



CASSPER-v2
A pixel based CryoEM particle picker
USER MANUAL

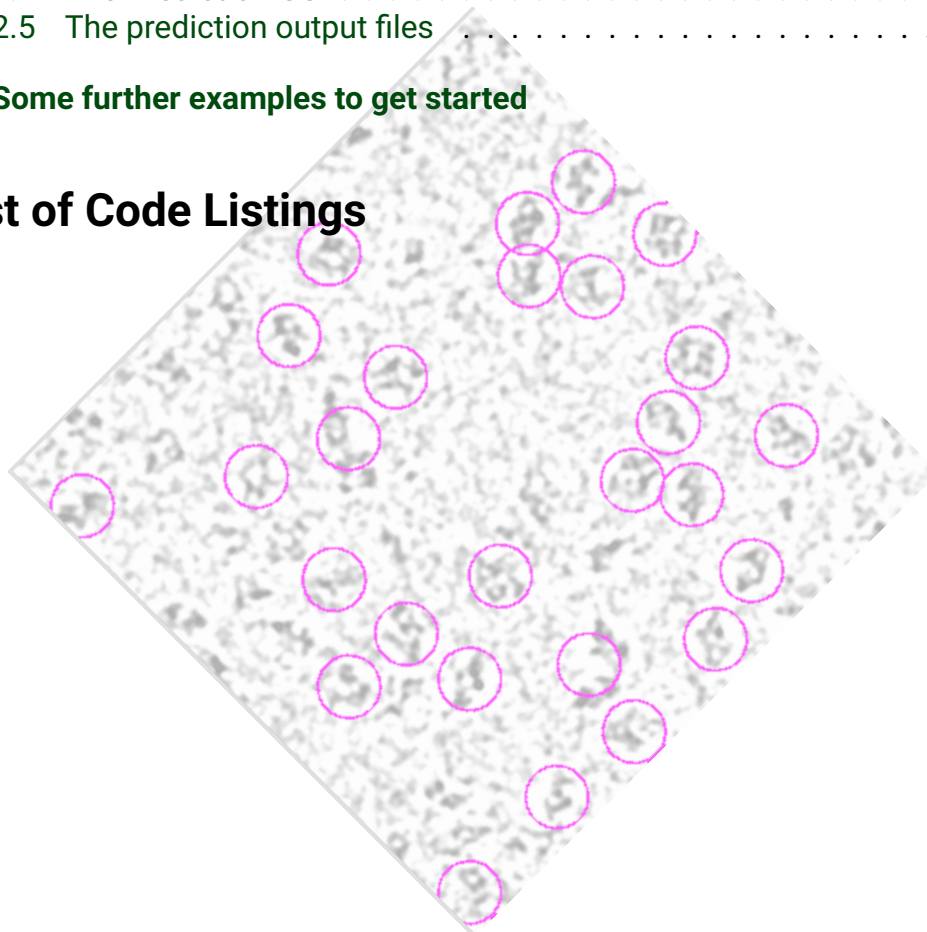
Version 2.3.1

Ninan Sajeeth Philip
Artificial Intelligence Research & Intelligent Systems
Thelliyoor - 689544
Kerala, INDIA
March 6, 2022

Contents

1	Introduction	2
2	Let us simplify complexities	2
2.1	Installation	3
2.2	Testing your first micrograph	3
2.3	The File Structure	4
2.4	The Prediction GUI	8
2.5	The prediction output files	12
3	Some further examples to get started	12

List of Code Listings



1 Introduction

CASSPER is a Cryo-EM Semantic Segmentation Particle Picker developed jointly by the Inter University Centre for Astronomy and Astrophysics (IUCAA), Pune, INDIA, the National Centre for Cell Sciences (NCCS), Pune, INDIA, and the Artificial Intelligence Research and Intelligent Systems (airis4D), Thel-lyoor, INDIA. The original work was published in Nature Communications in February 2021 with Blesson George, Anshul Assaiya, Robin Jacob Roy, Ajit Kembhavi, Radha Chauhan, Geetha Paul, Janesh Kumar and Ninan Sajeeth Philip as authors.

