Deep learning methods for 3D prediction, identification and classification of macromolecular surfaces in cryogenic electron tomography (Cryo-ET)

Cellular environment contains numerous macromolecules in interaction and their understanding is critical to understand their function. Following the recent advances in the determination of atomic resolution molecular structures and assemblies [5], in protein structure prediction [15, 12], and in the increased accessibility of molecular dynamics simulation [1, 4], there is a profusion of molecular structural biology data that is available to the scientific community. This data profusion makes machine learning methods, and in particular, deep learning methods totally adequate to support molecular structure and assemblies determination in Cryo-EM and Cryo-ET. Cryo-EM and cryo-ET are able to resolve atomic resolution structures of supramolecular systems and could now address the drawbacks of X-Ray crystallography and Nuclear Magnetic Resonance. In addition, Cryogenic-electron Tomography (Cryo-ET) allows to analyze macromolecules in their cellular environment.

Although cryo-EM has progressed to average structure resolution of 3Å, usual low-resolution density map data affects the precise determination of small molecular objects (less than 150/300 kDa) due notably to the conformational heterogeneity of macromolecules [2], the noise in the images and the missing wedge of information due to the low range of angles [13]. As a consequence, more than half of the cryo-EM samples available in the EMDataResource have no atomic structure determined yet [6].

The reconstruction and identification of macromolecules in Cryo-electron tomograms is a challenge. The performance of existing methods has notably been evaluated during the 3D Shape Retrieval Challenge community benchmark [3] (SHREC). Two groups of methods have been developed : (i) the structure refinement based-approaches requiring predefined model (Rosetta-Ref, Flex-EM, iMODFIT, MDFF, Situs), and (ii) the *de-novo* modelling that may be based on deep learning (DL) (Rosetta-dn, CR-I-TASSER, DeepTracer, DeepMainmast). While the DL methods improve the performance in 3D reconstruction, the results are significantly improvable on low density maps [8, 16, 11]. Performance can be improved by two tricks : (i) using AlphaFold (AF) to reconstruct accurately missing local regions, and (ii) annotating proteins [11]. But, the performance on multiple chain complexes (entire EM map) can be improved as in [11].

The doctoral project proposal is based on : (i) the evaluation and optimisation of DL-based 3D reconstruction methods on our benchmarking dataset in order to explore and explain the strong and weak points, and (ii) the identification and annotation of these macromolecules based on predictive models (AlphaFold, Molecular Dynamics) and on macromolecule retrieval.

The workflow can be as follows: (i) image denoising and extraction of molecular objects [13], (ii) application of various 3D reconstruction methods to convert 2D images into 3D structures, (iii) conformational sampling of these 3D models with structure prediction and Molecular Dynamics, and (iv) convert them back into shapes for retrieval on the cryo-EM map. The identification and annotation will allow us to understand their functional role within the cellular environment and will increase the number of identified proteins in Cryo-electron tomograms, which will be useful for further development and evaluation of 3D DL-based classification methods.

This research project is multidisciplinary, involving computer vision, machine learning and structural and molecular biology. It will require : (i) the retrieval and construction of a challenging reference benchmarking CryoEM dataset based on public data and collaborators tomographic data. (ii) an exhaustive evaluation of geometric deep learning methods in identifying macromolecules in CryoEM densities, with support of the SHREC community benchmark for instance, (iii) deciphering the limitations and specificities of DL methods on our benchmarking dataset and its extension for tomographic data, (iv)) analyzing collaborators and public data to identify macromolecules in cellular tomograms, and (v) set up and distribute a complete open-source pipeline.

Expected candidate The ideal candidate should possess a strong academic background in Computer Science, structural bioinformatics, or a related field. This includes a Master's degree or equivalent in a relevant discipline.

The ideal candidate should display: 1. Excellent programming skills in python and/or c++; 2. Solid foundation in geometric deep learning techniques and algorithms; 3. Familiarity with popular deep learning frameworks like TensorFlow or PyTorch; 4. Experience with structural bioinformatics tools and databases and 5. Knowledge of protein structure prediction, molecular dynamics simulations, or protein-protein interaction analysis.

