doi:10.1017/S1431927621006164

# CryoDiscovery™: A cryo-EM AI/ML Heterogeneity Analysis for Structural Biology

Narasimha Kumar

Health Technology Innovations Inc, Portland, Oregon, United States

Structural Biology is an emerging area for disease research and drug discovery. This can be the basis for detecting novel biological threats and hence will help prepare the country's readiness. It should be noted that much of the SARS-CoV-2 virus that causes Covid-19 has been analyzed using structural biology [1].

The existing technologies such as X-ray crystallography and NMR have severe limitations when studying large and non-crystallizable biomolecules. In addition, understanding of biological structures in their native states, at the atomic level, is critical for disease research and drug discovery.

The field will be well-served if methods and instrumentation were developed that can enable atomic-scale reconstruction of a biomolecule or complex in its native environment. This will reduce degrees of abstraction from the relevant biological context and accelerate the design-build-test cycle for protein analysis, disease discovery, and therapeutic molecule development.

Most biological macromolecules experience functionally important internal motions, but these movements limit macromolecular structure determination, which relies on averaging properties across thousands or millions of individual molecules. However, unlike X-ray crystallography or NMR, cryo-EM involves imaging of individual molecules. The unique ability of cryogenic Electron Microscopy (cryo-EM) to disentangle different conformational states is a key to its growing importance in structural biology. However, the heterogeneity problem of mapping the structural variability of macromolecules is widely recognized as a computational challenge in cryo-EM, and still represents a major roadblock in many projects.

First, a brief background on cryo-EM and CryoDiscovery:

An EM image is the 2D projection of a 3D object (i.e., particle, protein, virus, etc.). Therefore, to obtain a 3D reconstruction, hundreds of thousands to millions of particles (in their 2D projections) are recorded in multiple images. These particles are present in different unknown orientations. The processing involves multiple steps as shown in the workflow steps figure [2].

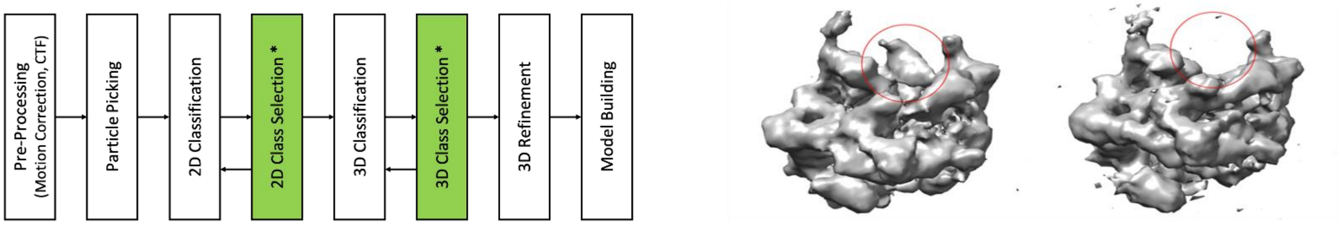
The currently available applications for cryo-EM image/data analysis are research-oriented, and require extensive, manual user involvement. Researchers will benefit greatly from a software solution that minimizes or eliminates manual steps in end-to-end processing and creates a more accurate dependable set of 3D models. To address this, Health Technology Innovations Inc. (HTI)[3] developed the early prototype, CryoDiscovery™, with NSF Phase I funding[4] for automatically sorting and evaluating images of individual molecules using simple Artificial Intelligence/Machine Learning (AI/ML) binary image classification techniques. HTI has presented the idea, AI/ML technology and current results of the prototype at 5 cryo-EM conferences in 2020, where it was accepted well.

Heterogeneity analysis:

Structural heterogeneity is the result of the dynamic nature of biological systems. It can be because the particles exist in different conformations. Examples: An enzyme can be present in its active form (active conformation), as well as in its inactive form (inactive conformation). A macromolecular complex can be present in its full form and in a partial form (due to equilibrium, damage, degradation, etc.). Particles can be present in a factor-bound and unbound form. As an example, the diagram shows two different structures (highlighted with a red circle) that will be a result of heterogeneity analysis [5].

Detection and construction of heterogeneous structures is the goal of heterogeneity analysis. The 2D and 3D Classification step in the previous section can be used to derive several class averages, each of which can be distinctly used to build 3D models with variations showing the differences in structures that contribute to heterogeneity. Currently, there are several mathematical methods [6] (such as Multivariate Statistical Analysis, Maximum Likelihood Estimation, 3D-Covariance, K-Means Clustering etc.) currently attempted with partial success. The introduction of AI/ML is new and has not been fully implemented in most of the applications. AI/ML would provide the probabilities of the various conformations of images. That would help sort and group images for better model construction. CryoDiscovery plans to extend its current AI/ML models for detecting the conformations and building the 3D model.

During the conference, this poster discussion would present the methods employed and how they fared in identifying the significant conformations, and how it compares with the existing methods.



CryoDiscovery components in green and marked with *

Structural Conformations

**Figure 1.** Figures 1 & 2

## References

[1] https://science.sciencemag.org/content/367/6483/1260
[2] https://www.embl-hamburg.de/biosaxs/courses/embo2017/slides/EMBO_cryoEM_2_Bhushan.pdf
[3] https://hti.ai/
[4] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1939142&HistoricalAwards=false
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4536832/pdf/nihms705308.pdf
[6] https://www.hindawi.com/journals/bmri/2017/1032432/

https://doi.org/10.1017/S1431927621006164 Published online by Cambridge University Press