CryoSPHERE: bridging the gap between AlphaFod and cryoEM experiments

Gabriel Ducrocq[1], **Lukas Grunewald**[2], **Sebastian Westenhoff**[2], **Fredrik Lindsten**[1]

[1] Department of Computer and Information science - Division of Statistics and Machine learning, Linköping University, Sweden

[2] Department of Chemistry - BMC, Biochemistry, Uppsala University, Sweden

Introduction

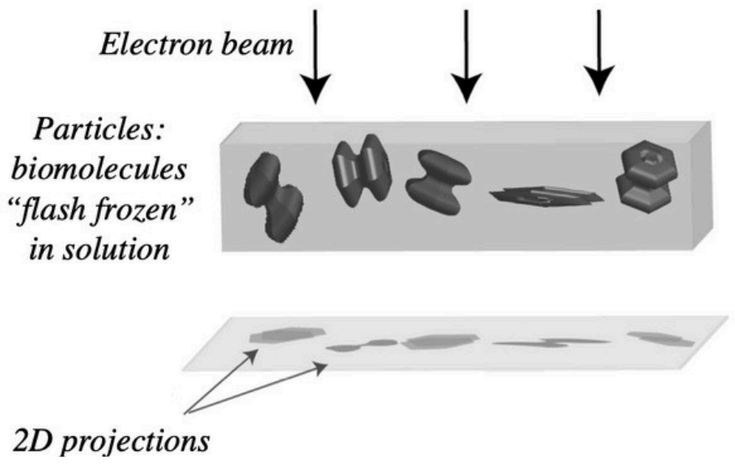


Figure 1: cryoEM acquisition process.

CryoEM reconstruction challenges:

- 100k to 1M images of copies of the same protein
- Orientation of the protein on each image is unknown
- Signal-to-noise ratio very low (0.1 to 0.001)
- On each image, the protein has different shape

We want to recover, for each image, the corresponding shape of the protein.

The current SOTA:

Too **low resolution** for very noisy datasets

Too **low resolution** for very mobile datasets

Do **not** use AlphaFold prior knowledge

Method

CryoSPHERE uses the prior knowledge on the protein structure brought by AlphaFold and learns concurrently:

- How to decompose the protein structure in N segments
- How to move these segments approximately rigidly to recover the conformations

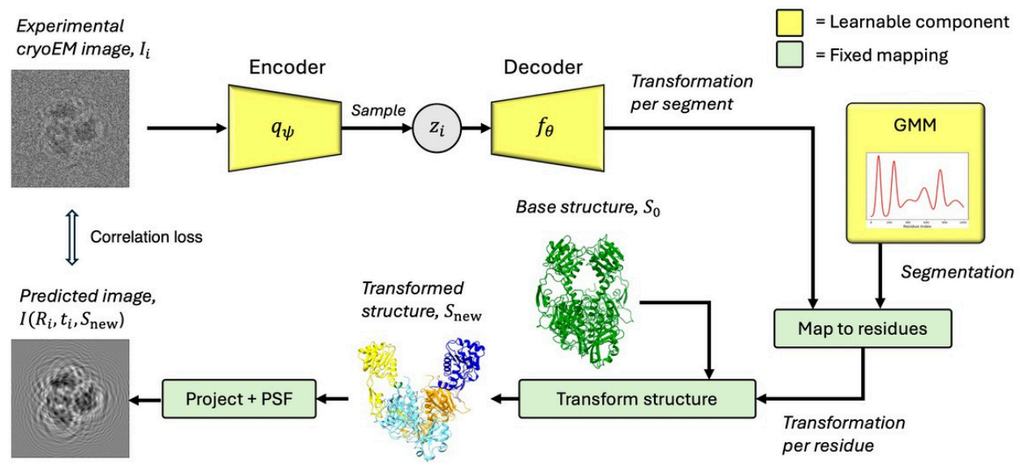


Figure 2: flow chart of cryoSPHERE

Experiments

Toy dataset:

