# CryoVirusDB: A Labeled Cryo-EM Image Dataset for AI-Driven Virus Particle Picking

Rajan Gyawali[1,2,†], Ashwin Dhakal[1,2,†], Liguo Wang[3], Jianlin Cheng[1,2,*]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
[2] NextGen Precision Health, University of Missouri, Columbia, Columbia, MO 65211, USA
[3] Laboratory for BioMolecular Structure (LBMS), Brookhaven National Laboratory, Upton, NY 11973, USA

*Corresponding author: Jianlin Cheng (chengji@missouri.edu)
[†]Joint first author

## Abstract

With the advancements in instrumentation, image processing algorithms, and computational capabilities, single-particle electron cryo-microscopy (cryo-EM) has achieved nearly atomic resolution in determining the 3D structures of viruses. The virus structures play a crucial role in studying their biological function and advancing the development of antiviral vaccines and treatments. Despite the effectiveness of artificial intelligence (AI) in general image processing, its development for identifying and extracting virus particles from cryo-EM micrographs (images) has been hindered by the lack of manually labelled high-quality datasets. To fill the gap, we introduce CryoVirusDB, a labeled dataset containing the coordinates of expert-picked virus particles in cryo-EM micrographs. CryoVirusDB comprises 9,941 micrographs of 9 different viruses along with the coordinates of 339,398 labeled virus particles. It can be used to train and test AI and machine learning (e.g., deep learning) methods to accurately identify virus particles in cryo-EM micrographs for building atomic 3D structural models for viruses.

## Background & Summary

Cryo-electron microscopy (cryo-EM) is a method of capturing 2D images of biological molecules and assemblies at extremely low (cryogenic) temperatures. Advancements in both instrumentation and computational methodologies have established cryo-EM as an essential tool for interrogating the structures and dynamics of biological macromolecular complexes including large virus particles [1]. Single particle experiment and data analysis in cryo-EM involves flash-freezing biological specimens, collecting micrographs of individual particles in an electron microscope (**Figure 1A**), followed by picking and extracting particle images (**Figure 1B**), applying image processing for correction and alignment (**Figure 1C**), and performing three-dimensional (3D) reconstruction of macromolecular complexes (**Figure 1D**) [1], [2].

In the realm of virology, cryo-EM has been instrumental in studying and determining the 3D structures and morphology of various viruses such as Polio, Ebola, HIV, and Corona viruses [3] [4]. Particularly during the COVID-19 pandemic, cryo-EM played a pivotal role in understanding the intricate structure of the SARS-CoV-2 spike protein [5] [6]. This knowledge has facilitated the development of highly effective vaccines. For instance, scientists have been able to design immunogens that mimic the spike protein's shape, eliciting targeted immune responses [7]–[9]. Moreover, cryo-EM has revolutionized epitope mapping, enabling the identification of specific binding sites [10] and facilitating the exploration of antibody mutations for the rapid discovery and development of precise vaccines and antiviral treatments [11] [12] [13].

To achieve high-resolution 3D reconstructions of virus structures, the initial step of accurately recognizing and extracting virus particles from 2D image projections (micrographs) is crucial. Currently, three virus particle picking approaches are employed: manual virus particle picking, template-based picking, and AI-based picking. Manual picking is laborious and time-consuming, requiring specialized expertise for precise identification, which cannot be used by regular users. Challenges in the manual picking arise from low single-to-noise ratios, low particle contrast, and the unpredictability of individual particle appearances due to orientation variations. Template-based virus particle picking requires experts to pick some initial particles as templates for software tools to search for more particles, which suffers from the presence of ice contamination, radiation damaged particles, carbon areas, and overlapping aggregated particles in micrographs. AI-based particle picking [14] [15] [16] has the best potential to automate the process and overcome the problems of the manual picking and template-based matching, but the development of sophisticated AI-based virus particle picking methods is largely hindered by the lack of high-quality labelled training and test data of virus particles.

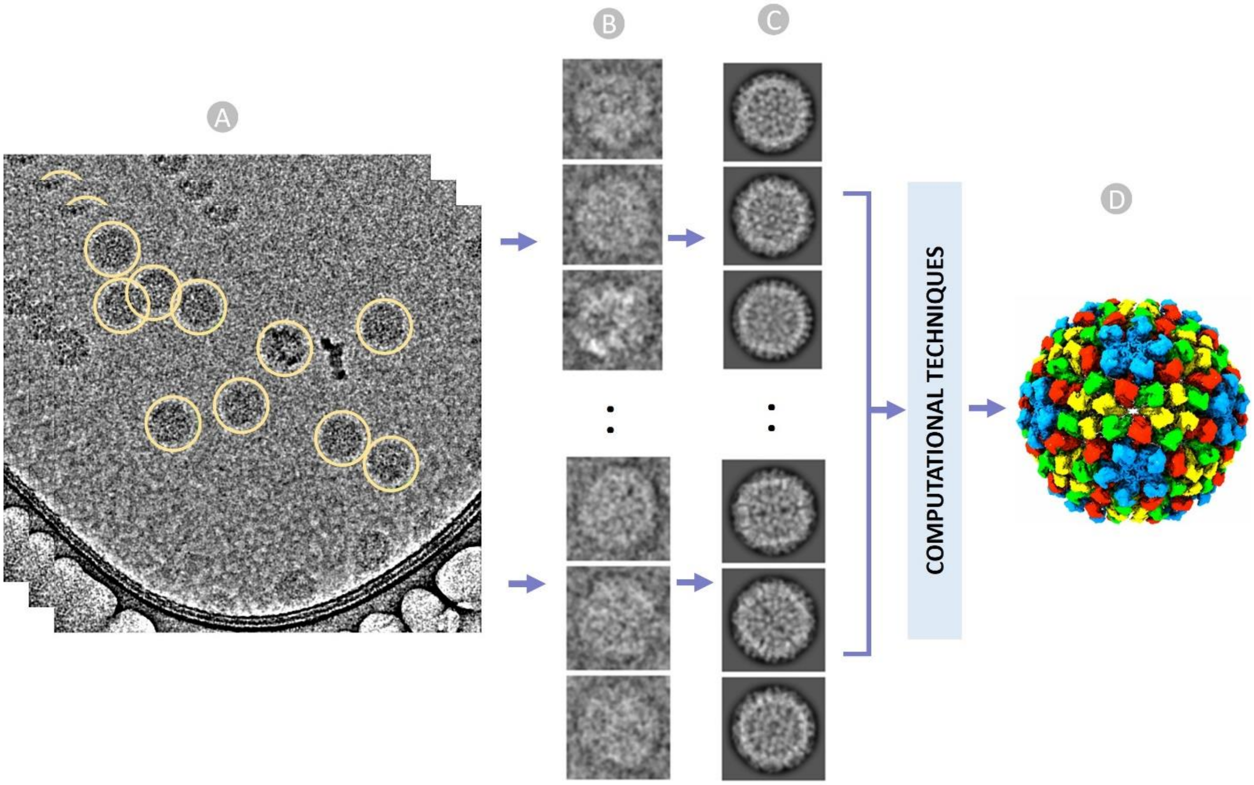

*Figure 1: An overview of cryo-EM single particle analysis from particle selection to 3D reconstruction of virus. (A) Stack of ideal micrographs where the true virus particles are picked (encircled yellow), (B) Extracted virus particles from micrographs with fixed box size. (C) Multiple 2D classes to facilitate stack cleaning and the removal of false particles. (D) Reconstructed 3D structure of the virus from 2D images using a series of computational techniques.*

