Spectral decomposition of atomic structures in heterogeneous cryo-EM

Carlos Esteve Yagüe

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

joint work with
Willem Diepeveen, Ozan Öktem and Carola-Bibiane Schönlieb

November 18, 2022

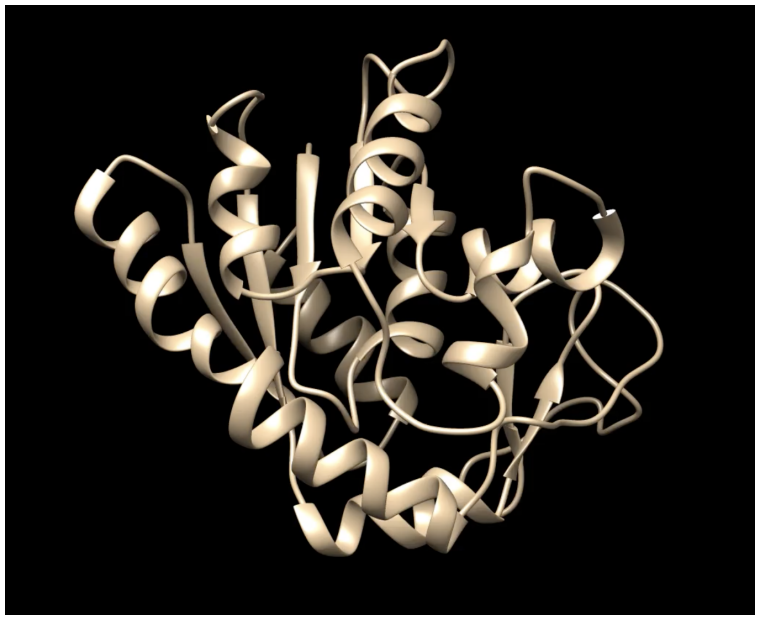
UNIVERSITY OF CAMBRIDGE

Goal

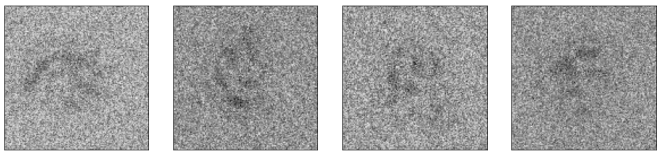
Determine the atomic structure of a macromolecule such as a protein.

Biological knowledge:

- Sequence of amino-acid residues and chemical composition.
- Extremely complex task due to the large number of atoms.
- Deep Learning techniques (AlphaFold).



Cryo-EM dataset:



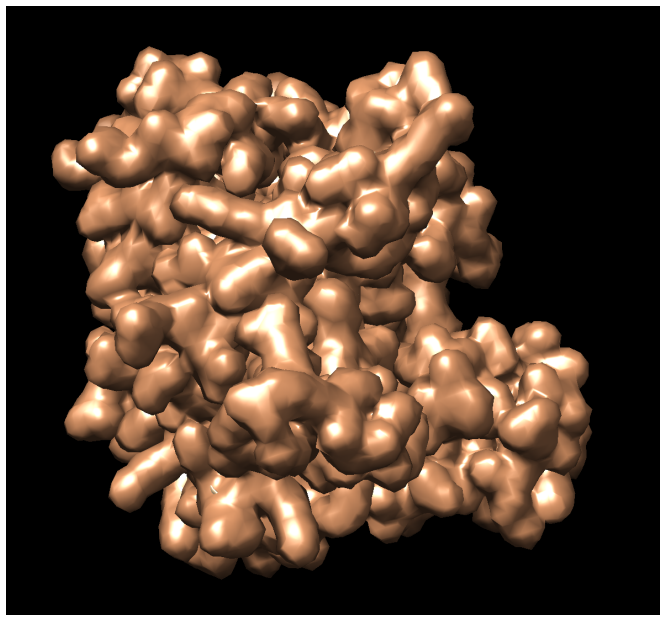
- Very noisy 2D images and unknown orientations.
- Maximum likelihood approach (Sigworth, 1998).
- RELION, Scheres, 2012.

Goal

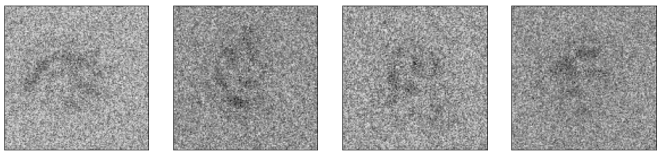
Determine the atomic structure of a macromolecule such as a protein.

Biological knowledge:

- Sequence of amino-acid residues and chemical composition.
- Extremely complex task due to the large number of atoms.
- Deep Learning techniques (AlphaFold).



Cryo-EM dataset:



- Very noisy 2D images and unknown orientations.
- Maximum likelihood approach (Sigworth, 1998).
- RELION, Scheres, 2012.

Heterogeneous structures

Many biological macromolecules are flexible and may be found in different conformations.

Goal: Determine the set of possible conformations $\mathcal{M}$ and construct a map

$$\Phi : \mathcal{M} \longrightarrow \mathcal{S}$$

where $\mathcal{S}$ is the space of molecular structures.

Many biological macromolecules are flexible and may be found in different conformations.

Goal: Determine the set of possible conformations $\mathcal{M}$ and construct a map

$$\Phi : \mathcal{M} \longrightarrow \mathcal{S}$$

where $\mathcal{S}$ is the space of molecular structures.

Many biological macromolecules are flexible and may be found in different conformations.

Goal: Determine the set of possible conformations $\mathcal{M}$ and construct a map

$$\Phi : \mathcal{M} \longrightarrow \mathcal{S}$$

where $\mathcal{S}$ is the space of molecular structures.

Continuous heterogeneity: $\mathcal{M} \subset \mathbb{R}^q$ is a connected manifold of dimension > 0.

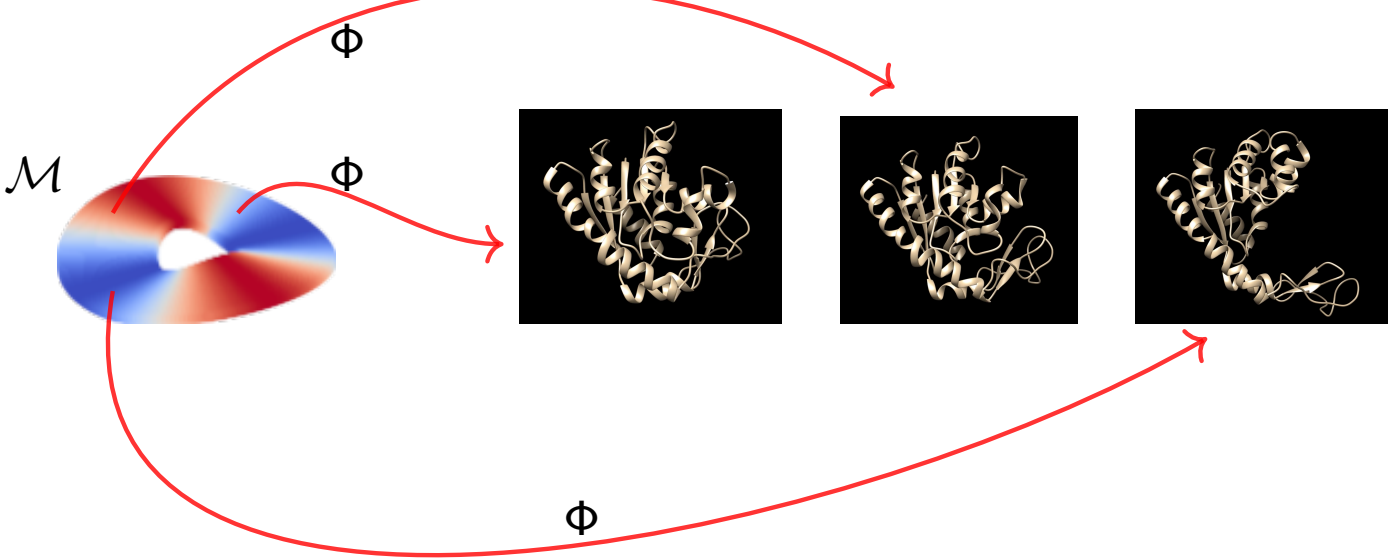
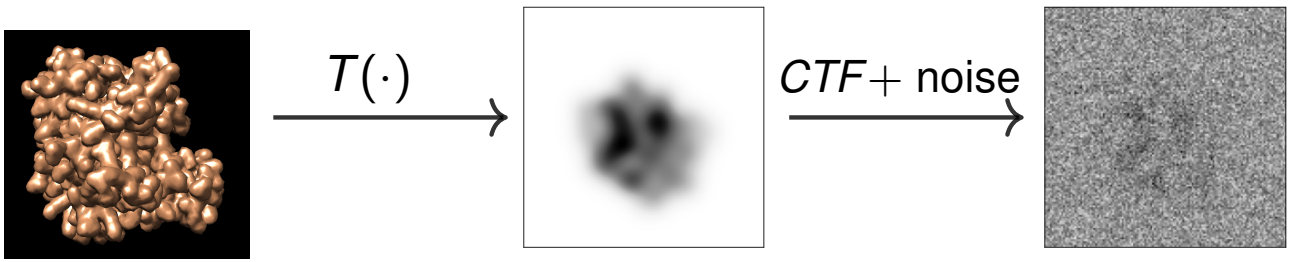


