Unsupervised deep learning approaches to extract atomic-scale conformational landscapes of biomolecules from cryo electron microscopy images

Supervisor: Slavica JONIC, IMPMC, UMR 7590 CNRS - Sorbonne Université, Paris (ED 130)
Co- supervisor: Catherine Vénien-Bryan, IMPMC, UMR 7590 CNRS - Sorbonne Université, Paris (ED 515)

1. Position of the project in relation to the state of the art

Studies of the structure and conformational transitions of biomolecular complexes are essential for understanding how the complexes function. Thanks to the so-called single-particle cryo electron microscopy (cryo-EM), multiple conformational states of the complexes can be obtained simultaneously from the same sample. Moreover, cryo-EM has a potential to collect the information about the full conformational variability of complexes (the so-called conformational landscapes), allowing to decipher continuous conformational transitions of the complexes (gradual transitions with many intermediate conformational states, i.e., a continuum of states instead of a few discrete states).

Single-particle cryo-EM is based on collecting parallel-beam projection images of different zones of a vitrified sample containing many copies of the same complex (referred to as particles) in random and unknown orientations, positions, and conformations, without tilting the sample in the microscope. The sample is exposed to a low electron dose to minimize radiation damage, which yields highly noisy images. Individual particle images are extracted from the collected images and advanced image processing algorithms and software are used to solve the heterogeneity of the particle orientations (three Euler angles), positions (shifts in x and y directions in the image plane), and conformations in the different particle images, in order to obtain 3D reconstructions of different conformational states. A very low signal-to-noise ratio (SNR) of cryo-EM images makes this task extremely difficult.

Currently, the majority of research labs, all over the world, still use cryo-EM image analysis methods that discretize the continuous conformational heterogeneity by classifying images into a given, small number of discrete classes, usually based on maximum likelihood classification (J Struct Biol 2012; J Struct Biol 2013), which may prevent from discovering the conformations potentially important because the output of these methods are class-average conformations instead of individual conformations. The development of cryo-EM image analysis methods that aim at deciphering individual conformations and full conformational landscapes of complexes by considering continuous conformational heterogeneity of the complexes started around 10 years ago and is currently a very dynamic field of research (Structure 2014; PNAS 2014; J Struct Biol 2015; Nat Methods 2021a; Nat Methods 2021b; Front Mol Biosci 2022; J Mol Biol 2023). Such continuous conformational analysis methods are usually tested and validated using cryo-EM images of large complexes, such as ribosomal complexes. For instance, in a recent work supervised by Dr. Jonic (PhD thesis of Rémi Vuillemot funded by the CNRS, thesis defense planned for September 2023), a method (MDSPACE) was developed for continuous conformational analysis that efficiently combines molecular dynamics simulation and single-particle cryo-EM image analysis (J Mol Biol 2023), and the method was tested using experimental cryo-EM data of large, yeast 80S ribosome (**Fig. 1**).

At present, there is no systematic study of the performance of methods using cryo-EM images containing conformational variability of small, membrane protein complexes. Once expressed in suitable quantities, membrane proteins need to be extracted from the membrane and purified in their native state using surfactants (e.g., detergent). They can also be reconstituted in model lipid environment such as nanodiscs (Q Rev Biophys 2021). Cryo-EM images of membrane proteins are difficult to analyze not only because of the low SNR, but also because of the occlusions due to the environment around the protein (surfactant or nanodisc) that hides the protein in the image (the environment projects onto the image together with the signal of interest, i.e., the protein). A high density of the environment and a low SNR of images make it difficult to distinguish the signal from the environment in the images (**Fig. 2**). Methods development specific to continuous conformational variability analysis of membrane proteins is currently lacking. Therefore, membrane proteins are usually studied using classical, discrete-classification-based methods, which result in a limited, small number of discovered conformational states and a loss of fine details of the dynamics of these proteins (Nature 2019, Sci Adv 2022).

This PhD thesis will be focused on image analysis methods development based on deep learning (DL) to decipher continuous conformational transitions of biomolecular complexes (under the supervision of Dr. Jonic), which will be followed by their validation using cryo-EM datasets of small, membrane protein complexes that are available at the IMPMC laboratory (under the supervision of Prof. Vénien-Bryan**)**. In particular, the tests will be performed using cryo-EM data of the human potassium channel Kir2.1. Kir2.1 is a membrane protein that selectively controls the permeation of K$^+$ ions at the cell membranes of a variety of tissues and regulates the membrane electrical excitability. The overall effect of rectification allows Kirs to play a key role in eukaryotic cells by driving the resting membrane potential to EK (K+ equilibrium potential) when the cell is at rest, regulating pancreatic insulin secretion, contributing to renal K$^+$ transport, controlling muscle contraction, and regulating pacing in cardiac cells and neurons. The physiological importance of the Kir channels is highlighted by the fact that genetically-inherited defects in Kir channels are responsible for a number of human diseases, such as Andersen's syndrome, Bartter's syndrome, neonatal diabetes, short QT syndrome, and atrial fibrillation (Ion Channels Disease 1999). To date, the available treatments are rather inefficient. How this channel gates is still unknown. In a previous study supervised by Prof. Vénien-Bryan, it has been shown that the human Kir2.1 is very dynamic and undergoes large conformational changes (Sci Adv 2022). These conformations are often impeded when the protein is mutated, which makes the channel dysfunctional. The knowledge of a fine detail of the dynamics of the protein, which will be studied using new methods, proposed to be developed here, will help to correct for the dysfunction of the protein.

