Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs

Tristan Bepler\textsuperscript{1,2}, Andrew Morin\textsuperscript{2,3}, Micah Rapp\textsuperscript{4,5}, Julia Brasch\textsuperscript{4,5}, Lawrence Shapiro\textsuperscript{4}, Alex J. Noble\textsuperscript{5*} and Bonnie Berger\textsuperscript{1,2,3*}

Cryo-electron microscopy is a popular method for the determination of protein structures; however, identifying a sufficient number of particles for analysis can take months of manual effort. Current computational approaches find many false positives and require ad hoc postprocessing, especially for unusually shaped particles. To address these shortcomings, we develop Topaz, an efficient and accurate particle-picking pipeline using neural networks trained with a general-purpose positive-unlabeled learning method. This framework enables particle detection models to be trained with few sparsely labeled particles and no labeled negatives. Topaz retrieves many more real particles than conventional picking methods while maintaining low false-positive rates, is capable of picking challenging unusually shaped proteins (for example, small, non-globular and asymmetric particles), produces more representative particle sets and does not require post hoc curation. We demonstrate the performance of Topaz on two difficult datasets and three conventional datasets. Topaz is modular, standalone, free and open source (http://topaz.csail.mit.edu).

Single-particle cryo-electron microscopy (cryo-EM) is a method capable of resolving high-resolution structures of proteins in near-native states. Cryo-EM projection images (micrographs) can contain hundreds or thousands of individual protein projections (particles). Given a sufficient number of particles, the three-dimensional (3D) structure of the protein can be determined\textsuperscript{1}. However, owing to the low signal-to-noise ratio (SNR) of cryo-EM images, large numbers of observations are required for accurate reconstruction. Studies show a log-linear relationship between the number of particles included and the inverse resolution of the reconstruction\textsuperscript{2}. The concentration of protein on EM grids, efficiency of data collection and completeness and accuracy of particle identification are factors determining the total number of particles available for downstream reconstruction and, hence, the achievable resolution. In particular, particle identification (particle picking) is a major bottleneck, often taking weeks or even months with current workflows for small or non-globular particles owing to variability in particle shapes and structured noise in micrographs.

A variety of methods have been developed for particle picking automation. The most widely used are difference of Gaussians (DoG) and template-based approaches\textsuperscript{3,4}. However, these methods are unable to detect unusually shaped particles and suffer from high false-positive rates necessitating post-picking curation. Most commonly, researchers use iterative 2D–3D classification and discard poor subsets by eye. These picking methods and downstream curation introduce bias into the final particle set, potentially removing rare particle views and conformations\textsuperscript{5–11}. Newer methods based on convolutional neural networks (CNNs) have been proposed\textsuperscript{12–14}, which use positive- and negative-labeled micrograph regions to train CNN classifiers, which then predict labels for the remaining regions. However, owing to factors like low SNR, structured background and the distribution of particle morphologies, researchers must label a large number of regions for training—a non-trivial and time-consuming task. Moreover, the diverse characteristics of negative data make it difficult to manually label a representative set of negative examples, and, hence, the number of labeled negatives must be an order of magnitude larger than the number of positives to achieve acceptable performance\textsuperscript{15}. This has limited adoption by the cryo-EM community and hand-labeling remains the gold standard.

To overcome the challenges inherent to current automatic particle-picking methods, we frame particle picking as a positive-unlabeled (PU) learning problem. In PU learning, we seek to learn a classifier of positives and negatives given a small number of labeled positive regions and the remaining unlabeled regions. This has proved to be an effective paradigm when working with partially labeled data in other domains (for example, document classification\textsuperscript{16}, time series classification\textsuperscript{17} and anomaly detection\textsuperscript{18}). Recent work has explored general-purpose PU learning for neural network models based on estimating the true positive–negative risk, but overfitting remains a challenge for PU learning\textsuperscript{19}. Therefore, we instead approach PU learning as a constrained optimization problem in which we wish to find classifier parameters to minimize classification errors on the labeled data subject to a constraint on the expectation over the unlabeled data. By imposing this constraint softly with a novel generalized expectation (GE) criteria\textsuperscript{20}, we are able to mitigate overfitting and train high-accuracy particle classifiers using very few labeled data points. Furthermore, by combining our PU learning method with autoencoder-based regularization, we can further reduce the amount of labeled data required for high performance.

Here we present Topaz, a pipeline for particle picking using CNNs with PU learning. Topaz retrieves many more particles than alternative methods, while maintaining a low false-positive rate. It substantially reduces the need for particle curation, removes systematic bias in particle picking introduced by conventional pickers and 2D–3D classification procedures and allows for robust and representative particle analysis and classification. Furthermore, Topaz...
is capable of reliably picking previously challenging particles (for example, small, non-globular and asymmetric particles) and avoiding aggregation, grid substrate and other background objects, while requiring minimal example particles.

We first demonstrate the capabilities of Topaz on a novel protein dataset for the Toll receptor—a ~105-kDa non-globular asymmetric particle. Despite aggregation and sparse labeling in the dataset, Topaz enables a 3.7-Å reconstruction and resolves secondary structures that could not be resolved with other methods. Topaz also decreases anisotropy by better detecting conventionally difficult particle views. Additionally, on three publicly available datasets, we find that by using Topaz with only 1,000 labeled training examples, we are able to retrieve many more real particles than were included in the published particle sets. This enables us to solve 3D structures of equal or greater quality to those found using the published particles. Remarkably, the Topaz results do not require any of the ad hoc postprocessing that is typically required for high-resolution structures; we feed Topaz particles directly into alignment and reconstruction. Finally, we compare our GE-based PU learning method against other off-the-shelf PU learning approaches and find that our method offers improvements over the current state-of-the-art methods (DoG7 and template picking followed by 2D class averaging) with particles picked using several other methods when applied to training particle-detection models. Topaz was a critical component in determining the single particle behavior of an elongated clustered protocadherin21.

Topaz source code is freely available (https://github.com/tbepler/topaz) and can be installed through Anaconda, pip, Docker, Singularity and SBGrid22. Topaz is designed to be modular, has been integrated into Appion23, is being integrated into Relion 24, CryoSparc25, EMAN26, Scipion27 and Focus28, and can easily be integrated into other cryo-EM software suites in the future. Topaz runs efficiently on a single GPU computer and includes a standalone graphical user interface (GUI)29 to assist with particle labeling.