Image formation in cryo-EM: The formation of the image in cryo-EM is often modelled as

$$Y_i = h_i * T(R_i u_i) + \xi_i \qquad \text{for } i = 1, 2, \ldots, N$$

with $N \sim 10^{4-7}$,

- $u_i \in L^2(\mathbb{R}^3)$ represents the electrostatic potential of a single particle in a specific conformation.
- $R_i \in SO(3)$ determines the orientation of the particle.
- $T : L^2(\mathbb{R}^3) \to L^2(\mathbb{R}^2)$ is the parallel beam ray transform.
- h_i* denotes the convolution with the point spread function (PSF).
- ξ_i is Gaussian noise.



Heterogeneous dataset: Each cryo-EM image contains the macromolecule in a different conformation.

Variational Autoencoders (VAE): the conformation of the particle is encoded in the so-called latent space.

- **CryoDRGN:** Zhong,Bepler,Berger,Davis, 2021

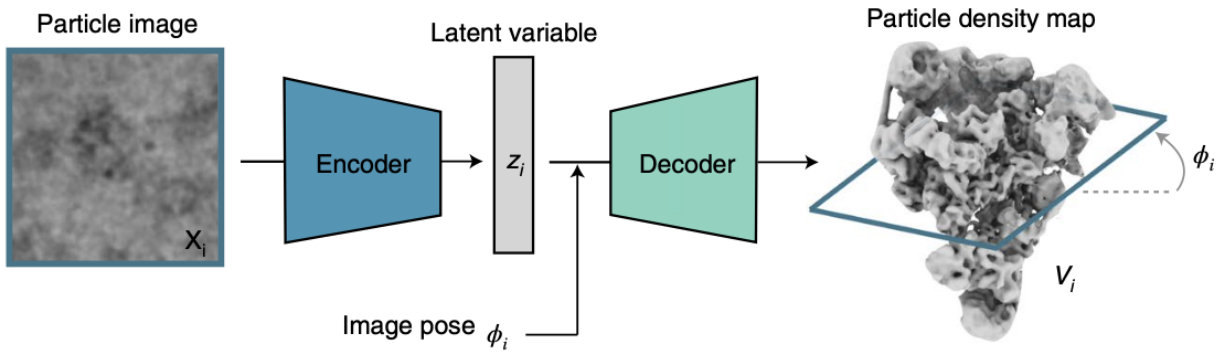


Image taken from Zhong et al., CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks, 2021.

- **Deep Mind:** Rosenbaum et al., 2021

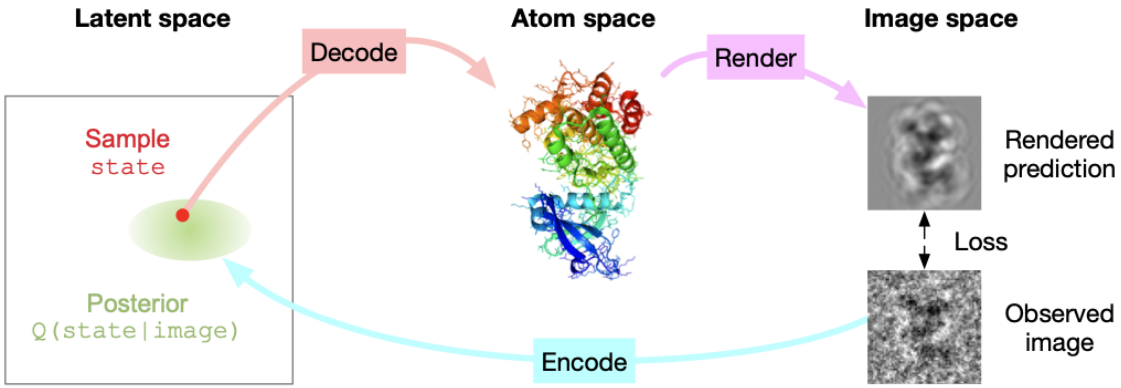


Image taken from Rosenbaum et al., Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs, 2021.

Principal component analysis reconstruction:

- P. A. Penczek. Variance in three-dimensional reconstructions from projections, 2002.
- P. A. Penczek, M. Kimmel, and C. M. Spahn. Identifying conformational states of macromolecules by eigenanalysis of resampled cryo-EM images, 2011.
- J. Andén, E. Katsevich, and A. Singer. Covariance estimation using conjugate gradient for 3D classification in cryo-EM, 2015.
- J. Andén and A. Singer. Structural variability from noisy tomographic projections, 2018.

- Estimate the mean volume $\hat{\mu} \in \mathbb{R}^{N\times N\times N}$ as the maximum likelihood estimator, using the row data.
- Estimate the covariance matrix $\Sigma \in \mathbb{R}^{N^3\times N^3}$ as

$$\hat{\Sigma} = \operatorname{argmin} \sum_{i=1}^{n} \left| (Y_i - P_i\hat{\mu})(Y_i - P_i\hat{\mu})^T - P_i\Sigma P_i^T - \Lambda \right|^2$$

- Compute the principal eigenvectors of $\hat{\Sigma}$ that we denote by

$$V_1, V_2, \ldots, V_q \in \mathbb{R}^{N\times N\times N}.$$

- Compute the coordinates for the PCA $\hat{\beta}_i \in \mathbb{R}^q$ of each particle as

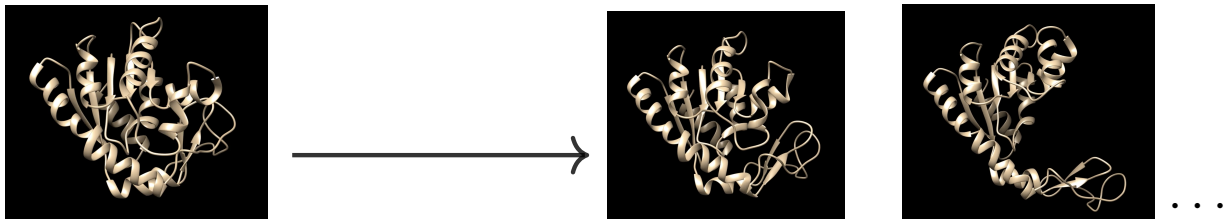
$$U_i = \hat{\mu} + \sum_{j=1}^{q} V_j\hat{\beta}_{i,j}$$

Problem formulation

What we have:

- Low-dimension representation of the conformation (PCA components, latent space ...).
- Pose estimation of the particles.
- Atomic structure of a conformation.
- Other structural properties like secondary structures.

