CryoSPHERE: Single-Particle HEterogeneous REconstruction from cryo EM

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

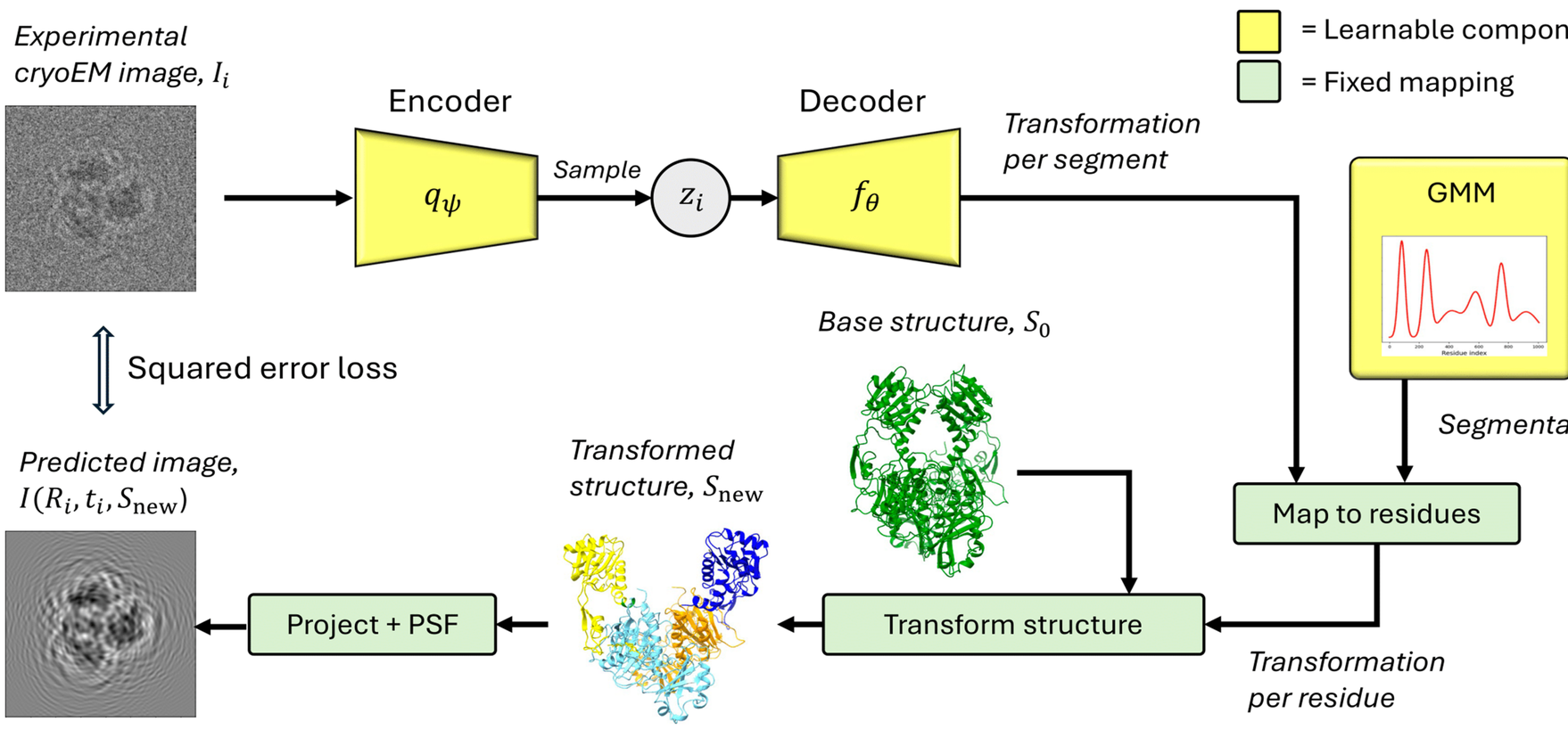
Abstract

Single-particle cryo electron microscopy is a powerful tool to solve the three-dimensional structure of a given protein. The three-dimensional structure of a protein plays a key role in determining its function. Methods like AlphaFold have revolutionized protein structure prediction based only on the amino-acid sequence. However, proteins often appear in multiple different conformations, and it is highly relevant to resolve the full conformational distribution.

