APPLICATION NOTE

Deep-learning-based particle picking of detergent-solubilized GPCRs using crYOLO Software

Matthew J. Belousoff, Denise Wootten, Patrick M. Sexton

Introduction

G protein-coupled receptors, or GPCRs, are an important drug-target family due to their association with a wide variety of diseases. Despite the promise of GPCRs, most remain undruggable due to the inherent difficulty of crystallizing and analyzing membrane proteins. Cryo-electron microscopy (cryo-EM) single particle analysis is ideally suited to address this problem as it is capable of producing high-resolution 3D structures without the need for sample crystals.

Despite this advantage, data collection and processing continue to be a challenging aspect of high-quality GPCR analysis. It was previously demonstrated that the highest-resolution results were achieved with high particle density in thin, vitreous ice.[1, 5] While this yields the best quality particle projections, it routinely causes sample crowding, where particles are essentially touching. Consequently, one of the most challenging steps of processing a GPCR dataset is particle picking for these densely packed grids. Additionally, GPCR complexes are generally less than 165 Å in diameter, and micrographs are typically collected between 0.5–0.9 µm objective defocus. This results in low phase contrast, making discernment between signal and noise subjectively harder. Even after bandpass filtering, it can still be difficult to determine individual particle positions in a micrograph (Figure 1).

Particle-picking methods

There are numerous particle-picking algorithms and programs for automated single-particle analysis, all of which have their own merits and drawbacks. Most critically, these programs must ensure that the center of mass is being picked for the particles. If this is not the case, it will become significantly more challenging for alignment programs to operate consistently and efficiently on the single-particle data set.

This application note describes a robust approach for challenging particle-picking as well as the process for exporting and re-importing data between the RELION workflow and external particle-picking software. We will demonstrate how to pick particle positions from densely packed micrographs and how to iteratively improve these results without having to exhaustively search for X,Y particle offsets during computationally expensive 2D or 3D refinements. This overall workflow is outlined in Figure 2.

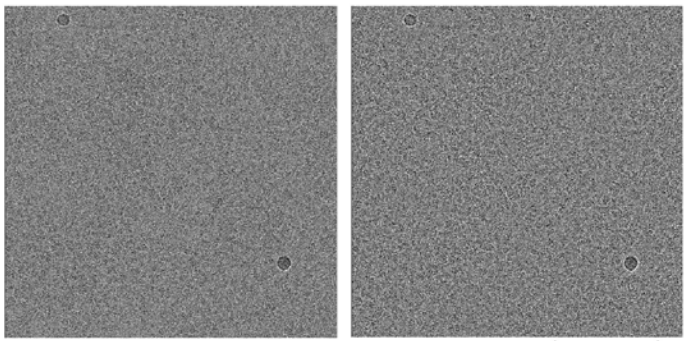

Figure 1. Motion-corrected micrograph of a highly crowded GPCR complex at 0.8 µm defocus that presents very good data but is difficult to interpret by eye. The micrograph on the left is unfiltered, whereas the micrograph on the right has been bandpass filtered (low pass 12 Å, high pass 300 Å).

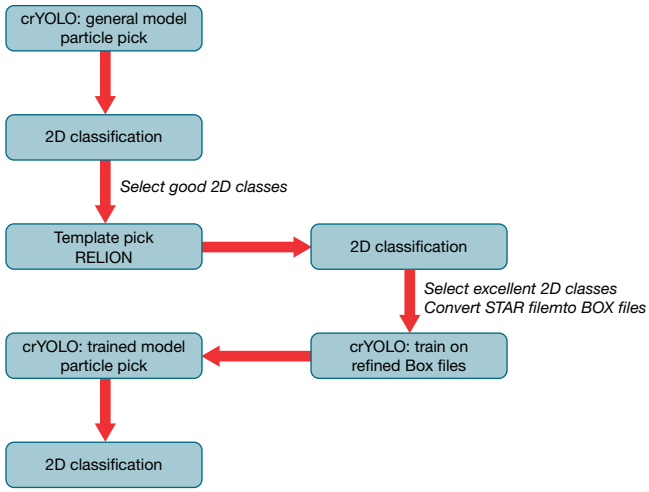

Figure 2. Flowchart of crYOLO training with a 2D classification used as the readout.