**Results**

**Topaz pipeline.** The Topaz particle-picking pipeline is composed of the following main steps (Fig. 1): (1) whole micrograph pre-processing with an optional mixture model newly designed to capture micrograph statistics (Methods; Supplementary Figs. 1, 2 and 3), (2) neural network classifier training with our PU learning framework, and (3) sliding window classification of micrographs and extraction of particle coordinates by non-maximum suppression.

**Classifier training from positive and unlabeled data.** We frame particle picking as a PU learning problem in which we seek to learn a classifier that discriminates between particle and non-particle micrograph regions given a small number of labeled particles and many unlabeled micrograph regions. CNN classifiers are trained using minibatched stochastic gradient descent with a novel objective function, GE-binomial (Methods), which explicitly models the sampling statistics of minibatches to regularize the posterior of the classifier over the unlabeled data. Combining this with an optional autoencoder module allows high-accuracy classifiers to be trained using very few positive examples. This approach allows us to overcome overfitting problems associated with recent PU learning methods developed for neural networks in domains other than cryo-EM analysis, and to effectively pick particles in challenging cryo-EM datasets.

**Micrograph region classification and particle extraction.** Given a trained CNN particle classifier, we extract predicted particle coordinates and their associated predicted probabilities. First, we calculate the per pixel predicted probabilities by applying the classifier to each micrograph region in a sliding window. Then, to extract coordinates from these dense predictions, we use the well-known non-maximum suppression algorithm which iteratively selects high-scoring pixels and removes their neighbors from consideration as particle centers. This yields a list of predicted particle coordinates and their associated model scores for each micrograph.

**Topaz picks challenging particles and orientations.** We explore the ability of Topaz to detect challenging particles of a small, asymmetric, non-globular and aggregated protein, a Toll receptor. To this end, we compared particles picked by Topaz (trained with 686 labeled particles) with particles picked using several other methods (DoG7 and template picking followed by 2D class averaging and manual filtering, and the CNN-based methods crYOLO30 and DeepPicker12) (Methods). The CNN-based methods were all trained following the software instructions with default settings and identical labeled particles.
After four rounds of 2D classification and filtering, DoG found 770,263 good particles from an initial stack of 1,599,638 and template picking found 627,533 good particles from an initial stack of 1,265,564. Using Topaz, after one round of 2D classification, we were left with 1,006,089 of an initial 1,010,937 particles, indicating that Topaz gives a remarkably low false-positive rate of only 0.5% on this data. We then compared the quality of the picked particles by taking each particle set through reconstruction (Fig. 2a–c). We found that particles picked using Topaz yield a structure with 0.731 sphericity at Fourier shell correlation (FSC)0.143 = 3.70 Å resolution, as compared to 0.706 sphericity at 3.92 Å for template-picked particles and 0.652 sphericity at 3.86 Å for particles picked using DoG. Furthermore, only the density map based on Topaz particles was of high enough quality to reliably resolve secondary structure (β-strands) and allow for model building. Other CNN-based picking methods, cryOLO and DeepPicker, were unable to find sufficient numbers of good particles for high-resolution reconstruction. cryOLO found 131,300 particles resulting in a 6.8 Å structure while DeepPicker failed to find any meaningful particles in this dataset (Supplementary Figs. 4–7).

We next quantified the ability of these methods to detect different particle views. The Toll receptor is strongly asymmetric and non-globular, thus it is important for picking methods to retrieve the full spectrum of view angles. By counting the number of particles assigned to each view in 2D class averages, we found that Topaz retrieved a much larger fraction of oblique, side and top views of the Toll receptor as compared to DoG and template-based methods (Fig. 2d). In addition, we note that these micrographs are challenging, containing junk and protein aggregation, yet Topaz is uniquely able to avoid these micrograph regions while picking only good particles (Fig. 2e and Supplementary Fig. 4).

Topaz enables high-resolution reconstruction with no post-processing. We next evaluated the full Topaz particle-picking pipeline by generating reconstructions for three cryo-EM datasets containing T20S proteasome (EMPIAR-10029), 80S ribosome (EMPIAR-10028) and rabbit muscle aldolase (EMPIAR-10215). Each of these datasets already had a curated set of particles yielding high-quality reconstructions, which we compared with particles predicted by Topaz (trained with 1,000 positives on the basis of reconstruction quality; Methods). We standardized the reconstruction procedure by using cryoSPARC homogeneous refinement on the raw Topaz particle sets (that is, no postprocessing was applied) and published particle sets with identical settings for each dataset. By considering the reconstruction resolution at decreasing probability thresholds (increasing numbers of particles) predicted by Topaz, we selected the particle set that optimized the resolution for each dataset.

We found that Topaz was able to retrieve substantially more good particles than were present in the curated particle sets, finding 3.22, 1.72 and 3.68 times more particles in EMPIAR-10025, EMPIAR-10028 and EMPIAR-10215, respectively. Furthermore, reconstructions from the Topaz particle sets were of equal or higher quality as compared to those given by the curated particles (Fig. 3). Topaz maps reached roughly equivalent resolution to the published structures for the 80S ribosome and rabbit muscle aldolase while improving the resolution by ~0.15 Å over the published structures for the T20S proteasome. Remarkably, this was achieved using only 1,000 labeled examples and no filtering of the particle set (for example, particle filtering with 2D or 3D class averaging or iterative reconstructions removing poor particles). We note that even though these labeled training particles are extremely sparse, PU learning enabled Topaz to pick with high precision as seen in example micrographs (Supplementary Figs. 8, 9 and 10). We verified that the additional particles found by Topaz were good particles by performing reconstructions using only the newly picked particles and find nearly identical structures (Fig. 3). For aldolase, although Topaz found many more particles than were in the published dataset, the Topaz, curated and the Topaz minus curated particle sets achieved the same reconstruction resolution (2.63 Å at FSC0.143), suggesting that the ~200,000 particles in the published set is already sufficient to reach the resolution limit of the data given standard reconstruction methods.

