Efficient Cryo-Electron Tomogram Simulation of Macromolecular Crowding with Application to SARS-CoV-2

Mayukhmali Das

Cryo-electron tomography enables the imaging of the three dimensional structure of macromolecular complexes at nano resolution and at native conditions . Some of the methods which can be used to analyse the Cryo-ET data are :

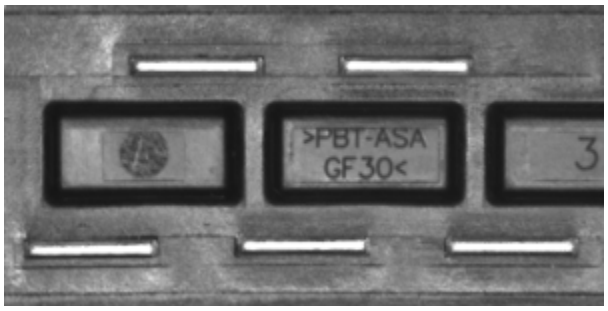
Template Matching :

Template matching is a typical approach for segmenting particles of known structures in the tomogram . Template Matching is a high-level machine vision technique that identifies the parts on an image that match a predefined template. Advanced template matching algorithms allow you to find occurrences of the template regardless of their orientation and local brightness.

Naive Template Matching

Imagine that we are going to inspect an image of a plug and our goal is to find its pins. We are provided with a *template image* representing the reference object we are looking for and the *input image* to be inspected.

Template image



Input image

We will perform the actual search in a rather straightforward way – we will position the *template* over the image at every possible location, and each time we will compute

some numeric measure of similarity between the template and the image segment it currently overlaps with. Finally we will identify the positions that yield the best similarity measures as the probable template occurrences.

One of the subproblems that occur in the specification above is calculating the *similarity measure* of the aligned template image and the overlapped segment of the input image, which is equivalent to calculating a similarity measure of two images of equal dimensions. This is a classical task, and a numeric measure of image similarity is usually called *image correlation*.

The fundamental method of calculating the image correlation is so-called *cross-correlation*, which essentially is a simple sum of pairwise multiplications of corresponding pixel values of the images.

Though we may notice that the correlation value indeed seems to reflect the similarity of the images being compared, the cross-correlation method is far from being robust. Its main drawback is that it is biased by changes in global brightness of the images - brightening of an image may sky-rocket its cross-correlation with another image, even if the second image is not at all similar.

$$\text{Cross-Correlation}(\text{Image1}, \text{Image2}) = \sum_{x,y} \text{Image1}(x, y) \times \text{Image2}(x, y)$$

We can apply this template matching with cross-correlation method to out tomogram :

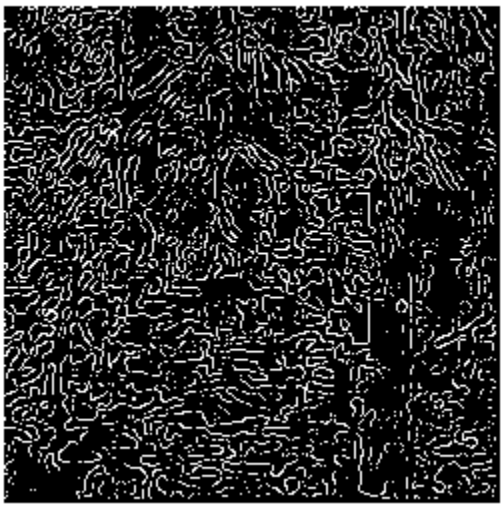
A template is cross-correlated over a tomogram to find locations and angles where the template matches the most

Difference of Gaussian method :