Supervision This doctoral research project will be directed by Prof. Matthieu Montes, CQSB, UMR7238 CNRS - Sorbonne Université and Dr. Nathalie Lagarde, laboratoire GBCM, EA7528 CNAM. The team organized several SHREC benchmarks on macromolecular shapes retrieval and co-authored 14 publications in computer science and structural bioinformatics linked to this project, in particular: [14, 10, 9, 7].

References

[1] Rommie Amaro, Johan Åqvist, et al. *The need to implement FAIR principles in biomolecular simulations*. arXiv:2407.16584 [q-bio]. Aug. 2024. URL: http://arxiv.org/abs/2407.16584 (visited on 10/14/2024).

[2] Xu Benjin and Liu Ling. "Developments, applications, and prospects of cryo-electron microscopy". In: *Protein Science* 29.4 (2020), pp. 872–882.

[3] Ilja Gubins et al. "SHREC 2021: Classification in cryo-electron tomograms". In: *Computers Graphics* 91 (2020), pp. 279–289. ISSN: 0097-8493.

[4] Adam Hospital and Modesto Orozco. "MD-DATA: the legacy of the ABC Consortium". en. In: *Biophysical Reviews* 16.3 (June 2024), pp. 269–271. ISSN: 1867-2450, 1867-2469. DOI: 10.1007/s12551-024-01197-3. URL: https://link.springer.com/10.1007/s12551-024-01197-3 (visited on 10/14/2024).

[5] John Jumper, Richard Evans, et al. "Atomic-resolution protein structure determination by cryo-EM". In: *Nature* 596 (2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[6] Catherine L. Lawson et al. "EMDataBank.org: unified data resource for CryoEM". In: *Nucleic Acids Research* 39.suppl1 (Oct. 2010), pp. D456–D464. ISSN: 0305-1048.

[7] Mohamed Machat et al. "Comparative evaluation of shape retrieval methods on macromolecular surfaces: an application of computer vision methods in structural bioinformatics". In: *Bioinformatics* 37.23 (July 2021), pp. 4375–4382.

[8] Jonas Pfab, Nhut Minh Phan, and Dong Si. "DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes". In: *Proceedings of the National Academy of Sciences* 118.2 (2021), e2017525118.

[9] Cyprien Plateau—Holleville et al. "Efficient GPU computation of large protein Solvent-Excluded Surface". In: *IEEE Transactions on Visualization and Computer Graphics* (2024), pp. 1–12.

[10] Léa Sirugue et al. "PLO3S: Protein LOcal Surficial Similarity Screening". In: *Computational and Structural Biotechnology Journal* 26 (2024), pp. 1–10. ISSN: 2001-0370.

[11] Genki Terashi et al. "DeepMainmast: integrated protocol of protein structure modeling for cryo-EM with deep learning and structure prediction". In: *Nature Methods* 21 (2024).

[12] Mihaly Varadi, Stephen Anyango, et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". In: *Nucleic Acids Research* 50.D1 (Nov. 2021), pp. D439–D444. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1061. eprint: https://academic.oup.com/nar/article-pdf/50/D1/D439/43502749/gkab1061.pdf. URL: https://doi.org/10.1093/nar/gkab1061.

[13] Simon Wiedemann and Reinhard Heckel. "A deep learning method for simultaneous denoising and missing wedge reconstruction in cryogenic electron tomography". In: *Nature Communications* 15 (2024).

[14] Taher Yacoub et al. "SHREC2024: Non-rigid Complementary Shapes Retrieval in Protein-protein Interactions". In: *Eurographics Workshop on 3D Object Retrieval*. Ed. by Silvia Biasotti et al. The Eurographics Association, 2024. ISBN: 978-3-03868-242-4.

[15] K.M. Yip, N. Fischer, et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 587 (2020), pp. 157–161. DOI: 10.1038/s41586-020-2833-4.

[16] Xi Zhang et al. "CR-I-TASSER: assemble protein structures from cryo-EM density maps using deep convolutional neural networks". In: *Nature Methods* 19 (2022).