Automatic particle picking with minimal supervision using positive-unlabeled neural networks

Massachusetts Institute of Technology

CSBi

Tristan Bepler
Berger Lab, MIT
NYSBC Deep Learning Workshop
April 10, 2018

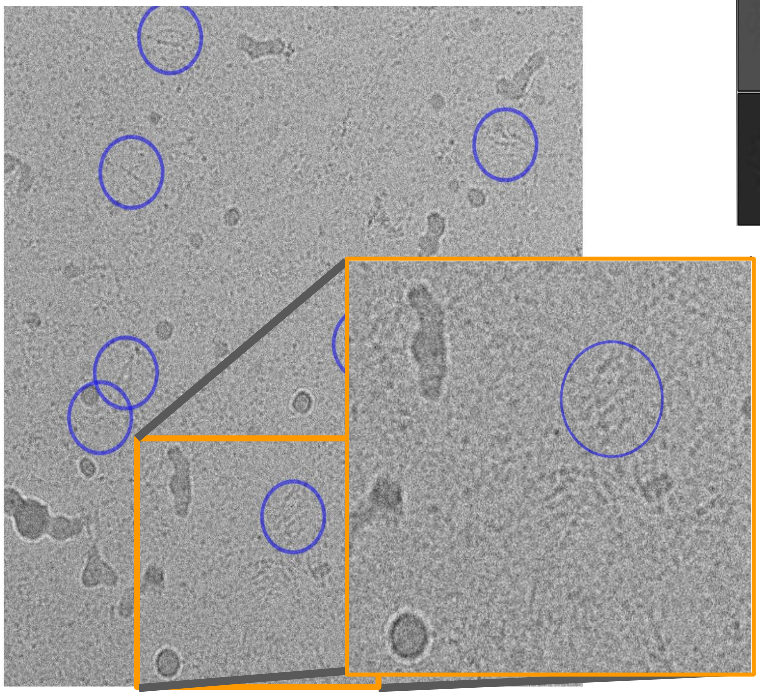
SIMONS ELECTRON MICROSCOPY CENTER

CSAIL

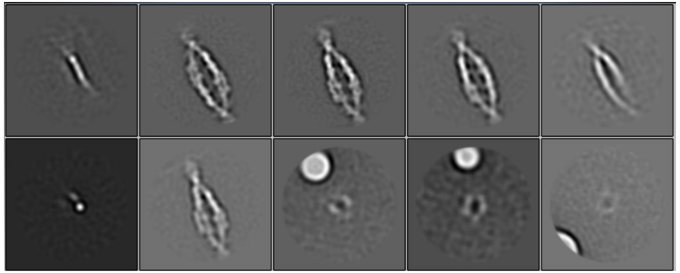
Single-particle cryo-electron microscopy

- Sample preparation
- 2D Imaging
- **Particle detection**
- 2D averaging
- 3D reconstruction

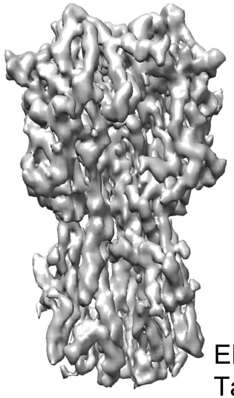
Particle picking



2D averages

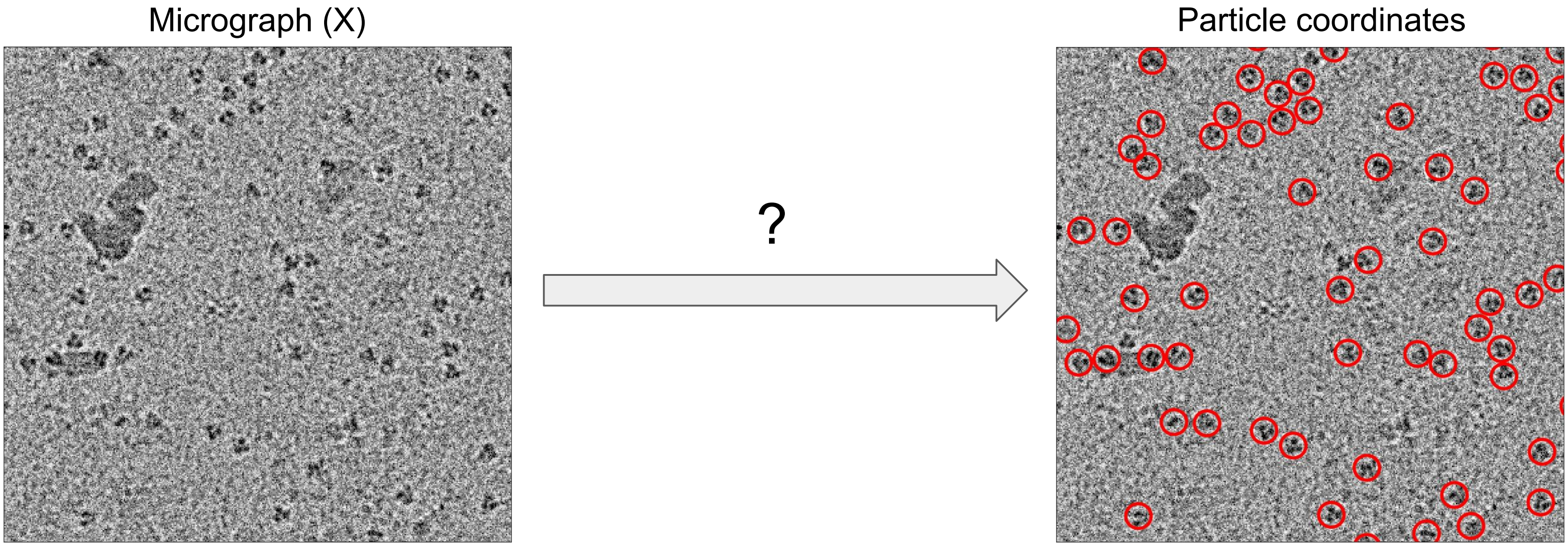


3D reconstruction
(not the same protein)