Topaz particle predictions are well-ranked and contain few false positives. We next quantified the quality of the particles predicted by Topaz over varying predicted probability thresholds by calculating the reconstruction resolution and estimating the number of false-positive particles on the basis of 2D class averaging. For each dataset, reconstructions were calculated using particles predicted by Topaz at decreasing probability cutoffs (Fig. 4a). The resolution of Topaz structures increased as we included more good particles and then dropped once the threshold became small and too many false positives were included, as demonstrated by the dip in resolution for the last threshold of EMPIAR-10025. Furthermore, we compared these curves with those obtained by randomly subsampling the published particle sets and found that Topaz particles quickly matched the resolution of the published particles for the proteasome and ribosome datasets. For the aldolase dataset, we saw that more Topaz particles were required to match and then exceed the resolution of the curated particle set. This could be because Topaz did not find enough side views of the particle until the probability was sufficiently lowered whereas the curated dataset had been filtered to be enriched for these views (Supplementary Fig. 11).
Fig. 3 | Single particle reconstructions from published particles, Topaz particles and Topaz particles with published particles removed. Published particles are on the left, Topaz particles in the middle, and Topaz particles with published particles removed are on the right. Below each reconstruction is the corresponding 3D FSC plot. a, T20S proteasome (EMPIAR-10025) using the provided aligned and dose-weighted micrographs. b, 80S ribosome (EMPIAR-10028). c, Rabbit muscle aldolase (EMPIAR-10215). Scale bars, 3 nm.
Fig. 4 | Reconstruction resolution and 2D class averages for Topaz particles at decreasing log-likelihood ratio thresholds. **a**, Number of particles versus reconstruction resolution for Topaz particles (increasing number of particles corresponds to decreasing log-likelihood threshold) and randomly sampled subsets of the published particle set. Resolution is as reported by cryoSPARC. For the published particle sets the mean of three replicates is marked with s.d. shaded in gray. **b**, Stacked bar plots show the quantification of the number of true and false positives at each threshold on the basis of 2D class averages. The decreasing threshold corresponds to an increasing number of predicted particles. True and false positives are shown in blue and orange, respectively. **c**, 2D class averages obtained at each score threshold for the T2OS proteasome (EMPIAR-10025). Number of particles (ptcls) and effective sample size (ess) for each class are reported by cryoSPARC. NaN is reported for classes without any particles assigned. Classes determined to be false positives are marked with orange boxes. Several classes which appear to be false positives at high score thresholds do not contain any particles and, therefore, are not highlighted.
We also classified the particle sets at each threshold into ten classes and manually examined the class averages to determine whether each class represented true particles or false positives. As expected, we found that as the probability threshold was decreased, the fraction of false positives increased (Fig. 4b), yet remained remarkably low even at relaxed thresholds. Furthermore, particles appear to be well-ranked, in that noisy or unusual particle classes only start to appear at low thresholds. For example, the T20S proteasome dataset was contaminated with gold particles, which appeared as dark spots in the micrographs. Particles in close proximity to gold were only selected when the probability threshold was decreased (Fig. 4). Similar trends can be observed in the ribosome (Supplementary Fig. 12) and aldolase (Supplementary Fig. 11) class averages. This can also be seen in the precision-recall curves for these datasets (Supplementary Figs. 13), in which Topaz maintains remarkably high precision even at high recall levels.

**GE-criteria-based PU learning method outperforms other general-purpose PU learning approaches.** Comparison of PU learning methods. We considered two GE-based approaches to PU learning, GE-KL and GE-binomial (Methods), and evaluated their effectiveness by benchmarking against the recent non-negative risk estimator approach of Kiryu et al.19 (NNPU) and the naive approach in which unlabeled data are considered as negative for classifier training (the naive approach is hereafter referred to as PN) on two additional cryo-EM datasets. This is important to keep the development of our PU learning methods separate from the full Topaz evaluation above. The first dataset, EMPIAR-10096, is a publicly available dataset containing influenza hemagglutinin trimer particles and the second, EMPIAR-10234 (clustered protocadherin), is a challenging dataset provided by the Shapiro lab containing a stick-like particle with low SNR (Supplementary Fig. 14). For purposes of comparison, we simulated positively labeled datasets of varying sizes by randomly subsampling the set of all positive examples within the training set of each dataset.

We found that across all experiments, classifiers trained with our GE-criteria-based objective functions dramatically outperformed those trained with the NNPU or PN methods. Generally, GE-binomial and GE-KL classifiers displayed similar performance with a few important exceptions where GE-binomial gave better results. For the dataset with more compact particles, EMPIAR-10096, GE-binomial gave significantly ($P < 0.05$, Student’s paired $t$ test) better test set average-precision scores than GE-KL when the number of data points was tiny (ten positive examples; Fig. 5a). At larger numbers of positives, both methods were statistically equivalent. On the challenging EMPIAR-10234 dataset, GE-binomial significantly outperformed GE-KL at 1,000 labeled positives ($P < 0.05$) whereas GE-KL gave better results ($P < 0.05$) within the 50–250 range of labeled examples. These results indicate that our GE-based PU learning approaches dramatically outperform previous PU learning methods, enabling particle picking despite few labeled positives on the challenging EMPIAR-10234 dataset and substantially improving picking quality on the easier EMPIAR-10096 dataset. Although GE-binomial and GE-KL performed similarly in this experiment, we did find that GE-binomial outperformed GE-KL in the two important cases of ten easy particles and 1,000 difficult particles.

**Augmentation with autoencoders.** We next considered whether classifier performance could be improved when few labeled data points...
are available by introducing a generator network with a correspond-
ing reconstruction error term in the objective to form a hybrid clas-
sifier with autoencoder network (Methods). We hypothesized that
including this reconstruction component would improve the gener-
alizability of the classifier when few labeled data points are available
by requiring that the feature vectors given by the encoder network
be descriptive of the input—acting as a sort of machine learning
 technique known as regularization.

We evaluated this hypothesis by training classifiers on the
EMPIAR-10096 and EMPIAR-10234 datasets with different settings
of the autoencoder weight, γ, and varying numbers of labeled data
points, N (Methods). We found that including the decoder network
with a reconstruction error term in the objective (γ = 1 and γ = ¼)
improved classifier performance in the regime with few labeled data
points (Fig. 5b). As the number of data points increased, the benefit
of using the autoencoder decreased and then hurt classifier perfor-
mance owing to over-regularization. Our results from both data-
s sets suggest that using the autoencoder with γ = ¼ gives best results
when N ≤ 250

Discussion
As our work originally appeared in RECOMB 2018 and as an
arXiv preprint, other works have followed on bioRxiv that propose alternative
CNN-based particle-picking methods. However, these methods follow the supervised learning paradigm (that is, some variant of PNN learning) and are limited by the associated assumptions. In the future, it may also be possible to provide par-
ticle-detection models pretrained on many publicly available data-
sets; however, we note that fully labeled ground-truth datasets are
presently unavailable and that these models are unlikely to general-
ize to new datasets with conventionally difficult particles, which we
focus on here. While it may seem difficult to provide labeled data
upfront, in practice we find that explicitly relaxing the requirement
to completely label micrographs eases this burden, and is a major
advantage of Topaz over other CNN-based methods. Users may also
‘bootstrap’ the labeling procedure using existing picking and cura-
tion methods, while remaining cautious against reintroducing bias.
We note that there may be some difference between randomly sam-
ping from a curated particle set and particles that would be labeled
by a user. However, the Toll receptor and clustered protodiginin
training sets were both provided by hand labeling and demonstrate
that labeling a small representative set of particles is easily achiev-
able even for conventionally difficult datasets.

