Spectral decomposition of protein structures in heterogeneous cryo-EM

Carlos Esteve Yagüe

Cambridge Image Analysis, DAMTP
University of Cambridge
ce423@cam.ac.uk

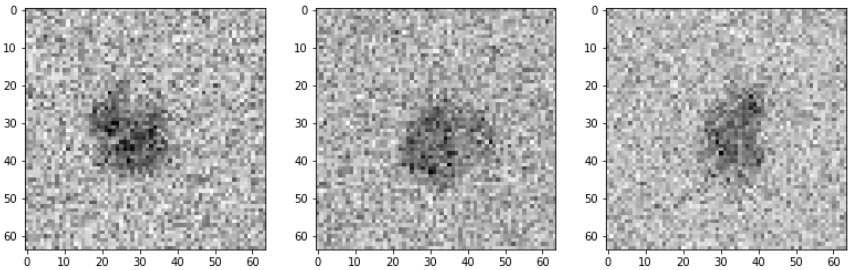
March 2022

The cryo-EM problem with continuous heterogeneity

We are given a dataset $\{Y_j\}_{j=1}^n$ with n images of the protein:

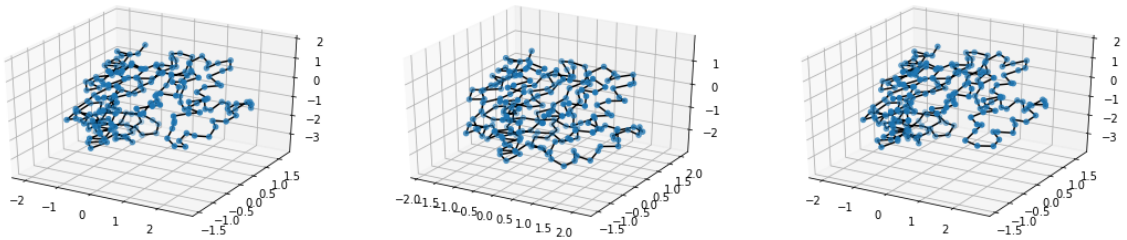
$$Y_j = \mathcal{F}(R_j u_j) + \xi_j \qquad \text{for each } j = 1, 2, \ldots, n,$$

where $\mathcal{F}$ is the forward operator, $R_j \in SO(3)$ and $u_j \in L^2(\mathbb{R}^3)$.



In each image the protein shows a different conformation.

Goal: Determine the atomic structure in each image.



Assumptions:

1. We assume that the set of possible **conformations of the protein forms a low-dimension compact manifold**, that we denote by $\mathcal{M}$.

2. We assume that we have solved the **homogeneous cryo-EM problem**:
   - We have the 3-D atomic structure of the average conformation.
   - We know the orientation of the protein in each image.

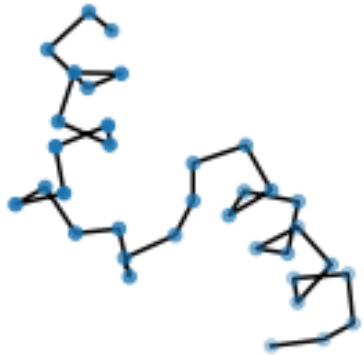
Goals:

1. Construct the manifold of conformations $\mathcal{M}$.
2. Determine the 3-D protein structure $u[m](\cdot) \in L^2(\mathbb{R}^3)$ corresponding to each conformation $m \in \mathcal{M}$.

We represent the 3-D structure of a protein by the spatial position of its C-α atoms:

$$(x_1, x_2, \ldots, x_L) \in \mathbb{R}^{3L},$$

where L is the number of C-α atoms in the protein.



From point clouds to densities

$$(x_1, x_2, \ldots, x_L) \in \mathbb{R}^{3L} \longmapsto u(z) = \sum_{i=1}^{L} \gamma e^{\frac{|z-x_i|^2}{2\sigma^2}}, \qquad \text{for } \gamma, \sigma > 0.$$

Parameter space of backbones

$$\mathcal{P} := \mathbb{R}^3 \times SO(3) \times [0, 2\pi)^{L-2} \times [0, \pi)^{L-2}.$$

- $x_0 \in \mathbb{R}^3$ determines the spatial location of the protein;
- $P_0 \in SO(3)$ determines the orientation of the protein;
- $(\Theta, \Psi) \in [0, 2\pi)^{L-2} \times [0, \pi)^{L-2}$ determines the conformation.