ThermoFisher SCIENTIFIC

The procedure will be shown in the context of a neural-network-based particle-picking software packaged called crYOLO,[2] however the overall concepts are equally applicable to other neural-network pickers such as Topaz[3] or the neural-network picker found in the WARP Software Suite.[4] In general, all these particle-picking algorithms leverage a machine-learning approach known as a convolutional neural network, which requires a set of training weights derived from a selection of micrographs. These weights can then be used to pick particle positions from a whole set of motion-corrected micrographs.

This app note assumes that you have the following software installed:

- crYOLO Software (https://cryolo.readthedocs.io/en/stable/installation.html)
- SPHIRE/EMAN Software (https://blake.bcm.edu/emanwiki/EMAN2/Install)
- RELION Software (https://github.com/3dem/relion)

within the working directory. To access the GUI, type the following command into a terminal emulator:

- cryolo_gui.py

This will bring up the crYOLO interface. Since the general model is used in this section, a "config_cryolo.json" file will need to be set up by providing project details in the "config" tab.

Of critical importance will be the box size (in pixels). (Since class-B GPCR particles are roughly 150 Å in diameter, a pixel size of 0.83 Å/pixel and a box size of 160 pixels will adequately represent the size of our particles.) Additionally, in the "Denoising options" it may be worth increasing the low-pass cutoff from 0.1 to 0.3.

Running this configuration job will output the required .json file. Setting up a prediction job will then require filling out the GUI as seen in Figure 3.

Input will consist of the micrographs folder within the "Motioncor" directory of your project. Create a new "boxfiles" folder in your working directory for the output coordinates.