The paper was a demonstration of the capability of Semantic Segmentation based pixel classification deep learning model for the accurate picking of protein and other particle projections seen in cryo-electron microscope micrograph images.

Since Semantic Segmentation is a pixel based classifier, it is able to correctly classify protein particles even from low Signal to Noise Ratio for the accurate determination and reconstruction of the 3D structure of the protein molecule confirmations. Secondly, it learns the association of pixels rather than the morphology of the molecules in Cryo-EM, meaning that, it can make reliable picking of even unseen protein particles. Finally, the model is light weight and can be trained easily to customise for new protein particles.

CASSPER-v2 is a complete rewrite of the original CASSPER code by the author to make it user friendly and inaccessible to non technical user community. This USER MANUAL is written with the intention to provide a comprehensive but nontechnical narration of the features available in the new release of the program.

2 Let us simplify complexities

The highest priority in rewriting CASSPER code was to simplify the steps involved. Installation of CASSPER-v2 also is simplified. More over, it can run on PCs, laptops or even high performance computing systems using GPU through the same interface. Most of the tasks are done in the background and user interface is requested only when necessary.

CASSPER-v2 can work as a python library. Though the code has over 3000 lines, this facility allows users to train and pick particles on micrographs with single line code segments without any worries about the functional details of the code. Attempts have been made to create fast and intuitive GUI interfaces to ensure consistency and reliability.

CASSPER-v2 was developed and tested on Linux OS but should work

on any OS that support python and machine learning libraries. For optimal performance, Linux platform is recommended.

2.1 Installation

Let us follow the best practices. Rather than installing packages as super user, it is always good to have a virtual environment for installation. To create a virtual environment named CASSPER, type in the following command:

```
python -m venv CASSPER
```

Once the virtual environment is created, it may be activated by the command:

```
source CASSPER/bin/activate
```

In Linux terminal, you may see the command line modified with (CASSPER) in front of the command prompt. This means, you are now ready to install CASSPER and any additional modules that you may want to run it. Besides avoiding any conflict with existing Python libraries in your system, this also ensures that any upgrades on the system will not affect your installation. All you need to do is to activate your virtual environment as we did above.

The installation of CASSPER-v2 can be done from GitHub directly with a single line pip install command from Python 3.x environment as:

```
pip install git+https://github.com/sajeethphilip/CASSPER2.git --upgrade
```

This might take a while depending on your internet speed and system performance. If you get upset while waiting, it is a good time to go and have a cup of coffee.

2.2 Testing your first micrograph

To familiarise with the features of CASSPER2, let us walk through an example. Cryo-EM maps can be downloaded from the Electron Microscopy Public Image Archive EMPIAR. The files are stored as numeric arrays with the extension "mrc" and hence are casually referred to as mrc files. Each protein molecule in EMPIAR database are assigned a unique number as Label. Thus 10017 refers to the protein molecule Beta-galactosidase and can be downloaded from:

<https://www.ebi.ac.uk/empair/EMPIAR-10017/>

The reason for selecting this molecule is the availability of manually selected coordinates for these micrograph images by Richard Henderson, one of the three Nobel prize winners for the discovery of Cryo-EM technology in 2017 for the study of biological protein structure. Comparing the predictions of CASSPER with authentic ground truth data is helpful to know the pros and cons of CASSPER tool. In the subsequent discussions, we will explore a few precautions to ensure good predictions using CASSPER even when using mrc files of entirely new protein molecules. See Figure 1. To download mrc files, click on the + symbol under download icon.


This will open up a long list of mrc files and the coordinate locations of the protein particles in them as shown in Figure 2. By default, all the files are selected. Since we want only a few for the time being, click the left top tick mark to unselect all files and individually select a few. In fact, if your internet connectivity is poor, you can just download one mrc file and its coord file. There is an option to preview the mrc before downloading by clicking on the mrc file. This is especially useful when one wishes to collect a heterogeneous set of images for training CASSPER.

Download the selected files by clicking on download. The coord files are simple text files that have the coordinates of the protein particles. CASSPER reads only the more common star files and thus we need to convert them to the star file format. "star" files usually have some additional information as metadata in them. Since we do not have that information, we will create pseudo "star" files from these coord files and use them with CASSPER.