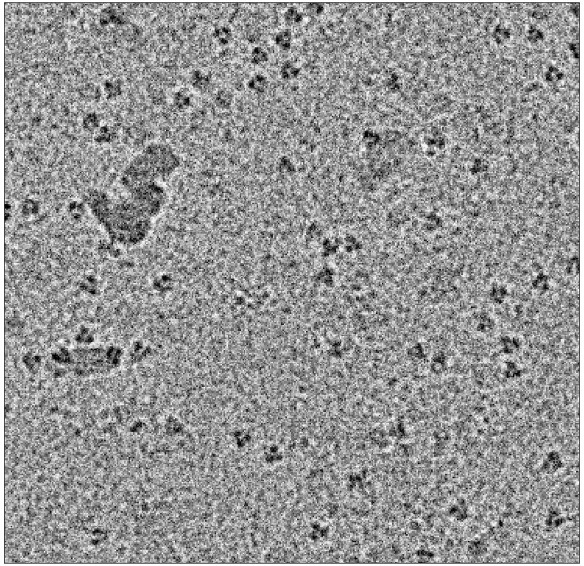
EMPIAR-10096
Tan et al., 2017

Particle picking

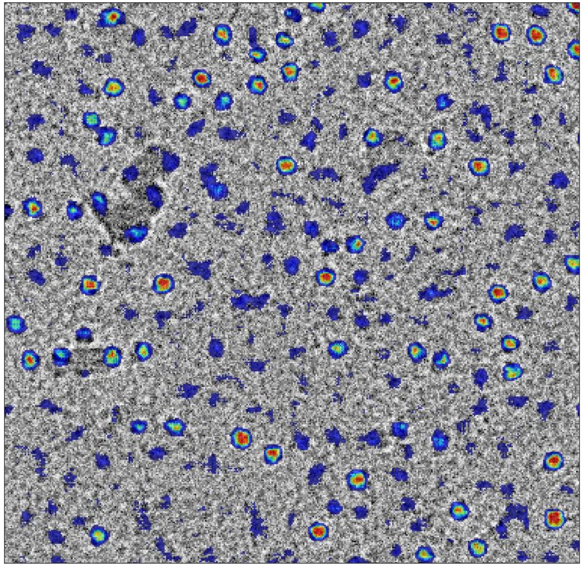


Particle picking

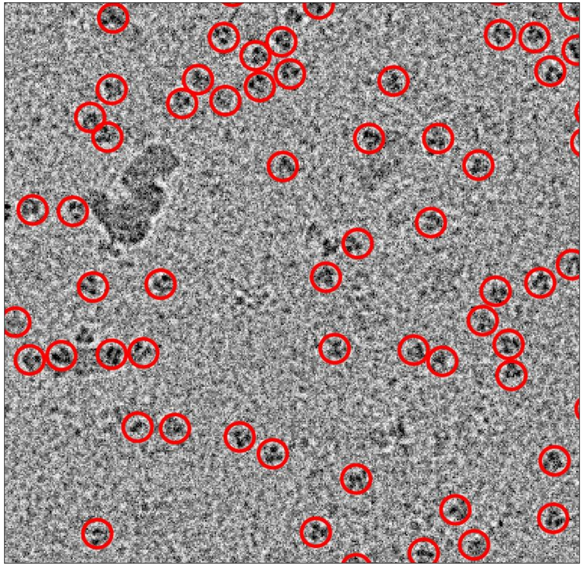
Micrograph (X)



Region scores ($g \ast X$)



Particle coordinates

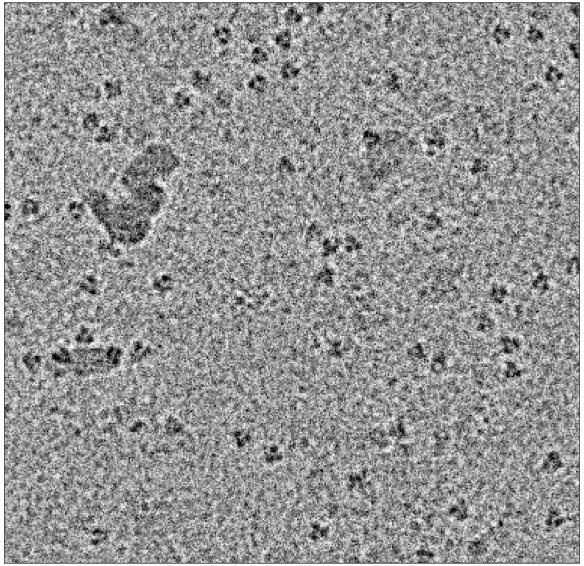


Given a scoring function, *g*, convolve it over the micrograph, *X*, to get per region scores

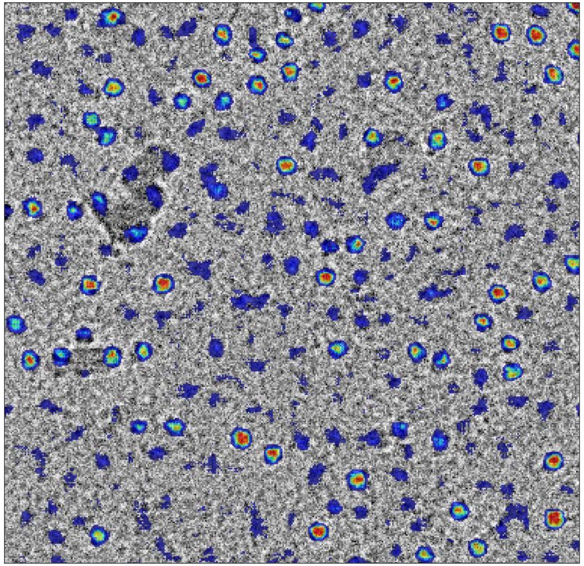
Extract coordinates by greedily selecting regions and removing nearby regions (non-maximum suppression)

Particle picking

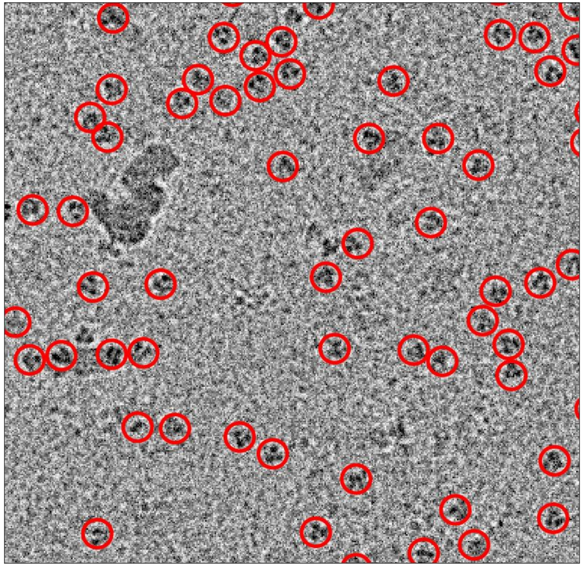
Micrograph (X)



Region scores ($g \ast X$)



Particle coordinates



What should g be?

Convolutional neural network

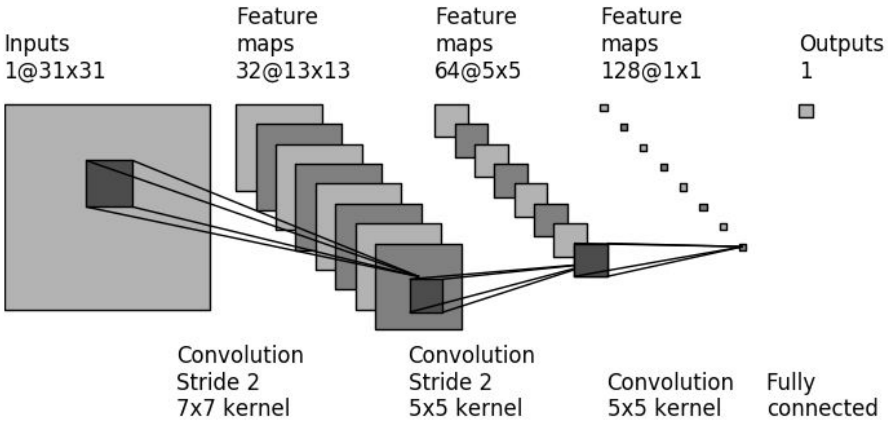
- Learn parameters, θ, of g from data

  Positives: X_P , y=1 , π_P

  Negatives: X_N , y=0 , 1 - π_P

  Loss function L

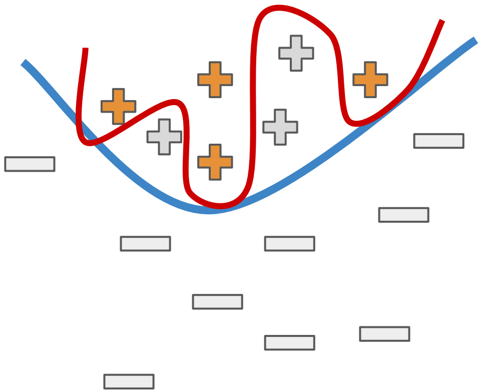
  $$\underset{\theta}{\operatorname{argmin}}\ \pi_P \underset{X_P}{E}[L(g(x,\theta),1)] + (1 - \pi_P)\underset{X_N}{E}[L(g(x,\theta),0)]$$