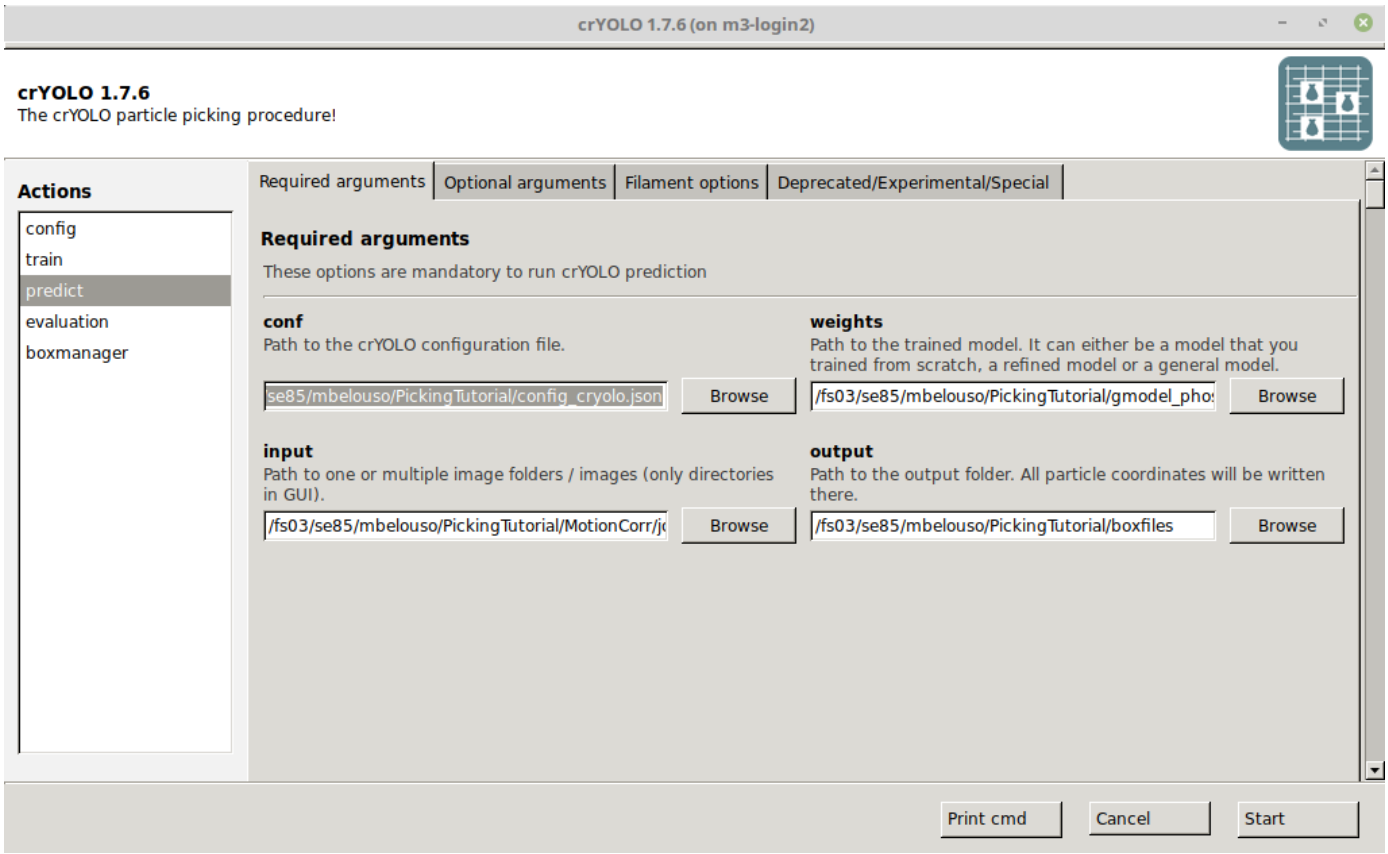

Figure 3. The crYOLO GUI used for particle-picking with the general-model deep neural network (DNN). The correct file and output locations have been defined in the "Predict" tab on the left. The input path for the image folders should point to the output directory of your dose-weighted motion-corrected averages.

Using crYOLO with pre-trained and trained models

1. Particle picking with crYOLO Software using a pretrained, general model*

A pre-trained set of weights can be a good starting point for the crYOLO neural network; up-to-date weights can be downloaded from the SPHIRE website. You can either make a symlink to a set of pretrained weights in your working directory or point to the file in the crYOLO graphical user interface (GUI), which can be found

NOTE: *If you have multiple GPUs available, ensure they are listed in the "Optional arguments" tab before initiating the particle picking.*

Once all required arguments are set, click "Start". The results can be assessed using:

- cryolo_boxmanager.py

They can also be overwritten by defining a different picking threshold.

**Micrographs in this application note consist of a class-B GPCR dataset that has been motion-corrected using RELION Software.*