2.3 The File Structure

CASSPER has a flexible file structure. However, for simplicity, we will use the default paths. The default path to store the mrc files is mrc/ folder in the working directory. Create a folder mrc/ and move the downloaded mrc files into it. Now create a folder StarFiles/ under mrc/ folder. This is where we will store the ground truth coordinate values as a pseudo star file. After creating the folder, move all coord files to the folder mrc/StarFiles/ and rename them to star files. Since these are text files, they can be opened in any text editor. Open them one by one and add 9 blank lines on top of each file. This is sufficient to convert them into pseudo star files that CASSPER can read. Save them with the same file name. It may be noted that the root name of both the mrc file and the newly created star file are the same. In fact, that is how the program identifies the star files corresponding to each mrc file.

We are now all set to run CASSPER for the first time. When you run CASSPER for the first time in a folder, it will create a few more folders and files that are necessary for it to run. However this is all automatically handled




[EMPIAR home](#)
[Deposition](#)
[REST API](#)
[FAQ](#)
[About EMPIAR](#)
[Policies](#)
[Feedback](#)
[Share](#)

EMPIAR-10017

Beta-galactosidase Falcon-II micrographs plus manually selected coordinates by Richard Henderson

Publication:

Semi-automated selection of cryo-EM particles in RELION-1.3.

Scheres SH 

J Struct Biol **189** 114-122 (2015)

PMID: 25486611

DOI: 10.1016/j.jsb.2014.11.010

Related EMD entry:

EMD-2824

Deposited:

2014-11-19

Released:

2014-11-19

Last modified:

2014-11-19

Dataset size:

5.3 GB


Dataset DOI:

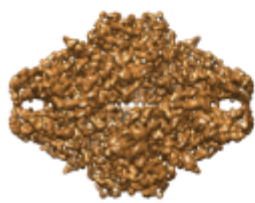
10.6019/EMPIAR-10017

Version history:

1	2015-10-28	Restructured the entry directory to have "data" and ".private".
---	------------	---

Contains:

 micrographs



Experimental metadata: [Download xml](#)



Image set


1. Beta-galactosidase micrographs + coordinates

Category:	micrographs - single frame
Image format:	MRC
No. of images or tilt series:	84
Image size:	(4096, 4096)
Pixel type:	32 BIT FLOAT
Pixel spacing:	(1.77 Å, 1.77 Å)
Details:	Particle positions in corresponding .coord text files (with X and Y coordinates for manually selected coordinates in 1st and 2nd column respectively). The micrographs are the average, without any realignment, of 24 raw movie












[Show more](#)

[Download](#)

  data 5.3 GB

License: 

Quick links

-  Talks and Tutorials
-  EMPIAR Quick tour
-  Statistics
-  Publications
-  Re-use case study
-  EMPIAR in the news
-  Contact us
-  EMDB
-  PDB
-  BioImage Archive
-  EMPIAR@PDB

EMPIAR citations

A biological nanofan: The wall of coniferous bisaccate pollen.
Cajocaru R, Mannix O, Capron M, Miller CG, Jouneau PH, Gallet B, Falconet D, Pacureanu A, Stukins S. (2022)

Protein-lipid interaction at low pH induces oligomerization of the MxA cytotoxin from *Vibrio cholerae*.
Nadeem A, Berg A, Pace H, Alam A, Toh E, Ådén J, Zlatkov N, Myint SL, Persson K, Gröbner G, Sjöstedt A, Bally M, Barandun J, Uhlin BE, Wai SN. (2022)

An archaeum filament composed of two alternating subunits.
Gambelli L, Isupov MN, Connors R, McLaren M, Bellack A, Gold V, Rachel R, Daum B. (2022)

Figure 1: The EMPIAR website provides publicly available Cryo-EM micrographs for downloading and testing.