- Problem: learning requires large amounts of labeled examples
  - Costly for a researcher to label enough particles
  - Can we learn θ from a small amount of labeled data and the rest of the unlabeled data?

Positive-unlabeled classification

- Learn parameters, θ, from positive, X_P, and unlabeled, X_U, data
- Unlabeled data contains both positive and negative examples
- Loss function: find parameters that minimize this function of the training data
- Assume we know $\boldsymbol{\pi}_P$



Naive:

$$\boldsymbol{\pi}_P \mathop{E}_{X_P}[L(g(x),1)] + (1 - \boldsymbol{\pi}_P)\mathop{E}_{X_U}[L(g(x),0)]$$

Unbiased estimator (du Plessis et al. 2016):

$$\boldsymbol{\pi}_P \mathop{E}_{X_P}[L(g(x),1)] - \boldsymbol{\pi}_P \mathop{E}_{X_P}[L(g(x),0)] + \mathop{E}_{X_U}[L(g(x),0)]$$

Non-negative estimator (Kiryo et al. 2017):

$$\boldsymbol{\pi}_P \mathop{E}_{X_P}[L(g(x),1)] + \max\{\, 0, \mathop{E}_{X_U}[L(g(x),0)] - \boldsymbol{\pi}_P \mathop{E}_{X_P}[L(g(x),0)] \,\}$$

Generalized expectation criteria (KL) (Mann and McCallum 2010):

$$\mathop{E}_{X_P}[L(g(x),1)] + \lambda\, \mathrm{KL}(\boldsymbol{\pi}_P, \mathop{E}_{X_U}[g(x)])$$

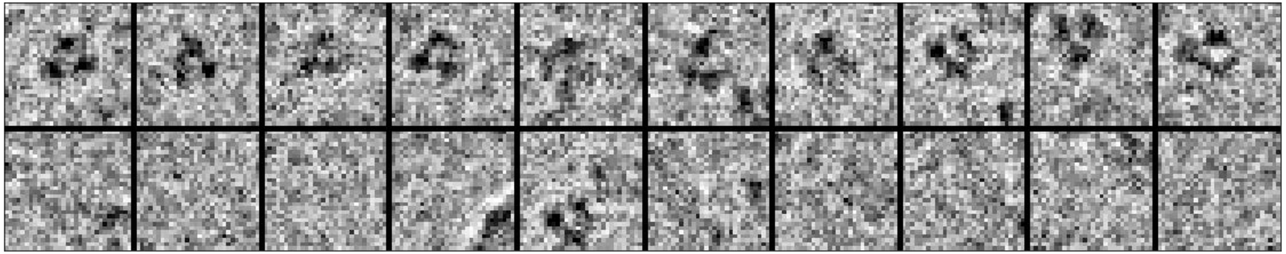
GE-binomial: a better GE criteria for positive-unlabeled learning with SGD

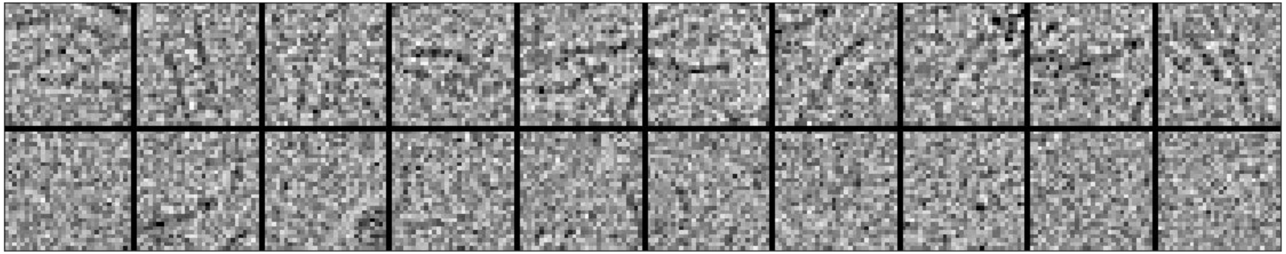
- Problem: neural network needs to be trained with **minibatch stochastic gradient descent (SGD)** - need to estimate gradient using samples of data
  - $\mathop{E}_{X_P}[L(g(x),1)] + \lambda\, KL(\pi_P, \mathop{E}_{X_U}[g(x)])$
- The number of positive data points, P, in an N data point minibatch follows the **binomial distribution with probability of success π_P**
  - $p_k = \text{binomial}(N, \pi_P)$
- Classifier predictions, $g(x_i)$ where x_i are unlabeled data points in the minibatch, also define a distribution over the number of positives - approximate this with a normal distribution
  - $\mu = \Sigma g(x_i)$ and $\sigma^2 = \Sigma g(x_i)(1-g(x_i))$
  - let q_k be the discretized probability of k positives given by this distribution
- Define a **new GE criteria (GE-binomial)** using these distributions: $\Sigma q_k \log(p_k)$

$$\mathop{E}_{X_P}[L(g(x),1)] + \Sigma q_k \log(p_k)$$

CryoEM datasets for evaluation

Dataset	Train		Test	
	Images	Particles	Images	Particles
EMPIAR-10096	347	100465	100	29535
Shapiro-lab	67	1167	20	373





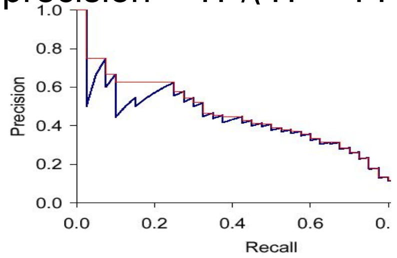
GE-binomial outperforms other positive-unlabeled learning objectives on cryoEM particle classification

	EMPIAR-10096			Shapiro-lab		
Model	**10**	**100**	**1000**	**10**	**100**	**1167**
classifier						
PN	0.072 ±0.029	0.187 ±0.038	-	0.012 ±0.005	0.036 ±0.012	-
NNPU	0.101 ±0.035	0.226 ±0.033	-	0.014 ±0.008	0.039 ±0.013	-
GE-KL	0.072 ±0.035	0.240 ±0.043	-	0.010 ±0.005	**0.062 ±0.017**	-
GE-binomial	**0.155 ±0.044**	**0.258 ±0.040**	**0.392 ±0.006**	**0.020 ±0.008**	0.061 ±0.010	**0.150 ±0.008**