Goal: Find a function

$$(\Theta, \Psi) : \mathcal{M} \longrightarrow [0, 2\pi)^{L-2} \times [0, \pi)^{L-2}$$

Approach:

$$\theta_i(m) := \sum_{k=0}^{r_i} \alpha_{i,k} \varphi_k(m) \quad \text{and} \quad \psi_i(m) := \sum_{k=0}^{r_i} \beta_{i,k} \varphi_k(m), \quad \text{for } i = 1, 2, \ldots, L-2,$$

where $\varphi_0(\cdot), \varphi_1(\cdot), \varphi_3(\cdot), \ldots$ are the first eigenfunction of the Laplace-Beltrami operator in $\mathcal{M}$.

Spectral decomposition of $\mathcal{M}$

For each $m \in \mathcal{M}$, we define

$$\theta_i(m) := \sum_{k=0}^{r_i} \alpha_{i,k}\varphi_k(m) \quad \text{and} \quad \psi_i(m) := \sum_{k=0}^{r_i} \beta_{i,k}\varphi_k(m), \qquad \text{for } i = 1, 2, \ldots, L-2,$$

where $\varphi_0(\cdot), \varphi_1(\cdot), \varphi_3(\cdot), \ldots$ are the first eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$.

Bad news: we do not know $\mathcal{M}$

However, we have many samples $\{m_j\}_{j=0}^{n} \subset \mathcal{M}$ in the dataset of cryo-EM images.

- We can approximate $\varphi_k(m_j)$ by the eigenvectors of the graph Laplacian.
- This relies on the possibility of being able to compare the conformation in each cryo-EM image (we can use low-resolution reconstructions).

Spectral decomposition of $\mathcal{M}$

For each $m \in \mathcal{M}$, we define

$$\theta_i(m) := \sum_{k=0}^{r_i} \alpha_{i,k}\varphi_k(m) \quad \text{and} \quad \psi_i(m) := \sum_{k=0}^{r_i} \beta_{i,k}\varphi_k(m), \qquad \text{for } i = 1, 2, \ldots, L-2,$$

where $\varphi_0(\cdot), \varphi_1(\cdot), \varphi_3(\cdot), \ldots$ are the first eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$.

Good news: we can use prior knowledge about the protein to choose r_i

- r_i is related to the variability of the parameters θ_i and ψ_i in $\mathcal{M}$.
- We may choose r_i small (even $r_i = 0$) at the C-α which are in stable parts of the protein (for instance at α-helices).

Let us define the maps

$$\begin{array}{rccc} X: & \mathcal{P} & \longrightarrow & \mathbb{R}^{3L} \\ & (x_0, P_0, \Theta, \Psi) & \longmapsto & (x_1, x_2, \ldots, x_L) \end{array}$$

and

$$\begin{array}{rccc} U: & \mathbb{R}^{3L} & \longrightarrow & L^2(\mathbb{R}^3) \\ & (x_1, x_2, \ldots, x_L) & \longmapsto & u(z) = \sum_{i=1}^{L} \gamma e^{\frac{|z-x_i|^2}{2\sigma^2}}, \end{array}$$

Assuming that we have already estimated the orientations $\{P_j\}_{j=1}^n$ of the images in the dataset, we obtain the coefficients $\alpha = \{\alpha_{i,k}\}$ and $\beta = \{\beta_{i,k}\}$ as

$$\underset{\alpha,\beta}{\text{minimize}} \sum_{j=1}^{n} \| Y_j - \mathcal{F}U(X(0, P_j, \Theta_j(\alpha), \Psi_j(\beta)))\|_{L^2}^2,$$

where $\Theta_j(\alpha) = \varphi_j(\alpha)$ and $\Psi(\beta) = \varphi_j(\beta)$ are obtained as the (approximated) spectral decomposition presented above.