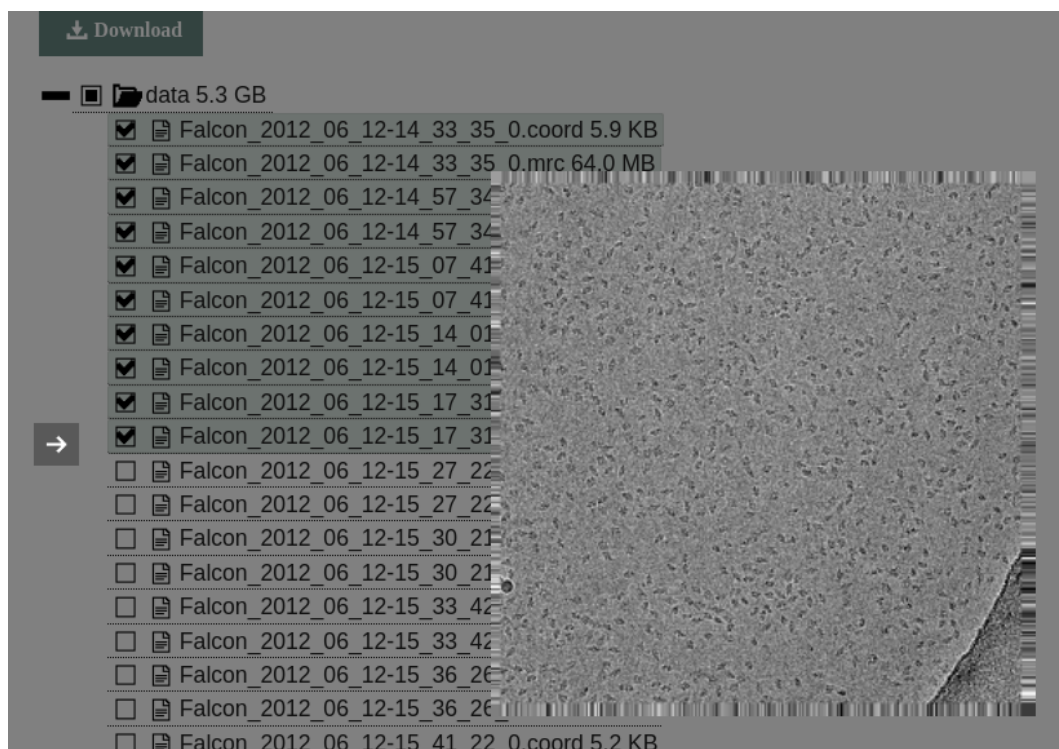


Figure 2: The list of mrc files and the list of coordinates of protein particles in them with extension coord are shown. One can download either the entire 5.2GB of data or select individual files for downloading. Clicking on the mrc file also display a snapshot of the image.

```

Working Directory ----->mrc/
|                               |____StarFiles/
|_____png/
|           |_____P_files/
|           |_____Predict_labels/
|           |_____Box/
|           |_____StarFiles/
|           |_____selected/
|_____models/
|           |_____checkpoints/
|_____Cmetadata/

```

Figure 3: File structure used by CASSPER during the Prediction round

in the background. Notice that CASSPER has to download the model files from the internet and so you need an active internet connection while you run CASSPER for the first time. The models are saved to a folder `models/` in the current working directory and in case you want to work in another directory or machine, it is fine to copy the `models/` folder to that directory or machine to save internet usage. Additionally, CASSPER in this round will create a folder `png/` in the current directory. This is where the results are stored. Four sub folders are created under the `png/` folder. The full directory structure used during the Prediction round of CASSPER is shown in Figure 3. The `P_files/` folder holds the contrast enhanced and histogram corrected png image of the micrograph that will be used for estimating the coordinates of the protein particles in the micrographs. The predicted label files are marked on a blue background and are placed under the folder `Predict_labels`. The colour codes used in these labels are defined in `class_dict.csv` file under the `models` folder. The box coordinates of the protein locations are given as a text "box" file in the folder `Box` and the same coordinates as a star file are placed in the `StarFiles/` folder with their image showing the protein locations in the sub folder `selected`.

The folder `Cmetadata` holds scaling and selection information relating to the files which will be described in Prediction GUI section. We shall now run CASSPER in the Prediction mode. Open a terminal and start Python and import CASSPER2 are shown in Figure 4.