Area under the precision-recall curve on the test set for models trained with subsets of positives from the training set

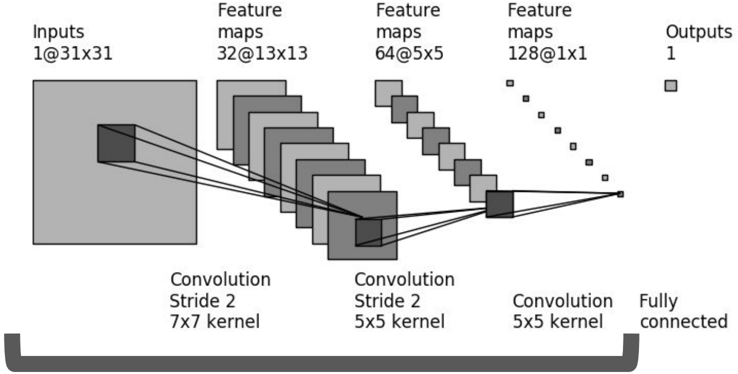
PN = naive
NNPU = non-negative estimator (Kiryo et al. 2017)
GE-KL = GE criteria with KL-divergence

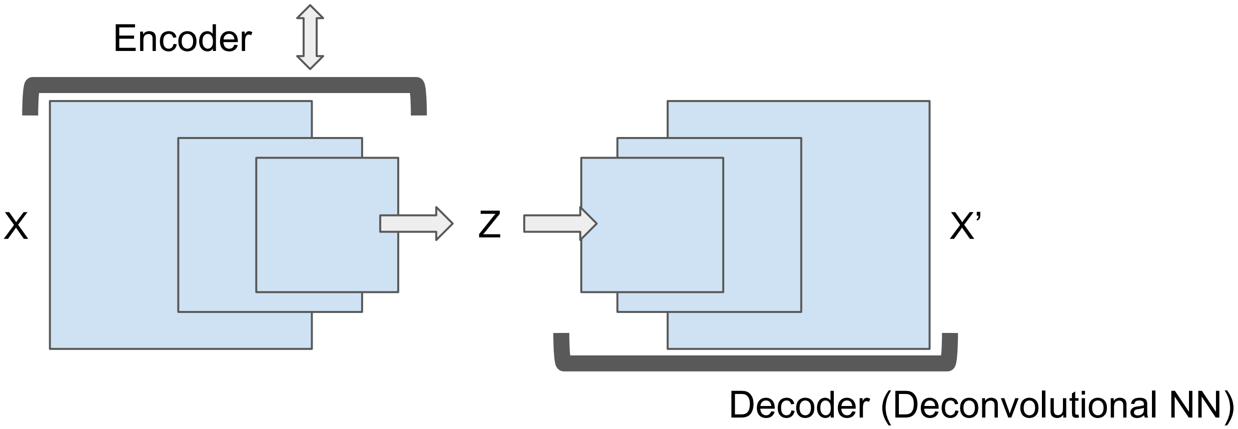
recall = TP/(TP + FN)
precision = TP/(TP + FP)



https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html

Hybrid classifier-autoencoder model





GE-binomial and autoencoder:

$$\underbrace{E_{X_P}[L(g(x),1)] + \Sigma q_k \log(p_k)}_{\text{GE-binomial loss}} + \underbrace{a||x - x'||^2}_{\text{reconstruction error}}$$

Including a decoder and reconstruction error improves generalization with few training examples

Model	EMPIAR-10096			Shapiro-lab		
	10	100	1000	10	100	1167
classifier						
PN	0.072 ±0.029	0.187 ±0.038	-	0.012 ±0.005	0.036 ±0.012	-
NNPU	0.101 ±0.035	0.226 ±0.033	-	0.014 ±0.008	0.039 ±0.013	-
GE-KL	0.072 ±0.035	0.240 ±0.043	-	0.010 ±0.005	**0.062 ±0.017**	-
GE-binomial	**0.155 ±0.044**	**0.258 ±0.040**	**0.392 ±0.006**	**0.020 ±0.008**	0.061 ±0.010	**0.150 ±0.008**
+autoencoder						
GE-binomial	0.260 ±0.016	0.324 ±0.017	0.368 ±0.013	0.029 ±0.011	0.078 ±0.018	0.120 ±0.006

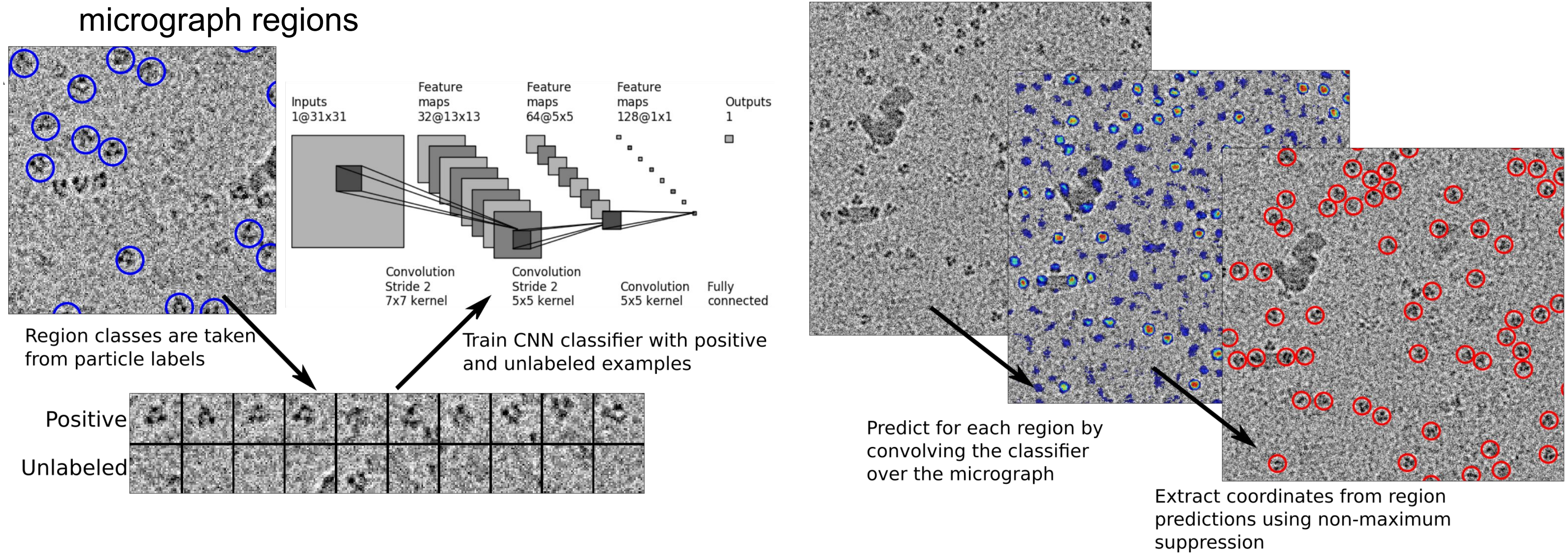
PN = naive
NNPU = non-negative estimator (Kiryo et al. 2017)
GE-KL = GE criteria with KL-divergence

