Meaningful Dimensionality Reduction of the Conformational Space

Axel Levy

TL;DR

I suggest to deform the conformational space such that the euclidian distances become meaningful in the deformed space. Computing the distance between two volumes is expensive so I suggest making use of amortization to decrease its evaluation cost.

What is the Problem?

CryoDRGN maps cryo-EM images to an abstract low-dimensional latent space (the *conformational space*). The problem is that distances in this space do not have a valuable meaning. In particular, points that are far apart in this space can correspond to very similar structures (up to a global rotation). This means that moving in the conformational space does not necessarily mean deforming the volume. Movements in the conformational space can represent pure rotations or no deformation at all. However, we want to be able to interpret distances in the conformational space: we want spread out points to represent continuous deformations and well separated clusters to represent discrete conformational changes.

Current Approach

Currently, the only way to interpret the landscape is to sample it, generate the associated volumes and analyze them by eye.

Suggested Solution

Importantly, this problem does not affect the quality of the reconstruction. The problem only concerns the *interpretability* of the conformational landscape, post-reconstruction. I therefore suggest a post-processing solution to address this interpretability issue. The idea is to (1) *learn* a meaningful distance in the latent space and (2) deform this latent space such that euclidian distances can be interpreted in a meaningful way.

Methods

Distance in Latent Space

The *distance* between two volumes is defined be the following function:

$$\delta : (V_1, V_2) \mapsto \min_{\phi \in \mathrm{SO}(3)} \|V_1 - \phi \cdot V_2\|_2. \tag{1}$$

V_1 and V_2 are voxel grids (elements of $\mathbb{R}^{n \times n \times n}$). $\phi \cdot V$ represents a rotated volume. It is crucial to factor out rotations as the conformational space can represent pure rotations (*entanglement problem*).

From this, we can define a distance in the latent space. Let's call V the function that transforms a latent z into a voxel grid,

$$V : z \in \mathbb{R}^d \mapsto \{\Gamma(x_k, z)\}_k \in \mathbb{R}^{n \times n \times n}, \tag{2}$$

where Γ is the hypervolume obtained with cryoDRGN. The distance in the latent space is then defined by

$$\Delta = \delta(V(.), V(.)) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}. \tag{3}$$

Objective Function

Ideally, we would like to have something like

$$\forall z_1, z_2 \in \mathbb{R}^d, \quad \|z_1 - z_2\|_2 \approx \Delta(z_1, z_2). \tag{4}$$

Each predicted latent z_i will be mapped to a *transformed* predicted latent $z'_i \in \mathbb{R}^{d'}$. The objective function we will aim at minimizing is

$$\mathcal{L}(\{z'_i\}) = \sum_{i,j=1}^{N} \left(\frac{\|z'_i - z'_j\|_2}{\Delta(z_i, z_j)} - 1 \right)^2. \tag{5}$$

over $\{z'_i\}$.

With a problem formulated like this, two challenges remain:

1. The number of unknowns (z'_i) scales linearly with the number of images.
2. The function Δ is (very) costly to evaluate, and needs to be evaluated a number of times that grows quadratically with the number of images.

Our Old Friend, Amortization

Amortization on the Number of Unknowns

Instead of looking for the set $\{z'_i\}$, we will look for a parameterized function

$$f_\xi : \mathbb{R}^d \to \mathbb{R}^{d'} \tag{6}$$

mapping predicted latents to transformed predicted latents $z' = f_\xi(z)$.

Amortization of Δ

Instead of evaluating Δ, we will learn an approximation from a small subset (smaller than $N \times N$). We first generate volumes from a subset of the predicted latents $\Omega = \{z_i\}_{i=\{1,\dots,L\}}$. Hopefully $L \ll N$ is enough. We compute $\Delta(z_1, z_2)$ for all pairs in Ω. Finally, we optimize a parameterized function h_ψ such that

$$\forall z_1, z_2 \in \Omega^2, \quad h_\psi(z_1, z_2) \approx \Delta(z_1, z_2). \tag{7}$$

Δ is a symmetric function, so this symmetry can be implicitly induced in h_ψ with the right parameterization.

Full Objective Function

Given a pretrained function h_ψ, the objective function we want to minimize is

$$\mathcal{L}(\xi) = \sum_{i,j=1}^{N} \left(\frac{\|f_\xi(z_i) - f_\xi(z_j)\|_2}{h_\psi(z_i, z_j)} - 1 \right)^2. \tag{8}$$

This sum contains $N \times N$ terms, which is too much for a cryo-EM dataset. However, it can be minimized with SGD. It is probably not necessary to use all the possible pairs to get a good idea of the function f_ξ.

Once we have f_ξ, we just generate and plot $\{z'_i\} = \{f_\xi(z_i)\}$ (possibly with a layer of PCA or UMAP on top).

Additional Ideas

- This can be seen as a dimension reduction method, similar to the Sammon mapping (`https://en.wikipedia.org/wiki/Sammon_mapping`), the high dimensional space being the space of voxel grids and the low dimensional one being the transformed latent space. The difference is that we have an intermediate space (the latent space) and we're looking for a function to deform that space, instead of directly optimizing the transformed predicted latents directly. Since this space is relatively low dimensional, f_ξ can probably generalize from a small number of training pairs.
- Instead of measuring the distance between voxel grids, we could measure the distance between pairwise atoms, if we fit an atomic model on top of generating the voxel grids. In some cases, the euclidian distance on potential is not the most indicative of chemical transformations (think about a tiny molecule that can bind to a big molecule in two places that are far apart). We could also use an optimal transport flow to define the distance on potentials.
- The method could be used to determine the intrinsic dimension of the movement. For each dimension d', we can measure the "volume" of the set of points $\{z'_i\}$. We gradually increase d'. When d' is higher than the intrinsic dimension of the movement, this "volume" should drop to 0. We need a way to properly measure volumes, but something like: do local PCAs and check that the last PCA component always has a small variance.
- If less step are needed to learn f_ξ than to learn h_ψ, the amortization of Δ might not be necessary.
- For large deviations, the euclidian distance on potential is not very meaningful. We could therefore replace (5) with

$$\mathcal{L}(\{z'_i\}) = \sum_{i,j=1}^{N} \left(\frac{1}{\Delta(z_i, z_j)} \left(\frac{\|z'_i - z'_j\|_2}{\Delta(z_i, z_j)} - 1 \right) \right)^2 . \tag{9}$$

- If $d = d'$, the parameterization of f_ξ could induce a bias towards the identity function (with normalization flows for example).