CASSPER will check your working setups and will issue some messages before returning back to the Python prompt. If there are no errors (warning are okay), you may type `CASSPER2.PredictMRC()` to initialise the prediction round. At this time, CASSPER will check for model files and if it does not find


```

-----
(CASSPER)$: python
              Python 3.8.10 (default, Nov 26 2021, 20:14:08)
              [GCC 9.3.0] on linux
>>> import CASSPER2

              -----Some messages from tensorflow-----

>>> CASSPER2.PredictMRC()
              ----Some message ----
-----

```

Figure 4: The commands to run the Prediction model in CASSPER2

them, will try to download them from the internet. Make sure that you have an active internet connection at this time. Notice that the () in PredictMRC command is empty. That means, we are running the Prediction round with default setup for CASSPER2. However, it is possible to change these default options by command line arguments as:

```

PredictMRC(image=None,model='default',
model_path='/models/',
            model_name='BestFr_InceptionV4_model_FRRN-B.ckpt',
            crop_height=512,crop_width=512,model='FRRN-B',
            dataset='./png/',mrcsrc='./mrc/',Savestar=True,
            CrossMatch=False,Box=True)

```

In the default mode, the model_name in command line is ignored and instead the Cross-model for mrc detection ("CrossPro.ckpt") described in the CASSPER Nature paper is used. This is to make sure that the model works well across different mrc files. The default option can be activated from command line by setting model='default'. If the model variable is not default, it will use whatever that is given as model_name as the name of the model. The given model_name "model_name='BestFr_InceptionV4_model_FRRN-B.ckpt'" is the name CASSPER2 assigns to the model that gets created during the training round that will be described later.

2.4 The Prediction GUI

Despite all precautions taken, the vitrification process in the creation of micrographs will introduce several types of artifacts into the image. It is thus