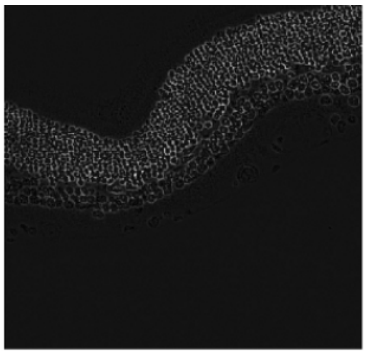
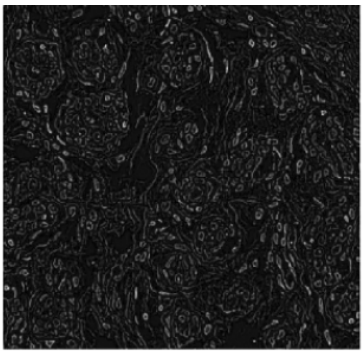
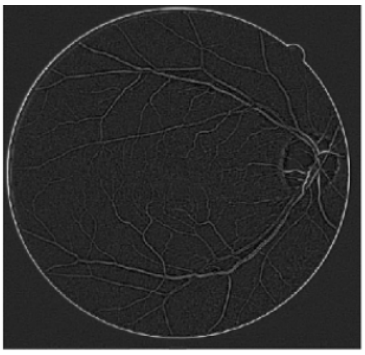
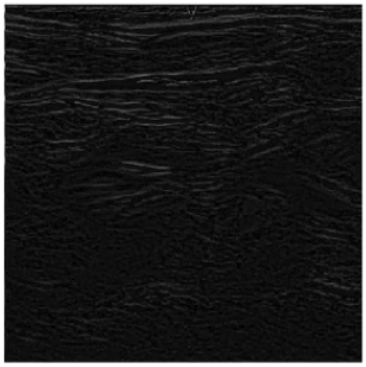
As the difference between two differently low-pass filtered images, the DoG is actually a band-pass filter, which removes **high frequency components representing noise**,

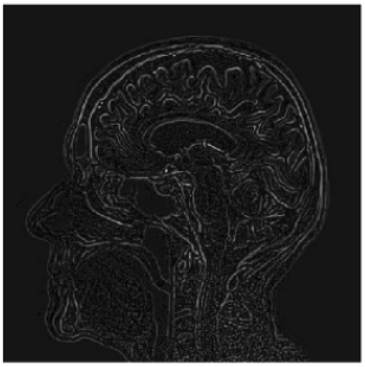
and also some low frequency components representing the homogeneous areas in the image. The frequency components in the passing band are assumed to be associated with the edges in the images. It is a method where features and edges can be easily detected .

Edge detection by DoG operator:



Some images generated using DOG process :



With the evolution of machine learning and deep learning , various models are invented which makes Cryo-ET analysis easier :

Some of the tasks while analysing are :

1. Localization (3D ResNet, U-net)
2. Classification (SVM , Kmean clustering etc)
3. End-to-end semantic segmentation

How to simulate Cryo-ET ????

The typical approach of simulating tomographic images is first to calculate the density map from atomic structures and then add distortions to the density map due to the missing wedge effect, interactions between electrons and the specimen, as well as introduced by image detectors

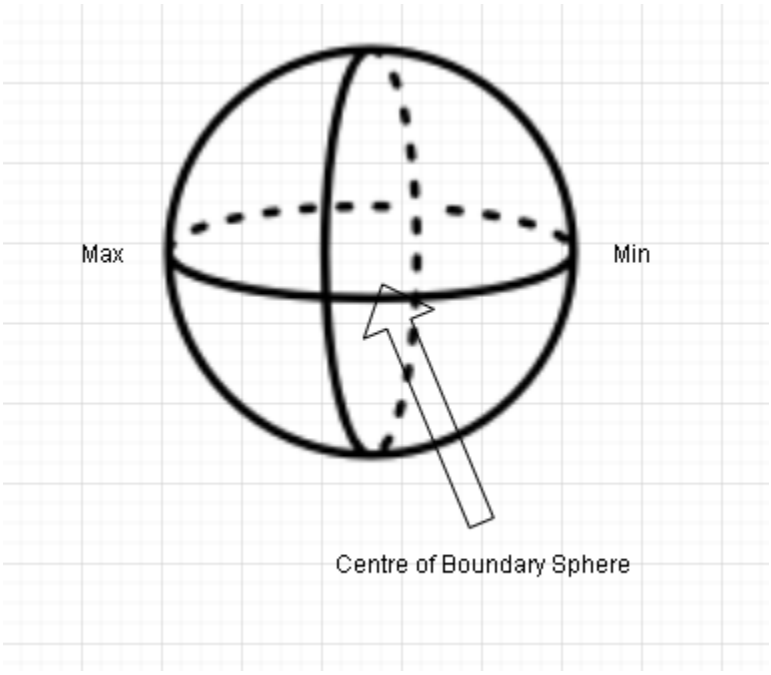
Efficient packing of Macromolecules for Cryo-ET Simulation :

To simulate cryo-ET efficiently, we need to first simplify all the macromolecules into single spheres. These spheres will be randomly placed in a box and packed together using the gradient descent method. The final coordinates will be used to synthesize the 3D density map with the help of single density maps of all the macromolecules.