To harness the power of cutting-edge AI technologies in automatic virus particle recognition and picking, we created a comprehensive and expert-labelled dataset – CroVirusDB - [17] in this work. This open-access dataset aims to expedite the development of automated virus particle picking workflows, and ultimately advance the research of viruses and the design of therapeutic interventions. CryoVirusDB includes 9,941 micrographs of 9 distinct viruses and the coordinates of 339,398 virus particles picked in them. The statistics of CroVirusDB is reported in **Table 1Table 2.**

***Table 1**: The statistics of micrographs and particles of 9 viruses in CryoVirusDB.*

| SN | EMPAIR ID | Virus Type | Number of Micrographs | Micrograph size | Particle Diameter (px) | Number of True Virus Particles |
|---|---|---|---|---|---|---|
| 1 | 10192 [18] | Feline calicivirus | 1000 | (4096, 4096) | 470 | 9660 |
| 2 | 11060 [19] | Nudaurelia capensis omega virus | 1276 | (4096, 4096) | 516 | 11916 |
| 3 | 10203 [20] | Macrobrachium rosenbergii nodavirus | 1000 | (3838, 3710) | 377 | 16601 |
| 4 | 10033 [21] | Human parechovirus 3 | 1000 | (4096, 4096) | 350 | 55732 |
| 5 | 10652 [22] | Coxsackievirus | 1127 | (3838, 3710) | 374 | 11144 |
| 6 | 10341 [23] | Bovine enterovirus | 1274 | (4096, 4096) | 376 | 22694 |
| 7 | 10193 [18] | Feline calicivirus | 1000 | (4096, 4096) | 516 | 96126 |
| 8 | 10205 [24] | Cowpea mosaic virus | 1000 | (4096, 4096) | 310 | 81037 |
| 9 | 10555 [25] | Nudaurelia capensis omega virus | 1264 | (4096, 4096) | 564 | 34488 |
|  |  | **Total** | **9941** |  |  | **339,398** |

# Methods

## 1. Raw Data Acquisition and Preprocessing

The metadata and cryo-EM virus micrographs from the EMPIAR web portal [26] were fetched using Python API and FTP scripts. The comprehensive metadata encompasses the EMPIAR ID for each cryo-EM dataset of a virus along with the corresponding identifiers such as Electron Microscopy Data Bank (EMDB) ID and Protein Data Bank (PDB) ID. Additionally, the dataset size, resolution, total number of micrographs, image specifications (size and type), pixel spacing, micrograph file extension, gain/motion correction file extension (if any), FTP and Globus paths for micrograph/gain files, and relevant publication information are meticulously recorded.

***Table 2:** Metrics of EM data acquisition and grid preparation utilized in importing micrographs for virus particle picking.*

| SN | EMPAIR ID | Micrograph Format | Pixel Spacing (Å) | Accl Voltage (kV) | Spherical Aberration (mm) | Electron Dose (e/A^2) | Defocus range (µm) | Microscope | Detector |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 10192 | mrc | 1.065 | 300 | 2.7 | 63 | NA | FEI TITAN KRIOS | FEI FALCON III (4k x 4k) |
| 2 | 11060 | mrc | 1.065 | 300 | 2.7 | 46 | 0.70 µm - 2.2 µm | FEI TITAN KRIOS | FEI FALCON III (4k x 4k) |
| 3 | 10203 | mrc | 1.06 | 300 | 2.7 | 36 | 1.0 µm - 2.5 µm | FEI TITAN KRIOS | GATAN K2 SUMMIT (4k x 4k) |
| 4 | 10033 | mrc | 1.14 | 300 | 2.7 | 36 | 0.42 µm - 2.34 µm | FEI TITAN KRIOS | FEI FALCON II (4k x 4k) |
| 5 | 10652 | mrc | 1.06 | 300 | 2.7 | 40 | 0.6 µm - 3.0 µm | TFS TALOS F200C | FEI FALCON III (4k x 4k) |
| 6 | 10341 | mrc | 1.065 | 300 | 2.7 | 49.5 | 0.75 µm - 3.5 µm | FEI TITAN KRIOS | FEI FALCON III (4k x 4k) |
| 7 | 10193 | mrc | 1.065 | 300 | 2.7 | 63 | NA | FEI TITAN KRIOS | FEI FALCON III (4k x 4k) |
| 8 | 10205 | mrc | 1.065 | 300 | 2.7 | 67.5 | NA | FEI TITAN KRIOS | GATAN K2 SUMMIT (4k x 4k) |
| 9 | 10555 | mrc | 1.0651 | 300 | 2.7 | 72 | 0.70 µm - 2.7 µm | FEI TITAN KRIOS | FEI FALCON III (4k x 4k) |

To ensure dataset diversity, we selected 9 representative EMPIAR virus datasets that encompassed a broad range of particle sizes, shapes, density distributions, noise levels, and variations in ice thickness and carbon areas to create CryoVirusDB. The datasets include viruses from different categories, such as Omage virus, Cowpea Mosaic virus, Feline calicivirus, and Human parechovirus, providing a comprehensive representation of the virus space.

For each EMPIAR virus dataset, we imported its raw micrographs. A meticulous analysis of the EM data acquisition descriptions and grid preparation details for each dataset was undertaken to gather essential information such as raw pixel size (Å), acceleration voltage (kV), spherical aberration (mm), and total exposure dose (e/Å 2) associated with the micrographs in the respective dataset as shown in **Table 2**.

## 2. Motion Correction and Patch based CTF Estimation of Micrographs

In our study, we used motion corrected micrographs as the starting point in CryoSPARC [27] for patch-based Contrast Transfer Function (CTF) estimation. Since CTF functions can vary substantially among micrographs and cannot be precisely predefined, accurately identifying CTF parameters for each micrograph is crucial.  This precision is necessary for proper corrections and achieving high-resolution 3D reconstructions. Two stages (estimating the CTF and correcting it) are applied to the CTF analysis.

We employ the patch-based Contrast Transfer Function (CTF) to generate output micrographs containing information about their average defocus and the defocus landscape. Upon particle extraction, this information is automatically utilized to allocate a local defocus value to each particle based on its position in the landscape. The one-dimensional search across defocus values for a micrograph is shown in **Figure 2A**.

The distinctive characteristic of the Contrast Transfer Function (CTF) is its oscillating pattern, easily observable as Thon rings in the power spectra of images (**Figure 2B**). Thon rings exhibit more frequent oscillations with larger defocus values and fewer oscillations with smaller defocus values. This connection between defocus and Thon rings forms the foundation for both manual and automated methods of fitting the CTF.