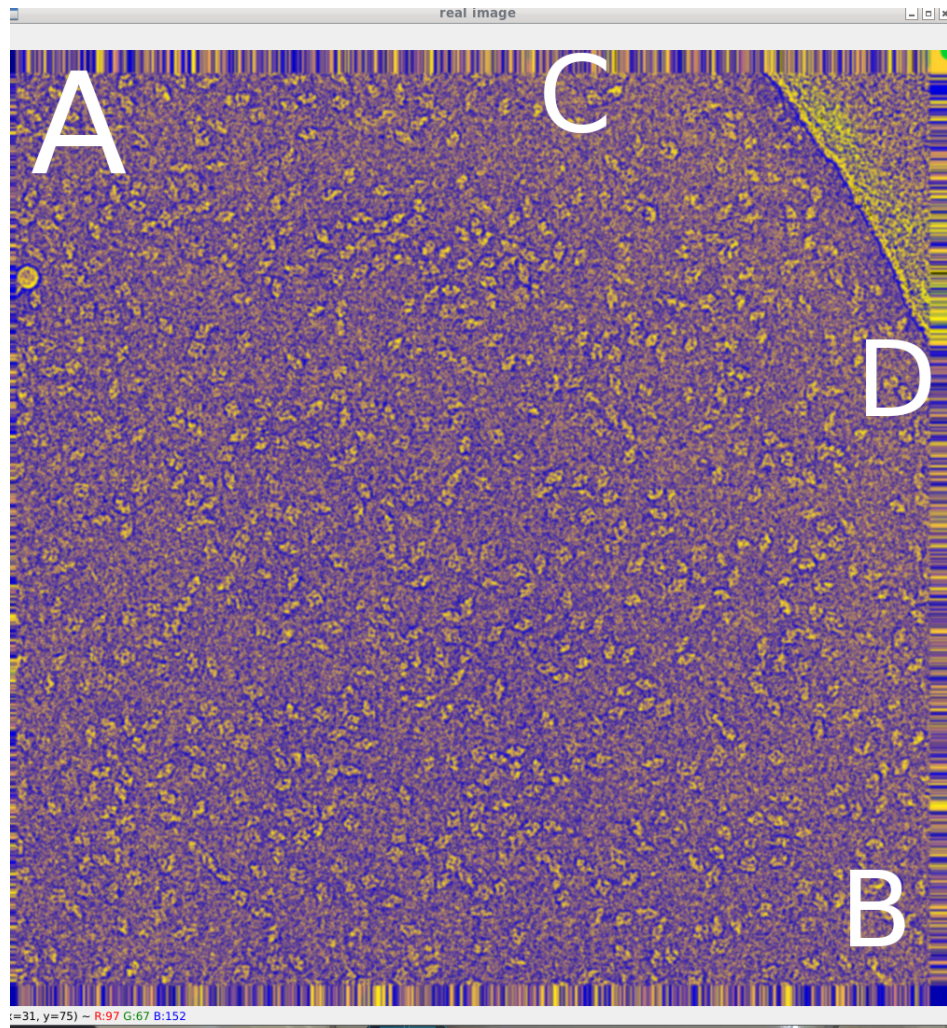


Figure 5: The micrographs are presented to the user for selecting and rejecting regions of interest. We want to reject the borders that have some artifacts. For that, drag the mouse with the Left mouse button pressed from point A to B. This will display another image that has the rectangular region with A and B as corners. Since we want to use this region, press "a" key for add. That will close the new window and you will return to the first image. It is also possible to select multiple regions and add them in case where you have a highly deformed micrograph. CASSPER can recognise the presence of ice and graphite contaminants. But to be safe, we may want to remove some distortions near the ice region on left top of the micrograph. For this, drag the mouse with left button pressed from region C to D. Note that you need to always drag mouse from left to right. This will display the rectangular region between C and D. Since we want to reject it, press "r". Now press "q" to exit from the selection procedure.

desirable to minimise the possibility that they may result in the wrong identification of the protein particles. Since humans are best in detecting artifacts, a GUI interface is presented to the user. Since the micrographs might have a resolution much higher than the screen resolution, the display might go out of screen. Press the +/- keys to zoom in and zoom out the image to fit the screen. See the instructions given under Figure 5. On the terminal the name of the micrograph and the the set of key bindings that are associated with the image displayed. This GUI will be presented for the subsequent images also.

Testing image Falcon_2012_06_12-14_33_35_0.png

```
-----
|           Press "q" key to quit           |
|      "a" key to append to selection      |
|    "r" key to remove from selection    |
|    +/- to zoom in and out image        |
|-----|
```

Once the region selection process is completed, a second GUI will now show up with a set of slide bars and a few green circles around some probable protein particles in a portion of the micrograph. See Figure 6. This GUI appears only for the first micrograph and its purpose is to redefine protein particles that are to be considered. An option like this is often helpful to get rid of protein fragments or other unwanted elements that look like the protein in the micrograph. As above, a brief description on key bindings will now appear on the terminal as below:

Reading Image: ./png/Predict_labels/Falcon_2012_06_12-14_33_35_0.png

```
-----Verify the selection-----
Move frame up,down,left or right by pressing u,d,r,l
Adjust radius and other parameters as required
Select kernel type by pressing k
To save selection press s after each setting
Press q to quit
-----
```

A zoomed version of the micrograph is shown to ease the selection procedure. It is possible to navigate up and down through the micrograph by typing "u" for up, "d" for down, "r" for right and "l" for left. The selected particles are those coming under the circles and it will change as you change any of the slide bars. Adjust them so that the circles are over the desired protein particles. It is not necessary to have all the protein particles selected in one go. You can press "s" key to save your selection at any time and go on to