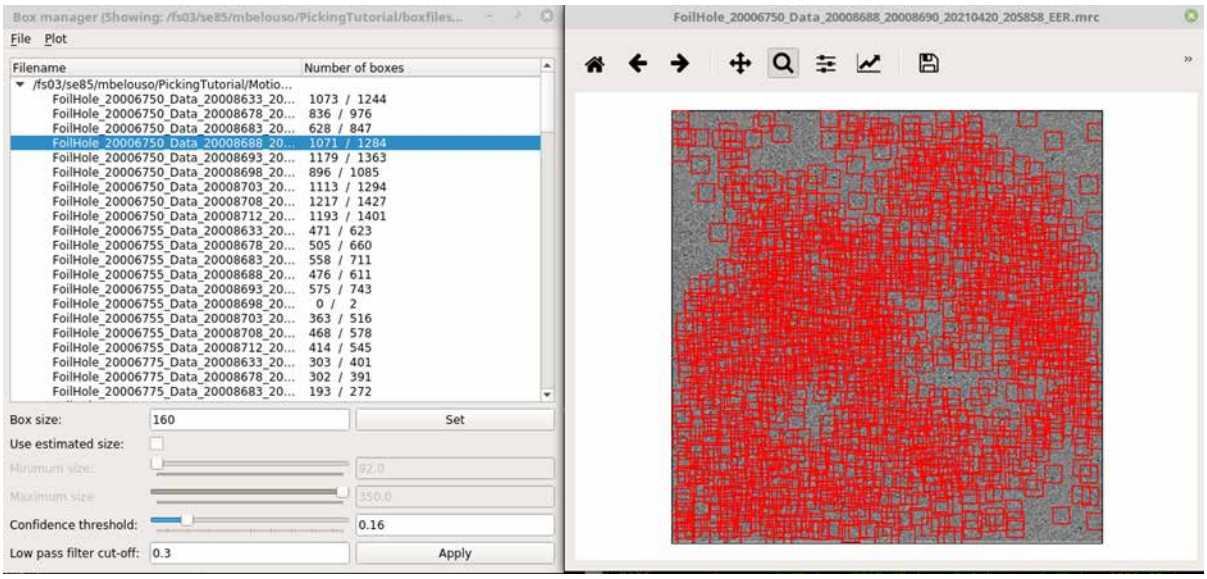

Figure 4. The crYOLO boxmanager. In this example, the image folder is located at MotionCorr/job001/Micrographs and the boxfiles are imported from boxfiles/CBOX.

It is important to tune the confidence threshold in order to avoid including too many "overpicked" particles (Figure 4). Once an appropriate threshold is found, the STAR files containing the particle positions for each micrograph can be rewritten through the *File > Save* option in the GUI.

These files must be renamed in order be imported into RELION Software. A simple way to do so is through a perl renamer, using the following command lines:

- cd boxfiles/STAR
- rename 's/\.star$/_manualpick.star/' *.star

Subsequently, a manual pick job must be set up in RELION Software in order to copy the crYOLO results into the corresponding folder. Picking a single particle from one micrograph will create the appropriate directory structure. Take note of the job number of your "ManualPick" so you can copy the crYOLO picks into the RELION working directory:

- cp boxfiles/STAR/*.star ManualPick/job0##/Micrographs
  - Here, ## represents the appropriate job number in the RELION directory structure

RELION Software can now be used to extract particles and begin 2D classification. After Fourier scaling the particles at the same time of extraction (to save computational time) a reference-free 2D classification is performed in RELION Software. Figure 5 shows the 2D classification of the extracted particles. Note that not all classes look like GPCRs and many of the classes may be poorly picked or uncentered.

