doi:10.1017/S143192761900566X
Microsc. Microanal. 25 (Suppl 2), 2019

TOPAZ: A Positive-Unlabeled Convolutional Neural Network CryoEM Particle Picker that can Pick Any Size and Shape Particle

Tristan Bepler[1,2], Andrew Morin[2,3], Micah Rapp[4], Julia Brasch[4], Lawrence Shapiro[4], Alex J Noble[5,*], and Bonnie Berger[2,3,**]

[1.] Computational and Systems Biology, MIT, Cambridge, MA, USA.
[2.] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA.
[3.] Department of Mathematics, MIT, Cambridge, MA, USA.
[4.] Department of Biochemistry and Molecular Biophysics, Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, NY, NY, USA.
[5.] National Resource for Automated Molecular Microscopy, Simons Electron Microscopy Center, New York Structural Biology Center, NY, NY, USA.
* Corresponding author for cryoEM: anoble@nysbc.org
** Corresponding author: bab@mit.edu

Single particle cryoEM projects are often hampered by ad-hoc post-processing steps to remove junk particle picks and to attempt to recover all real particles in micrographs. This process can be particularly difficult or even insurmountable for significantly non-globular, small, and/or aggregated proteins. Not maximizing the number of true positive picks in a dataset may cause the resulting 3D structures to not be representative of the data. Moreover, conventional post-processing strategies often involve user-bias.

To overcome these and other common issues, we present Topaz [1, 2], an efficient and accurate particle picking pipeline using convolutional neural networks that requires only 100-1,000 sparsely-picked positive particles for training (or possibly as few as 10 training particles). Topaz uniquely uses positive-unlabeled training workflow which in effect removes the assumption that the manually picked training micrographs are fully picked, and does not use explicit particle negatives (Figure 1). These advances both help address the hard problem of not having a fully labeled training dataset and the requirement of accurately labelling negatives.

Here we show that this novel positive-unlabeled framework allows for Topaz to pick small, non-globular particles, to pick significantly more particles and views than all other pickers we tested, and to avoid junk, aggregated particles, and grid substrate. These advances 1) enable conventionally difficult single particle projects to move forward, 2) significantly decrease ad-hoc user post-processing (e.g. particle filtering/2D classification), 3) make classification more robust and representative given the significantly larger number of real particles picked, and 4) increase collection and processing efficiency. Moreover, we show that the Topaz model score for each particle accurately ranks particles versus junk. We have exemplified these advances with five samples: Clustered protocadherin [3], a Toll receptor, 80S ribosome (EMPIAR-10025), T20S proteasome (EMPIAR-10028), and Rabbit muscle aldolase.

For two conventionally difficult particles due to their shape (clustered protocadherin and the Toll receptor) and their size (Toll receptor), we found that Topaz picks significantly more than hand-picking (Figure 2a) and significantly more than DoG picking (Figure 2b) and crYOLO (not shown) [4] when trained on 1,000 – 1,500 particles. For the Toll receptor which was amenable to 3D processing, these additional particles improved the resolution and isotropy critically enough to allow for model building.

For the larger and conventionally tractable particles, we show that Topaz picks 1.7x – 3.7x more real particles compared to the curated datasets (Figure 3a). We also found that as the Topaz model score for each particle was relaxed further, only then did junk particles begin to be picked (Figure 3b).

References:

[1] T Bepler et al., arXiv (2018).
[2] T Bepler, AJ Noble, and B Berger, Topaz and a standalone GUI, https://github.com/tbepler/topaz
[3] J Brasch et al., Nature (2019).
[4] T Wagner et al., bioRxiv (2018).

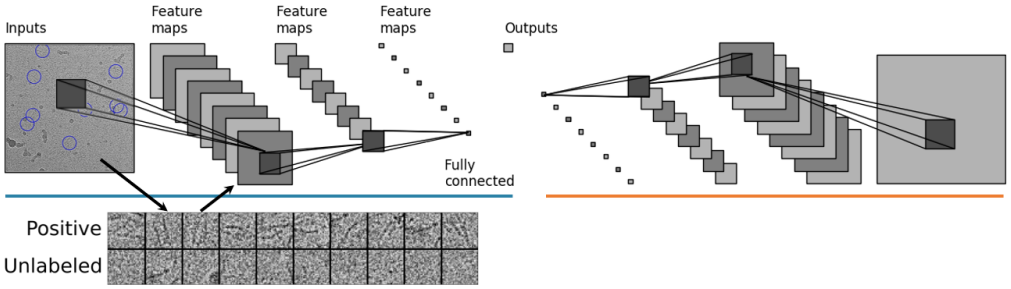

Figure 1. Topaz's CNN uses sparse positive picks, unlabeled areas, a GE-binomial loss function (blue), and an optional autoencoder (orange).