Goal: Determine how the given 3D structure is deformed into the other conformations.



Structural assumption

Let us assume that the macromolecule of interest can be modelled as a chain, or discrete curve.

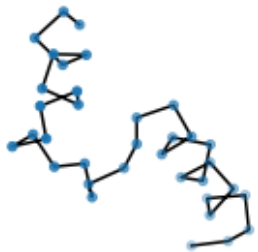
The distance between consecutive (pseudo)atoms is constant. (1)

For instance a protein backbone with a single chain of amino-acid residues.

We use a Gaussian model to estimate the molecule density

$$\mathcal{Z} := (z_1, z_2, \ldots, z_m) \in \mathbb{R}^{3m} \longmapsto U(\mathcal{Z}) := \sum_{i=1}^{m} \gamma_i G(z_i, \sigma_i) \in L^2(\mathbb{R}^3).$$

where m is the number of atoms in the chain.



Structural assumption

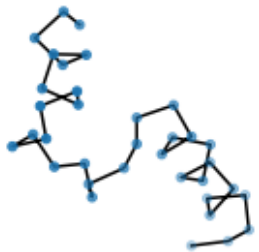
Let us assume that the macromolecule of interest can be modelled as a chain, or discrete curve.

The distance between consecutive (pseudo)atoms is constant. (1)

For instance a protein backbone with a single chain of amino-acid residues.

We use a Gaussian model to estimate the molecule density

$$\mathcal{Z} := (z_1, z_2, \ldots, z_m) \in \mathbb{R}^{3m} \longmapsto U(\mathcal{Z}) := \sum_{i=1}^{m} \gamma_i G(z_i, \sigma_i) \in L^2(\mathbb{R}^3).$$



where m is the number of atoms in the chain.

Structural assumption

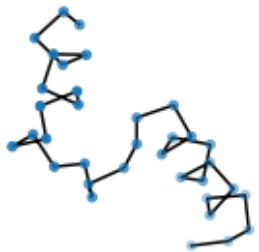
Let us assume that the macromolecule of interest can be modelled as a chain, or discrete curve.

The distance between consecutive (pseudo)atoms is constant. (1)

For instance a protein backbone with a single chain of amino-acid residues.

We use a Gaussian model to estimate the molecule density

$$\mathcal{Z} := (z_1, z_2, \ldots, z_m) \in \mathbb{R}^{3m} \longmapsto U(\mathcal{Z}) := \sum_{i=1}^{m} \gamma_i G(z_i, \sigma_i) \in L^2(\mathbb{R}^3).$$



where m is the number of atoms in the chain.

Problem:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{N} \| T(U(\mathcal{Z}_i)) - Y_i \|_{L^2} \\ &\text{s.t. } \mathcal{Z}_i \text{ satisfies (1) } \forall i. \end{aligned}$$

Structural assumption

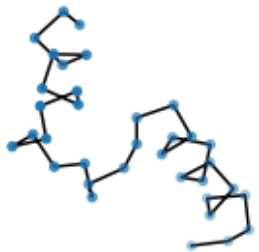
Let us assume that the macromolecule of interest can be modelled as a chain, or discrete curve.

The distance between consecutive (pseudo)atoms is constant. (1)

For instance a protein backbone with a single chain of amino-acid residues.

We use a Gaussian model to estimate the molecule density

$$\mathcal{Z} := (z_1, z_2, \ldots, z_m) \in \mathbb{R}^{3m} \longmapsto U(\mathcal{Z}) := \sum_{i=1}^{m} \gamma_i G(z_i, \sigma_i) \in L^2(\mathbb{R}^3).$$



where m is the number of atoms in the chain.

This doesn't work

- Too many parameters to estimate: $3mN$.
- The images Y_i are too noisy.
- We are not using the relation between the different structures.

Relation between structures

Let us assume that we have access to the atomic model of a conformation of the molecule $\mathcal{Z}_0 \in \mathbb{R}^{3m}$.

Any other structure $\mathcal{Z}_i$ in the dataset can be obtained from $\mathcal{Z}_0$ as

$$\mathcal{Z}_i = (R_i \circ D_i)\mathcal{Z}_0$$

where

- R_i is a rigid transformation, that we know from the pose estimation.
- D_i is a deformation of the structure, which preserves property (1), and we don't know.

For any point cloud $\mathcal{Z} = (z_1, \ldots, z_m) \in \mathbb{R}^{3m}$ satisfying the discrete curve condition (1) there exists $\delta > 0$ and

$$(\widehat{z}, T_1, T_2, \ldots, T_{m-1}) \in \mathbb{R}^3 \times (\mathbb{S}^2)^{m-1}$$

such that

$$\begin{cases} z_{j+1} = z_j + \delta T_j & j \in \{1, \ldots, m-1\} \\ \text{with } z_{j_0} = \widehat{z} \in \mathbb{R}^3 \end{cases}$$

- Any point cloud $\mathcal{Z} = (z_1, \ldots, z_m) \in \mathbb{R}^{3m}$ can be represented by $(\widehat{z}, T_1, T_2, \ldots, T_{m-1})$.
- Any translation can be written as

$$(\widehat{z}, T_1, T_2, \ldots, T_{m-1}) \longrightarrow (\widehat{z} + \Delta z, T_1, T_2, \ldots, T_{m-1})$$

- **What about rotations and deformations?**

For any point cloud $\mathcal{Z} = (z_1, \ldots, z_m) \in \mathbb{R}^{3m}$ satisfying the discrete curve condition (1) there exists $\delta > 0$ and

$$(\widehat{z}, T_1, T_2, \ldots, T_{m-1}) \in \mathbb{R}^3 \times (\mathbb{S}^2)^{m-1}$$

such that

$$\begin{cases} z_{j+1} = z_j + \delta T_j & j \in \{1, \ldots, m-1\} \\ \text{with } z_{j_0} = \widehat{z} \in \mathbb{R}^3 \end{cases}$$

- Any point cloud $\mathcal{Z} = (z_1, \ldots, z_m) \in \mathbb{R}^{3m}$ can be represented by $(\widehat{z}, T_1, T_2, \ldots, T_{m-1})$.
- Any translation can be written as

$$(\widehat{z}, T_1, T_2, \ldots, T_{m-1}) \longrightarrow (\widehat{z} + \Delta z, T_1, T_2, \ldots, T_{m-1})$$

- **What about rotations and deformations?**

Let $x(\cdot) \in C^2([0, L]; \mathbb{R}^3)$ be a smooth curve in $\mathbb{R}^3$ parametrized by its arclength.

Then $x(\cdot)$ solves

$$\begin{cases} x'(s) = T(s) & s \in [0, L] \\ F'(s) = r(s)F(s) & s \in [0, L] \\ \text{with } x(s_0) = \widehat{z} \in \mathbb{R}^3 \text{ and } F(s_0) = \widehat{F} \in SO(3). \end{cases}$$

where

$$F(s) = \begin{bmatrix} T(s) \\ N(s) \\ B(s) \end{bmatrix} \in SO(3) \qquad \widehat{F} = \begin{bmatrix} x'(s_0) \\ x''(s_0)/\|x''(s_0)\| \\ T(s_0) \times N(s_0) \end{bmatrix}$$

and

$$r(s) = \begin{bmatrix} 0 & \kappa(s) & 0 \\ -\kappa(s) & 0 & \tau(s) \\ 0 & -\tau(s) & 0 \end{bmatrix} \in \mathfrak{so}(3) \qquad \text{for some } (\kappa(\cdot), \tau(\cdot)) : [0, L] \to \mathbb{R}^2.$$

Frenet-Serret formulas

Let $x(\cdot) \in C^2([0, L]; \mathbb{R}^3)$ be a smooth curve in $\mathbb{R}^3$ parametrized by its arclength.