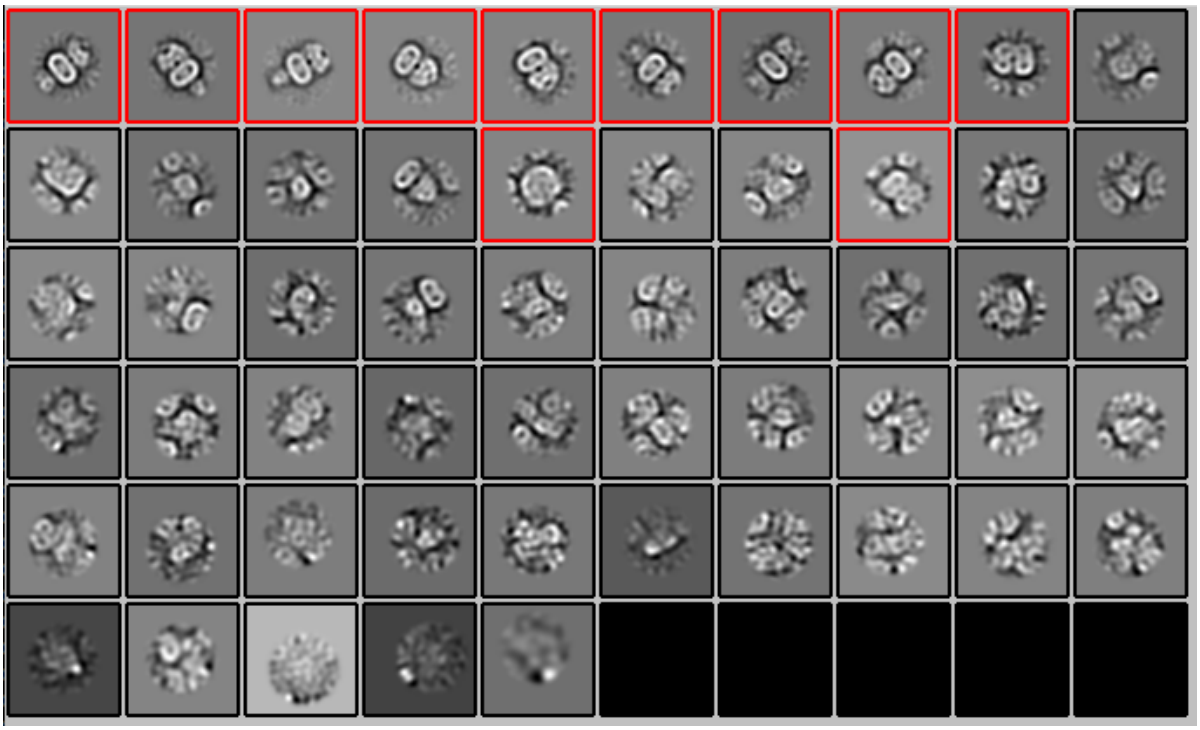

Figure 5. 2D classes from the crYOLO general model. The classes outlined in Red were selected for template picking. From this selection, only 18k of the initial 62k picked particles are clearly the molecule of interest.

In this example, a subset selection job was performed on the 2D classification and only visually appropriate classes were used as the 2D template in an effort to improve centered picking. An “Auto-picking” job was then performed in RELION Software with a low picking threshold (0.01), ensuring that as many particles as possible are included at this stage.

After particle extraction (using the same parameters as the crYOLO general model: 320pix box, Fourier scaled to 64 pixels), another round of 2D classification was performed (Figure 6). While there are fewer clear classes than in Figure 5, there are more projections captured by this classification. A subset selection was then performed on this data, in an effort to capture as many different projections as possible.