The plot in **Figure 2C** serves primarily as a verification for the successful execution of background subtraction and envelope function fitting. The X-axis represents frequency in inverse angstroms. The radially averaged power spectrum is depicted in black, where high values correspond to the bright portions of the Thon rings and low values to the dark regions. The orange curve represents the envelope function, aiming to model the expected falloff of Thon rings up to the Nyquist resolution [28], accounting for aberrations. Lastly, the fitted Contrast Transfer Function (CTF), scaled by the envelope function, is presented in blue. This oscillating plot is crucial for confirming the proper execution of background subtraction and envelope fitting procedures. In the plot in **Figure 2D**, we assess the background strength (depicted by the black line) within the area where thicker ice leads to an augmented background referred to as relative ice thickness.

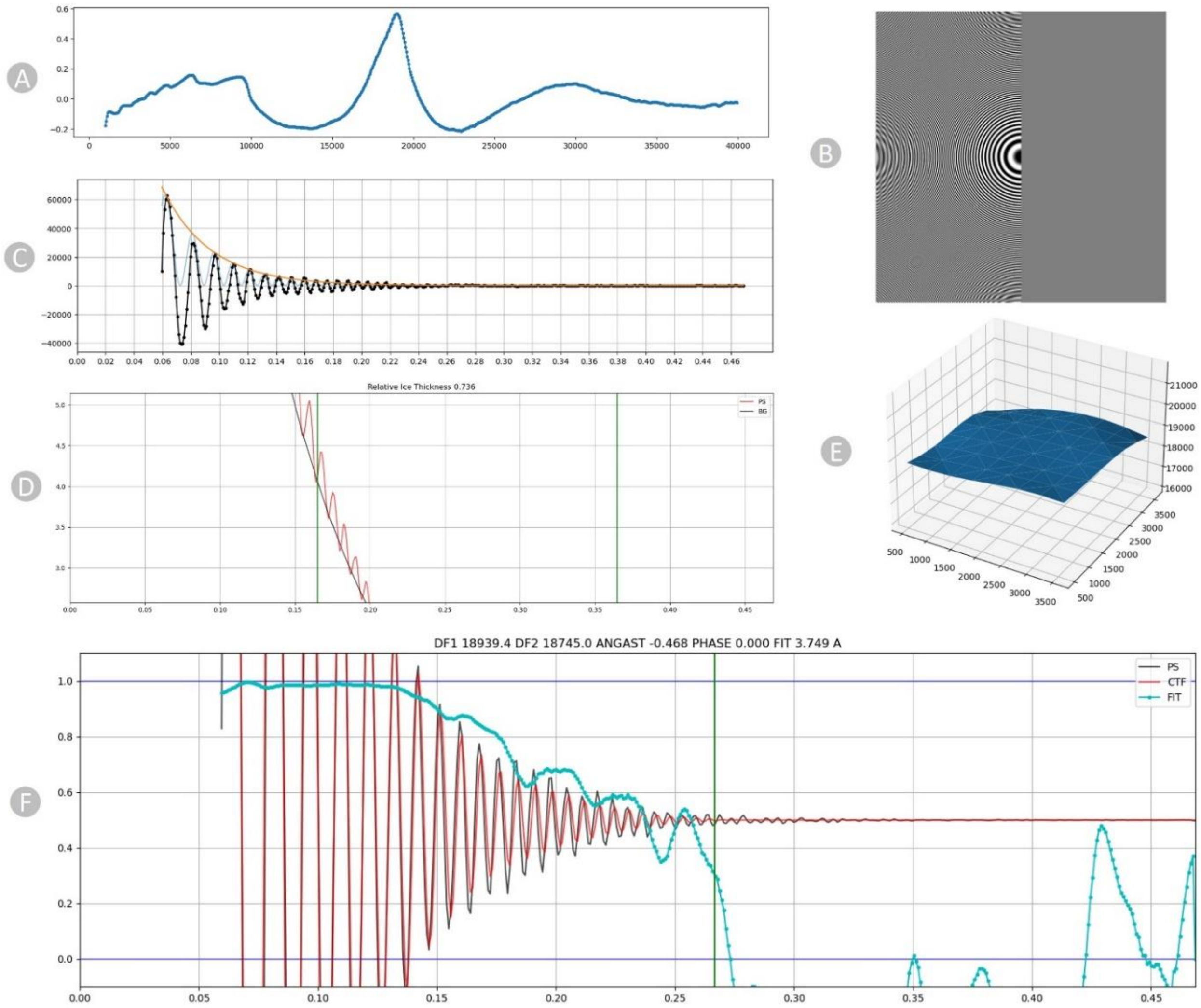

***Figure 2:*** *Diagnostic plots of CTF for EMPAIR 11060. (A) 1D search over varying defocus values. (B) Thon rings visible in the Fourier transform. (C) Fitted envelope function diagnostic plot. (D) Power Spectrum showing the relative ice thickness. (E) 2D Patch result. (F) CTF fit plot. The frequency, measured in inverse angstroms ($Å^{-1}$), is represented on the X-axis, while the correlation metric between the power spectrum (PS) and CTF value is shown on the Y-axis. The black line corresponds to the observed experimental power spectrum, the red line represents the calculated CTF, and the cyan line indicates the cross-correlation (fit).*

The 3D surface plot (**Figure 2E**) shows the local defocus estimated throughout the micrograph. The surface plot in blue illustrates the defocus that has been fitted for each position along the micrograph. The x- and y-coordinates align with the micrograph's coordinates, while the z-coordinate represents the defocus values. The X, Y, and Z axes are all expressed in Angstrom units. The Contrast Transfer Function fit plot, illustrated in **Figure 2F**, depicts the alignment between the simulated and observed Thon rings in the micrograph, accounting for variations in defocus and astigmatism. The cyan curve indicates the cross-correlation fit level. The CTF fit resolution (3.749 angstroms) is the resolution at which this value drops below a threshold. The vertical green line in the plot signifies the frequency at which the fit deviates from cross-correlation threshold of 0.3, indicating a successful fit.

## 3. Manual Particle Picking and 2D Class Formation

Following the CTF estimation, we manually identified and selected true virus particles interactively from aligned and motion-corrected micrographs with the aim of generating some particle templates. We specify the particle diameter based on the virus particles' size and shape. Picking particles directly from raw noisy micrographs is challenging (**Figure 3A)**. So we adjusted the 'Contrast Intensity Override' using low pass filter while inspecting micrographs to achieve the most distinct view for particle selection (**Figure 3B)**. Additionally, we employed a visual guide to encircle virus particles (**Figure 3C)**, ensuring that the chosen particles are well-centered for improved results in subsequent 2D alignment steps.

Manually selecting particles from raw micrographs with smaller defocus values proves to be quite challenging. To create a comprehensive set of ground-truth templates covering a broad range of defocus values, we manually picked particles from numerous micrographs exhibiting diverse defocus and CTF fit values. Given the time-intensive nature of manual picking, we chose a small subset of micrographs (around 15% of the micrographs) specifically for generating templates. The detailed information about the manually picked particles and the micrographs considered for the manual picking can be found in **Supplementary Table S1**.