2. Approach, Supervisors, Candidate, Time plan

The Jonic team started developing DL approaches for extracting information on conformational landscapes from cryo-EM data in 2019, as one of the first groups in the world trying to use DL for solving this difficult inverse problem (Nat Methods 2021a; Nat Methods 2021b). They have published one supervised DL method (DeepHEMNMA, Front Mol Biosci 2022) that combines a ResNet 34 convolutional neural network (CVPR 2016) with a multilayer perceptron block. DeepHEMNMA learns 3 Euler angles, 2 in-plane shifts, and a small number of conformational parameters for each particle image in a training dataset obtained by HEMNMA image analysis (Structure 2014) using a linear combination of vectors (the so-called normal modes) that simulate the potential motion directions of a given, reference conformational model (the conformational parameters are determined by the coefficients of the linear combination of normal modes). DeepHEMNMA was shown to be multiple tens of times faster than its non-DL version HEMNMA (at least 40 times faster in the tests with synthetic data shown in Front Mol Biosci 2022). With experimental data, DeepHEMNMA revealed that the conformational heterogeneity of yeast 80S ribosome is larger than the one previously found with discrete-classification methods (Front Mol Biosci 2022). The work on DeepHEMNMA is part of the PhD thesis of Ilyes Hamitouche, successfully defended on March 29th, 2023, funded by the ANR (project EMBioMolMov, 10/2019-03/2024). In the same PhD thesis framework, the continuous conformational variability problem was also addressed using vision transformers (IEEE Trans PAMI 2023), by setting up one supervised and one unsupervised network (cryo-ViT networks, Link to manuscript). This will serve as the basis for the new PhD thesis work, as explained hereafter.

The supervised cryo-ViT learns the relationships between each particle image in the training dataset and its corresponding large number of atomic model coordinates (e.g., obtained by MDSPACE image analysis, i.e., 3 times N coordinates, where N is the number of atoms). The unsupervised cryo-ViT determines the atomic coordinate displacement for each particle image, with respect to a given reference atomic model, based on the similarity between the particle image and the image generated (physics-based simulator) from the displaced atomic coordinates. The latter approaches have shown encouraging results (Link to manuscript), but they require training and testing on larger datasets and improvements regarding speed, accuracy, uncertainty due to the low SNR and artefacts observed in experimental images (minimize the risk of the network hallucinations due to imperfect data), generalization when using different datasets for training and tests such as data collected under different imaging conditions (e.g., different microscopes), and occlusions (e.g., detergent around membrane proteins). The PhD thesis proposed here will tackle all these aspects. The current cryo-ViT approaches use the original architecture of the vision transformer (number of transformer blocks, number and size of image patches, ICLR 2021), which will be studied in the proposed PhD thesis to define the minimum size architecture with an improved efficiency and reduced redundancy of different layers for this cryo-EM task. The generalization problem could be addressed using techniques based on the domain adaptation principle Adv Data Sci Inform Eng. 2021 (imposing invariance to changes in data distribution), including batch normalization Pattern Recognition 2018, and adversarial techniques CVPR 2017. The approaches to address the uncertainty problem include training using simulated data (simulating noise and artefacts), such as in the case of Generative Adversarial Networks (GANs) NIPS 2014 and the use of the energy of conformation in the loss function to constrain the network to learn only those structures that are physically plausible (physics-based motions).

This work will mainly be focused on unsupervised DL approaches, including the improvement of unsupervised cryo-ViT and the development of a GAN-based approach. These approaches will be compared with strongly and weakly supervised approaches. The methods adaptation to membrane proteins in different environments could involve masking out the environment around the protein using hand-made masks or the masks inferred from images by DL.

Each supervisor has unique expertise in France: development of innovative cryo-EM image analysis methods (Dr. Jonic, ED 130) and functional and structural studies of potassium channels at high resolution using cryo-EM (Prof. Vénien-Bryan, ED 515). The methods development and tests with typical test complexes such as ribosomes and other large complexes available in the public cryo-EM experimental-data archive EMPIAR will be supervised by Dr. Jonic. The validation with experimental cryo-EM data of small membrane protein complex Kir2.1 (human potassium channel) will be supervised by Prof Venien-Bryan. The experimental cryo-EM images of Kir2.1 are already available in the team of Prof Venien-Bryan. The feedback from the data analysis of Kir2.1 in different environments will help improve the methods and the improved methods will help better understand the differences in the amplitude of the conformational change of Kir2.1 in the different environments observed with classical, discrete-classification methods.