Single-particle cryo-electron microscopy (cryo EM) is a powerful tool for capturing a large number of images of a given protein, frequently in different conformations. The images are, however, very noisy projections of the protein, and traditional methods for cryo EM reconstruction are limited to recovering a single, or a few, conformations.

Here we introduce cryoSPHERE, a deep learning method that takes as input a nominal protein structure, e.g. from AlphaFold, learns how to divide it into segments, and how to move these as approximately rigid bodies to fit the different conformations present in the cryo EM dataset. This formulation is shown to provide enough constraints to recover meaningful reconstructions of single protein structures.

Network Flowchart



CryoSPHERE is based on a **variational auto-encoder** (VAE) that, from a input structure and a set of cryo EM images:

- Learns how to divide the amino acid chain into segments, given a user defined maximum number of seqgments (see Fig. 2). The input structure can for instance be obtained by AlphaFold.
- Learns for each image approximately rigid transformations of the identified segments of the input structure to fit different conformations on an image-by-image (single particle) basis.

Figure 1: CryoSPHERE flow chart. The learnable parts of the model are the encoder, decoder and Gaussian mixture model (GMM).

Learned GMM divides the protein perfectly into flexible and rigid segments

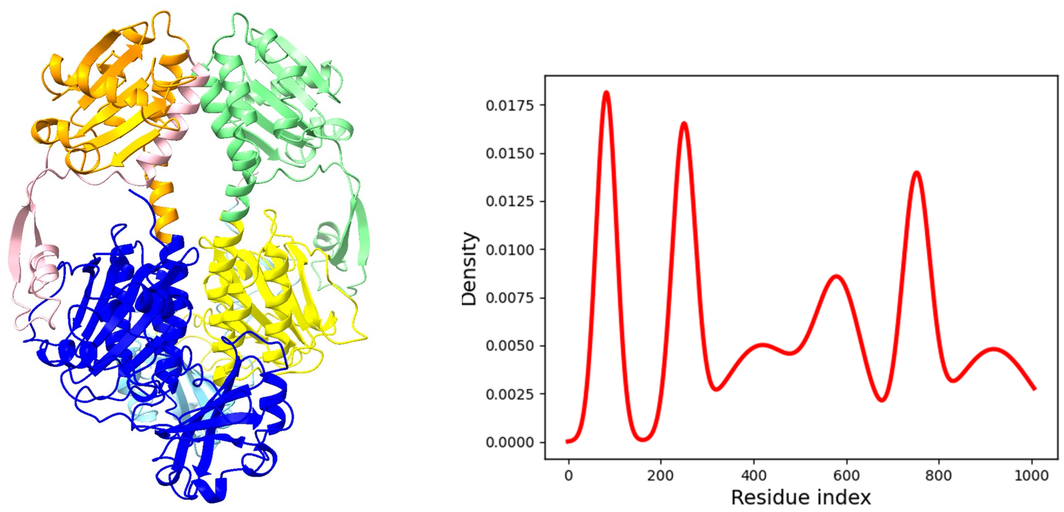


Figure 2: Example of recovered segments with a Gaussian mixture of 6 components (in different colors).

CryoSPHERE recovers continuous conformational heterogeinity of a protein dissociation process