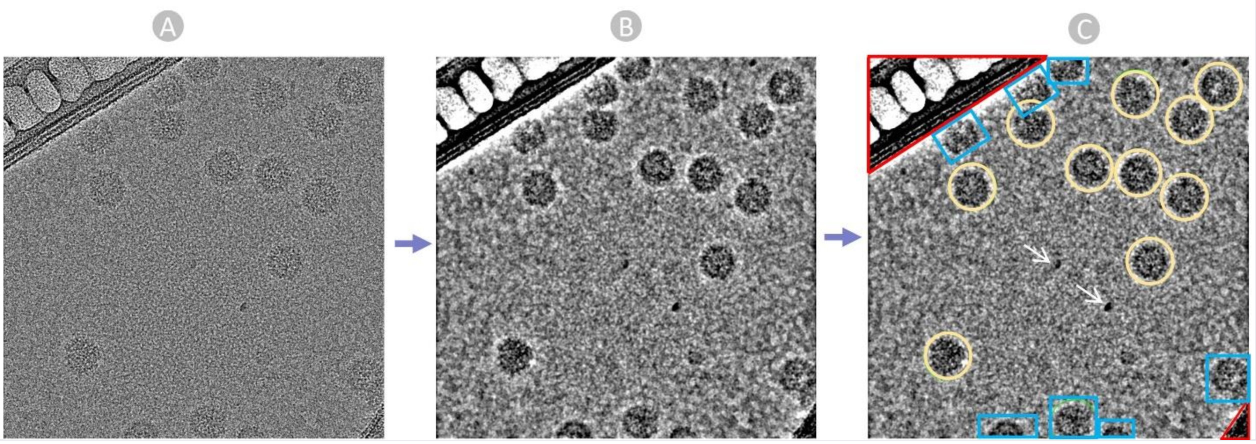

*Figure 3: The manual Picking Process. (A) Raw micrograph obtained from EMPIAR. (B) Preprocessed Micrograph with Low Pass filter: 28 A to ease particle recognition and picking. (C) Manually picked true virus particle encircled in yellow, carbon region colored in Red, ice patches and artifacts pointed by white arrows, and cut particles in edges colored in blue.*

After manually picking the virus particles, the coordinates of particle centers and the designated box size are used to extract particles from the original micrographs. During this process, the box size is defined to provide ample padding, usually ranging from 25% to 50% extra space around the particles. The manually selected particles undergo a 2D classification step, where we categorized and chose the most favorable classes. This classification step organized particles into distinct 2D classes, streamlining the cleaning of the particle stack and removal of undesirable particles. Finally, we assessed the quality of the particles and eliminated classes containing unwanted particles. The remaining particle classes are used by the template-based picking for the identification of high-quality particle classes.

## 4. Template-based Picking

After exporting the optimal particle classes, we employed templates created in the ‘2D Class Formation’ step in CryoSPARC. We followed an iterative approach, wherein the output from 'template-based picking and inspection' is once again utilized in the ‘2D Class Formation' step to select only the high quality 2D particles discarding the false positives. This cycle was repeated until we obtained high-resolution particles that encompass all possible viewing directions of the virus particle.

Using CryoSPARC's Template Picker job, we employed the high-resolution templates to precisely pick virus particles that align with the geometry of the target structure. We set specific constraints, such as the Particle diameter in angstrom and a minimum distance between the particles for generating templates based on the SK97 sampling algorithm [29].

## 5. Manual Particle Inspection and Extraction

The acquired particles above underwent the manual inspection, in which we scrutinized and refined the picked particles using different thresholds. We fine-tuned parameters such as the lowpass filter, normalized cross-correlation (NCC), and power threshold (**Figure 4A**) to eliminate false positives. The 2D colored histogram plots were employed to carefully analyze the median pick scores of micrographs against defocus, aiding in the extraction of coordinates for high-quality virus particles as depicted in **Figure 4B**.

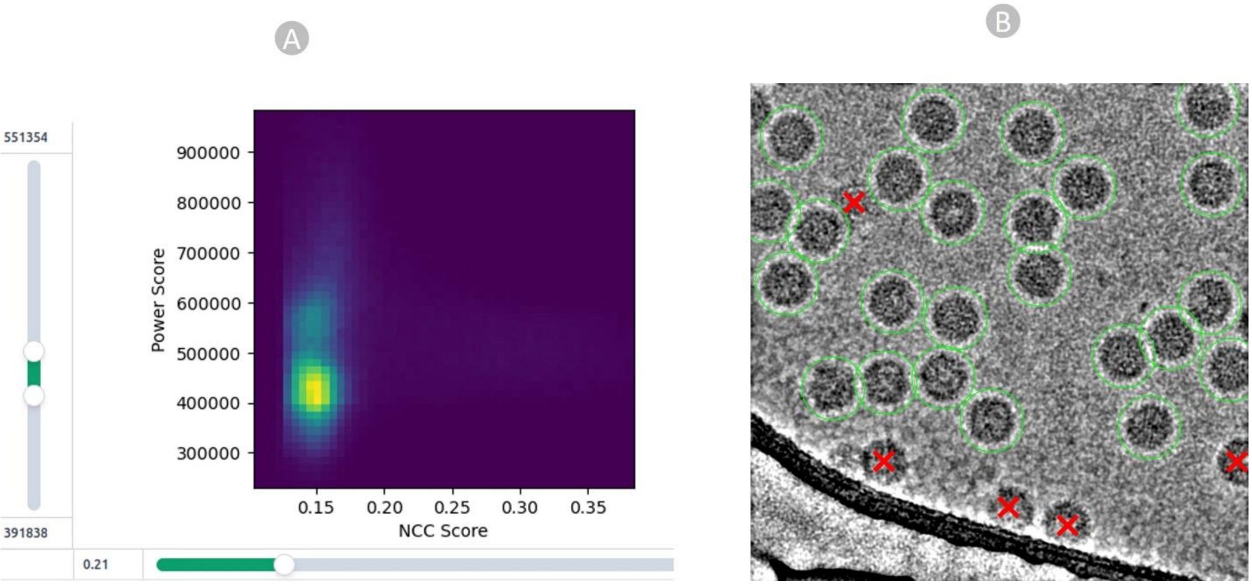

***Figure 4:*** *Particle Quality inspection. (A) particle filtration achieved by manipulating the NCC score on the X-axis and local power score on the Y-axis for EMPIAR 11060. (B) high quality true virus particles (depicted by green circles) chosen through the template-based picking process and eliminated radiation-damaged, cut, and false positive particles, represented by red crossed markers.*

Ultimately, we applied a 2D Classification step to perform a final examination of the selected particles. The Select 2D job categorized particles into various 2D classes (usually 50 in our case), aiding in stack cleaning and the elimination of undesired particles. This process is valuable not only for assessing particle quality before entering the 3D reconstruction phase but also for qualitatively exploring the distribution of views within the dataset. Following 2D Classification, certain classes are identified as "junk" classes, representing non-particle images, ice crystals, or instances of two particles being conjoined. Consequently, we filtered out the particles associated with these "junk" classes from the picked particles. More information about the overall intermediate metadata and the final set of true virus particles can be found in **Supplementary Table S1**.