Experiment in 2-D

- We aim to reconstruct 2-d images from 1-d noisy tomographic projections.
- The structure has two moving arms and a square that can bend sidewise.
- We use 2000 tomographic projections of the structure taken from arbitrary directions.

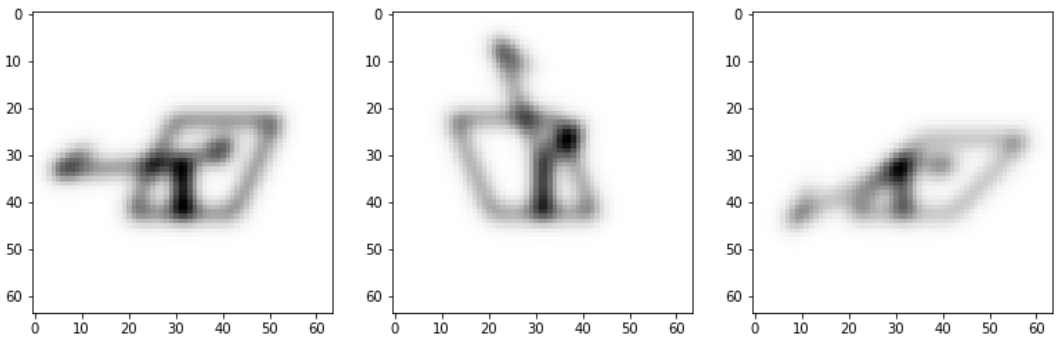


Figure: Original images

Experiment in 2-D

- We aim to reconstruct 2-d images from 1-d noisy tomographic projections.
- The structure has two moving arms and a square that can bend sidewise.
- We use 2000 tomographic projections of the structure taken from arbitrary directions.

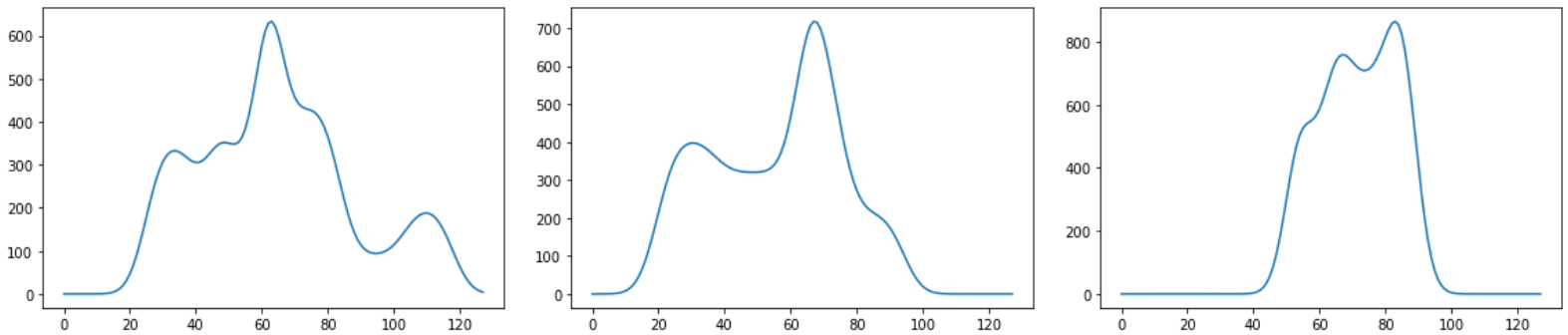


Figure: Clean tomographic projections

Experiment in 2-D

- We aim to reconstruct 2-d images from 1-d noisy tomographic projections.
- The structure has two moving arms and a square that can bend sidewise.
- We use 2000 tomographic projections of the structure taken from arbitrary directions.