Steps :

1. Simplify Macromolecule to Sphere

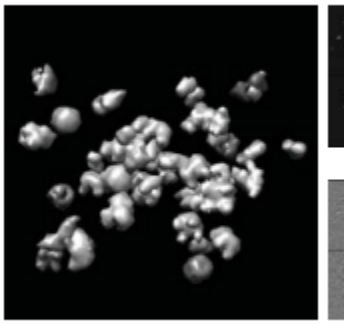
All the **macromolecules** are represented by minimum boundary spheres with different radii. For a macromolecule P, the center of the boundary sphere is obtained by the mean value of the maximum and minimum atoms coordinate on the X, Y, Z axis.



2. Initialization

A target macromolecule and several random neighbor macromolecules are selected to generate the subtomogram.

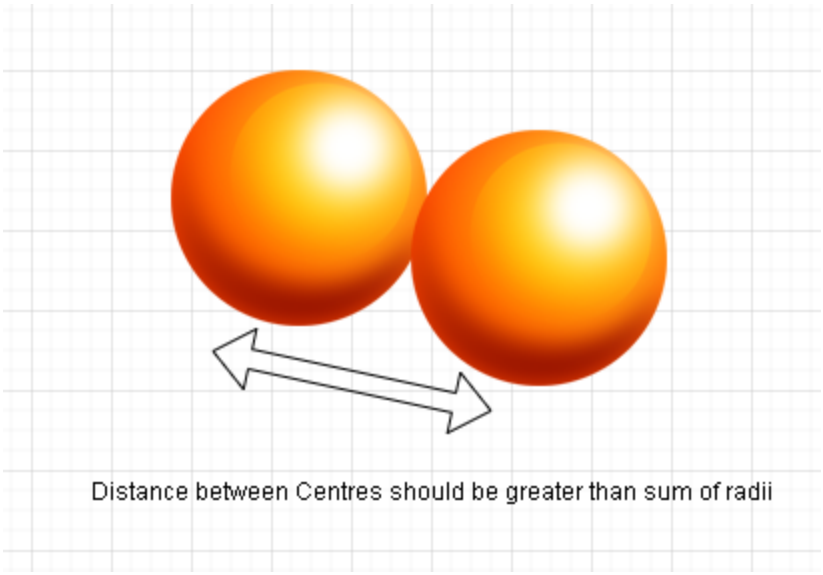
The value of l of the cubic simulation scene is generated using a particular formula .



(a) 3D visulization

We then initialize all macromolecules with different locations by setting macromolecules' centers randomly according to the box size.

The overlap is avoided by forcing the Euclidean distance between two sphere centers to be greater than the sum of their radii