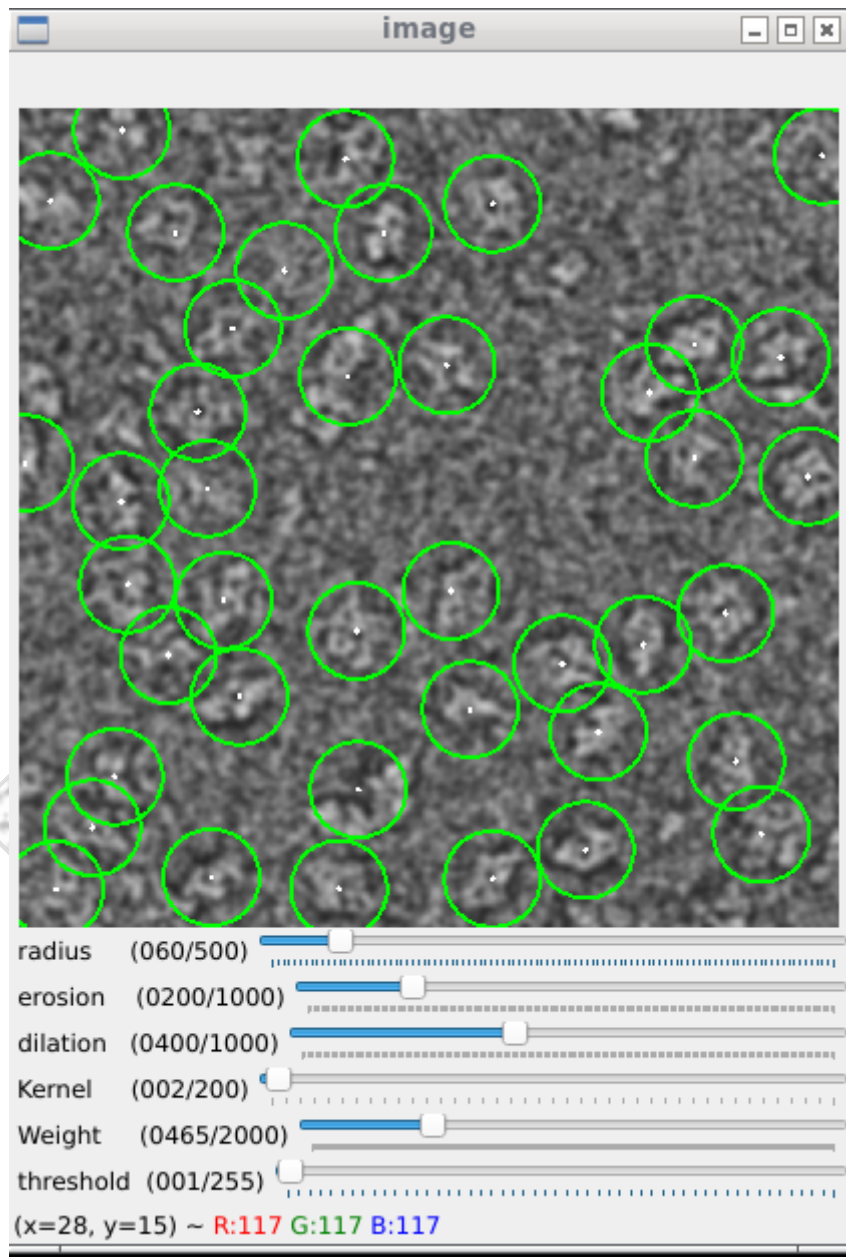


Figure 6: The purpose of this GUI is to isolate potential protein particles from fragmented protein and other contaminants in the micrograph images. This can be done by adjusting the slide bars to maximise the number of genuine protein particles under the circles. To accommodate varying protein structures due to projection effects, one can press the "s" key to save intermediate selections and go on adjusting the track bars to select new shapes. A white dot appears at the centre of the proteins already selected to help target the missing ones in subsequent steps. Navigate through the micrograph using "u","d","l" and "r" keys to ensure that all regions are covered. Finally press "q" key to quit.

select the missing ones. A white dot will appear on top of all proteins that are selected already so that you know what to look for. You can tolerate some wrong labeling here as those are most likely rejected when the protein alignment is done later, after particle picking. Make sure that you select all the genuine candidates.

CASSPER will now display your selections somewhat as below and will move on to display the next micrograph. However, in this case only the first GUI will be available and whatever selection criteria that was used in the case of the first micrograph will be used for the subsequent images. **That means, all the micrographs in one run should be of the same kind. You should remove the png/ folder before using CASSPER for a different protein.**

Your required Radius are [60, 21, 21]
Your required erosion count are [200, 200, 200]
Your required dilation count are [400, 400, 400]
Your required kernels are [1, 1, 1]
Your required weights are [23, 53, 53]
Your required thresholds are [0, 0, 0]

2.5 The prediction output files

The outputs from CASSPER is available for each micrograph as a labeled image as well as in a star file and box file format that it may be readily integrated with any existing 3D reconstruction software packages for structure determination. As explained earlier, all results are stored in sub folders in the png/ folder.

3 Some further examples to get started