10k structures obtained by MD simulation

- 150k images in total with SNR 0.001

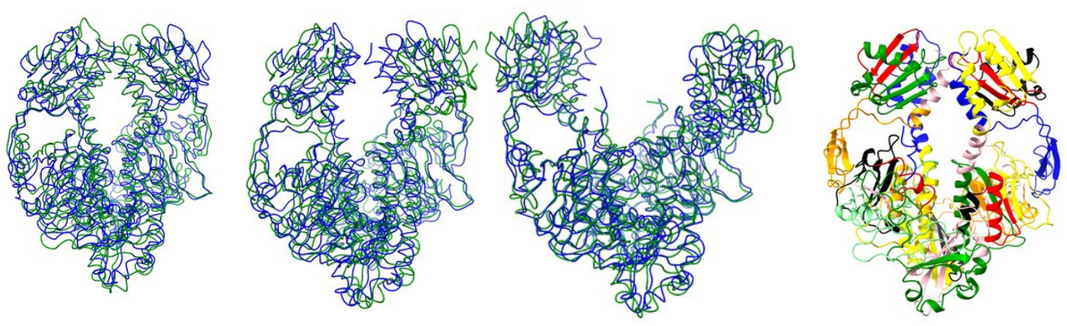
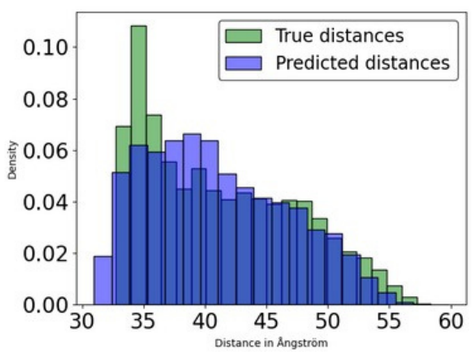


Figure 3: in blue, selection of predicted structures by cryoSPHERE. In green, corresponding ground truth. Rightmost: base structure with colored segments learned by cryoSPHERE.



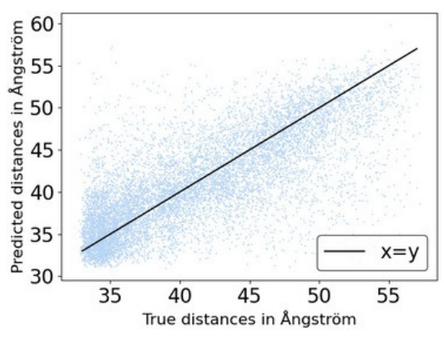


Figure 4: left: histograms of the ground truth and predicted distances between the two upper domains of the protein. Right: predicted vs true distances.

EMPIAR 10180:

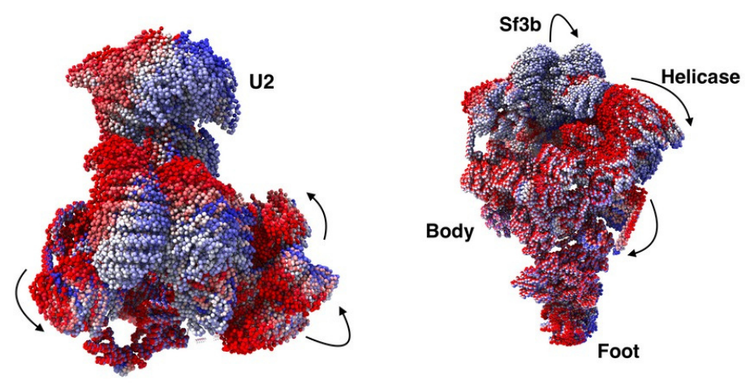


Figure 4: traversal of the latent space, from red to blue. The Sf3b domain gets incurvated down while the helicase move closer to the foot of the protein.

Contact informations:

Email: gabriel.ducrocq@liu.se

Website: https://gabriel-ducrocq.github.io/cryosphere.github.io/