Then $x(\cdot)$ can be represented by the initial condition

$$(\widehat{z}, \widehat{F}) \in \mathbb{R}^3 \times SO(3)$$

and the curvature and the torsion

$$\kappa : [0, L] \to \mathbb{R} \qquad \text{and} \qquad \tau : [0, L] \to \mathbb{R}.$$

- Rigid transformations of $x(\cdot)$ can be obtained by

$$(\widehat{z}, \widehat{F}) \longmapsto (\widehat{z} + t, R\widehat{F}) \qquad \text{for some } t \in \mathbb{R}^3 \text{ and } R \in SO(3)$$

- Any length-preserving deformation of $x(\cdot)$ of the curve can be obtained by

$$(\kappa(\cdot), \tau(\cdot)) \longmapsto (\kappa(\cdot) + \Delta\kappa(\cdot), \tau(\cdot) + \Delta\tau(\cdot))$$
$$\text{for some } (\Delta\kappa(\cdot), \Delta\tau(\cdot)) : [0, L] \to \mathbb{R}^2.$$

Frenet-Serret formulas

Let $x(\cdot) \in C^2([0, L]; \mathbb{R}^3)$ be a smooth curve in $\mathbb{R}^3$ parametrized by its arclength.

Then $x(\cdot)$ can be represented by the initial condition

$$(\widehat{z}, \widehat{F}) \in \mathbb{R}^3 \times SO(3)$$

and the curvature and the torsion

$$\kappa : [0, L] \to \mathbb{R} \qquad \text{and} \qquad \tau : [0, L] \to \mathbb{R}.$$

- Rigid transformations of $x(\cdot)$ can be obtained by

$$(\widehat{z}, \widehat{F}) \longmapsto (\widehat{z} + t, R\widehat{F}) \qquad \text{for some } t \in \mathbb{R}^3 \text{ and } R \in SO(3)$$

- Any length-preserving deformation of $x(\cdot)$ of the curve can be obtained by

$$(\kappa(\cdot), \tau(\cdot)) \longmapsto (\kappa(\cdot) + \Delta\kappa(\cdot), \tau(\cdot) + \Delta\tau(\cdot))$$
$$\text{for some } (\Delta\kappa(\cdot), \Delta\tau(\cdot)) : [0, L] \to \mathbb{R}^2.$$

Let $\mathcal{Z} = (z_1, \ldots, z_m) \in \mathbb{R}^{3m}$ be a point cloud satisfying

$$\|z_{i+1} - z_i\| = \delta > 0 \qquad \forall i \in \{1, \ldots, m\},$$

then

$$\begin{cases} z_{j+1} = z_j + \delta T_j & j \in \{1, \ldots, m-1\} \\ F_{j+1} = R_j F_j & j \in \{1, \ldots, m-2\} \\ \text{with } z_{j_0} = \widehat{z} \in \mathbb{R}^3 \text{ and } F_{j_0} = \widehat{F} \in SO(3). \end{cases}$$

where

$$F_j = \begin{bmatrix} T_j \\ N_j \\ B_j \end{bmatrix} \in SO(3)$$

with

$$T_j = \frac{z_{j+1} - z_j}{\delta} \quad B_j = \frac{T_j \times T_{j+1}}{\|T_j \times T_{j+1}\|} \quad N_j = \frac{B_j \times T_j}{\|B_j \times T_j\|}.$$

Any $\mathcal{Z} = (z_1, \ldots, z_m) \in \mathbb{R}^{3m}$ satisfying (1) can be written as

$$\begin{cases} z_{j+1} = z_j + \delta e_3 F_j & j \in \{1, \ldots, m-1\} \\ F_{j+1} = R_j F_j & j \in \{1, \ldots, m-2\} \\ \text{with } z_{j_0} = \widehat{z} \in \mathbb{R}^3 \text{ and } F_{j_0} = \widehat{F} \in SO(3). \end{cases}$$

for some $(\widehat{z}, \widehat{F}) \in \mathbb{R}^3 \times SO(3)$ and sequence of rotation matrices $\{R_j\}_{j=1}^{m-2} \in SO(3)^{m-2}$, that we parametrize by using the Euler angles

$$R_j = R(\theta_j, \psi_j) := \begin{bmatrix} \cos\psi_j \cos\theta_j & \cos\psi_j \sin\theta_j & -\sin\psi_j \\ -\sin\theta_j & \cos\theta_j & 0 \\ \sin\psi_j \cos\theta_j & \sin\psi_j \sin\theta_j & \cos\psi_j \end{bmatrix}$$

Parameter space

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

- $(\widehat{z}, \widehat{F}) \in \mathbb{R}^3 \times SO(3)$ determines the pose of the particle;
- $(\Theta, \Psi) \in [-\pi, \pi]^{m-2} \times [-\pi, \pi]^{m-2}$ determines the conformation.

Spectral decomposition

Goal: Construct a map from the manifold of conformations to the dihedral angles.

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

Approach:

$$\Theta(m) := \Theta_0 + \sum_{k=0}^{K-1} \mathbf{a}_k \phi_k(m) \quad \text{and} \quad \Psi(m) := \Psi_0 + \sum_{k=0}^{K-1} \mathbf{b}_k \phi_k(m), \quad \text{for } m \in \mathcal{M},$$

where

- $\phi_0(\cdot), \phi_1(\cdot), \phi_3(\cdot), \ldots$ are the first eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$.
- The vectors $\Theta_0 \in [-\pi, \pi]^{m-2}$ and $\Psi_0 \in [-\pi, \pi]^{m-2}$ are the rotation angles of the given known conformation.

Goal: Construct a map from the manifold of conformations to the dihedral angles.

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

Approach:

$$\Theta(m) := \Theta_0 + \sum_{k=0}^{K-1} \mathbf{a}_k \phi_k(m) \quad \text{and} \quad \Psi(m) := \Psi_0 + \sum_{k=0}^{K-1} \mathbf{b}_k \phi_k(m), \quad \text{for } m \in \mathcal{M},$$

where

- $\phi_0(\cdot), \phi_1(\cdot), \phi_3(\cdot), \ldots$ are the first eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$.
- The vectors $\Theta_0 \in [-\pi, \pi]^{m-2}$ and $\Psi_0 \in [-\pi, \pi]^{m-2}$ are the rotation angles of the given known conformation.

Under the assumption that $\mathcal{M}$ is a compact connected manifold, $\{\phi_k(\cdot)\}_{k=0}^{\infty}$ form an orthonormal basis of $L^2(\mathcal{M})$.

Spectral decomposition

Goal: Construct a map from the manifold of conformations to the dihedral angles.

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

Approach:

$$\Theta(m) := \Theta_0 + \sum_{k=0}^{K-1} \mathbf{a}_k \phi_k(m) \quad \text{and} \quad \Psi(m) := \Psi_0 + \sum_{k=0}^{K-1} \mathbf{b}_k \phi_k(m), \quad \text{for } m \in \mathcal{M},$$

where

- $\phi_0(\cdot), \phi_1(\cdot), \phi_3(\cdot), \ldots$ are the first eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$.
- The vectors $\Theta_0 \in [-\pi, \pi]^{m-2}$ and $\Psi_0 \in [-\pi, \pi]^{m-2}$ are the rotation angles of the given known conformation.

Using the eigenfunctions associated to the smallest eigenvalues allows us to only capture low-frequency deformations.

Goal: Construct a map from the manifold of conformations to the dihedral angles.

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

Approach:

$$\Theta(m) := \Theta_0 + \sum_{k=0}^{K-1} \mathbf{a}_k \phi_k(m) \quad \text{and} \quad \Psi(m) := \Psi_0 + \sum_{k=0}^{K-1} \mathbf{b}_k \phi_k(m), \quad \text{for } m \in \mathcal{M},$$

where

- $\phi_0(\cdot), \phi_1(\cdot), \phi_3(\cdot), \ldots$ are the first eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$.
- The vectors $\Theta_0 \in [-\pi, \pi]^{m-2}$ and $\Psi_0 \in [-\pi, \pi]^{m-2}$ are the rotation angles of the given known conformation.