These final true particles are exported in the form of particle stacks, star files and csv files, which include a lot of information about the particles in micrographs like: X-coordinate, Y-coordinate, Angle-Psi, Origin X (Ang), Origin Y (Ang), Defocus U, Defocus V, Defocus Angle, Phase Shift, CTF B Factor, Optics Group, and Class Number.

## Data Records

CryoVirusDB includes 9 virus subsets (each including approximately 1200 cryo-EM micrographs) along with the labelled coordinates of the virus particles in the micrographs. The total size of the CryoVirusDB database is 634 GB. The organizational structure of the directories of CryoVirusDB is depicted in **Figure 5**.

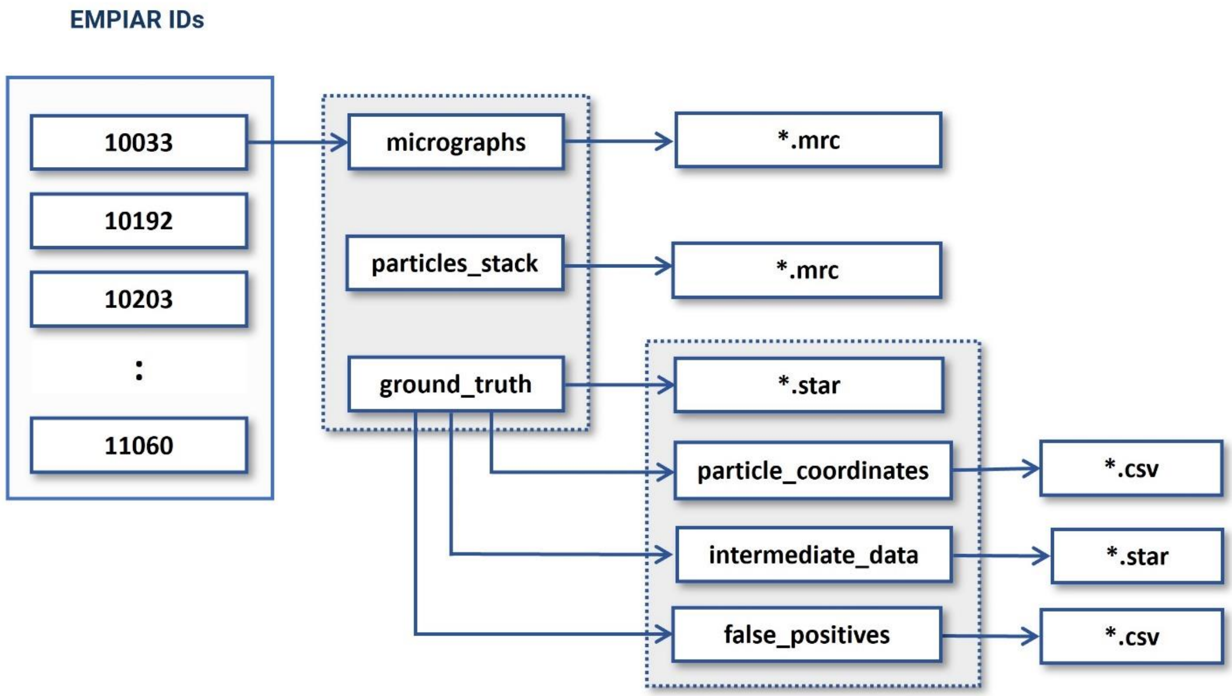

***Figure 5:*** *The directory structure of CryoVirusDB. The numbers in the blocks on the left side are the respective EMPIAR IDs.*

### 1. Motion Corrected Micrographs

These are the two-dimensional images captured by the microscope during the imaging process. All the micrographs in CryoVirusDB are stored in *.mrc* image format. Each sub-dataset (named with EMPIAR ID) in CryoVirusDB contains around 1200 micrographs.

### 2. Virus Particle Stack

The particle stack consists of *.mrc* files, each named after the corresponding micrograph's filename, containing ground truth virus particles. These files form a three-dimensional grid of voxels, where each voxel value corresponds to electron density, essentially forming a stack of 2D images. To view and inspect the particle stacks, one can use EMAN2 [30] or UCSF Chimera [31] / ChimeraX [32] .

### 3. Ground Truth Labels (Coordinates)

The ground truth directory includes particle coordinates (in .csv format), false positives (in .csv format), intermediate data (in .star file), and the collective star file of all ground truth particles. The false positives contain viruses like particles that are actually ice contaminations, aggregates, radiation damaged particles, and false particles over carbon regions.

## Technical Validation

### 1. 2D Particle Class Validation

We compared our picked virus particles with a popular AI-based particle picking method, Topaz [33], considering factors such as the total number of classes, number of picked particles, 2D resolution, and visual orientation. Our manually picked particles have a better 2D class resolution than Topaz. It's noteworthy that a higher particle count alone does not ensure higher resolution. Instead, selecting a substantial number of high-quality particles across a broad angular distribution is crucial for achieving both high 2D and 3D resolution. The 2D class comparison for two databases: EMPIAR 10205 and EMPIAR 10193 (each containing 1000 micrographs) are shown in **Table 3** and **Figure 6**. In both cases, Topaz picked many more particles than CryoVirusDB but it had a worse 2D class resolution.

*Table 3: 2D classification result comparison for EMPIAR 10205 and EMPIAR 10193*

| **EMPIAR 10205** | | |
|---|---|---|
| | **2D Particle Class Statistics (Topaz)** | **2D Particle Class Statistics (CryoVirusDB)** |
| Number of Picked Particles | 155,953 | 81,037 |
| Weighted Average Resolution of 2D classes (N=50) | 9.41 Å | 6.59 Å |
| Weighted Average Resolution of 2D classes (N=10) | 13.42 Å | 10.96 Å |

| **EMPAIR 10193** | | |
|---|---|---|
| | **2D Particle Class Statistics (Topaz)** | **2D Particle Class Statistics (CryoVirusDB)** |
| Number of Picked Particles | 239,852 | 96,126 |
| Weighted Average Resolution of 2D classes (N=50) | 18.52 Å | 15.02 Å |
| Weighted Average Resolution of 2D classes (N=10) | 23.68 Å | 21.72 Å |

We also assessed the density projections derived from the intermediate output during the ab initio reconstruction phase, as depicted at the bottom of each block in **Figure 6**. The plot illustrates the integrated density values along the perpendicular direction to that plane. The heatmap's color scheme represents scalar density values at each voxel, with the intensity of color indicating the magnitude of density. This indicates the high quality of the virus particles in CryoVirusDB.