Although we use a simple CNN architecture with reasonable
default hyperparameters, and show that it performs well on these
data sets, any model architecture that can be trained with gradient
descent can use our GE criteria objective functions to learn from
positive and unlabeled data. Furthermore, additional hyperparame-
ter tuning, such as L2 or dropout regularization, can improve model
performance. The only hyperparameters introduced by our objective
function is the unknown positive class prior, π, and the constraint
strength, λ. Although the positive class prior could also be chosen
by cross validation, we observed that our results were relatively insensitive to its choice (Supplementary Fig. 15). Furthermore, we
do not find that λ needs to be changed from the default setting. Our
proposed GE-binomial PU learning method could also have wide-
spread utility for object detection in other domains where positive
labels are frequently incomplete, for example, in light microscopy or
medical imaging. Additionally, although we proposed GE-binomial
for positive-unlabeled learning, it is straightforward to extend to the
typical semisupervised case (where some labeled negative regions
are provided) by taking the expectation of the loss over all labeled
data in the first term.

Topaz particle probability thresholding allows particles to be
included iteratively until the reconstruction resolution stops
improving. It is possible for reconstruction algorithms to explicitly
take these probabilities into account when determining 3D structures
in the future.

Topaz requires researchers to label very few particles to achieve
high quality predictions. It performs well independently of particle
shape, opening automated picking to a wide selection of proteins
previously too difficult to locate computationally. In addition, our
pipeline is computationally efficient—training in a few hours on
a single GPU and producing predictions for hundreds of micrographs
in only minutes. Furthermore, once a model is trained for a specific
particle, it can be applied to new imaging runs of the same parti-
cle. Topaz greatly expedites structure determination by cryo-EM,
enabling particle picking for previously difficult datasets, reducing
the manual effort required to achieve high-resolution structures,
and thus increasing the efficiency of cryo-EM workflows and the
completeness of particle analytics.

Online content
Any methods, additional references, Nature Research reporting
summaries, source data, statements of code and data availability
and associated accession codes are available at https://doi.org/10.1038/s41592-019-0575-8.

Received: 18 September 2018; Accepted: 14 August 2019;
Published online: 7 October 2019

References
1. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A primer to single-particle
cryo-electron microscopy. Cell 161, 438–449 (2015).
2. Stagg, S. M., Noble, A. J., Spilman, M. & Chapman, M. S. ResLog plots as an
empirical metric of the quality of cryo-EM reconstructions. J. Struct. Biol. 185, 41–426 (2014).
3. Rosenthal, P. B. & Henderson, R. Optimal determination of particle
orientation, absolute hand, and contrast loss in single-particle electron
cryo microscopy. J. Mol. Biol. 333, 721–745 (2003).
4. Scheres, S. H. W. Semi-automated selection of cryo-EM particles in
RELION-1.3. J. Struct. Biol. 189, 114–122 (2015).
5. Tang, G. et al. EMAN2: an extensible image processing suite for electron
microscopy. J. Struct. Biol. 157, 38–46 (2007).
6. Rosman, A. M. Particle finding in electron micrographs using a fast local
correlation algorithm. Ultramicroscopy 94, 225–236 (2003).
7. Voss, N. R., Yoshiooka, C. K., Radermacher, M., Potter, C. S. & Carragher, B.
DoG Picker and TiltPicker: software tools to facilitate particle selection in
single particle electron microscopy. J. Struct. Biol. 166, 205–213 (2009).
8. Zhang, K., Li, M. & Sun, E. Gautomatch: an efficient and convenient
gpu-based automatic particle selection program. https://www.mrc-lmb.cam.
ac.uk/kzhang/ (2011).
9. Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy:
Einstein from noise. Proc. Natl Acad. Sci. USA 110, 18037–18041 (2013).
10. Subramaniam, S. Structure of trimeric HIV-1 envelope glycoproteins. Proc. Natl Acad. Sci. USA 110, E4172–E4174 (2013).
11. van Heel, M. Finding trimeric HIV-1 envelope glycoproteins in random
noise. Proc. Natl Acad. Sci. USA 110, E4175–E4177 (2013).
12. Wang, F. et al. DeepPicker: a deep learning approach for fully automated
particle picking in cryo-EM. J. Struct. Biol. 195, 325–336 (2016).
13. Zhu, Y., Ouyang, Q. & Mao, Y. A deep convolutional neural network
approach to single-particle recognition in cryo-electron microscopy. BMC Bioinformatics 18, 348 (2017).
14. Xiao, Y. & Yang, G. A fast method for particle picking in cryo-electron
micrographs based on fast R-CNN. AIP Conf. Proc. 1836, 020080 (2017).
15. Chen, M. et al. Convolutional neural networks for automated annotation of
cellular cryo-electron tomograms. Nat. Methods 14, 983–985 (2017).
16. Li, X.-L. & Liu, B. in Machine Learning. ECML 2005 (eds Gama, J. et al.)
178–189 (Springer, 2005).
17. Nguyen, M. N., Li, X.-L. & Ng, S.-K. Positive unlabeled learning for time
series classification. IJCAI 11, 1421–1426 (2011).
18. Zhang, L., Wang, Z., Yuan, J. & Tan, Y.-P. Positive and unlabeled learning for
anomaly detection with multi-features. in Proc. 2017 ACM on Multimedia
Conference 854–862 (ACM, 2017).
19. Kiryo, R., Niu, G., du Plessis, M. C. & Sugiyama, M. in Advances in
Neural Information Processing Systems 30 (eds Guyon, I. et al.)
1675–1685 (Curran Associates, 2017).
20. Mann, G. S. & McCallum, A. Generalized expectation criteria for semi-
supervised learning with weakly labeled data. Proc. Natl Acad. Sci. USA
110, 18037–18041 (2013).
21. Brasch, J. et al. Visualization of clustered protocadherin neuronal self-recognition complexes. *Nature* **569**, 280–283 (2019).
22. Morrow, A. et al. Cutting edge: collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
23. Lander, G. C. et al. Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166**, 95–102 (2009).
24. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
25. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
26. de la Rosa-Trevín et al. Scipion: a software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* **195**, 93–99 (2016).
27. Biyani, N. et al. Focus: the interface between data collection and data processing in cryo-EM. *J. Struct. Biol.* **198**, 124–133 (2017).
28. Dutta, A. & Zisserman, A. The VIA annotation software for images, audio and video. Preprint at https://arxiv.org/abs/1904.10699 (2019).
29. Wagner, T. et al. SPHIRE-cyYOLO: a fast and well-centering automated particle picker for cryo-EM. *Comm. Biol.* **2**, 218 (2019).
30. Bepler, T. et al. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. in *Proc. 22nd Annual International Conference on Research in Computational Molecular Biology.* (ed. Raphael, B. J.) 245–247 (Springer, 2018).
31. Tegunov, D. & Cramer, P. Real-time cryo-EM data pre-processing with Warp. Preprint at https://doi.org/10.1101/338558 (2018).