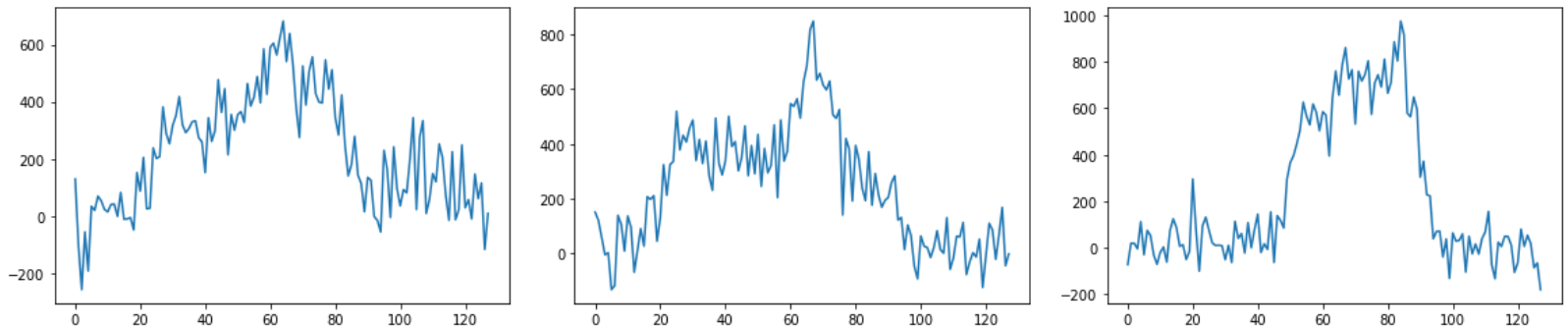
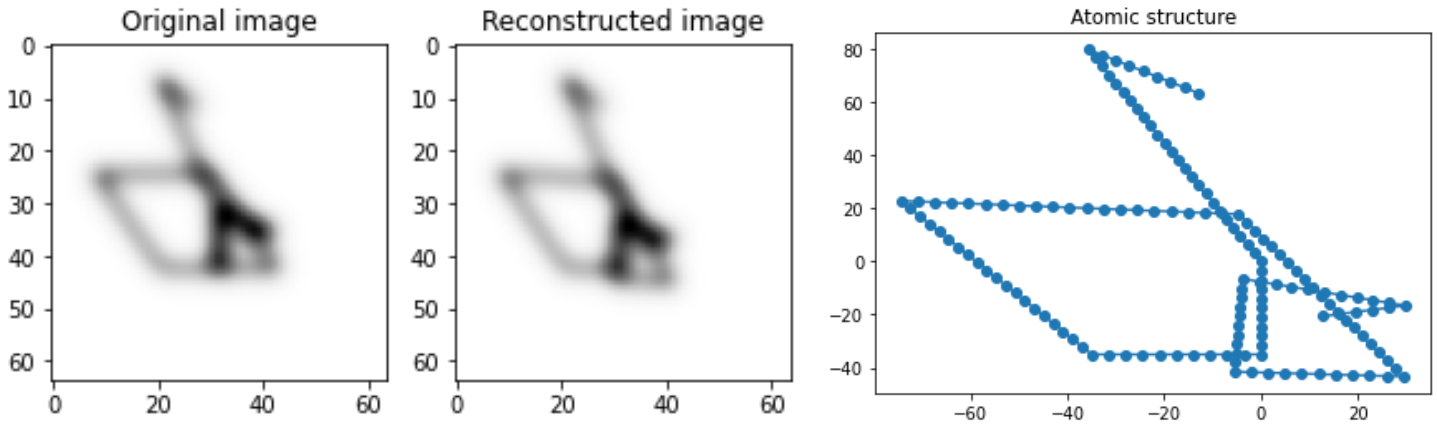