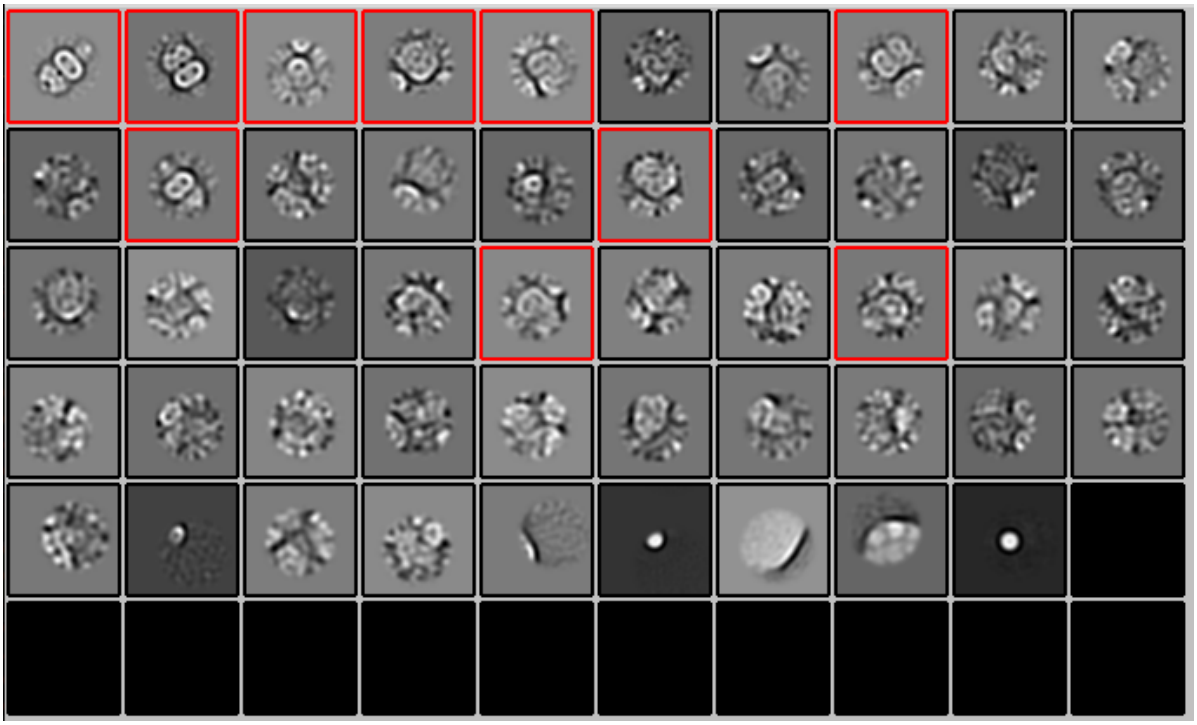

Figure 6. 2D classification of particles picked by the RELION template picker. Selected classes are shown in red.

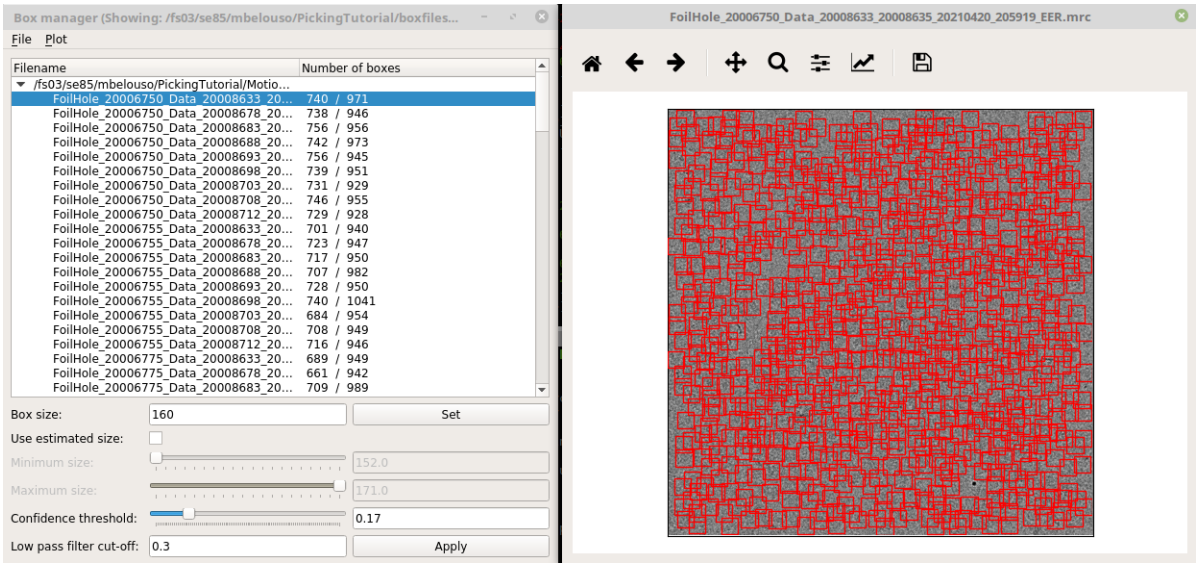

Figure 7. crYOLO boxmanager view of the results produced with the trained picking model.