**Acknowledgements**

The authors wish to thank Simons Electron Microscopy Center (SEMC) OPs for the aldolase sample preparation and collection, Y. Z. Tan for SPA discussion, and the Electron Microscopy Group at the New York Structural Biology Center (NYSBC) for performing the Toll receptor. We would also like to thank T. Jakkola for his valuable feedback on the machine learning methods. We thank the developers of Relion, cryoSPARC, Appion, EMAN2, Scipion and Focus for their efforts in integrating Topaz. The Topaz GUI is based on VGG Image Annotator (VIA), which is developed and maintained with the support of EPSRC program grant Seebibyte: Visual Search for the Era of Big Data (EP/M013774/1). T.B., A.M. and B.B. were supported by NIH grant R01-GM081871. M.R. was supported by NSF GRFP (DGE-1644869). L.S. was supported by NIH grant R01-1MH1114817. A.J.N. was supported by a grant from the NIH National Institute of General Medical Sciences (NIGMS) (F32GM128303). The cryo-EM work was performed at the SEMC and National Resource for Automated Molecular Microscopy located at NYSBC, supported by grants from the Simons Foundation (SF349247), NYSTAR and the NIH NIGMS (GM103310) with additional support from the Agouron Institute (F00316) and NIH (OD019994).

**Author contributions**

T.B., A.M. and B.B. conceived the project. T.B. developed the PU learning methods and implemented Topaz, processed and analyzed single particle datasets, and carried out the computational experiments under the guidance of B.B. M.R. prepared and collected the Toll receptor dataset. J.B. prepared and collected the clustered protocadherin dataset. A.J.N. analyzed the single particle cryo-EM reconstructions. A.J.N. developed the Topaz GUI based on VIA. T.B., A.M., M.R., J.B., L.S., A.J.N. and B.B. designed the experiments. T.B., M.R., A.J.N. and B.B. wrote the manuscript.

**Competing interests**

The authors declare no competing financial interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0575-8.

Correspondence and requests for materials should be addressed to A.J.N. or B.B.

**Peer review information** Allison Doerr was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

**Dataset description.** Aligned and summed micrographs and star files containing published particle sets were retrieved from the Electron Microscopy Public Image Archive (EMPIAR) for datasets EMPIAR-10025 (ref. 1), EMPIAR-10028 (ref. 2) and EMPIAR-10096 (ref. 3). Aligned and summed micrographs and hand-labeled particle coordinates were provided by the Shapiro lab for the EMPIAR-10234 dataset. Aligned and summed micrographs and a curated in-house particle set were provided by the New York Structural Biology Center for the EMPIAR-10215 dataset. Micrographs for each dataset were downsampled to the resolution specified in Table 1 and normalized as described in the following section. Each dataset was then split into training and test sets at the micrograph level. The number of micrographs and labeled particles in each split are also reported in Table 1. To demonstrate the utility of our Gaussian mixture model (GMM) normalization method, we also retrieved micrographs for EMPIAR-10261 (ref. 10) from EMPIAR.

**Micrograph normalization.** Images were normalized using a per-image scaled two component GMM. Given K images, each pixel is modeled as being drawn from a two component GMM, parameterized by \( \rho \), the mixing parameter and \( \mu_1, \sigma_1, \mu_2, \sigma_2 \), the means and standard deviations of the Gaussian distributions, with a scalar multiplier for each image, \( \alpha \). Let \( x_{ij,k} \) be the value of the pixel at position \( ij \) in image \( k \), it is distributed according to:

\[
x_{ij,k} | z_{ij,k} \sim \text{Gaussian}(\alpha \mu_{z_{ij,k}}, \sigma_{z_{ij,k}})
\]

where \( z_{ij,k} \) is a random variable denoting the component membership of the pixel. The maximum likelihood values of the parameters \( \rho, \mu_1, \sigma_1, \mu_2, \sigma_2 \) are found by expectation-maximization for each dataset. Then, the pixels are normalized by first dividing by the image scaling factor and then standardizing to the dominant mixture component. Let \( \mu', \sigma' \) be \( \mu_{1}, \rho \cdot \sigma_{1} \) if \( \rho < 0.5 \) and \( \mu_{2}, \sigma_{2} \) otherwise, then the normalized pixel values \( x'_{ij,k} \) are given by:

\[
x'_{ij,k} = \frac{x_{ij,k} - \mu'}{\sigma'}
\]

We positively contrasted this normalization with standard affine normalization of micrographs (Supplementary Figs. 1, 2, and 3). In affine normalization, micrographs are transformed by subtracting the mean and dividing by the s.d. of all pixel values in each micrograph.

**PU learning baselines.** Let \( P \) be the set of labeled positive micrograph regions (centered on a particle), and \( U \) be the set of unlabeled micrograph regions where \( x \) is the fraction of positive examples within \( U \). Then, the task is to learn a classifier \( g \) that discriminates between positive and negative regions given \( P \) and \( U \). When \( x \) is small, treating the unlabeled examples as negatives for the purposes of classifier training with the following standard loss minimization objective, given cost function \( L \) can be effective (referred to as PN)

\[
f_{\text{PN}} = \mathbb{E}_{x \sim U}[L(g(x), 1)] + (1 - x)f_{\text{U}}[L(g(x), 0)]
\]

However, in general, this approach suffers from overfitting owing to poor specification of the classification objective—it is minimized when positives are perfectly separated from unlabeled data points. To address this, Kiryoo et al. recently proposed an unbiased estimator of the true positive–negative classification objective for positive and unlabeled data with known \( \pi \) and a non-negative estimator (PU), which is shown to reduce overfitting still present in the unbiased estimator.