Prior knowledge about the rigidity of certain parts of the macromolecule (secondary structures) may be used to set some of the coefficients $(\mathbf{a}_k, \mathbf{b}_k)$ equal to zero.

We need to estimate two elements:

- The eigenfunctions $\phi_k(\cdot)$ of the Laplace-Beltrami operator on $\mathcal{M}$.
- The coefficients $\mathbf{a}_k \in \mathbb{R}^{m-2}$ and $\mathbf{b}_k \in \mathbb{R}^{m-2}$ for all $k = 0, 1, \ldots, K-1$.

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

We use a known technique in manifold learning, used in Moscovich et al. , Inverse problems, 2020.

We use the low-dimension representation of the conformation in each particle

$$\{\beta_i\}_{i=1}^{n} \subset \mathbb{R}^q, \qquad \text{where } n \text{ is the number of cryo-EM images.}$$

to construct a symmetric weighted graph with n nodes and weights given by

$$w_{ij} := \gamma \exp\left(-\frac{\|\beta_i - \beta_j\|^2}{2\sigma^2}\right) \qquad \text{for all } (i,j) \in \{1,2,\ldots,n\}^2.$$

We need to estimate two elements:

- The eigenfunctions $\phi_k(\cdot)$ of the Laplace-Beltrami operator on $\mathcal{M}$.
- The coefficients $\mathbf{a}_k \in \mathbb{R}^{m-2}$ and $\mathbf{b}_k \in \mathbb{R}^{m-2}$ for all $k = 0, 1, \ldots, K-1$.

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

We use a known technique in manifold learning, used in Moscovich et al. , Inverse problems, 2020.

We use the low-dimension representation of the conformation in each particle

$$\{\beta_i\}_{i=1}^n \subset \mathbb{R}^q, \qquad \text{where } n \text{ is the number of cryo-EM images.}$$

to construct a symmetric weighted graph with n nodes and weights given by

$$w_{ij} := \gamma \exp\left(-\frac{\|\beta_i - \beta_j\|^2}{2\sigma^2}\right) \qquad \text{for all } (i,j) \in \{1,2,\ldots,n\}^2.$$

Approximated spectral decomposition

Once we have the similarity matrix $W \in (\mathbb{R}^+)^{n \times n}$ we construct the associated normalized graph Laplacian

$$L := D^{-1/2}(D - W)D^{-1/2}$$

where D is the $n \times n$ diagonal matrix given by $D_{ii} = \sum_{j=1}^{n} W_{ij}$.

Let $\phi^{(0)}, \phi^{(1)}, \ldots \in \mathbb{R}^n$ be the ordered eigenvectors of the Laplacian matrix L.

Known result about the graph Laplacian

Let $\mathcal{M} \subset \mathbb{R}^q$ be a connected compact manifold and for any $n \in \mathbb{N}$, let $\{\beta_i\}_{i=1}^n$ be a sampling of a uniformly distributed random variable on $\mathcal{M}$.
Then, for any $k \geq 0$, the eigenvector $\phi^{(k)}$ converges in probability to the k-th eigenfunction $\phi_k(\cdot)$ of a linear differential operator in $\mathcal{M}$, i.e.

$$\sup_{i=1,\ldots,n} |\sqrt{n}\phi_i^{(k)} - \phi_k(\beta_i)| \to 0 \qquad \text{almost surely as } n \to \infty.$$

(see (von Luxburg, Belkin, Bousquet, 2004) and (Belkin, Niyogi, 2008)

Approximated spectral decomposition

Once we have the similarity matrix $W \in (\mathbb{R}^+)^{n \times n}$ we construct the associated normalized graph Laplacian

$$L := D^{-1/2}(D - W)D^{-1/2}$$

where D is the $n \times n$ diagonal matrix given by $D_{ii} = \sum_{j=1}^{n} W_{ij}$.

Let $\phi^{(0)}, \phi^{(1)}, \ldots \in \mathbb{R}^n$ be the ordered eigenvectors of the Laplacian matrix L.

Approximated spectral decomposition

$$\Theta_i(A) = \Theta_0 + \sum_{k=0}^{K-1} \mathbf{a}_k \phi_i^{(k)} = \Theta_0 + A\Phi_i \quad \text{and} \quad \Psi_i(B) = \Psi_0 + \sum_{k=0}^{K-1} \mathbf{b}_k \phi_i^{(k)} = \Psi_0 + B\Phi_i$$

where

$$A = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{K-1}] \in \mathbb{R}^{(m-2)\times K} \quad \text{and} \quad B = [\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_{K-1}] \in \mathbb{R}^{(m-2)\times K}$$

and the vectors $\Phi_i \in \mathbb{R}^K$ for $i = 1, 2, \ldots, n$ are given by

$$\Phi_i = (\phi_i^{(0)}, \phi_i^1, \ldots, \phi_i^{(K-1)}) \in \mathbb{R}^K$$

We need to estimate two elements:

- The eigenfunctions $\phi_k(\cdot)$ of the Laplace-Beltrami operator on $\mathcal{M}$.
- The coefficients $\mathbf{a}_k \in \mathbb{R}^{m-2}$ and $\mathbf{b}_k \in \mathbb{R}^{m-2}$ for all $k = 0, 1, \ldots, K-1$.

We formulate a minimisation problem in which we compare the atomic model predicted for each particle with the cryo-EM images.

Recall that the atomic model can be represented by the parameters

$$(\Theta, \Psi, \widehat{z}, \widehat{F}) \in [-\pi, \pi]^{m-2} \times [-\pi, \pi]^{m-2} \times \mathbb{R}^3 \times SO(3)$$

- $\{(\widehat{z}_i, \widehat{F}_i)\}_{i=1}^n \in (\mathbb{R}^3 \times SO(3))^n$ can be obtained from the pose estimation of the particles.
- $\Theta_i \in [-\pi, \pi]^{m-2}$ and $\Psi_i \in [-\pi, \pi]^{m-2}$ are estimated as

$$\Theta_i(A) = \Theta_0 + A\Phi_i \quad \text{and} \quad \Psi_i(B) = \Psi_0 + B\Phi_i$$

with $A \in \mathbb{R}^{(m-2)\times K}$ and $B \in \mathbb{R}^{(m-2)\times K}$ are matrices to be estimated.

We need to estimate $2(m-2)K$ parameters.

For every $A \in \mathbb{R}^{(m-2)\times K}$ and $B \in \mathbb{R}^{(m-2)\times K}$ we define

$$\Gamma_i(A, B) := \mathcal{Z}(\Theta_i(A), \Psi_i(B), \widehat{z}_i, \widehat{F}_i) = [z_1, z_2, \ldots, z_m] \in \mathbb{R}^{3m}$$

where $\mathcal{Z}(\Theta, \Psi, \widehat{z}, \widehat{F})$ is the solution to the discrete dynamical system introduced above.

For any $i = 1, 2, \ldots, n$, the estimated density of the i-th particle is given by

$$\widehat{U}_i(A, B) = \sum_{z_j \in \Gamma_i(A,B)} \gamma_j G(z_j, \sigma_j).$$

Finally, we project apply the forward operator to the estimated densities

$$\widehat{Y}_i(A, B) := CTF \circ T(\widehat{U}_i(A, B)$$

The matrices of coefficients A and B are estimated by means of the following minimisation problem:

$$[\widehat{A}, \widehat{B}] = \underset{A,B}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \| \widehat{Y}_i(A, B) - Y_i \|_2^2.$$

where $\widehat{Y}_i(A, B) = \mathcal{F}[\widehat{U}_i(A, B)]$ is the cryo-EM forward operator applied to the estimated electrostatic potential.

