550 particles. Micrographs for each dataset were scaled from 0.98 Å to 7.84 Å resolution and normalized with the mean/standard deviation normalization [6]. The ResNet16 model was fine-tuned on randomly selected 10, 100, 250, 500, and 1000 particles from only 2 randomly selected micrographs from our dataset (Table 3). We fine-tuned the entire ResNet16 model and only the final linear classifier layer, which is responsible for rendering the ultimate classification decisions. The optimization during these training scenarios was conducted using the Adam optimizer, amalgamating binary cross-entropy and L2 loss functions, following the methodology outlined in [6]. For the active learning approach, we trained the model by iteratively correcting 1 or 10 particles at a time selected by an algorithm based on the confidence score. The active learning model was trained on the same 2 micrographs from our dataset (Table 3) by labeling 10, 25, 50, and 100 particles indicated by the algorithm.

Evaluation. In order to evaluate the robustness and efficiency of an active learning approach, we employed the pre-trained ResNet16 model from Topaz as our baseline model. As the baseline we used standard pre-trained Topaz ResNet16 model that is reguallry used for 2D particle picking. To evaluate how our active learning method improves particle prediction, all models were evaluated using the average precision (AP) score [6]. Additionally, we also reported an average precision score at the recall of 90% [AP90] to verify how well our model perform on selecting only valid target particles in comparison to our baseline and fine-tuned models. The AP score quantifies the capability of the model to rank the predicted regions of a micrograph based on the likelihood of containing a particle. It corresponds to the area beneath the precision-recall curve. The AP is derived by iterating through the ranked micrograph regions, calculating the precision at 'k' predictions, and multiplying it by the change in recall from the preceding rank. Consequently, the AP encapsulates the models proficiency in discerning and ranking micrograph regions by their particle presence probability. For this, we first extracted predicted particle coordinates with a non-maximum suppression algorithm [6] and their associated predicted probabilities. This allowed us to extract coordinates for both the pre-trained and fine-tuned Topaz ResNet16 model as well as prediction after each active learning iteration. We also measure the true/false positive ratio for each threshold of predicted probability by counting the number of true positive particles with centers within the particle radius of a particle center in the EMPIRE-10025 dataset.

Experiment	0	10	25	50	100	250	500	1000
Baseline	0.624	-	-	-	-	-	-	-
Topaz [fine-tuned model]	-	0.577	-	-	0.660	0.773	0.776	0.768
Topaz [fine-tuned classifier]	-	0.626	-	-	0.671	0.773	0.774	0.774
BLR model [Batch = 1]	-	0.662	0.711	0.853	0.849	-	-	-
BLR model [Batch = 10]	-	0.762	0.794	0.816	**0.863**	-	-	-

Table 1: Comparison of an average precision metric against a number of labeled particles. We compared the baseline model (pre-trained ResNet16 model from Topaz) with fine-tuned Topaz models. We could observe that fine-tuned models gradually improve AP scores with the increase of labeled particles. We further compared our baseline with the BLR model train by selecting 1 or 10 particles at each training iteration. We could observe that the BLR model AP increases with only 10 labeled particles which was equivalent to re-training the ResNet16 model with 100-250 labeled particles.

Benchmark results. The metric of the true/false positive is shown in the figure 3. Our proposed active learning method allowed for on average $\sim 50\%$ reduction of false positive predictions compared to the pre-trained and fine-tuned ResNet16 models. This was achieved by labeling only 100 particles which were selected using an active learning approach. We also measure how our approach compares to the fine-tuned Topaz ResNet16 model by training it with a different number of labeled particles. We could observe that in both cases where we fine-tuned the entire ResNet16 and only the last classifier layer, we could observe a gradual increase in AP (Table 1) and AP90 (Table 2) corresponding to fine-tuning with a higher number of label particles. Interestingly we could observe that when we fine-tuned ResNet16 with only 10 particles, a model showed a reduction in AP and AP90 which in our opinion could be related to the over-fitting of the fine-tuned ResNet16 model. Next, we evaluated how our active learning approach compares against the pre-trained ResNet16 model from Topaz. For our active learning framework, when we train the model by labeling 1 particle at a time, we observe that the AP increases from the baseline after only labeling 10 particles, we could observe this trend continuing till we labeled 100 particles which yielded $\sim 26\%$ increase in the AP metric and $\sim 10\%$ in AP90. Moreover, we could observe that to achieve similar results with fine-tuned ResNet16, we would need to re-train it with up to 1000 particles.