Figure: Noisy tomographic projections

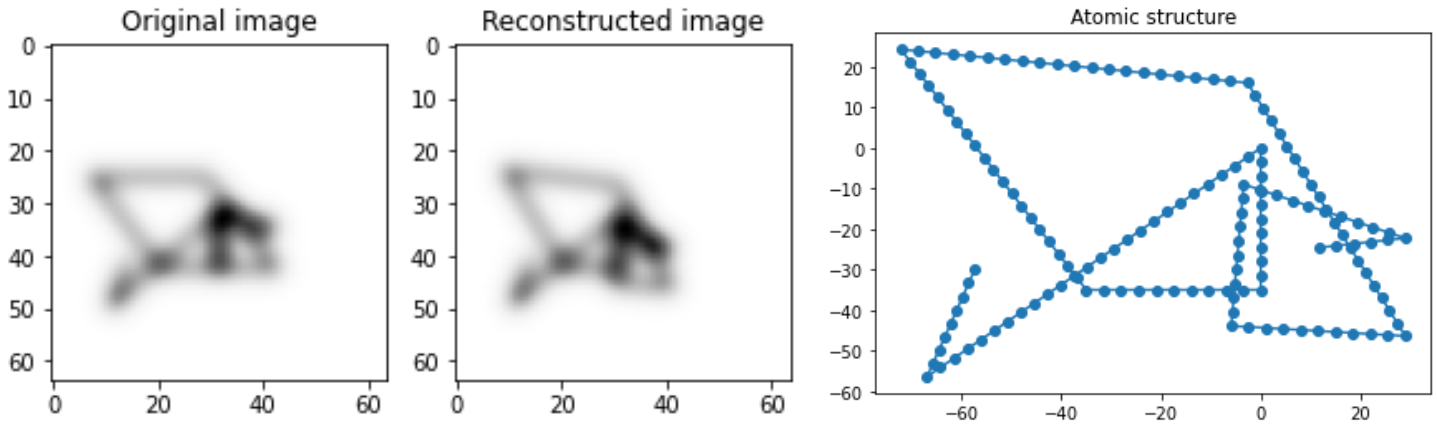
Experiment in 2-D

We not only reconstruct the image, but also obtain the atomic structure



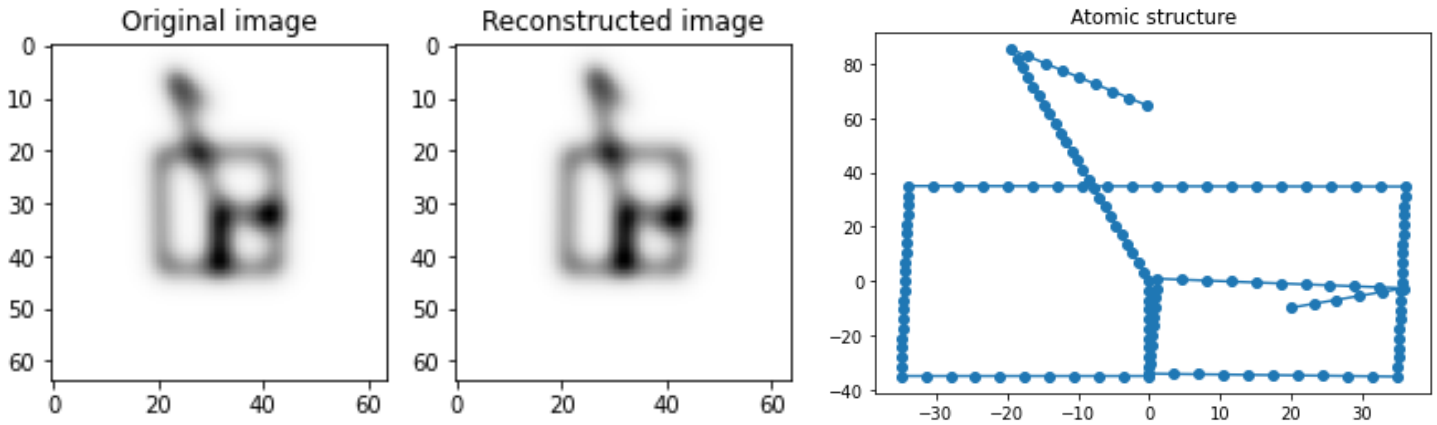
Experiment in 2-D

We not only reconstruct the image, but also obtain the atomic structure



Experiment in 2-D

We not only reconstruct the image, but also obtain the atomic structure



Thanks for the attention!