For each particle $i \in \{1, \ldots, n\}$, the parameters $(\Theta_i, \Psi_i) \in [-\pi, \pi]^{2(m-2)}$ determining the conformation are then estimated as

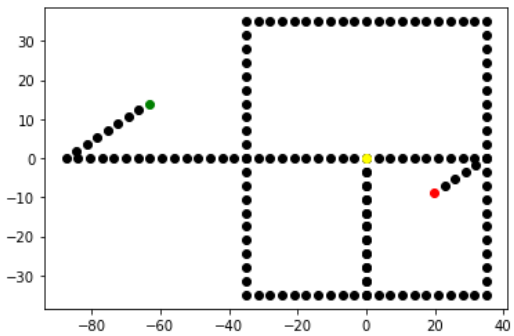
$$\Theta_i(\widehat{A}) = \Theta_0 + \widehat{A}\Phi_i \qquad \text{and} \qquad \Psi_i(\widehat{B}) = \Psi_0 + \widehat{B}\Phi_i.$$

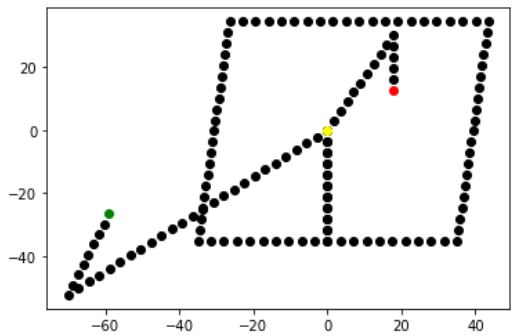
- We use SGD to approximate a solution.
- We initialise the parameters A and B by setting them equal to zero, so that the prediction of the atomic model is the same for all the particles (the given known conformation).

Numerical experiments

We consider a two-dimensional structure consisting of a discrete curve with $m = 149$ points and inter-atomic distance $\delta = 4$.

$$\begin{cases} z_{j+1} = z_j + \delta e_2 F_j & j \in \{1, \dots, m-1\} \\ F_{j+1} = R(\theta_j) F_j & j \in \{1, \dots, m-2\} \\ \text{with } z_{j_0} = \widehat{z} \in \mathbb{R}^2 \text{ and } F_{j_0} = \widehat{F} \in SO(2). \end{cases}$$

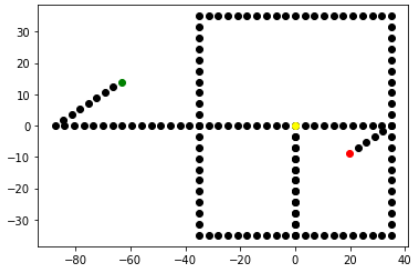




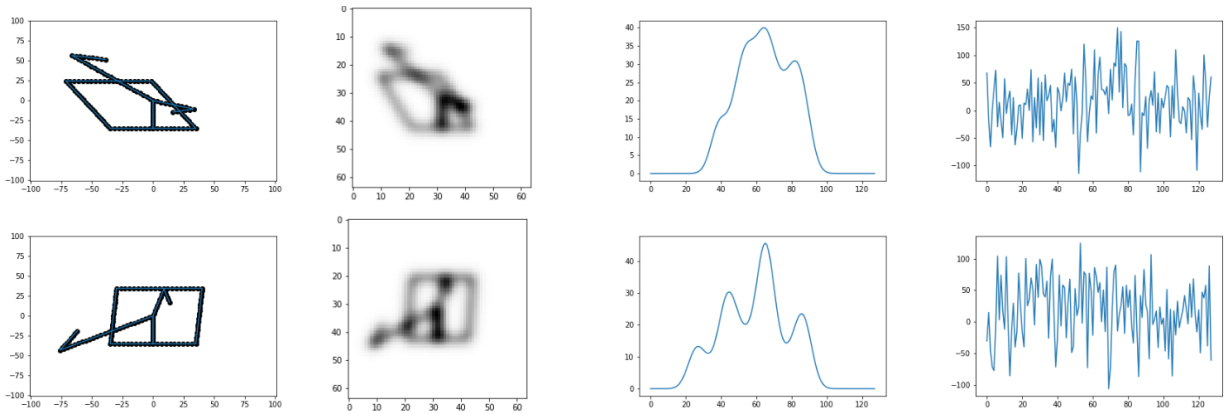
The structure consists of a flexible box and two moving arms.

Given information:

- A known conformation



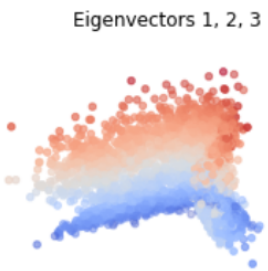
- A cryo-EM dataset with 1D noisy tomographic projections: 4000 images



$$SNR = \frac{182.63}{2500} = 0.073.$$

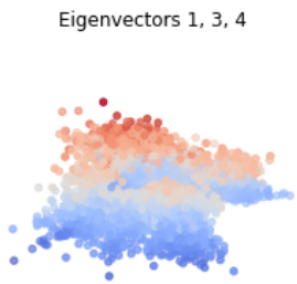
In order to construct the graph Laplacian we use the low-resolution (64×64) 2D images of the dataset.

We can therefore construct the graph Laplacian and compute the first eigenvectors $\phi^{(0)}, \phi^{(1)}, \ldots, \phi^{(K-1)} \in \mathbb{R}^{4000}$.

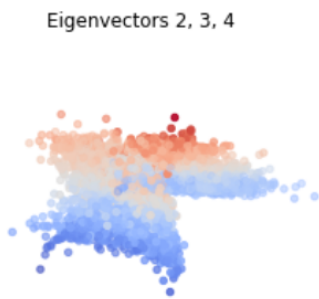


(A) $[\phi^{(1)}, \phi^{(2)}, \phi^{(3)}]$

(B) $[\phi^{(1)}, \phi^{(2)}, \phi^{(4)}]$



(C) $[\phi^{(1)}, \phi^{(3)}, \phi^{(4)}]$



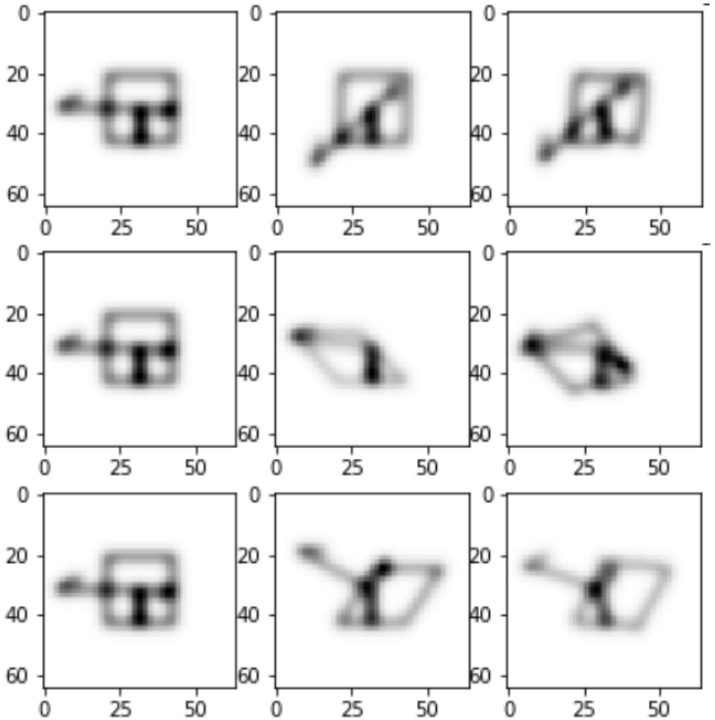
(D) $[\phi^{(2)}, \phi^{(3)}, \phi^{(4)}]$

We now need to compute the coefficients in the matrix A

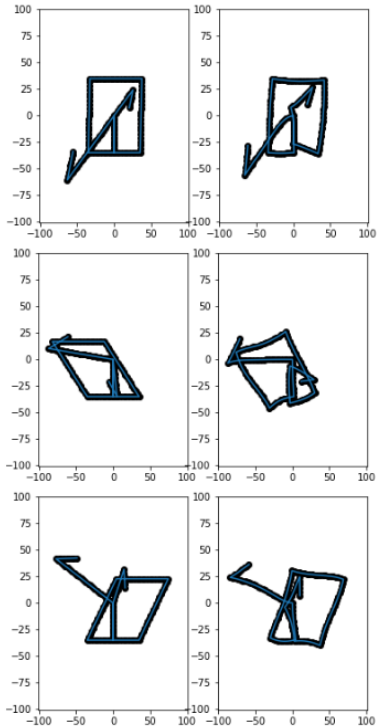
$$\Theta_i(\widehat{A}) = \Theta_0 + \widehat{A}\Phi_i$$

here Θ_0 are the rotation angles of the known conformation and Φ_i are i-th component of the first K eigenvectors.