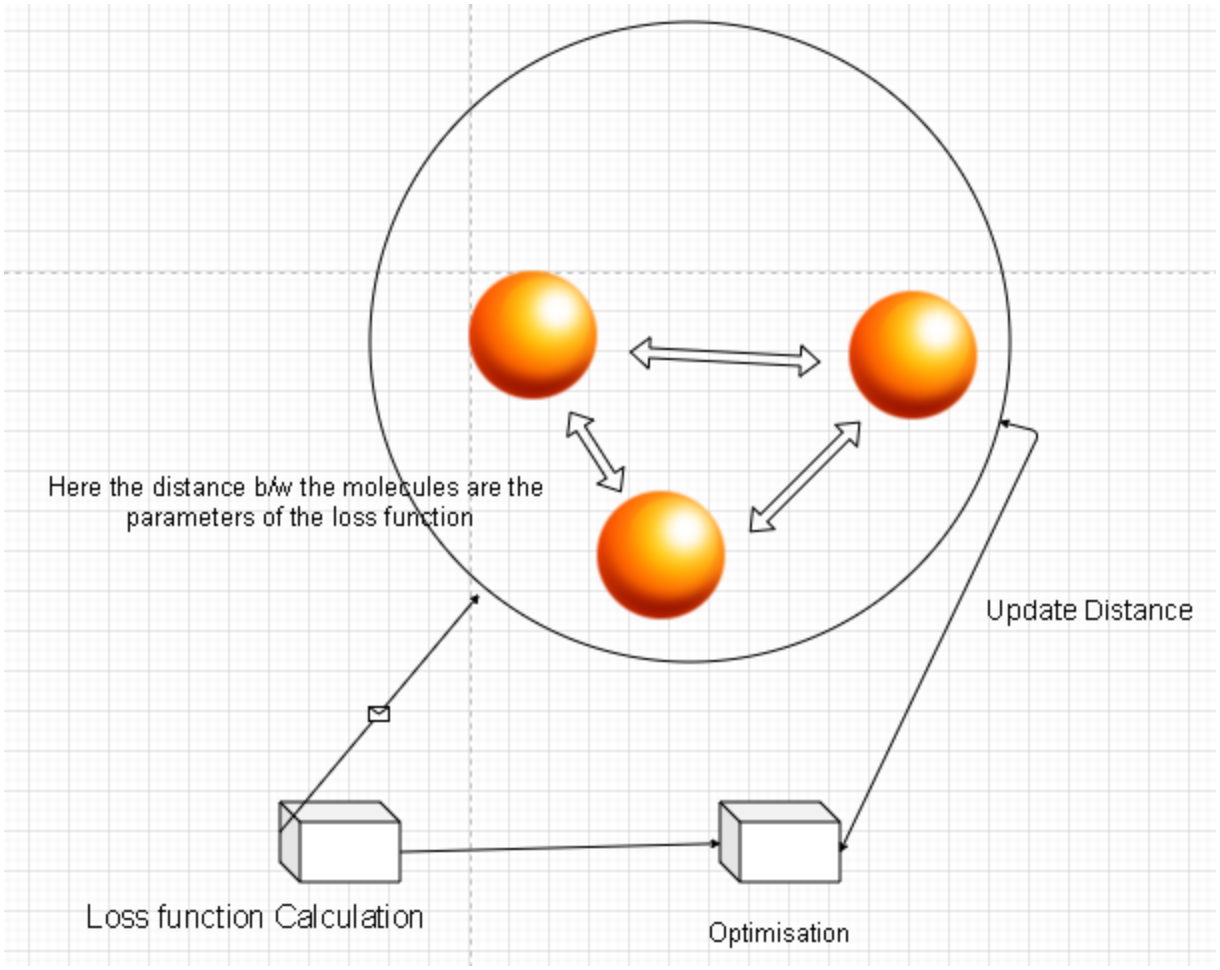
3. Packing Macromolecules to Macromolecular Crowding

We need to reduce the distance between the macromolecules . So if we define a loss function which has the distance term in it , then in the optimization step the loss function will be reduced as well as the distance will be reduced , thus the molecules will be packed .

$$Loss_{P_k} = \sum_{i}^{N} (x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2$$

An important thing to keep in mind is that the macromolecules should not overlap , so checks must be done after each backpropagation and optimisation step .

Thus this step will continue until all loss function converges and we will get the most packed output



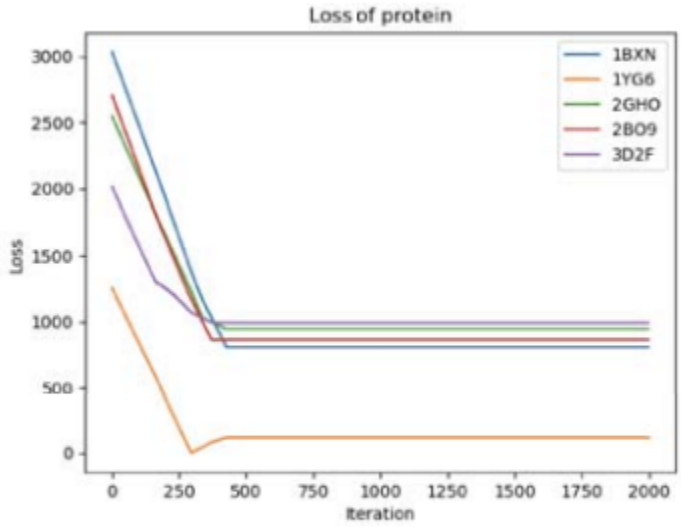
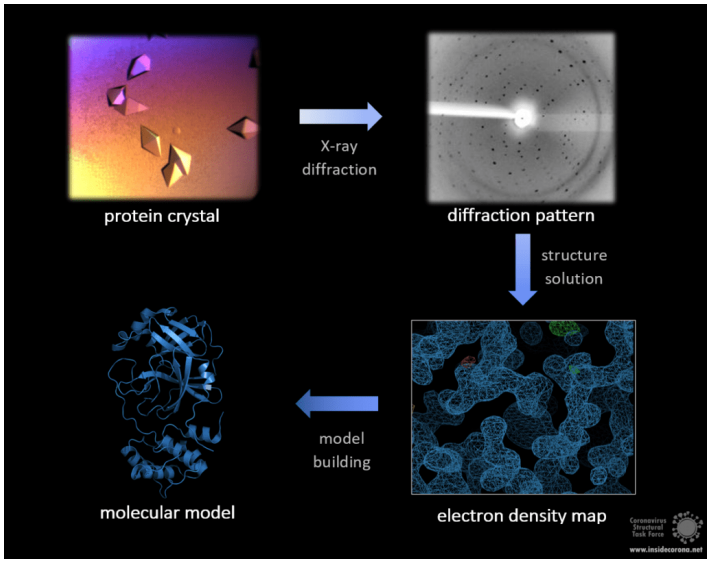
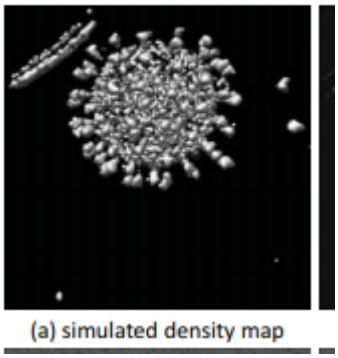