4 Conclusion

We propose active learning approach with the logistic regression model to improve particle picking prediction. Leveraging the foundation of pre-trained CNN models from Topaz, we have established a proof-of-concept that integrates human expertise directly into the refining process. Through this approach, we achieved a commendable $\sim 26\%$ enhancement in AP and $\sim 10\%$ in AP90 metrics with minimal iterations and significantly reduced human intervention, by improving model prediction and reducing the number of false positive predictions. Moreover, our approach also significantly reduces the time needed for labeling particles and re-training the pre-trained ResNet16 model. We are planning to further develop this method and include active learning with the logistic regression model with template matching to further improve accuracy as well as to equip our model to simultaneously predict diverse particle types. In our observations, the AP90 metric exhibited noticeable variability with the increment in the number of labeled particles. This phenomenon is hypothesized to stem from a diversity deficit, potentially linked to the BCE term in our loss function, which predominantly selects high-entropy samples. To enhance the robustness and effectiveness of our BLR model, future endeavors will concentrate on a more strategic exploitation of the diversity inherent in unlabeled data. Moreover, the development of the Napari plugin improves the current method for the generation of human-curated data by allowing for the quick correction and refinement of the picked particles. We expect that this approach will allow us to generate curated data with a human-in-the-loop approach and will be instrumental in addressing the challenges in particle picking as well as improving the reconstruction results. With this paper, we are releasing the code and the Napari plugin which will allow users to use our approach for 2D particle picking. We are working on extending this approach to 3D particle picking in the near future.

References

[1] J. L. Vilas, J. M. Carazo, and C. O. S. Sorzano. Emerging themes in cryoem single particle analysis image processing. *Chemical Reviews*, 122(17):13915–13951, 2022.

[2] V. Abrishami, A. Zaldívar-Peraza, J. M. de la Rosa-Trevín, J. Vargas, J .and Otón, R. Marabini, Y. Shkolnisky, J. M. Carazo, and C. O. S. Sorzano. A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs. *Bioinformatics*, 29(19):2460–2468, 2013.

[3] D. Kimanius, B. Forsberg, S. Scheres, and E. Lindahl. Accelerated cryo-em structure determination with parallelisation using gpus in relion-2. *eLife*, 5(e18722):e18722, 2016.

[4] F. Wang, H. Gong, G. Liu, M. Li, C. Yan, T. Xia, X. Li, and J. Zeng. Deeppicker: A deep learning approach for fully automated particle picking in cryo-em. *Journal of Structural Biology*, 195(3):325–336, 2016.

[5] T. Wagner, F. Merino, M. Stabrin, T. Moriya, C. Antoni, A. Apelbaum, P. Hagel, O. Sitsel, T. Raisch, D. Prumbaum, D. Quentin, D. Roderer, S. Tacke, B. Siebolds, E. Schubert, T. Shaikh, P. Lill, C. Gatsogiannis, and S. Raunser. Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. *Communication Biology*, 2(218), 2019.

[6] T. Bepler, A. Morin, M. Rapp, J. Brasch, L. Shapiro, A. J. Nobel, and B. Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 27(2):1153–1160, 2019.

[7] A. Al-Azzawi, A. Ouadou, H. Max, T. Duan, J. Tanner, and J. Cheng. Deepcryopicker: fully automated deep neural network for single protein particle picking in cryo-em. *BMC Bioinformatics*, 21(509), 2020.

[8] X. Zhang, T. Zhao, J. Chen, Y. Shen, and X. Li. Epicker is an exemplar-based continual learning approach for knowledge accumulation in cryoem particle picking. *Nature Communications*, 13(1), 2022.

[9] J. Ahlers, D. Althviz Moré, O. Amsalem, A. Anderson, G. Bokota, P. Boone, J. Bragantini, G. Buckley, A. Burt, M. Bussonnier, A. Can Solak, C. Caporal, D. Doncila Pop, K. Evans, J. Freeman, L. Gaifas, C. Gohlke, K. Gunalan, H. Har-Gil, M. Harfouche, K. I. S. Harrington, V. Hilsenstein, K. Hutchings, T. Lambert, J. Lauer, G. Lichtner, Z. Liu, L. Liu, A. Lowe, L. Marconato, S. Martin, A. McGovern, L. Migas, N. Miller, H. Muñoz, J. H. Müller, C. Nauroth-Kreß, J. Nunez-Iglesias, C. Pape, K. Pevey, G. Peña-Castellanos, A. Pierré, J. Rodríguez-Guerra, D. Ross, L. Royer, C. T. Russell, G. Selzer, P. Smith, P. Sobolewski, K. Sofiiuk, N. Sofroniew, D. Stansby, A. Sweet, W. M. Vierdag, P. Wadhwa, M. Weber Mendonça, J. Windhager, P. Winston, and K. Yamauchi. napari: a multi-dimensional image viewer for python. *Zendo*, 2019.