2. Training your own set of weights for the crYOLO picker
Training a new set of weights for a neural network picker requires a training set of picked micrographs where most of the particle positions are correctly chosen. This data can be adapted from the crYOLO general model or templated picked 2D classifications that were obtained in the previous section. Subset selection and 2D classification were already performed, and a "particles.star" file of the selected particle positions was generated. The only additional step required is the extraction of particle positions from the monolithic "particles.star" file into separate particle-position files for each micrograph.

In this example, a software routine in the SPHIRE/EMAN2 Software Package was used on the 2D classification of the template-picked particles.

NOTE: The aim here is to create a .box file for each micrograph, which consists of a series of X,Y coordinates for each particle in the micrograph that share the same name with the file. These files can be read by RELION, EMAN2, SPHIRE, cryoSPARC, and crYOLO Software.

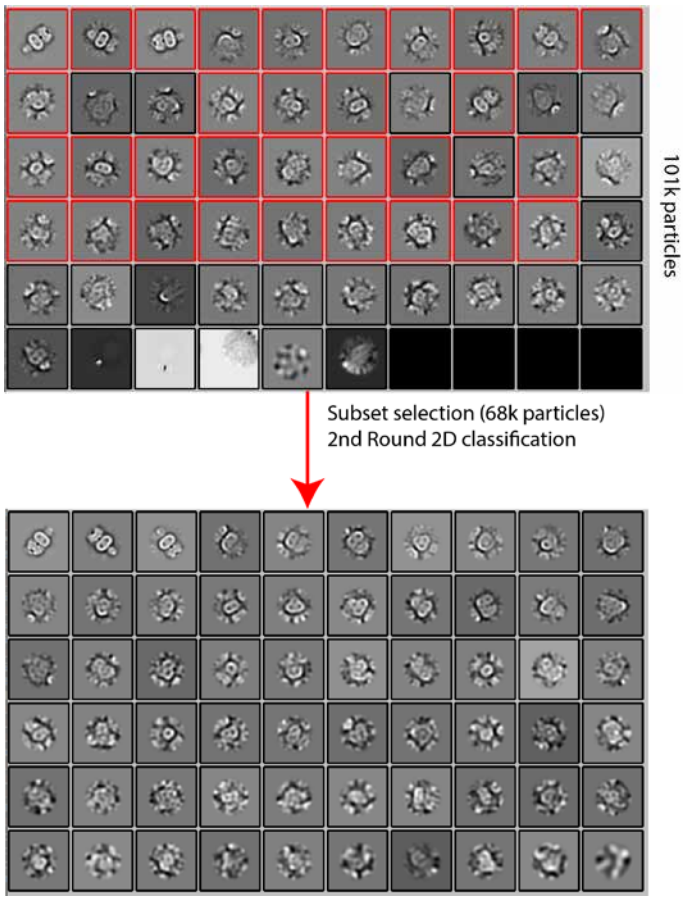

Figure 8. 2D classification results from the particles picked by a trained crYOLO neural network. The top panel is the first round of 2D classification. Selected classes that underwent another round of 2D classification are highlighted in red; the results are shown in the bottom panel. Note that the particles are well centered.

Using the following command in the working directory will create a new directory called "boxfiles_After2D," within which there will be a subfolder called "Coordinates." These will be used as the training input for crYOLO particle picking.

- sp_relion2sphire.py Select/job0##/particles.star boxfiles_After2D --relion_project_dir=./ --star_section=data_particles --box_size=160 --do_not_create_stack

In the config panel of the crYOLO GUI a new config (.json) file must be created to train a new set of weights. The "train_annot_folder" must also point to the SPHIRE-routine coordinates while the "train_image_folder" must point to the "motioncorrection" folder containing the dose-weighted averages.