Adding a decoder ($a = N/10$) can further improve classification performance when very few labeled data points are available

Topaz particle picking pipeline

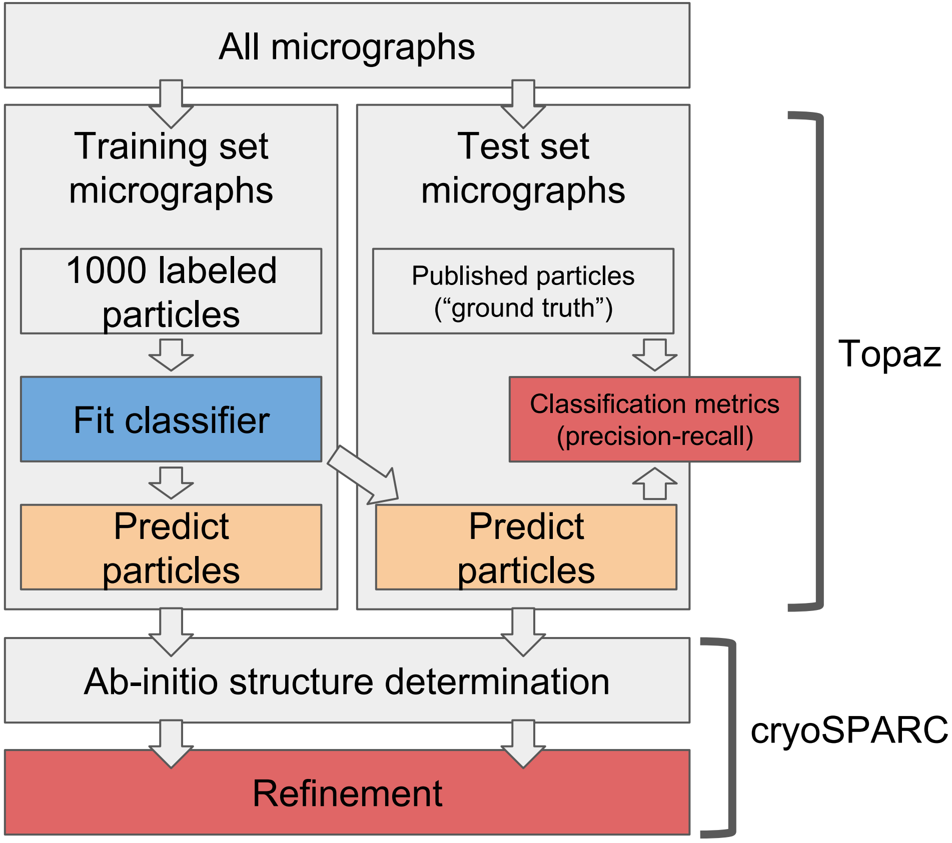
Train classifier with positive and unlabeled micrograph regions

Score micrograph regions and extract predicted particle coordinates

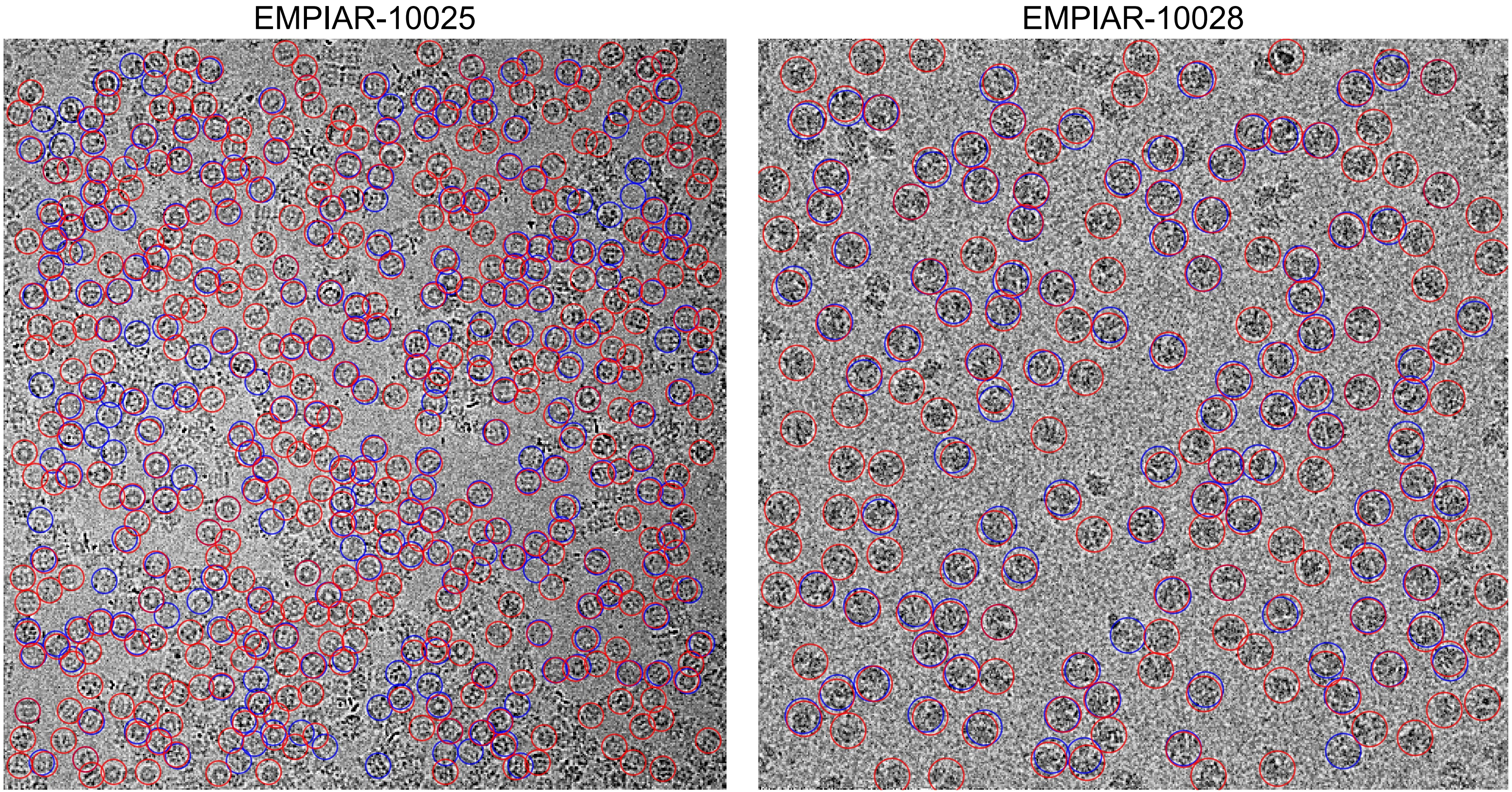


Structure determination with predicted particles

- 2 new datasets: EMPIAR-10025 (T20S proteasome) and EMPIAR-10028 (80S ribosome)
- 20% of micrographs held out for testing particle detection
- 1000 labeled training particles
- Predicted particles selected at decreasing score thresholds
- Ab-initio structure determination and refinement performed with each particle set with cryoSPARC
- *No post-processing of predicted particles (no 2D/3D classification)*