Fig. 4. The loss function value in each step of the five selected proteins.

4. Simulate Density Map :

Density mapping is simply a way to show how macromolecules are concentrated in a given area .



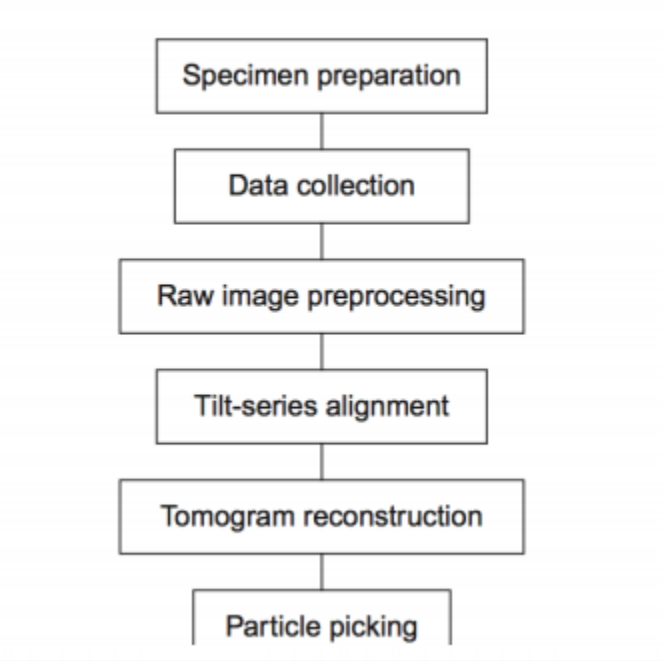


Now the macromolecules are optimally packed . Now the final Density map is generated by combining the single density maps of all single macromolecules based on their PDB file, the final coordinate and a random orientation .

PDB file is Protein Data Bank File , an example is :