The "train" panel is then pointed to the new .json file and, under "optional arguments," GPU 0 1 is specified. The training can then be initiated by hitting RUN.

NOTE: This can take up to 2-3 hours for 100 micrographs on a system with 2 GPUs. However, it can be sped up with either more resources (GPUs) or training on less micrographs. 20 well-picked micrographs, with various levels of defocus and featuring some ice/contamination, will often yield the same results.

Once training is completed, you will have a new *.h5 file containing your trained weights, which can be used as input for a new picking job. At this point a "predict" job can be run again in crYOLO Software using the newly trained weights instead of the general model. As is seen in Figure 7, the picking results are more consistent and seem to cover the micrograph more evenly.

Figure 8 shows the 2D-classification results from the particles picked by the trained neural network. Not only were more useable particles picked, but they were also well centered, and more projection angles were apparent in the 2D classes.

Discussion

The deep-learning-based particle picker crYOLO is an example of machine learning that increases the productivity and robustness of cryo-EM workflows, which is particularly important for pharma companies. It is often advantageous to train your own weights for crYOLO particle picking, as they routinely outperform a general model in specific use cases like GPCRs.

In practice, a random 20–40 micrograph subset is first chosen from the dataset. These micrographs are then over-picked using the built-in RELION Laplacian of Gaussian picker. A 2D classification is performed to clean up the over-picked results, and the coordinates are extracted into a new set of boxfiles using SPHIRE Software. These boxfiles are used for training in crYOLO Software.

Alternatively, to be even more accurate, the final 3D-refinement STAR file could be used and the boxfiles would be extracted from that. This could be useful if large amounts of data will be collected for many similar proteins. Generally, weights trained on an exemplar GPCR (e.g. the class B GLP-1 receptor) can produce good results for other GPCRs, including those from a different subclass (such as class-A GPCRs).

NOTE: it can also be useful to have a different set of weights for each direct electron detector.

Speaking broadly, these procedures ensure that particle positions can be accurately picked from packed grids. This has multiple advantages, as having more particles in a single micrograph allows more accurate determination of the contrast transfer function and lets the user identify more particles with less time on a microscope.

Beyond this, neural-network-based particle picking enables researchers to confidently collect closer to focus (e.g. 0.5–1.0 µm), as visual identification of particles is less critical with a particle-picking algorithm. Taken together, this increases the probability of determining higher-resolution structures using less data and instrument time.

thermo scientific

About the authors

Matthew J Belousoff, Denise Wootten and Patrick M Sexton are researchers at the Monash Institute of Pharmaceutical Sciences at Monash University and the ARC Industrial Transformation Training Centre for Cryo-electron Microscopy of Membrane Proteins (CeMMP). Thermo Fisher Scientific is a member of the CeMMP and supports the goal of training scientists in cryo-EM of membrane proteins.

References

1. Belousoff, M.J., Danev, R., and Sexton, P.M.; *Grid preparation for detergentsolubilized GPCR samples*, Thermo Fisher Scientific, 2020.
2. Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., et al. (2019); *SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM*, Commun Biol, 2019 2, 218.
3. Bepler, T., Kelley, K., Noble, A.J., and Berger, B.; *Topaz-Denoise: general deep denoising models for cryoEM and cryoET*, Nature Communications, 2020 11, 5208.
4. Tegunov, D., and Cramer, P.; *Real-time cryo-electron microscopy data preprocessing with Warp* Nat Methods, 2019, 16, 1146-1152.
5. Danev, R., Belousoff, M., Liang, Y.-L., Zhang, X., Wootten, D., and Sexton, P.M.; *Routine sub-2.5 Å cryo-EM structure determination of B-family G protein-coupled receptors*, bioRxiv, 2020.

Find out more at
thermofisher.com/PharmaceuticalResearchusingCryoEM

ThermoFisher SCIENTIFIC