The candidate will have a background in Computer Science, Applied Mathematics or a related field.

The thesis will be organized according to the following time plan:

1) Develop unsupervised DL approaches to simultaneously learn angles, shifts, and conformations from a large number (~10^6) of cryo-EM particle images in the context of continuous conformational heterogeneity (**Year 1-2**)
2) Adapt the new approaches to membrane proteins in different environments (**Years 2-3**)
3) Test the new approaches with cryo-EM images of typical test complexes such as ribosomes and other large complexes available in the public data archive EMPIAR (**Years 1-3**)
4) Test the new approaches with cryo-EM images of small human membrane protein complex Kir2.1 that are available at the IMPMC laboratory (**Years 1-3**)

3. References of the supervisors related to the project:

1. Vuillemot R, …, **Jonic S** (**2023**) MDSPACE: Extracting continuous conformational landscapes from cryo-EM single particle datasets using 3D-to-2D flexible fitting based on Molecular Dynamics simulation. *J Mol Biol*: 167951. doi: 10.1016/j.jmb.2023.167951. Link
2. Harastani M, …, **Jonic S** (**2022**) ContinuousFlex: Software package for analyzing continuous conformational variability of macromolecules in cryo electron microscopy and tomography data. *J Struct Biol* 214:107906. doi: 10.1016/j.jsb.2022.107906. Link
3. Hamitouche I, **Jonic S** (**2022**) DeepHEMNMA: ResNet-based hybrid analysis of continuous conformational heterogeneity in cryo-EM single particle images. *Front Mol Biosci* 9:965645. doi: 10.3389/fmolb.2022.965645. Link
4. Fernandes CAH, …, **Vénien-Bryan C** (**2022**) Cryo-electron microscopy unveils unique structural features of the human Kir2.1 channel. *Sci. Adv*. 8(38):eabq8489. doi: 10.1126/sciadv.abq8489. Link
5. Fagnen C, …, **Vénien-Bryan C** (**2021**) Integrative Study of the Structural and Dynamical Properties of a KirBac3.1 Mutant: Functional Implication of a Highly Conserved Tryptophan in the Transmembrane Domain. *Int. J. Mol. Sci.* 23, 335. Link
6. **Vénien-Bryan C**, …, Boutin J (**2017**).Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. *Acta Cryst F Struct Biol Commun.* 73, 174–183. doi: 10.1107/S2053230X17003740. Link

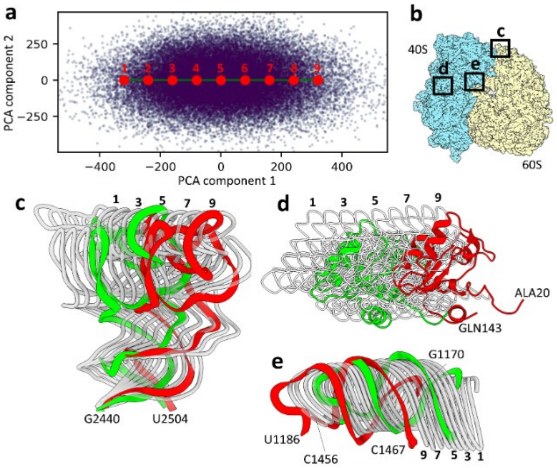

Fig. 1: MDSPACE analysis of cryo-EM images of yeast 80S ribosome (Vuillemot et al., 2023)

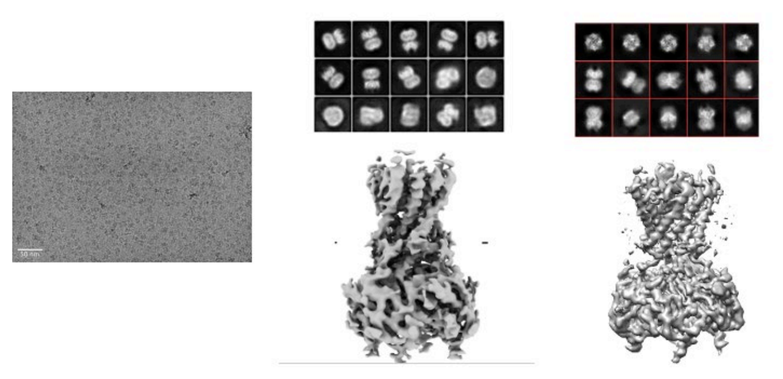

Fig. 2:Cryo-image of the Kir Channel (left panel), 2D classification and 3D reconstruction of Kir channel in 2 environments, in detergent DDM (middle panel, Fernandes et al., 2022) and in Apols (right panel)