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

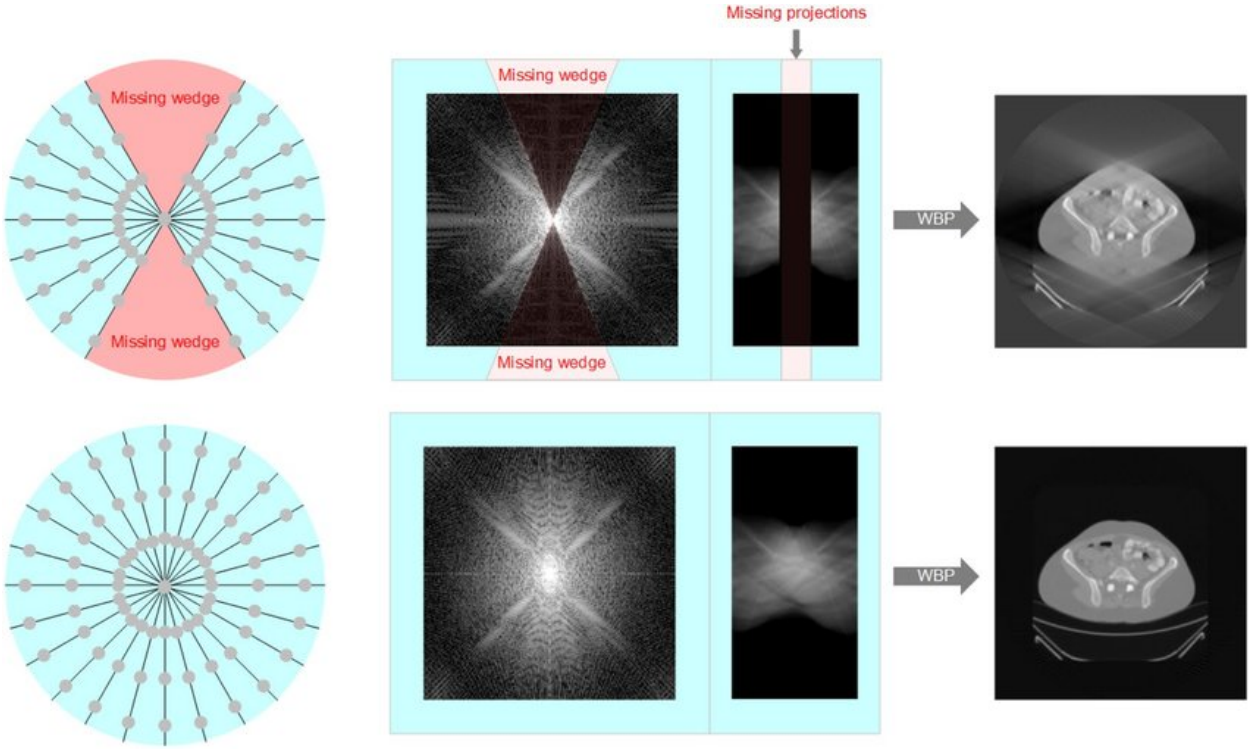
5. Simulate Tomogram :



According to the actual tomographic image reconstruction process, a simulated cryo-ET is generated by adding noise, tomographic distortions and electron optical factors.

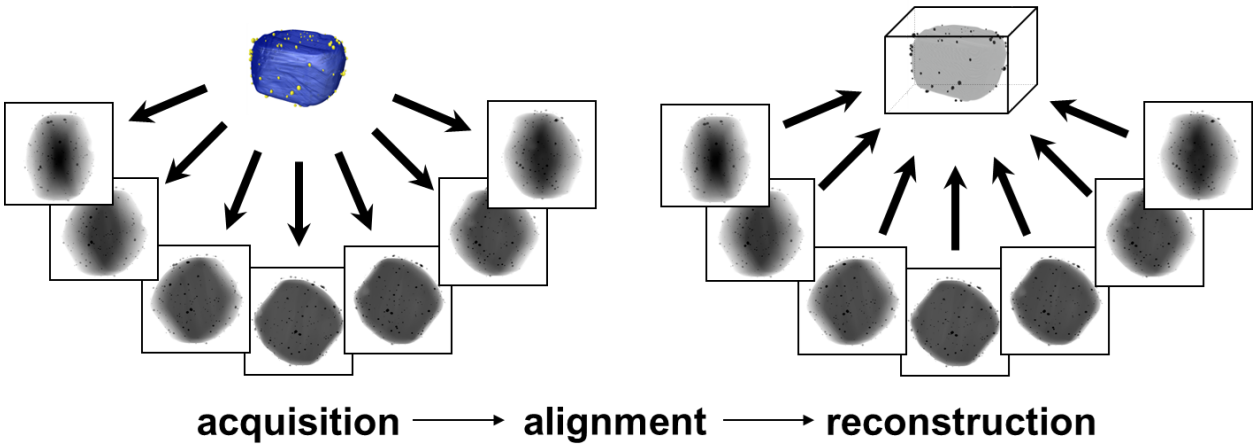
Noise was added to achieve different SNR levels.

The tomographic distortions are caused by missing wedge effect, which is due to the limited tile angle with range typically [−60, 60]. We simulated 2D projection electron micrographs of the simulated sample using a tilt angle range from -60 to 60 degrees with step increments of 2 degrees.



