Novel machine learning methods for determination of protein structural ensembles from single molecule cryo electron microscopy

Summary: Accurate determination of a protein's structure and dynamics are key to fundamental and applied advances in life science. Today, single-particle cryo electron microscopy (EM) can resolve the structure of large protein complexes and artificial intelligence (AI) can predict protein structure. Combining these approaches offers a large untapped potential for determining more accurate protein structure and dynamics from a much wider range or target proteins.

Here, we propose to integrate AI protein structure prediction with instance-specific experimental cryo EM data. This will require fundamental advances in machine learning to steer the predictions through auxiliary constraints at inference time and to rigorously propagate probabilities through this inference process. We will refine hybrid (experimental/predicted) protein structures for multidomain complexes and for conformational ensembles of the proteins.

The mathematical advances can be transferred to numerous other applications of AI, not least in the natural and life sciences, including materials discovery, drug design, and personalized medicine. At the same time, the project will result in methodological progress in structural biology, exemplified for proteins important for plant photoreception and the circadian rhythm. We will make the methods widely available to the Swedish Life Science sector.

Project plan

Background: Determination of biomolecular structures has led to enormous scientific breakthroughs, innovations, and new treatments for diseases (>15 Nobel prizes in structural biology alone). Initially protein structures were determined by crystallography, but over time single-particle cryo electron microscopy has become equally important. Here, images are recorded of thousands of single copies of proteins, but unfortunately state-of-the-art analysis requires averaging of most of the particles prior to structure determination. Thus, the information on conformational heterogeneity between single copies of the proteins are mostly lost. This means that the dynamics of biomolecules in solution, or rare but important conformations, such as those in activated states, cannot be determined.

The recent advent of deep-learning-based systems for structure prediction reached a climax when AlphaFold 2 (**Jumper, Nature, 2021**) resulted in the first ever solution to the protein folding problem. However, unsolved challenges remain: proteins with shallow alignment depths or multiple domains cannot be predicted well and the reliability of the predictions and their associated uncertainty measures are unclear. As a result, prediction of conformational heterogeneity is currently not possible. Single-particle cryo electron microscopy can provide crucial information with respect to these shortcomings. However, to realize this potential for molecular biology, machine learning and experiment have to be integrated in a systematic way.

Predictive machine learning models are commonly trained from large amounts of data, but once trained they are unable to adapt to constraints or auxiliary, instance-specific data during inference. This is a particular problem when reliable uncertainty quantification (UQ) regarding their predictions are required. Novel machine learning methods need to be developed to overcome this challenge. In this project, we will address this problem and develop new AI methods, applying them to protein structure predictions.

Project goals and implementation

Goal 1: Protein ensembles from ensemble-averaged cryo EM data and protein structure predictions.
First, we will explore statistically sound methods to refine predicted conformational ensembles of proteins to ensemble-averaged cryo EM maps. We will use Bayesian statistics and fully consider the reliability of the predicted structures in the ensemble to achieve this. We will generate large ensembles of proteins with AlphaFold, which is now possible by tuning the so-called drop out parameters. We will use Westenhoff's existing cryo EM data on a bacterial phytochrome (**Wahlgren et al., Nat Comm, 2022**) as a test case and develop a refinement script based on Bayesian fitting of the conformational ensembles. We will propagate fully the reliability of the predictions, which is important in order not to over-interpret the data, and which requires mathematical insight. While this goal does not (yet) realize the full potential of the cryo EM data, it will generate a robust method and will be a solid start for the student to get into the mathematical and applied sides of the topic. Expected time to completion: 6 months.

Goal 2: Adaption of predictive models of protein structure to experimental constraints.
At the core of the project is the development of new AI methodology, enabling the adaptation of (pre-trained) predictive models during inference, by guiding their predictions based on instance-specific experimental data or similar constraints. We will have to consider the underlying machine learning models, inference algorithms, and statistical methods to be able to consistently incorporate the constraints. The resulting methods for model adaptation will be a generic contribution to the AI field. As a concrete application of the new methodology and with very high scientific value in itself, we will attempt to use the new prediction method to obtain more accurate models of protein complexes that are transiently present during the circadian rhythm in mammalian cells. These samples are hard to produce in microgram quantities, which limits the applicability of cryo EM - at best one can expect a low-resolution structure. Our new AI/experimental hybrid structure can overcome this bottleneck, opening up for structure determination of thousands of important protein targets, which suffer from low abundance. We expect the method to be widely applicable. Expected time to completion: 2 years.