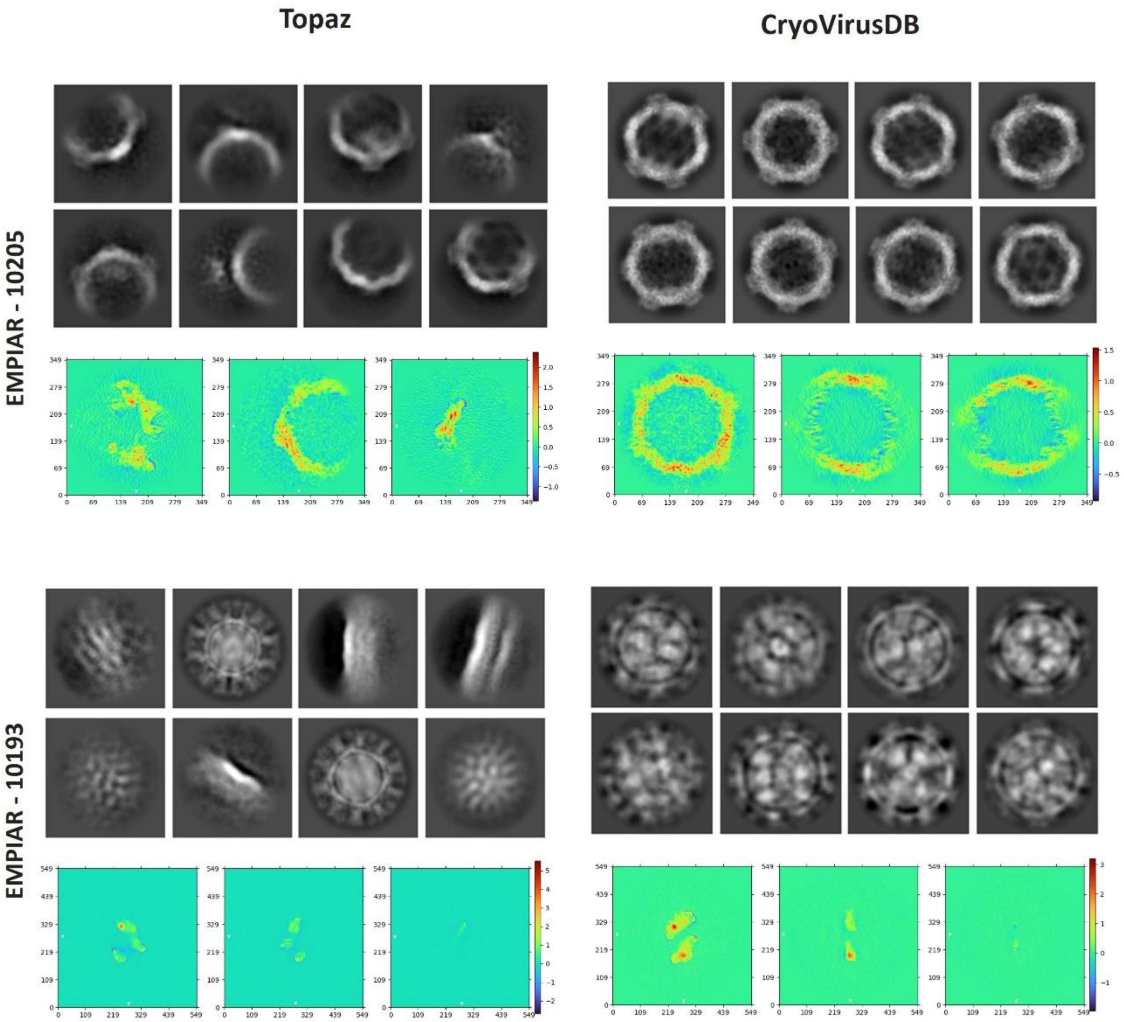

*Figure 6: The comparison of 2D particle classification and density projections from the intermediate output of the ab initio reconstruction phase for EMPIAR 10205 and EMPIAR 10193. For each EMPIAR ID, the 2D classes are visualized at the top and the density projects are visualized at the bottom. The color scheme in the heatmap corresponds to the scalar density values at each voxel.*

## 2. 3D Density Map Validation

We reconstructed 3D density maps from the particles in CryoVirusDB and from those picked by Topaz for two datasets: EMPIAR 10205 and EMPIAR 10193, each comprising 1000 micrographs. The ab-initio density map reconstruction and homogenous refinement were carried out in CryoSPARC using the generated star files that included the selected particles. To ensure an unbiased evaluation, we repeated the ab-initio 3D reconstruction experiment with three distinct random seeds for each method.

**Figure 7** presents a comparison of the resolution and distribution direction of the reconstructed 3D density maps. The Fourier Shell Correlation (FSC) plots include a 'loose mask' curve that utilizes an automatically generated mask with a 15 Å falloff, and a 'tight mask' curve that employs an auto-generated mask with a falloff of 6 Å for all FSC plots.

In the case of EMPIAR 10205, Topaz picked around 75,000 more particles compared to our manual picking. Despite this, the density map reconstructed from particles selected by CryoVirusDB achieved a resolution of 4.34 Å, substantially better than Topaz's resolution of 6.48 Å. This indicates the high quality of the picked particles in CryoVirusDB. For EMPIAR 10193, the resolution of CryoVirusDB is 5.16 Å, also better than 5.74 Å of Topaz.

The heightened intensity of the red color in the direction distribution shown in the lower section of each block in **Figure 7** corresponds to an increased number of particles in the elevation vs azimuth plots. CryoVirusDB demonstrated superior particle picking by capturing a substantial number of particles with a wide angular distribution, evident in the red coloration on the heatmap for both validation cases.