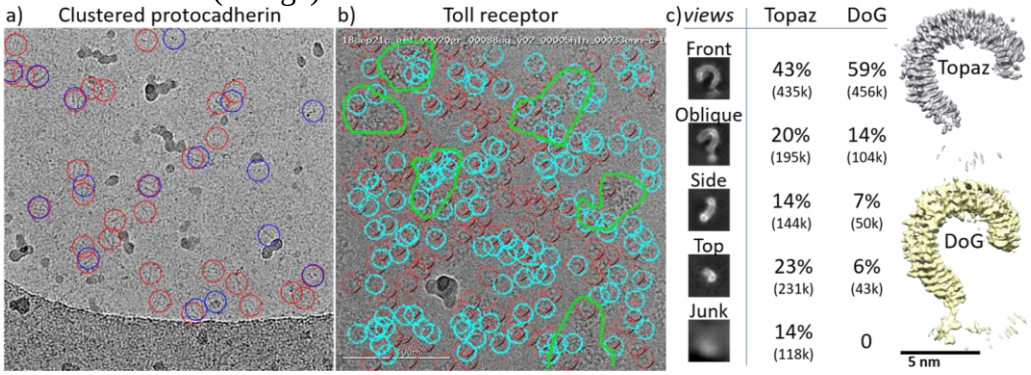

Figure 2. a) Clustered protocadherin hand-picks (blue) and Topaz picks (red). b) Toll receptor with DoG picks (blue), Topaz picks (red), and aggregation (green). c) Toll receptor Topaz and DoG picks.

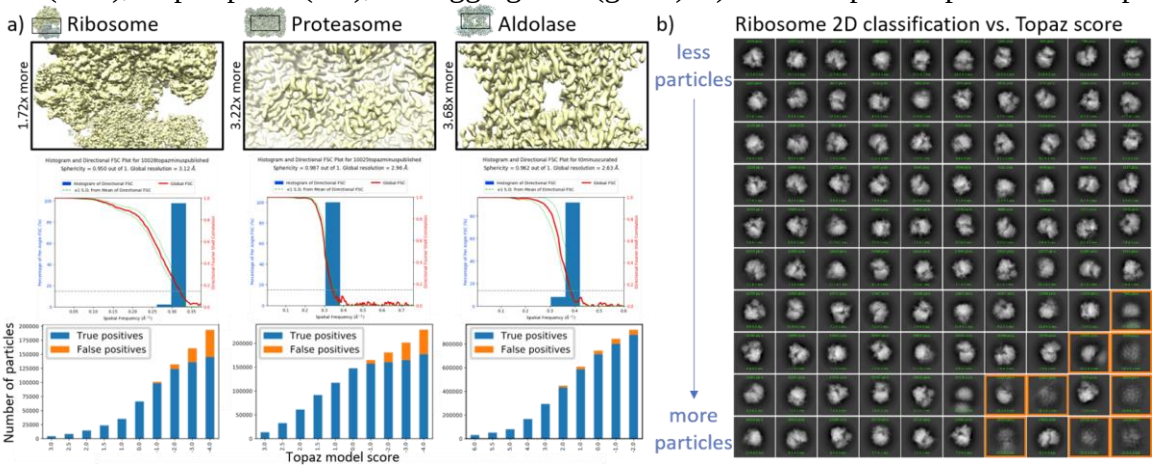

Figure 3. a) Additional real particles (top), 3D FSC (middle), True/false positives vs. Topaz score (bottom). b) 2D classifications (rows) of ribosome particles for different thresholds (orange = junk).

https://doi.org/10.1017/S143192761900566X Published online by Cambridge University Press