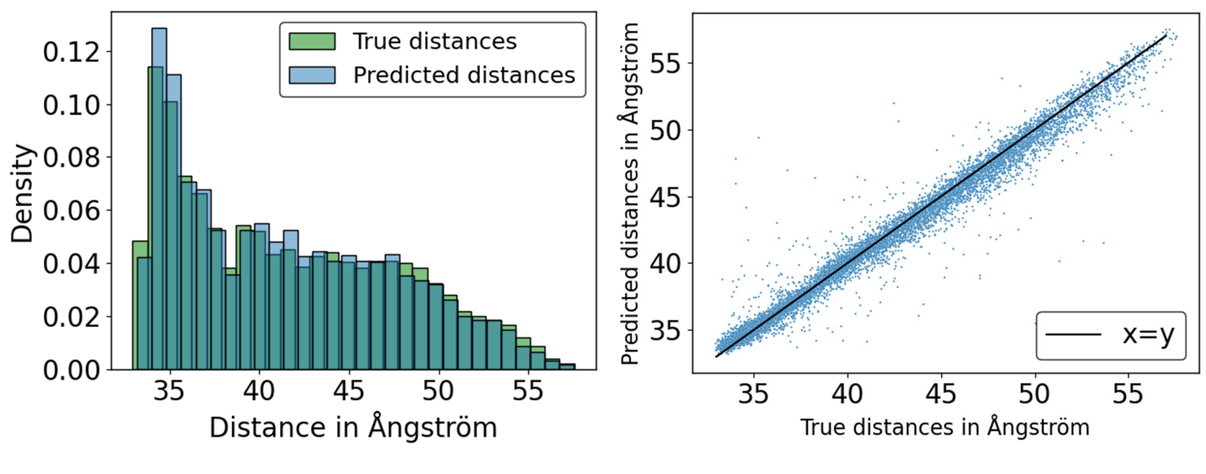


Figure 3: Simulated dissociation process of the upper two domains of the protein, which is depicted in Fig. 2. Left: Histograms of the distances of the two upper domains. The true distances of the dissocation process are in green. The recovered distances are in blue. Right: Predicted against true distances in Ångstrom of the dissociation process.

CryoSPHERE predictions of dissociated structures

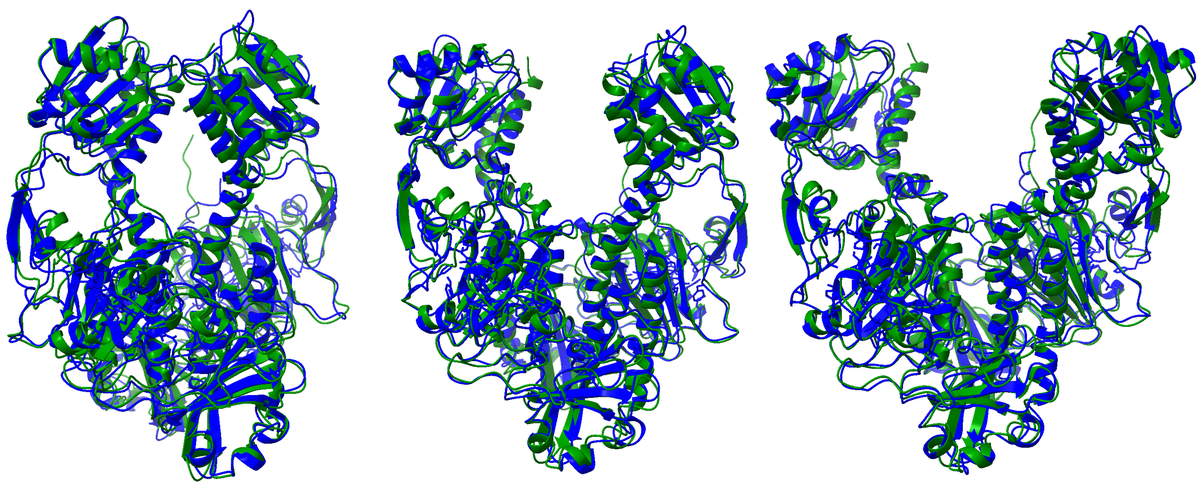


Figure 4: From left to right: true structures (blue) and cryoSPHERE predicted structures (green). We see an excellent agreement between ground truth and prediction.

Summary

- CryoSPHERE is interpretable. We understand how each segment is deformed to obtain a specific conformation.
- CryoSPHERE reduces the dimensionality of the problem: instead of learning a transformation for each residue, it targets one transformation per segment, where the number of segments is small compared to the number of residues.
- CryoSPHERE is memory efficient compared to methods acting on an explicit or implicit grid.
- CryoSPHERE is resilient to noise: by constraining the motion to be approximately rigid, we restrict the motion to low frequency movements, less polluted by noise.