- We use SGD to estimate the matrices of parameter $[A, B] \in \left[\mathbb{R}^{(m-2)\times K}\right]^2$, initializing with $[A, B] = [0, 0]$.
- In this case we do estimate all the angles θ_j in the structure, even those which are constant over the conformations.

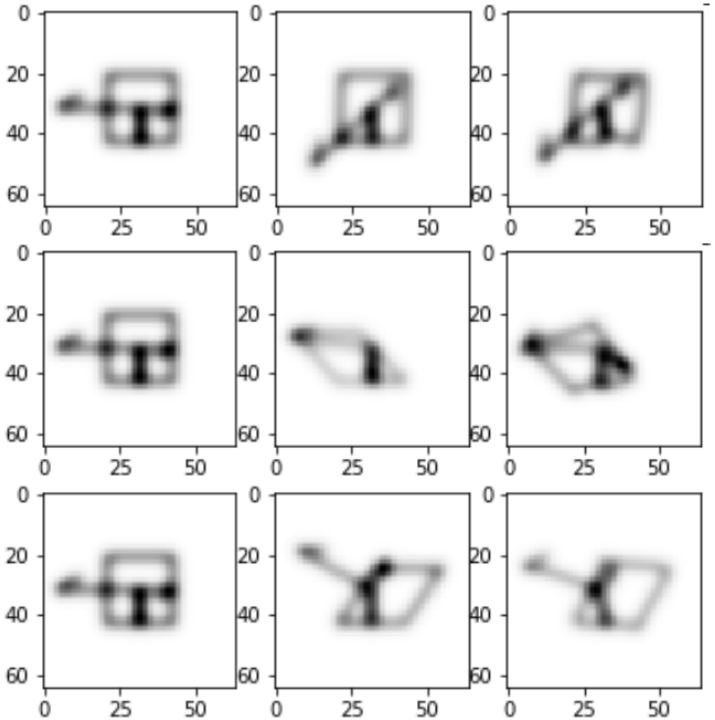


(A) Reconstruction of the 2D images using our method. At the left we see the structure in the given conformation, in the middle the ground true and at the right the reconstruction.

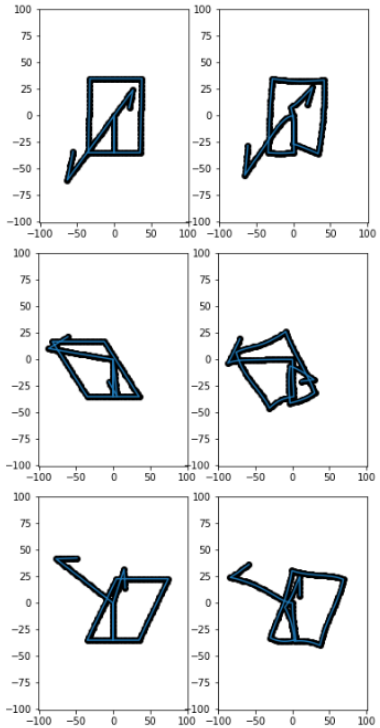


(B) At the left the 2D structure (ground truth) and at the right the structure predicted by means of our method.

- We use SGD to estimate the matrices of parameter $[A, B] \in \left[\mathbb{R}^{(m-2)\times K}\right]^2$, initializing with $[A, B] = [0, 0]$.
- In this case we do estimate all the angles θ_j in the structure, even those which are constant over the conformations.



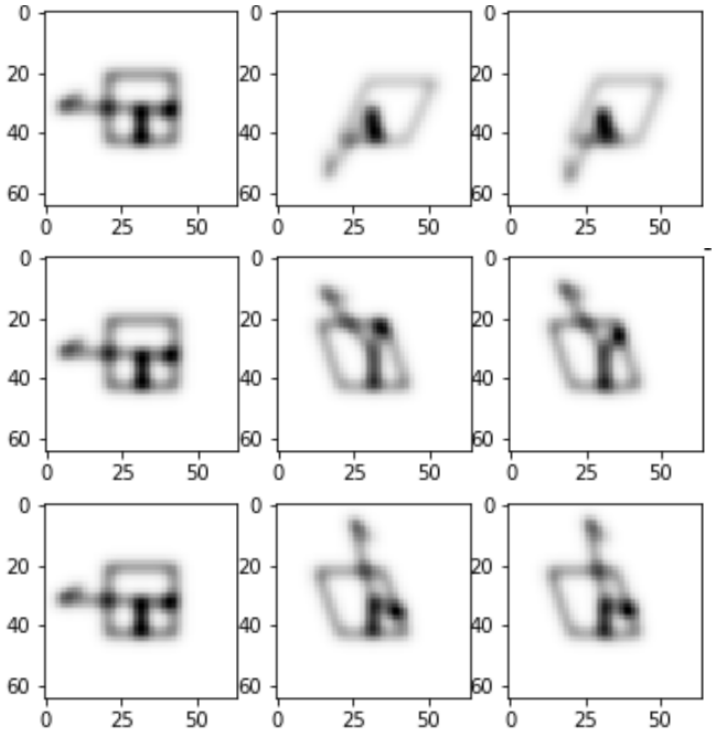
(A) Reconstruction of the 2D images using our method. At the left we see the structure in the given conformation, in the middle the ground true and at the right the reconstruction.



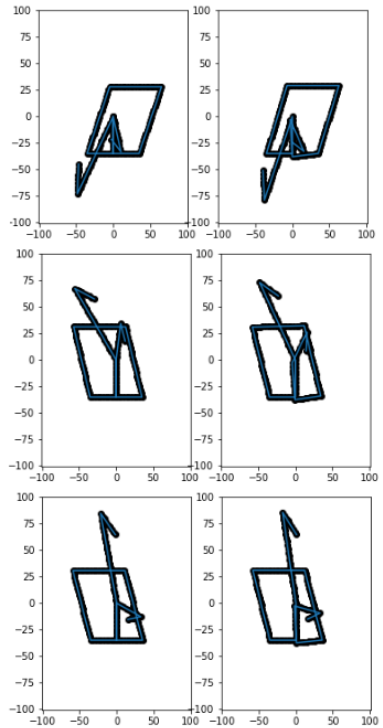
(B) At the left the 2D structure (ground truth) and at the right the structure predicted by means of our method.

Using the knowledge about the length of the arms and the sides of the box:

In this case we only need to estimate the angles θ_j which are not constant in all the conformations. The rows in A corresponding to the other angles are set to 0.



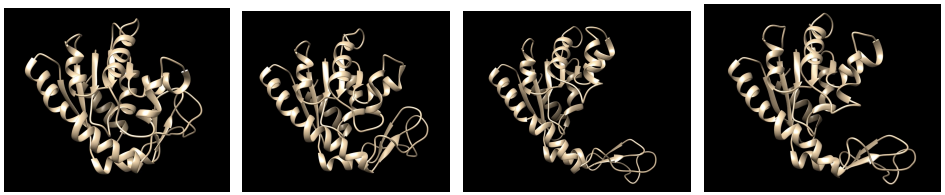
(A) Reconstruction of the 2D images using our method. At the left we see the structure in the given conformation, in the middle the ground true and at the right the reconstruction.