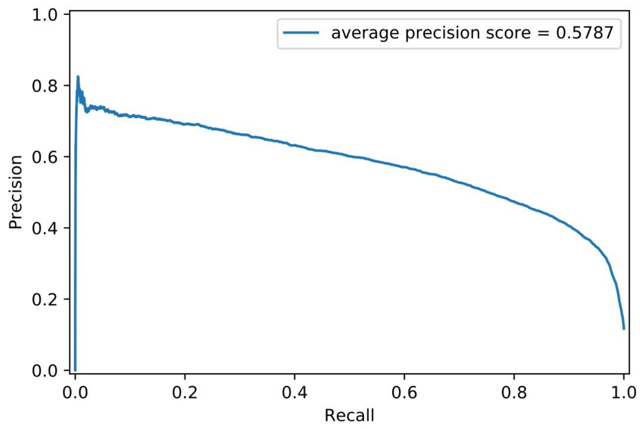
Example test set micrographs show extra predicted particles are true particles

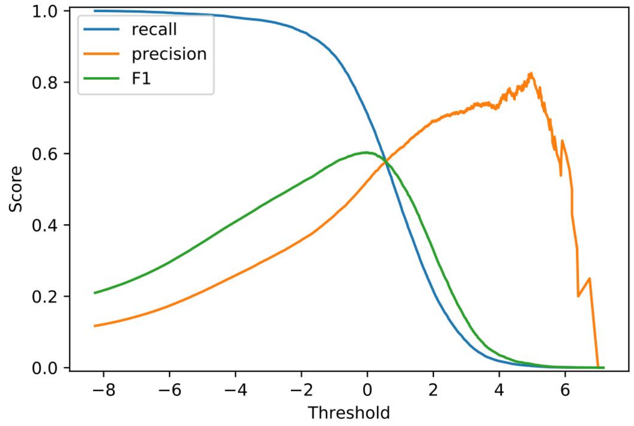


Red: predicted particles, Blue: published particles

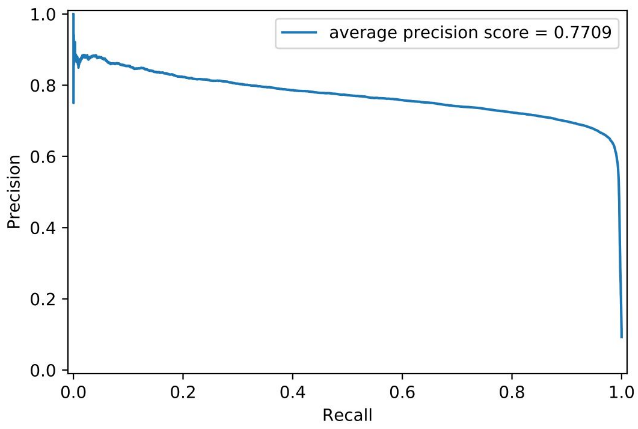
Models detect test set particles with good average precision scores

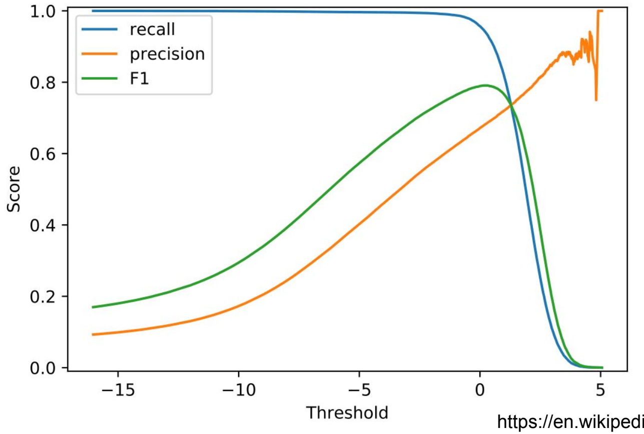
EMPIAR-10025





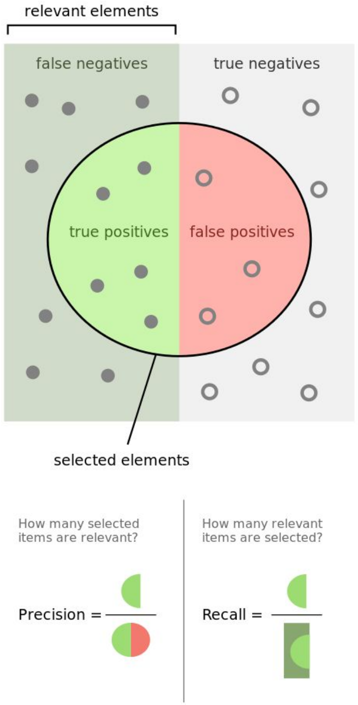
EMPIAR-10028





recall = TP/(TP + FN)

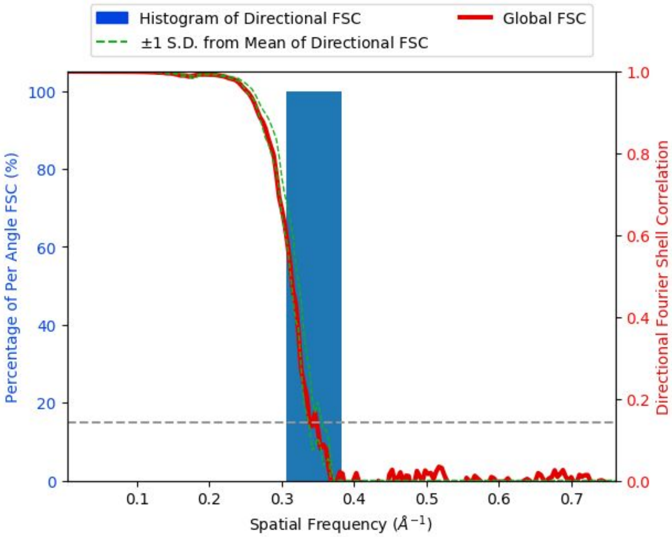
precision = TP/(TP + FP)



https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg

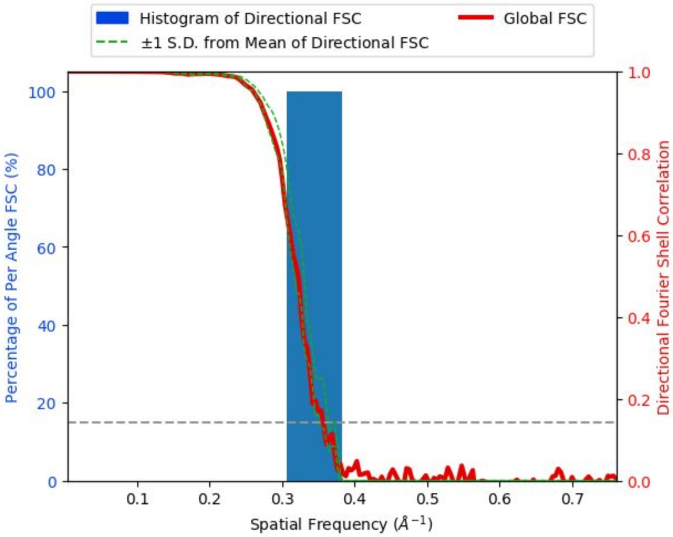
2.8 Å reconstruction of EMPIAR-10025
(without dose weighting)