Missing wedge

Then we reconstructed the cryo-ET via a back projection algorithm :



Some algorithms for back projection :

- Weighted back-projection (WBP)
  - Most widely used
  - High-resolution structures obtained in cryo-ET and subtomogram averaging are from WBP-tomograms

- Iterative reconstruction algorithms
  - Algebraic reconstruction technique (ART)
  - Simultaneous iterative reconstruction technique (SIRT)
  - Simultanous algebraic reconstruction technique (SART)
  - Give improved contrast over WBP and are more often used in cellular tomography

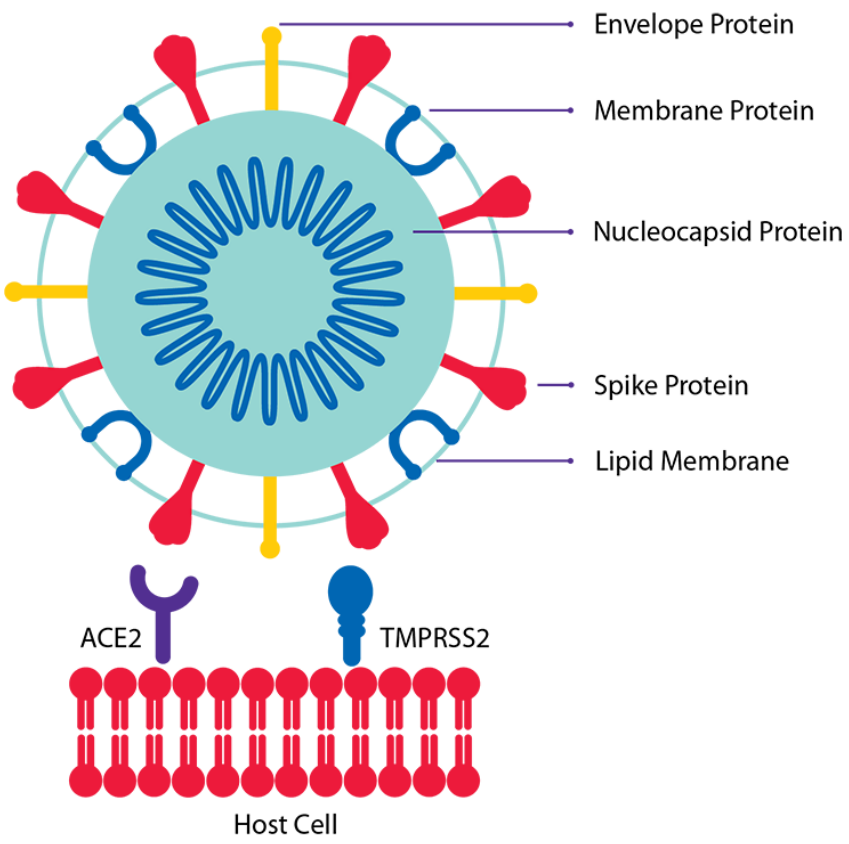
SARS-COV-2 TOMOGRAM SIMULATION :

1. A SARS-CoV-2 virion and its constituent proteins are randomly selected as the primary content of subtomograms.

2. We exploit the **proposed packing algorithm** to calculate these particles' random locations. This method is discussed earlier .

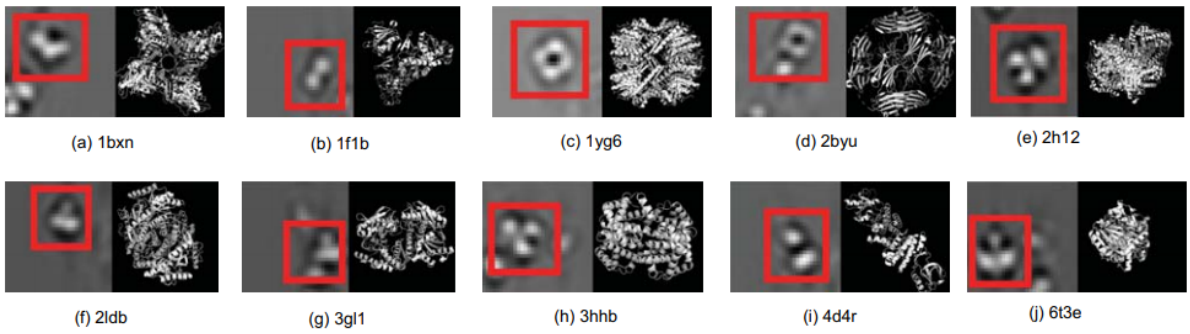
3. The particles are then placed according to their locations after a random rotation is used to generate the **density map** of SARS-CoV-2 surrounded by its constituent macromolecules.