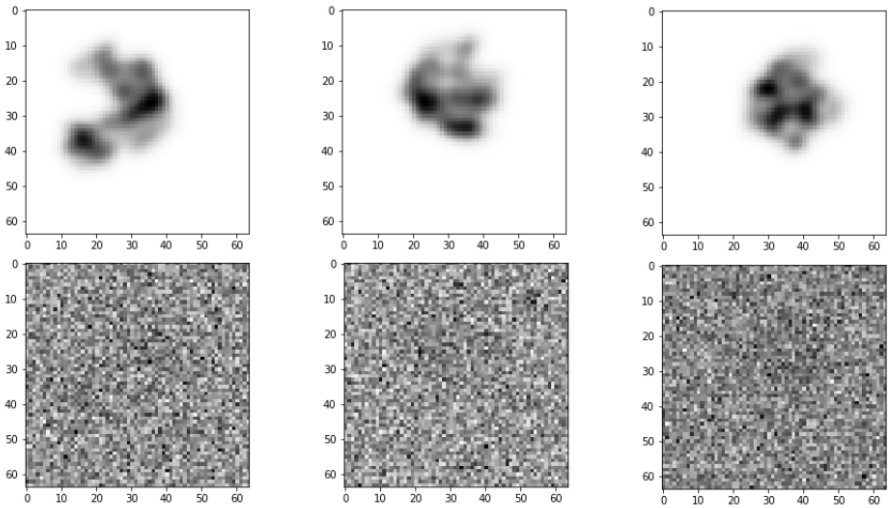
(B) At the left the 2D structure (ground truth) and at the right the structure predicted by means of our method.

3D structure

We consider a protein backbone with 214 C-α atoms. We use an MD trajectory of the adenylate kinase with 102 frames.



We generated a dataset with 4000 particles randomly selected from the 104 frames, each one rotated by an element of $SO(3)$ randomly selected.

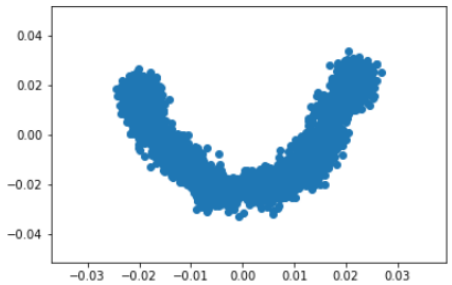


The cryo-EM contains 4000 noisy tomographic projections of the particles.

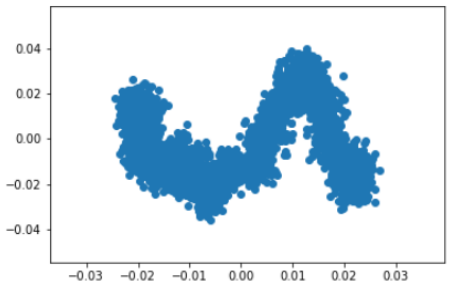
$$SNR \approx 0.01$$

In order to construct the graph Laplacian we use the low-resolution $(16 \times 16 \times 16)$ 3D volumes of the particles in the dataset.

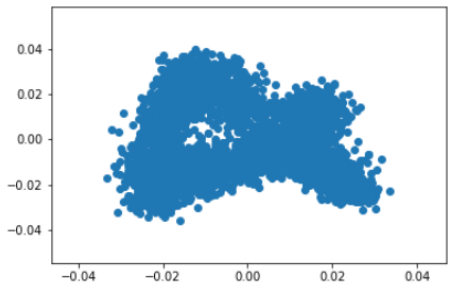
We can therefore construct the graph Laplacian and compute the first eigenvectors $\phi^{(0)}, \phi^{(1)}, \ldots, \phi^{(K-1)} \in \mathbb{R}^{4000}$.



(A) $[\phi^{(1)}, \phi^{(2)}]$



(B) $[\phi^{(1)}, \phi^{(3)}]$

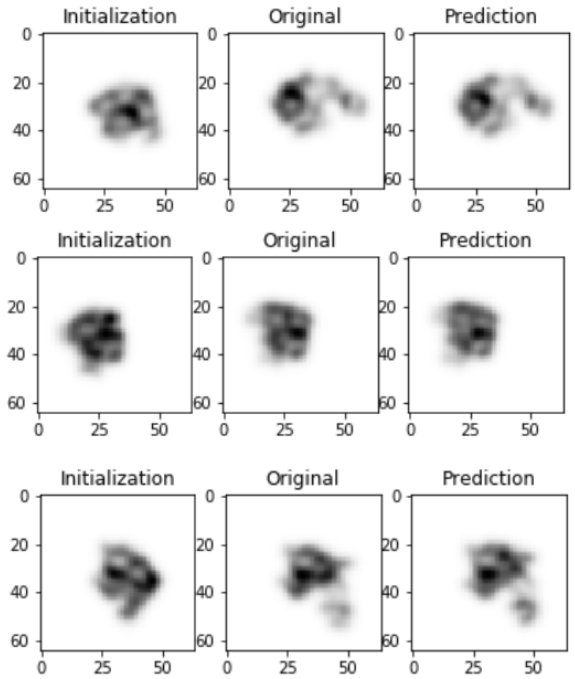


(C) $[\phi^{(2)}, \phi^{(3)}]$

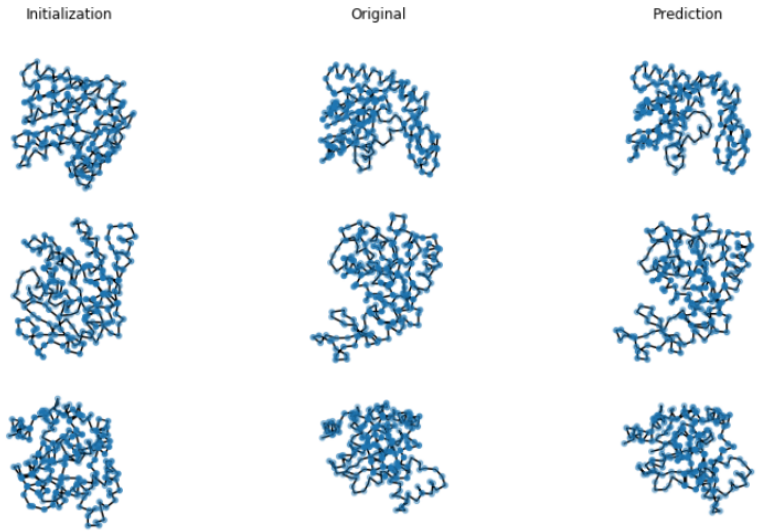
We now need to compute the coefficients in the matrices A and B

$$\Theta_i(\widehat{A}) = \Theta_0 + \widehat{A}\Phi_i \quad \text{and} \quad \Psi_i(\widehat{B}) = \Psi_0 + \widehat{B}\Phi_i$$

here Θ_0 and Ψ_0 are the rotation angles of the known conformation (first conformation in the molecular trajectory) and the vectors Φ_i are formed by the i-th component of the first K eigenvectors.



(A) Clean 2D tomographic projections of the reconstructed 3D structures. At the left we see the structure in the given conformation, in the middle the ground true and at the right the reconstruction.



(B) At the left we see the structure in the known conformation (after a random rotation), in the middle we see the particle in its specific conformation (ground truth) and at the right the 3D atomic structure predicted by means of our method.

Conclusions:

- We propose a strategy to combine a low-dimension representation of the conformations and prior knowledge about the structure to recover the estimate the deformations of a given atomic structure.
- In our approach we can exploit knowledge about secondary structures to reduce the number of parameters to be estimated.
- Numerical experiments show that the method works in toy examples.

Problems and future perspectives:

- Further develop the method to be applicable to more realistic scenarios.
- The conformation may affect the pose estimation.
- Study the limitations of the method, since applying SGD over $SO(3)$ may converge to local minima.
- Adapt the method to more complex structures (no necessarily a single chain).

Thanks for the attention!

and thanks to my collaborators:

Willem Diepeveen, Ozan Öktem and Carola-Bibiane Schönleib.