Published particle set



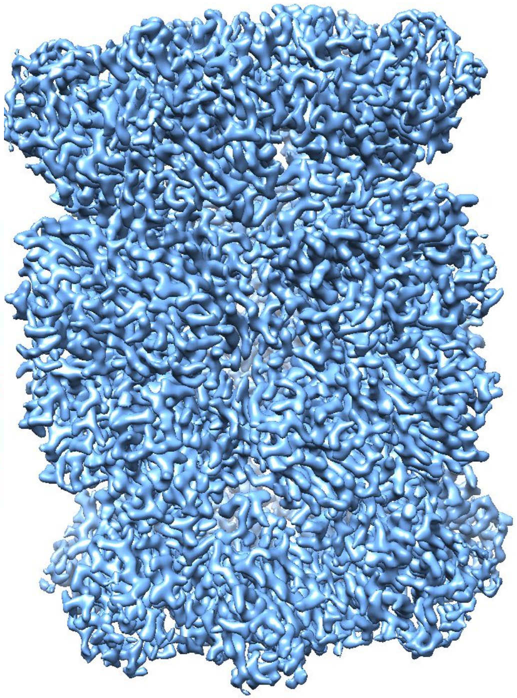
Sphericity = 0.988
Global resolution = 2.99 Å
Number of particles = 49954

Predicted particle set
(best threshold)



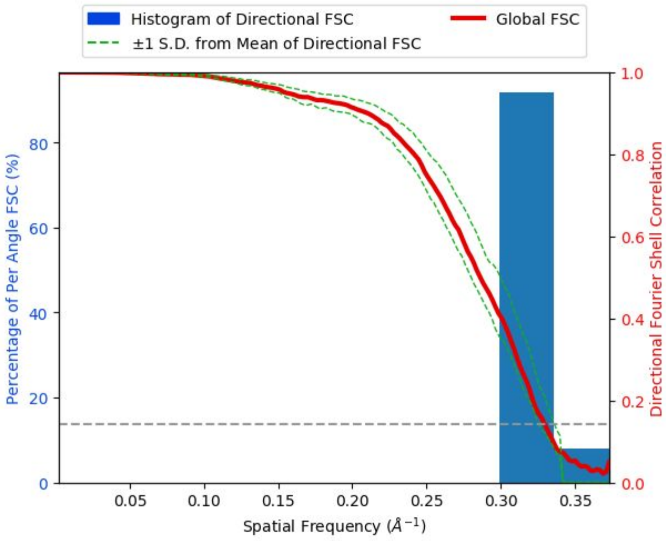
Sphericity = 0.982
Global resolution = 2.83 Å
Number of particles = 160658

Predicted 3D structure



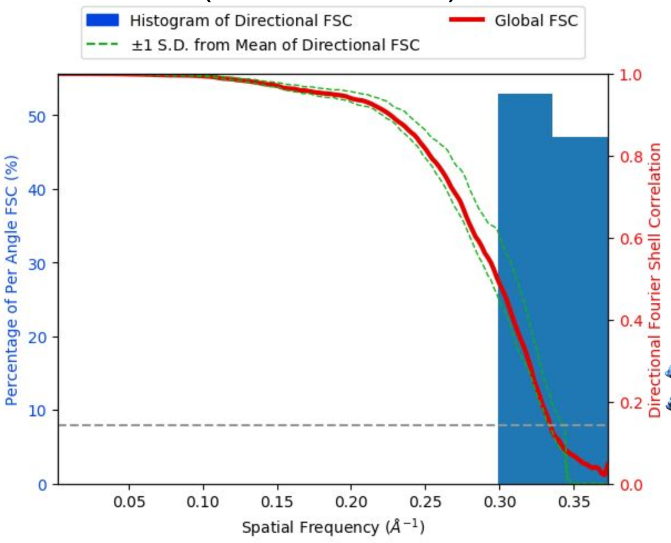
3.0 Å reconstruction of EMPIAR-10028

Published particle set



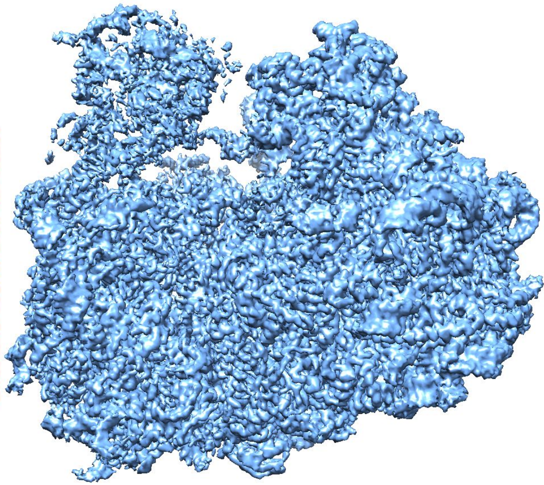
Sphericity = 0.958
Global resolution = 3.05 Å
Number of particles = 105247

Predicted particle set
(best threshold)