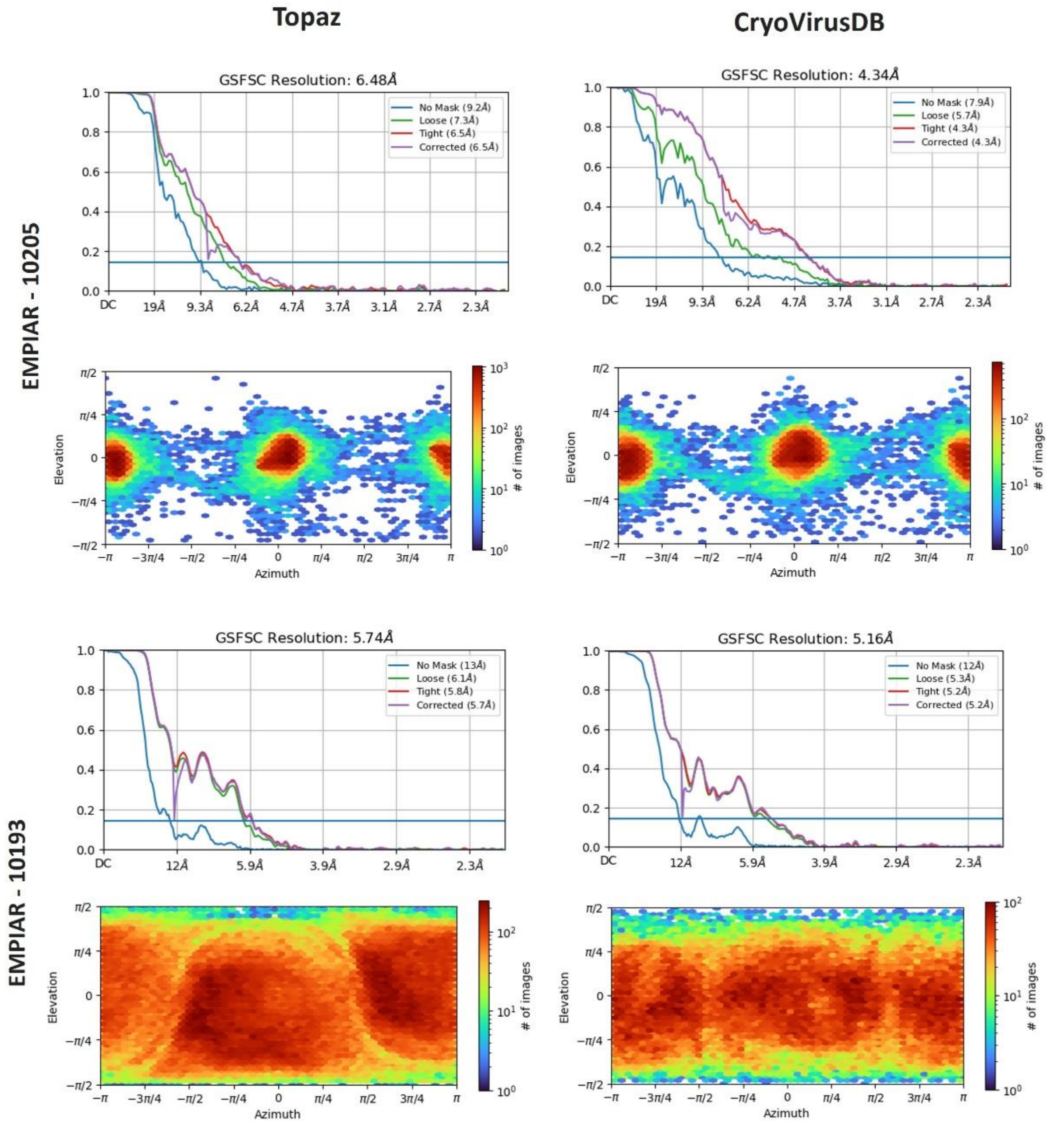

*Figure 7: The comparison of 3D density resolution and direction distribution obtained by Topaz and CryoVirusDB on EMPIAR 10205 and EMPIAR 10193.*

*Table 4: 3D density map result comparison for EMPIAR 10205 and EMPIAR 10193. Bold fonts highlight the resolution of the best of the three trials for each method.*

| **EMPIAR 10205** | | | | | | |
|---|---|---|---|---|---|---|
| | **3D Density Map Statistics (Topaz)** | | | **3D Density Map Statistics (CryoVirusDB)** | | |
| Number of Picked Particles | 155,953 | | | 81,037 | | |
| GSFSC Resolution (Å) | Trial 1 | Trial 2 | Trial 3 | Trial 1 | Trial 2 | Trial 3 |
| | 6.97 | 6.59 | **6.48** | **4.34** | 4.40 | 4.47 |
| No Mask Resolution (Å) | 9.3 | 9.5 | **9.2** | **7.9** | 8.8 | 8.1 |
| Loose Mask Resolution (Å) | 7.7 | 7.2 | **7.3** | **5.7** | 7.1 | 6.3 |
| Tight Mask Resolution (Å) | 6.8 | 6.6 | **6.5** | **4.3** | 4.5 | 4.4 |
| Corrected Mask Resolution (Å) | 7 | 6.6 | **6.5** | **4.3** | 4.4 | 4.5 |

| **EMPAIR 10193** | | | | | | |
|---|---|---|---|---|---|---|
| | **3D Density Map Statistics (Topaz)** | | | **3D Density Map Statistics (CryoVirusDB)** | | |
| Number of Picked Particles | 239,852 | | | 96,126 | | |
| GSFSC Resolution (Å) | Trial 1 | Trial 2 | Trial 3 | Trial 1 | Trial 2 | Trial 3 |
| | 5.86 | **5.74** | 5.82 | **5.16** | 5.22 | 5.18 |
| No Mask Resolution (Å) | 12 | **13** | 11 | **12** | 9.4 | 9.1 |
| Loose Mask Resolution (Å) | 5.9 | **6.1** | 5.8 | **5.3** | 5.8 | 5.5 |
| Tight Mask Resolution (Å) | 5.8 | **5.8** | 5.7 | **5.2** | 5.2 | 5.2 |
| Corrected Mask Resolution (Å) | 5.9 | **5.7** | 5.6 | **5.2** | 5.2 | 5.3 |

The detailed comparison of the 3D density map reconstruction of three trials for CryoVirusDB and Topaz is provided in **Table 4**. The density maps constructed from the labeled particles in CryoVirusDB consistently exhibit a higher quality than Topaz in terms of multiple resolution metrics, even though the number of particles in CryoVirusDB is much smaller than the number of particles picked by Topaz, indicating that Topaz may pick quite some false positives and/or miss some true positives representing different views.

## Code Availability

The GitHub repository: https://github.com/BioinfoMachineLearning/CryoVirusDB contains all the scripts used in every stage of data curation. It also provides instructions on how to download and use the data.

## Author Contributions

J.C. conceived the research. R.G., A.D., and J.C. designed the methodology and experiment. R.G. and A.D. wrote the scripts and codes for dataset preprocessing. R.G. and A.D. curated the data. J.C. and L.W. conceptualized data validation. R.G. and A.D. drafted the manuscript.  J.C and L.W. revised manuscript. All authors participated in result discussions, data analysis, and made contributions to the final manuscript.

## Competing Interests

The authors declare no competing interests.

## Funding

This work was supported by National Institute of Health grant (grant #: R01GM146340) to JC and LW.

## References

[1] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng, "A large expert-curated cryo-EM image dataset for machine learning protein particle picking," *Sci. Data*, vol. 10, no. 1, pp. 1–22, 2023, doi: 10.1038/s41597-023-02280-2.

[2] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng, "CryoPPP : A Large Expert-Labelled Cryo-EM Image Dataset for Machine Learning Protein Particle Picking Background & Summary I . Cryo-EM Grid Preparation and Image Acquisition II . Cryo-EM Micrographs and Single Particle Analysis," 2023.

[3] C. F. Hryc, D. H. Chen, and W. Chiu, "Near-atomic resolution cryo-EM for molecular virology," *Curr. Opin. Virol.*, vol. 1, no. 2, pp. 110–117, 2011, doi: 10.1016/j.coviro.2011.05.019.

[4] W. Jiang and L. Tang, "Atomic cryo-EM structures of viruses," *Curr. Opin. Struct. Biol.*, vol. 46, pp. 122–129, 2017, doi: 10.1016/j.sbi.2017.07.002.