PU learning with generalized expectation criteria. Here we adopt an alternative approach to positive-unlabeled learning that is not based on estimating the PN misclassification risk. Instead, we observe that unlabeled data with known \( \pi \) can be used to constrain a classifier such that it minimizes the classification loss on the labeled data and matches the expectation \( \langle x \rangle \) over the unlabeled data. In other words, we wish to find the classifier, \( g \), that minimizes \( \mathbb{E}_x[L(g(x), 1)] \) subject to the constraint \( \mathbb{E}_x[g(x)] = \pi \). This constraint can be imposed 'softly' through a regularization term in the objective function with weight \( \lambda \) (referred to as GE-KL)

\[
f_{\text{GE-KL}} = \mathbb{E}_{x \sim U}[L(g(x), 1)] + \lambda KL(\mathbb{E}_{x \sim U}[g(x)] || \pi)
\]

In this objective function, we impose a constraint through the KL-divergence between the expectation of the classifier over the unlabeled data and the known fraction of positives, which is minimized when these terms are equal. This approach is an instance of a general class of posterior regularization called GE criteria, as specifically proposed by Mann and McCallum. However, because we wish for our classifier to be a neural network and to optimize the objective using minibatched stochastic gradient descent, the gradient of the objective must be approximated using samples from the data. Estimates of the gradient of the GE-KL objective from samples are biased, which could cause stochastic gradient descent to find a suboptimal solution.

To address this issue, we propose an alternative GE criteria, GE-binomial, defined so as to minimize the difference between the distribution over the number of positives in the minibatch and the binomial distribution parameterized by \( \pi \). The number of positive data points, \( k \), in a minibatch of \( N \) samples from \( U \) follows the binomial distribution with parameter \( \pi \). Furthermore, the classifier \( g \) also describes a distribution over the number of positives in the minibatch as

\[
q(k) = \sum_{y \in \{0, 1\}} \binom{N}{y} \pi^y (1 - \pi)^{N-y}
\]

where \( x \) is a micrograph region, \( y \) is an indicator vector \( y \in \{0, 1\} \) denoting which data points are positive \( (y = 1) \) and negative \( (y = 0) \) and \( \{0, 1\} \) is the set of all such vectors summing to \( k \). This allows us to define the new GE criteria as the cross entropy between these two distributions \( \sum_{k=1}^{N} q(k) \text{log}(q(k)) \) giving the full GE-binomial objective function

\[
f_{\text{GE-binomial}} = \mathbb{E}_{x \sim U}[L(g(x), 1)] + \lambda \sum_{k=1}^{N} q(k) \text{log}(q(k))
\]

In practice, because computing exact \( q(k) \) is slow, we make a Gaussian approximation with mean \( \sum_{i=1}^{N} \hat{g}(x_i) \) and variance \( \sum_{i=1}^{N} \hat{g}(x_i)(1 - \hat{g}(x_i)) \) and substitute the Gaussian probability density function with these parameters for \( q \) in the above equation.

**Autoencoder-based classifier regularization.** When including the autoencoder component, we break our classifier network into the following two components: an encoder network composed of all layers except the final linear layer and the linear classifier layer. We denote these networks \( A \) and \( f \), respectively, with the full network, \( g \), being given by \( g(x) = f(A(x)) \). Furthermore, we introduce a deconvolutional (also called transposed convolutional; see next section) decoder network, \( d \), which takes the output of the feature extractor network and returns a reconstruction of the input image, \( x = d(f(x)) \). The objective function is then modified to include a term penalizing the expected reconstruction error over all images in the dataset, \( D \), with weight \( \gamma \)

\[
f_{\text{AE-Adopt}} = \mathbb{E}_{x \sim D}(L(g(x), 1)) + \lambda \sum_{i=1}^{N} q(k) \text{log}(q(k)) + \gamma \mathbb{E}_{x \sim D}[\|x - d(f(x))\|_2^2]
\]

This forms the full GE-binomial objective function with autoencoder component used in Topaz.

| Dataset       | Protein               | Original (Angstroms per pixel) | Downsampled (Angstroms per pixel) | Particle radius (pixels) | Training radius (pixels) | \( \pi \) | Number of micrographs | Number of particles | Number of micrographs | Number of particles |
|---------------|-----------------------|--------------------------------|----------------------------------|------------------------|-------------------------|----------|---------------------|-------------------|--------------------|-------------------|
| EMPIAR-10025  | T20S proteasome       | 0.98                           | 15.7                             | 7                      | 3                      | 0.035    | 156                 | 39,653            | 40                 | 10,301            |
| EMPIAR-10028  | 80S ribosome          | 1.34                           | 10.7                             | 12                     | 3                      | 0.012    | 831                 | 80,701            | 250                | 24,546            |
| EMPIAR-10096  | Hemagglutinin trimer  | 1.31                           | 5.24                             | 10                     | 4                      | 0.035    | 347                 | 100,465           | 100                | 29,535            |
| EMPIAR-10215  | Rabbit muscle aldolase| 0.832                          | 6.64                             | 10                     | 3                      | 0.1      | 865                 | 163,758           | 200                | 39,347            |
| EMPIAR-10234  | Clustered protocadherin| 1.061                          | 8.49                             | 15                     | 4                      | 0.015    | 67                  | 1,167             | 20                 | 373               |
| Toll receptor | Toll receptor         | 0.832                          | 3.328                            | 25                     | 5                      | 0.035    | 30                  | 686               | 14                 | 362               |

**Table 1 | Summary of cryo-EM datasets and hyperparameters used for classifier training on each; dataset was downsampled and split into training and test sets at the whole micrograph level**
Classifier and autoencoder architectures and hyperparameters. We use a simple three-layer convolutional neural network with striding, batch normalization\(^{26}\) and parametric rectified linear units (PReLU) as the classifier in this work. The model is organized as 32 conv5x5 filters with batch normalization and PReLU, stride by 2, 64 conv5x5 filters with batch normalization and PReLU, and a final fully connected layer with a single output. We use sigmoid activation on this output to convert it into the predicted probability of a region being from the positive class (that is, the output is interpreted as the log-likelihood ratio between positive and negative classes).

When augmenting with an autoencoder, we use a decoder structure similar to that of DCGAN\(^{37}\). The \(d\)-dimensional representation that is output by the final convolutional layer of the classifier network is projected to a representation with small spatial dimensions but large feature dimensions. This is repeatedly projected into representations with larger spatial dimensions and smaller feature dimensions until the final output is of the original input image size. Specifically, this model is structured as repeated transpose convolutions with batch normalization and leaky ReLU activations. Let \(z\) be the representation output by the final convolutional layer of the classifier and \(X\) be the image reconstruction given by the decoder, the decoder structure is \(z \rightarrow \text{transpose conv}(4 \times 4) \rightarrow \text{conv}(32 \times 4, \text{stride} 2) \rightarrow \text{conv}(1 \times 1) \rightarrow X\).