Sphericity = 0.974
Global resolution = 3.00 Å
Number of particles = 180723

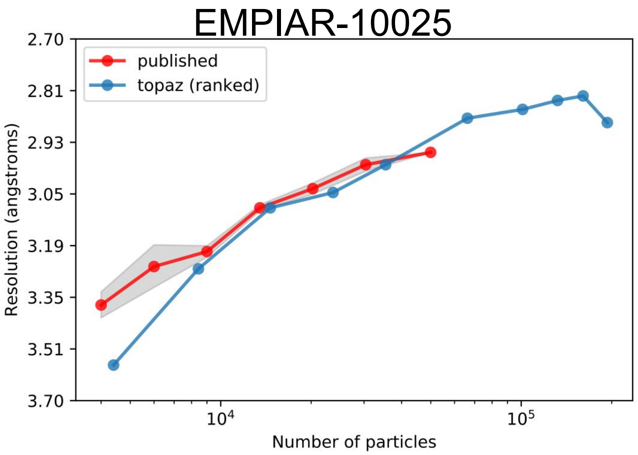
Predicted 3D structure

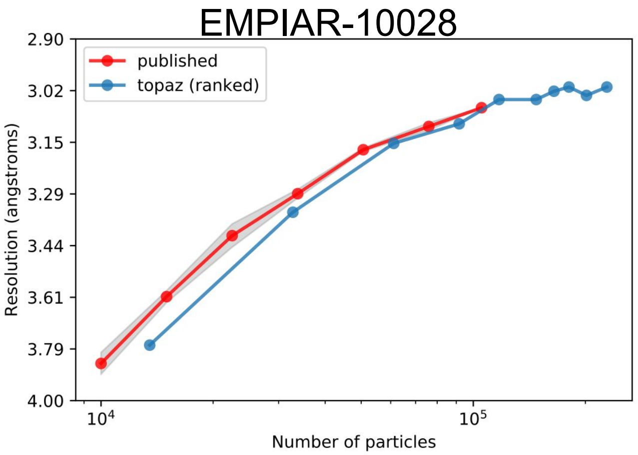


** EM-map challenge best resolution 3.10 Å

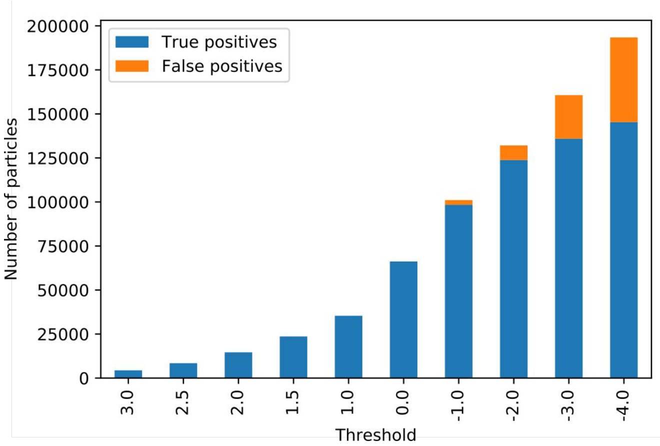
Predicted particles are well-ranked

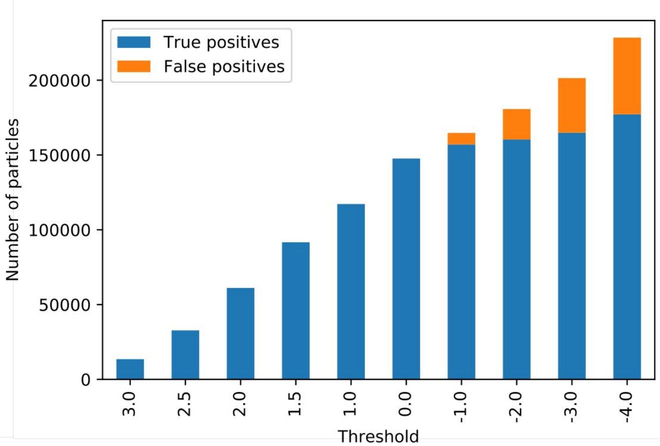
3d reconstruction resolution



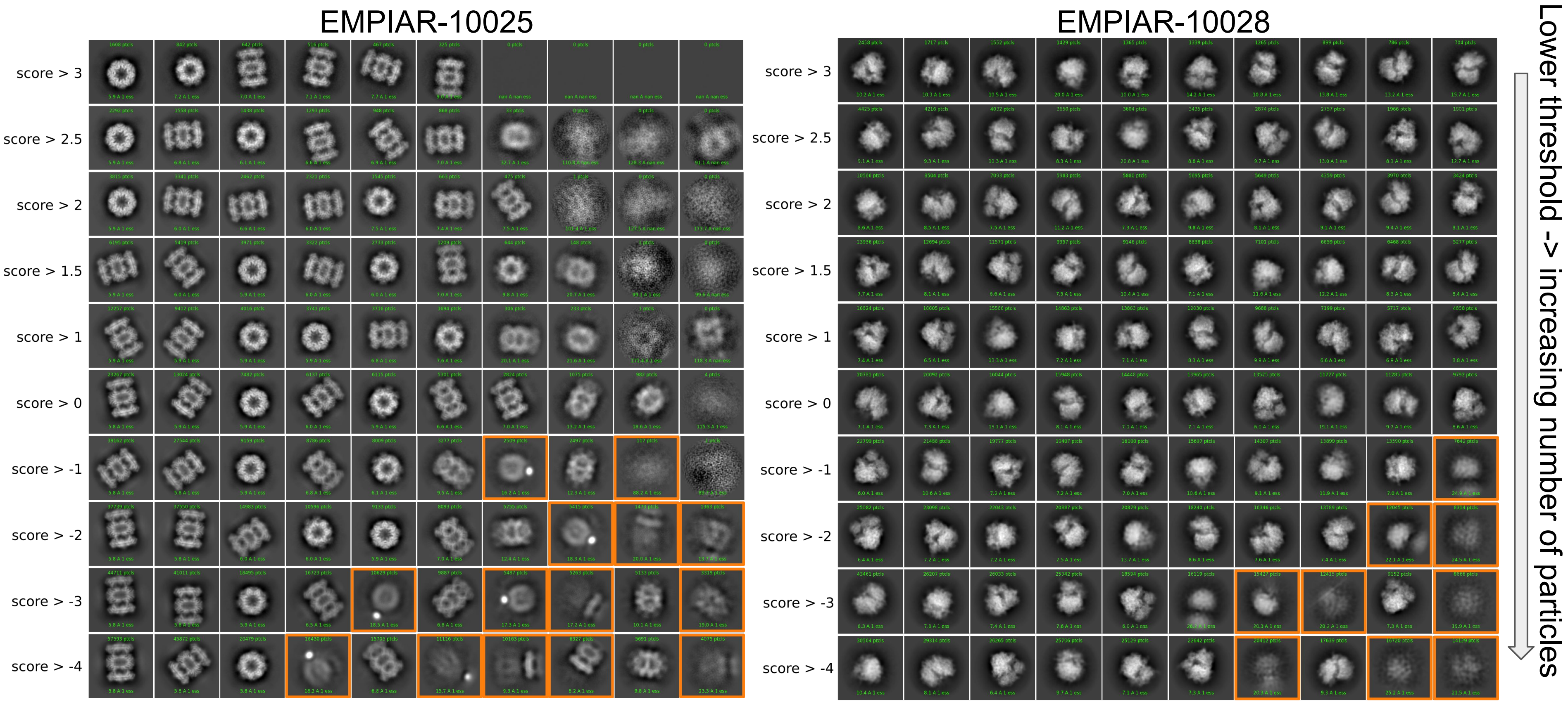


2d class quantification





2d class averages with decreasing particle threshold



Summary

1. **We proposed the GE-binomial loss function** and showed that neural network classifiers trained to minimize this loss on positive and unlabeled micrograph regions **outperform classifiers trained with other positive-unlabeled learning objective functions** on 2 challenging cryoEM datasets (Shapiro-lab, EMPIAR-10096)
2. We showed that creating a joint training scheme in which the classifier is trained together with a decoder to form a **hybrid classifier+autoencoder can further improve performance** when few labeled data points are available
3. We developed an object detection pipeline for picking particles using classifiers trained from positive and unlabeled examples.
4. We showed that particles predicted by Topaz (with only 1000 labeled training examples and no postprocessing) give **state-of-the-art reconstructions** on 2 additional datasets (EMPIAR-10025, EMPIAR-10028)

Topaz - our implementation of this particle picking pipeline - is available at https://github.com/tbepler/topaz

Manuscript in preparation - preprint can be found at https://arxiv.org/abs/1803.08207

Acknowledgements

Massachusetts Institute of Technology

- **Bonnie Berger**
- **Andrew Morin**
- Hoon Cho
- Tommi Jaakkola

New York Structural Biology Center

- **Alex Noble**
- Bridget Carragher
- Clint Potter

Columbia University

- **Julia Brasch**
- **Larry Shapiro**