[5] J. Zhang, T. Xiao, Y. Cai, and B. Chen, "Structure of SARS-CoV-2 spike protein," *Curr. Opin. Virol.*, vol. 50, pp. 173–182, 2021, doi: 10.1016/j.coviro.2021.08.010.

[6] X. Xia, "Domains and functions of spike protein in sars-cov-2 in the context of vaccine design," *Viruses*, vol. 13, no. 1, pp. 1–16, 2021, doi: 10.3390/v13010109.

[7] B. M. Hauser *et al.*, "Rationally designed immunogens enable immune focusing following SARS-CoV-2 spike imprinting," *Cell Rep.*, vol. 38, no. 12, p. 110561, 2022, doi: 10.1016/j.celrep.2022.110561.

[8] E. Ong, X. Huang, R. Pearce, Y. Zhang, and Y. He, "Computational design of SARS-CoV-2 spike glycoproteins to increase immunogenicity by T cell epitope engineering," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 518–529, 2021, doi: 10.1016/j.csbj.2020.12.039.

[9] K. M. Castro, A. Scheck, S. Xiao, and B. E. Correia, "Computational design of vaccine immunogens," *Curr. Opin. Biotechnol.*, vol. 78, p. 102821, 2022, doi: 10.1016/j.copbio.2022.102821.

[10] A. Dhakal, C. McKay, J. J. Tanner, and J. Cheng, "Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions," *Briefings in Bioinformatics*, vol. 23, no. 1. 2022, doi: 10.1093/bib/bbab476.

[11] L. A. Earl and S. Subramaniam, "Cryo-EM of viruses and vaccine design," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 32, pp. 8903–8905, 2016, doi: 10.1073/pnas.1609721113.

[12] M. Arista-Romero, S. Pujals, and L. Albertazzi, "Towards a Quantitative Single Particle Characterization by Super Resolution Microscopy: From Virus Structures to Antivirals Design," *Front. Bioeng. Biotechnol.*, vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fbioe.2021.647874.

[13] A. Dhakal, R. Gyawali, and J. Cheng, "Predicting Protein-Ligand Binding Structure Using E(n) Equivariant Graph Neural Networks," *bioRxiv*, p. 2023.08.06.552202, 2023, [Online]. Available: http://biorxiv.org/content/early/2023/08/07/2023.08.06.552202.abstract.

[14] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng, "CryoTransformer: A Transformer Model for Picking Protein Particles from Cryo-EM Micrographs," 2023, doi: 10.1101/2023.10.19.563155.

[15] R. Gyawali, A. Dhakal, L. Wang, and J. Cheng, "Accurate cryo-EM protein particle picking by integrating the foundational AI image segmentation model and specialized U-Net," 2023, doi: 10.1101/2023.10.02.560572.

[16] F. He *et al.*, "Adapting Segment Anything Model (SAM) through Prompt-based Learning for Enhanced Protein Identification in Cryo-EM Micrographs," *ArXiv*, 2023, doi: 10.48550/arXiv.2311.16140.

[17] R. Gyawali, A. Dhakal, L. Wang, and J. Cheng, “CryoVirusDB,” *Zenodo*, 2023. https://zenodo.org/records/10397742.

[18] M. J. Conley *et al.*, “Calicivirus VP2 forms a portal-like assembly following receptor engagement,” *Nature*, vol. 565, no. 7739, pp. 377–381, 2019, doi: 10.1038/s41586-018-0852-1.

[19] L. G. Castells-Graells R , Hesketh EL , Johnson JE , Ranson NA , Lawson DM, “Decoding virus maturation with cryo-EM structures of intermediates,” *EMPIAR*, 2022. https://www.ebi.ac.uk/empiar/EMPIAR-11060/.

[20] K. L. Ho *et al.*, “Structure of the Macrobrachium rosenbergii nodavirus: A new genus within the Nodaviridae?,” *PLoS Biol.*, vol. 16, no. 10, pp. 1–20, 2018, doi: 10.1371/journal.pbio.3000038.

[21] S. Shakeel *et al.*, “Multiple capsid-stabilizing interactions revealed in a high-resolution structure of an emerging picornavirus causing neonatal sepsis,” *Nat. Commun.*, vol. 7, pp. 1–8, 2016, doi: 10.1038/ncomms11387.

[22] J. W. Flatt, A. Domanska, A. L. Seppälä, and S. J. Butcher, “Identification of a conserved virion-stabilizing network inside the interprotomer pocket of enteroviruses,” *Commun. Biol.*, vol. 4, no. 1, pp. 1–8, 2021, doi: 10.1038/s42003-021-01779-x.

[23] R. Chandler-Bostock *et al.*, “Assembly of infectious enteroviruses depends on multiple, conserved genomic RNA-coat protein contacts,” *PLoS Pathog.*, vol. 16, no. 12, pp. 1–23, 2020, doi: 10.1371/journal.ppat.1009146.

[24] R. F. Thompson, M. G. Iadanza, E. L. Hesketh, S. Rawson, and N. A. Ranson, “Collection, pre-processing and on-the-fly analysis of data for high-resolution, single-particle cryo-electron microscopy,” *Nat. Protoc.*, vol. 14, no. 1, pp. 100–118, 2019, doi: 10.1038/s41596-018-0084-8.

[25] R. Castells-Graells *et al.*, “Plant-expressed virus-like particles reveal the intricate maturation process of a eukaryotic virus,” *Commun. Biol.*, vol. 4, no. 1, pp. 1–12, 2021, doi: 10.1038/s42003-021-02134-w.

[26] A. Iudin *et al.*, “EMPIAR: the Electron Microscopy Public Image Archive,” *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1503–D1511, 2023, doi: 10.1093/nar/gkac1062.

[27] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, “CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination,” *Nat. Methods*, vol. 14, no. 3, pp. 290–296, 2017, doi: 10.1038/nmeth.4169.

[28] T. Nyquist and H. Nyquist, “Nyquist – Shannon sampling theorem,” no. May, pp. 1–7, 2019.

[29] P. R. Baldwin and P. A. Penczek, “The Transform Class in SPARX and EMAN2,” *J. Struct. Biol.*, vol. 157, no. 1, pp. 250–261, 2007, doi: 10.1016/j.jsb.2006.06.002.

[30] G. Tang *et al.*, “EMAN2: An extensible image processing suite for electron microscopy,” *J. Struct. Biol.*, vol. 157, no. 1, pp. 38–46, 2007, doi: 10.1016/j.jsb.2006.05.009.

[31] E. F. Pettersen *et al.*, “UCSF Chimera - A visualization system for exploratory research and analysis,” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004, doi: 10.1002/jcc.20084.

[32] E. F. Pettersen *et al.*, “UCSF ChimeraX: Structure visualization for researchers, educators, and developers,” *Protein Sci.*, vol. 30, no. 1, pp. 70–82, 2021, doi: 10.1002/pro.3943.

[33] T. Bepler *et al.*, “Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs,” *Nat. Methods*, vol. 16, no. 11, pp. 1153–1160, 2019, doi: 10.1038/s41592-019-0575-8.