[10] C. G. Broyden. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6(1):76–90, 1970.

[11] R. Fletcher. A new approach to variable metric algorithms. *Computer Journal*, 13(3):317–322, 1970.

[12] D. Goldfarb. A family of variable metric updates derived by variational means. *Mathematics of Computation*, 7(1):23–26, 1970.

[13] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(1):647–656, 1970.

[14] M. G. Campbell, D. Veesler, A. Cheng, C. S. Potter, and B. Carragher. 2.8 a resolution reconstruction of the thermoplasma acidophilum 20s proteasome using cryo-electron microscopy. *eLife*, 4(e06380), 2015.

5 Supplementary Material

Experiment	0	10	25	50	100	250	500	1000
Baseline	0.825	-	-	-	-	-	-	-
Topaz [fine-tuned model]	-	0.806	-	-	0.822	0.969	0.943	**0.998**
Topaz [fine-tuned classifier]	-	0.801	-	-	0.840	0.870	0.878	0.883
BLR model [Batch = 1]	-	0.876	0.948	0.867	0.891	-	-	-
BLR model [Batch = 10]	-	0.907	0.876	0.853	0.888	-	-	-

Table 2: Comparison of an average precision with recall 90% metric against a number of labeled particles. This metric was selected to visualize how well the presented model predicts the confidence score for each particle. We could observe that re-trained Topaz models show gradual improvement with the increased number of labeled particles. We further compared our baseline with the BLR model train by selecting 1 or 10 particles at each training iteration. We could observe that the BLR model AP90 score increased by 7% and 10% with only labeled 10 particles, respectively. This was equivalent to re-training the ResNet16 model with $\sim$ 250-1000 particles.

Table 3: Summary of cryo-EM EMPIAR-10025 datasets used for refining Topaz ResNet16 model and evaluation of active learning method. For this small dataset, we selected random 8 micrographs. We chose to train and evaluate on such a small dataset to evaluate how well our BLR model performs in comparison to human-annotated ground truth data. This micrograph consists of around $\sim$ 550 particles with $\sim$ 1 A resolution.

Micrograph name	Number of particles	Split
14sep05c-c000024sq-00003hl-00002es	576	Train
14sep05c-c00024sq-00003hl-00005es	559	Train
14sep05c-c00024sq-00006hl-00003es	459	Evaluation
14sep05c-c00003gr-00014sq-00002hl-00005es	557	Evaluation
14sep05c-c00003gr-00014sq-00004hl-00004es	511	Evaluation
14sep05c-c00003gr-00014sq-00005hl-00002es	617	Evaluation
14sep05c-c00003gr-00014sq-00005hl-00003es	581	Evaluation
14sep05c-c00003gr-00014sq-00005hl-00005es	576	Evaluation