PU learning benchmarking. To compare classifiers trained with the different objective functions, we simulated hand-labeling with various amounts of effort by randomly sampling varying numbers of particles from the training sets to treat as labeled. All other particles were considered unlabeled. We used cross validation; the folds were determined by randomly sampling varying numbers of labeled data points, \(N\) and \(\gamma\), and varying numbers of labeled positive examples. Statistical significance of performance differences between methods at each number of labeled positive examples was assessed using a two-sided \(t\) test.

Classifier evaluation. Classifiers were evaluated by average-precision score. This score is a measure of how well ranked the micrograph regions were when ordered by the predicted probability of containing a particle, and corresponds to the area under the precision-recall curve. It is calculated as the sum over the ranked micrograph regions of the precision at \(k\) elements times the change in recall

\[
\sum_{k=1}^{N} \text{Pr}(k) \cdot \text{Re}(k) - \text{Re}(k-1)
\]

where precision (Pr) is the fraction of predictions that are correct and recall (Re) is the fraction of labeled particles that are retrieved in the top \(k\) predictions. Let \(TP(k)\) be the number of true positives in the top \(k\) predictions, then Pr and Re are given by

\[
\text{TP}(k) = \sum_{i=1}^{k} \text{Pr}(i)
\]

\[
\text{Pr}(k) = \frac{k}{N}
\]

\[
\text{Re}(k) = \frac{TP(k)}{TP(A)}
\]

This measure is commonly used in information retrieval.

Non-maximum suppression algorithm for extracting particle coordinates. Non-maximum suppression chooses coordinates and the corresponding predicted probabilities of being a particle greedily starting from the highest scoring region. To prevent nearby pixels from also being considered particle candidates, all pixels within a second user-defined radius are excluded when a coordinate is selected. We set this radius to be the half major-axis length of the particle; however, smaller radii may give better results for closely packed, irregularly shaped particles.

Micrograph preprocessing. For EMPIAR-10025 and EMPIAR-10028, the aligned and summed micrographs along with contrast transfer function (CTF) estimates were taken directly from the public data release on EMPIAR. For EMPIAR-10028 and EMPIAR-10261, frames were aligned and summed without dose compensation using MotionCor2\(^{38}\). Whole micrograph CTF estimates provided with the public release were used for this dataset.

For the clustered protocolaR dataset (EMPIAR-10234), single particle micrographs were collected on a Titan Krios electron microscope (Thermo Fisher Scientific) equipped with a K2 counting camera (Gatan); the microscope was operated at 300 kV with a calibrated pixel size of 1.061 Å. Ten-second exposures were collected (40 frames per micrograph) for a total dose of 68 e\(^{-}\) Å\(^{-2}\) with a defocus range of 1–4 μm. A total of 896 micrographs were collected using Legion\(^{41}\). Frames were aligned using MotionCor2\(^{39}\). A total of 1,540 particles were picked manually using Appion Manual Picker\(^{42}\) from 87 micrographs and used as a training dataset for Topaz.

The rabbit muscle aldolase dataset (EMPIAR-10215) was collected on a Titan Krios electron microscope (Thermo Fisher Scientific) equipped with a K2 counting camera (Gatan); the microscope was operated at 300 kV with a calibrated super-resolution pixel size of 0.416 Å. Six-second exposures were collected (30 frames per micrograph) for a total dose of 70.32 e\(^{-}\) Å\(^{-2}\) with a defocus range of 1–2 μm. A total of 1,052 micrographs were collected using Legion\(^{41}\). Frames were aligned, Fourier binned by a factor of 2 and dose compensated using MotionCor2\(^{38}\). Whole-image CTF estimation was performed using CTFFIND4\(^{40}\).

3D reconstruction procedure. Reconstruction was performed using cryoSPARC\(^{43}\). For each particle set, we first generated an ab initio structure with a single class. These structures were then refined using the homogeneous refinement option of cryoSPARC with symmetry specified depending on the dataset (T20S proteasome, D\(_{2}\); 80S ribosome, C\(_{2}\); and aldolase, D\(_{4}\)). For the aldolase dataset, we used C\(_{2}\) symmetry for ab initio structure determination. Otherwise, all other parameters were left in the default setting. When evaluating the quality of Topaz particle sets for reconstructing a high resolution template, each particle set was selected by taking all particles predicted by the Topaz model with scores greater than or equal to the given threshold. Reconstructions were calculated for each of these sets independently as described above.

Removal of overlapping particles. To evaluate the quality of the extra particles predicted by Topaz, we removed particles from the Topaz particle set that were also included in the published particle set. This was done by removing all Topaz particles with centers within the particle radius of a particle center in the published particle set.

2D class averages (EMPIAR-10025, EMPIAR-10028 and EMPIAR-10215). Class averages were calculated using the cryoSPARC 2D classification option. All settings were left as default except the number of 2D classes, which was set to ten for every particle set.

3D structure analysis (EMPIAR-10025, EMPIAR-10028 and EMPIAR-10215). The final 3D reconstructions were analyzed visually in UCSF Chimera\(^{15}\) and 3DFSC\(^{14}\). In Chimera, the previous 3D reconstruction was first loaded (with the fitted Protein Data Bank structure, if available), the newly processed 3D reconstruction was then aligned to the previous reconstruction. The structures were visually compared and representative areas were chosen for display in Fig. 4. The 3DFSC reconstructions were calculated using the public server (https://3dfsc.salk.edu), which compares Fourier shell components for several solid angles to determine the range of resolutions and the amount of anisotropy in the reconstruction.

Toll receptor particle picking. A total of 1,599,638 particles were picked using DoG Picker 2 (ref. 1) from 8,974 micrographs and imported into cryoSPARC for all subsequent processing. After particle curation using 2D classification described below, the particle picks from 44 micrographs were visually inspected. Picks in areas of obvious particle aggregation were removed, and lower SNR particles corresponding to views typically missed by DoG Picker were selected. The remaining 1,048 particles were split into 686 training and 362 testing particles at the micrograph level. Topaz was then trained on the training particles and applied with the default score threshold of 0 for particle prediction. The ‘oblique’, ‘side’, and ‘top’ 2D classes (Fig. 3d) were lowpass filtered to 15 Å and used for template correlation with FindEM\(^{16}\) implemented in the Appion\(^{2}\) software package.