Goal 3: Machine-learned protein structure ensembles and dynamics from single-molecule cryo EM data and proteins structure prediction
Cryo EM data of proteins consists of thousands of images of single proteins, however, each image is at low resolution. Only when the images are combined, a high-resolution three-dimensional structure can be obtained. Here we propose to overcome this current limitation by learning the ensemble of conformationally disordered proteins from the single-particle images directly. This final goal will require a combination of the Bayesian refinement (explored in goal 1) and the restriction of the prediction at inference time (goal 2), all integrated into a machine learning scheme, in which the proteins ensemble will be learned from the single-particle images (constraints) and from the protein data base.
We will apply the new method to the phytochrome proteins and possibly other targets. Westenhoff has recently solved partial structures of the protein using cryo EM (**Wahlgren et al., Nat. Comm, 2022**). The structures show a high degree of conformational flexibility in approximately 50% of the protein, which makes them an ideal test case for this study. Expected time to completion: 4 years.

Supervisors /roles:

Professor S. Westenhoff (Biochemistry, Department of Chemistry-BMC, University of Uppsala) will be the main supervisor. Westenhoff is an expert in protein structure determination. He has pioneered time-resolved methods for determination of protein structural changes **(Takala et al., Nature 2014; Björling et al. Science Advances 2016; Claesson et al., elife, 2020; Dods et al., Nature, 2021)**. Westenhoff's research group is actively solving proteins structures in the fields of photoreception of plants and bacteria and the circadian clock.

Assoc. Prof F. Lindsten (Computer and Information Science, Linköping University), will be the co-supervisor. Prof Lindsten is an expert in uncertainty estimation for AI, where state-of-the-art involves probabilistic modeling **(Sidén and Lindsten, ICML 2020),** approximate inference **(Naesseth et al., NeurIPS 2020),** calibration evaluation **(Widmann et al., ICLR, 2021)**, and ensembles **(Naesseth et al., FnTML 2019).**

Dr. Gabriel Ducrocq (Computer and Information Science, University of Linköping), is a shared PostDoc between Westenhoff and Lindsten. He will overlap with the new student during the first 9 month and therefore be able to transfer knowledge to the new student.

Interdisciplinarity

The student will be placed at the Department of Chemistry - BMC, University of Uppsala. S/he will be an integral part of Westenhoff's team and Westenhoff's and Lindsten's ongoing WASP/DDLS project on this topic. The team currently consists of a project-funded PostDoc (Ducrocq; computer science) and a PhD student in Westenhoffs group (Monroy, experimental structural biology). The new student will become a vital link in this small team, bridging mathematics, machine learning, and biochemistry, and will profit from the regular team meetings (2/term).

On the mathematics side, this PhD project requires deep knowledge of machine learning models, inference algorithms, and statistical methods as well as some programming skills. Support is primarily provided by Lindsten at Linköping University, who was employed at UU until recently. The student will have regular supervisions on zoom and part of this work may be carried out during placements in Lindsten's group at Linköping University. The student will also have access to the extensive network in AI at the University of Uppsala, the AI4Research initiative, and the CIM PhD school.

Regarding the biophysical aspects, the project requires a solid understanding of the theory and biochemical processes underlying protein folding, (2) the infrastructure and know-how required for obtaining cryo EM data (which can be supplied by Westenhoff's group members), and (3) some biological expertise of plant photodetection and circadian rhythm. These expertise areas are provided by Westenhoff, who has an excellent track record in biophysical chemistry. Westenhoff has also previously developed computational tools for time-resolved structural biology **(Björling et al., J Chem Theory and Computation, 2015)**.

Co-financing. The position will be financed 50% by CIM, 25% by the Department of Chemistry -BMC and 25% by Westenhoff's WASP/DDLS project.