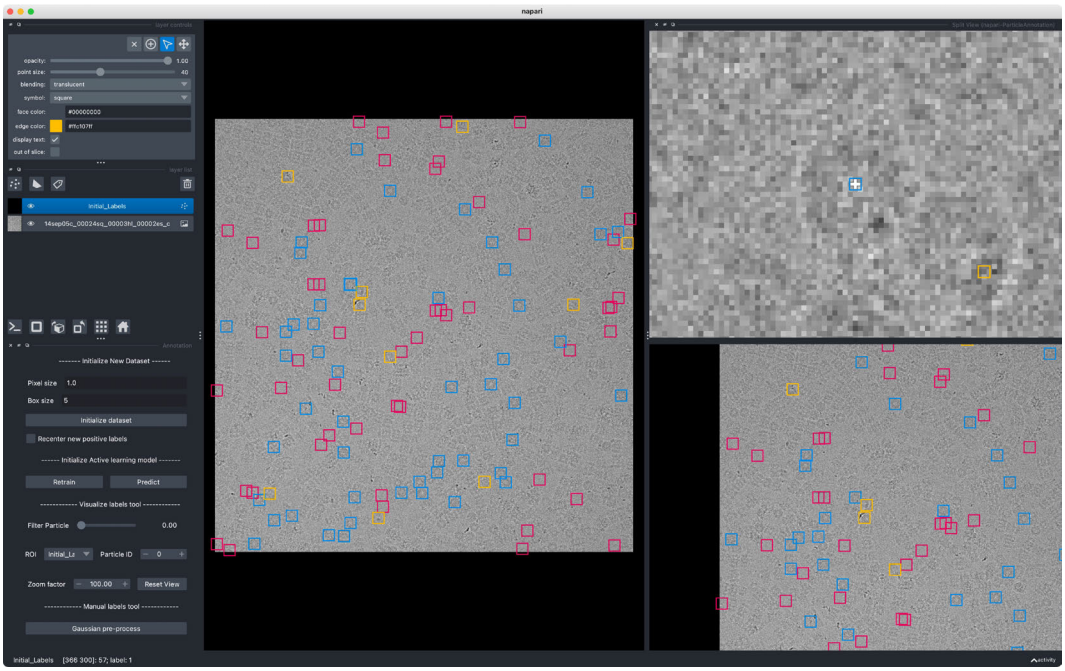

Figure 2: The screenshot from Napari software, demonstrating our Active learning plugin for particle picking. Blue and red squares show true and false positive examples, respectively. Yellow squares show examples of which model was least certain, and the user is asked to correct it. The user can iterate several times re-training the BLR model. At the end of this process by pressing the 'predict' button user can retrieve a fine-tuned prediction about the particle positions. Additionally to this, the user can load and save the pre-trained model and use it to immediately predict particle positions.

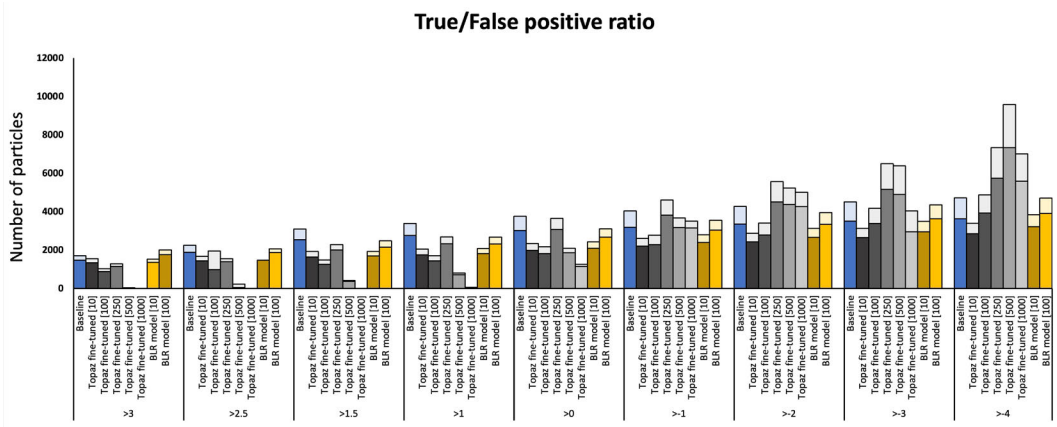

Figure 3: True/False positive ratio for particle prediction from Topaz ResNet16, active learning approach, and fine-tuned Topaz model with different particle numbers. In the bar plot, we represent the number of true positives with a darker color and false positives with a lighter color. The blue bars are associated with our baseline, the gray bars are correlated with the fine-tuned Topaz models, and the orange color is associated with our BLR model.

Equation 1: Binary Cross-Entropy Loss

$$L_{\text{binary}} = -\sum_i [y_i \log(x_i) + (1 - y_i) * \log(1 - x_i)] \tag{1}$$

Where (σ):

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Equation 2: L2 Regularization

$$L_{\text{reg\_L2}} = \frac{\lambda}{2} \sum_i w_i^2 \tag{2}$$

Equation 3: Penalty on Expected π

$$L_\pi = \text{pi\_weight} \times (\log(p) - \log(1 - p) - \text{pi\_logit})^2 \tag{3}$$

Where p and $1 - p$ are the probabilities of the positive and negative classes respectively.

Equation 4: Combined loss function

$$L = \frac{L_{\text{binary}} + L_{\text{reg\_L2}} + L_\pi}{n} \tag{4}$$

Equation 5: Entropy-based acquisition function

$$H(logits) = \text{-}\sigma(logits) * \log(\sigma(logits)) - \sigma(\text{-}logits) * \log(\sigma(\text{-}logits)) \tag{5}$$