The cryoYOLO\(^{17}\) network was trained on the complete set of 1,048 labeled particles with 20% held out for validation by default. Micrographs were filtered and training was performed as described in the cryoYOLO tutorial. Picking was performed at the default threshold of 0.3.

The DeepPicker\(^{17}\) network was also trained on the complete set of 1,048 particles. Though no micrograph processing is required in the DeepPicker tutorial, micrographs were binned in Fourier space and lowpass filtered to 10 Å using EMAN2\(^{18}\). Even with a threshold of 0, no particles were predicted by DeepPicker.

Toll receptor 3D reconstruction. All reconstructions were performed using cryoSPARC\(^{43}\). For all particle-picking approaches, we performed 2D classification with default parameters and 100 2D classes, then removed obvious non-particles.
For the DoG dataset, four rounds of 2D classification yielded 770,263 particles from an initial stack of 1,599,638. For the template dataset, four rounds of 2D classification yielded 627,533 particles from an initial stack of 1,265,564. For the Topaz dataset, one round of 2D classification yielded 1,006,089 particles from an initial stack of 1,010,937. For the crYOLO dataset, one round of 2D classification yielded 131,300 particles from an initial stack of 133,644. For all datasets, ab initio reconstruction was used to generate an initial model, and the structures were further refined using homogeneous refinement with $C_1$ symmetry, followed by non-uniform refinement. All parameters were left in their default setting. Unfiltered half maps and masks were used to calculate 3DFSC reconstructions using the public server (https://3dfsc.salk.edu).

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

Single-particle half maps, full sharpened maps and masks for T20S proteasome, 80S ribosome, rabbit muscle aldolase and the Toll receptor (DoG, template and Topaz picks) have been deposited in the Electron Microscopy Data Bank (EMDB) under accessions EMD-9194, EMD-9201, EMD-9206, EMD-9207, EMD-9208, EMD-9209, EMD-9210, EMD-9211, EMD-20529, EMD-20531 and EMD-20532. The full rabbit muscle aldolase dataset has been deposited in the Electron Microscopy Pilot Image Archive (EMPIAR) under accession EMPIAR-10215.

**Code availability**

Source code for Topaz is publicly available via Code Ocean and on GitHub at https://github.com/tbepler/topaz. Updates to Topaz will be posted at http://topaz.csail.mit.edu. Topaz is licensed under the GNU General Public License v.3.0.

**References**

32. Campbell, M. G., Veesler, D., Cheng, A., Potter, C. S. & Carragher, B. 2.8 Å resolution reconstruction of the *Thermoplasma acidophilum* 20S proteasome using cryo-electron microscopy. *eLife* 4, e06380 (2015).
33. Wong, W. et al. Cryo-EM structure of the *Plasmodium falciparum* 80S ribosome bound to the anti-protozoan drug emetine. *eLife* 3, e03080 (2014).
34. Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* 14, 793–796 (2017).
35. Xu, H. et al. Structural basis of Nav1.7 inhibition by a gating-modifier spider toxin. *Cell* 176, 702–715 (2019).
36. Ioffe, S. & Szegedy, C. in Proc. 32nd International Conference on Machine Learning (eds Bach, F. & Blei, D.) 448–456 (PMLR, 2015).
37. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. Preprint at https://arxiv.org/abs/1511.06434 (2015).
38. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* 14, 331–332 (2017).
39. Carragher, B. et al. Leginon: an automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol.* 132, 33–45 (2000).
40. Rohou, A. & Grigorieff, N. CTFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* 192, 216–221 (2015).
41. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612 (2004).
42. Roseman, A. M. FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. *J. Struct. Biol.* 145, 91–99 (2004).
43. Repler, T. et al. Topaz: positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Code Ocean* https://doi.org/10.24433/CO.1911124.v1 (2019).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a  Confirmed

☐  The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐  A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐  The statistical test(s) used AND whether they are one- or two-sided

☐  Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐  A description of all covariates tested

☐  A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐  A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐  For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

☐  Give P values as exact values whenever suitable.

☐  For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐  For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐  Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Legimom (Suloway, 2005) was used for cryoEM single particle data collection.

Data analysis

Topaz was developed for particle picking as described in this work. Topaz is available on GitHub at https://github.com/tbepler/topaz. Structure determination and 2d class averaging was performed with cryoSPARC v0.6.5. For additional data processing and analysis, software used was MotionCor2 v1.2.1, Appion v3.3beta, CTFFIND4, UCSF Chimera v1.13.1, D3FSC, DoG Picker 2, FindEM v2, cRYOLO v1.1.3, DeepPicker, and EMAN2.22.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Single particle half maps, full sharpened maps, and masks for T205 proteasome, 80S ribosome, rabbit muscle aldolase, and the Toll receptor (DoG, template, and Topaz picks) have been deposited to the Electron Microscopy Data Bank (EMDB) with accession codes EMD-9194, EMD-9201, EMD-9202, EMD-9206, EMD-9207, EMD-9208, EMD-9209, EMD-9210, EMD-9211, EMD-20529, EMD-20531, and EMD-20532. The full rabbit muscle aldolase dataset has been deposited to the Electron Microscopy Pilot Image Archive (EMPIAR) with accession code EMPIAR-10215.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☑ Life sciences       ☐ Behavioural & social sciences       ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-list.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

When comparing PU learning methods, we randomly sampled training particles and fit models ten times. No sample-size calculations were performed. This number was chosen to get the best mean and variance estimates within a reasonable runtime.

Data exclusions

No data were excluded.

Replication

Particle picking with topaz and reconstruction was replicated across three cryoEM datasets. As discussed in the manuscript, performance of Topaz was well reproduced across all three datasets. We also reproduced the PU learning comparison on two cryoEM datasets where we confirmed for both that the generalize-expectation criteria approaches outperformed the non-negative risk estimator and naive baselines.

Randomization

Training particles were sampled randomly.

Blinding

Blinding was not relevant. No data was collected and analysis was performed using well established metrics without investigator intervention.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|----------------------------------|---------|
| n/a                             | n/a     |
| ☑ Involved in the study         | Involved in the study |
| ☐ Antibodies                    | ☐ ChIP-seq |
| ☐ Eukaryotic cell lines         | ☐ Flow cytometry |
| ☑ Palaeontology                 | ☑ MRI-based neuroimaging |
| ☐ Animals and other organisms   |         |
| ☐ Human research participants   |         |
| ☑ Clinical data                 |         |