4. To give the virion a more realistic appearance, we fill the virus with nucleocapsid proteins



5. To mimic the scene where the virus infects cells, we overlay the simulated cell membrane onto the resulting density map, which is part of a sphere with radius Rcell.

6. The **general density map** of the virus and cell scene is obtained through a combination of the density maps of the packing result including the virus and its constituent macromolecules, the N-Proteins inside the virus, the cell membrane, and ACE2 on the cell surface.



7. Fig. 5. The 10 macromolecules in the simulated dataset used for assisting cryo-ET classification method. In each subfigure, an example of subtomogram slice is shown in left, and the pdb result is shown in the right.

Now there are various other steps which are mentioned in the paper while generating the Tomography simulation . But as our main focus is Data analysis lets focus on that part after this basic introduction .

Quoting a paragraph from the paper :

“ Both S-Proteins and ACE2 can be clearly seen in our simulated images and can be traced with their position label. Our proposed approach can automatically generate enough scenes with the virus and the cell membrane, which can be used as a benchmark dataset for testing developing cryo-ET analysis algorithms or as labeled samples to train macromolecule detection and classification framework. “

Classification :

1. ## Using Squeezenet :

The SqueezeNet model that is traditionally used for the classification of images and was extended to classify the 3-D sub tomograms present in the dataset. We choose the SqueezeNet model because of its high performance on regular computer-vision-based tasks in various domains. The novelty of SqueezeNet lies in the fact that it uses Fire modules

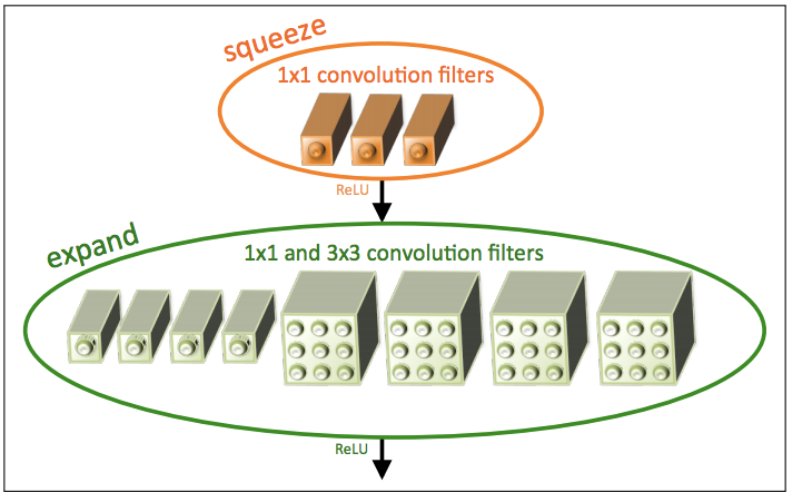


Figure 1: Microarchitectural view: Organization of convolution filters in the **Fire module**. In this example, $s_{1x1} = 3$, $e_{1x1} = 4$, and $e_{3x3} = 4$. We illustrate the convolution filters but not the activations.

2. Using CNN

The CNN architecture consists of two blocks. Each block consists of two convolution layers and one max pool layer. The first block uses 32 filters on each convolution layer, whereas the second uses 64 filters on each of the two convolution layers. All convolution layers use ‘ReLU’ activations. These blocks are followed by two fully connected layers of 512 neurons respectively. Finally, the output layer of 10 units uses a ‘Softmax’ activation. We separated each of the three sets of subtomograms

3. Comparison of results :

TABLE I
COMPARISON OF VALIDATION ACCURACY BETWEEN THE SQUEEZENET MODEL AND CNN METHOD ON SIMULATED DATASETS WITH DIFFERENT SNRS

Methods \ SNRs	0.03	0.05	$+\infty$
SqueezeNet	72.20%	76.00%	79.20%
CNN	87.10%	88.20%	90.60%

Improvement that can be done :

Using VGG 19 or Resnext , and then the model can be made fully connected . We should use leaky relu instead of Relu . Noise can be reduced by implementing an autoencoder architecture . We can also implement transformer networks .

We can increase the efficiency of the close packing algorithm by modifying the loss function formula :

$$Loss_{P_k} = \sum_{i}^{N} (x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2$$

We can add a term in this :

ƛ (xi * yi * zi) as a punishment term such that the convergence is faster .