Learning Riemannian Metric Preserving Diffeomorphisms in Protein Dynamics

Friso de Kruiff[1,2] Willem Diepeveen[3] Erik Bekkers[4]
Ozan Öktem[2]

[1]Delft University of Technology [2]KTH Royal Institute of Technology
[3]Cambridge University [4]University of Amsterdam

The Problem of Euclidean Representation Learning

The **manifold hypothesis** states that data lie on a lower dimensional manifold, i.e. $\boldsymbol{x}_i \in \mathcal{M} \subset \mathbb{R}^n$ for $i = 1, \dots, N$ with $\dim(\mathcal{M}) = d << n$. The problem of representation learning is how to model the distribution of the latent space given the data,

$$p(\boldsymbol{z}_i|\boldsymbol{x}_i) \text{ for } \boldsymbol{x}_i \in \mathbb{R}^n \text{ and } \boldsymbol{z}_i \in \mathbb{R}^d \text{ for } d << n \text{ and } i = 1, \dots N. \tag{1}$$

Standard techniques such as Variational Auto-Encoders (VAEs) assume the data space and latent space to be Euclidean. However, interpreting the latent space as Euclidean in case of the manifold hypothesis leads to poor interpretability since the distance in the latent space does not correspond to distance in the data space. To accurately represent the data, representation learning should account for its non-linear geometry.

Riemannian Geometry-based Representation Learning

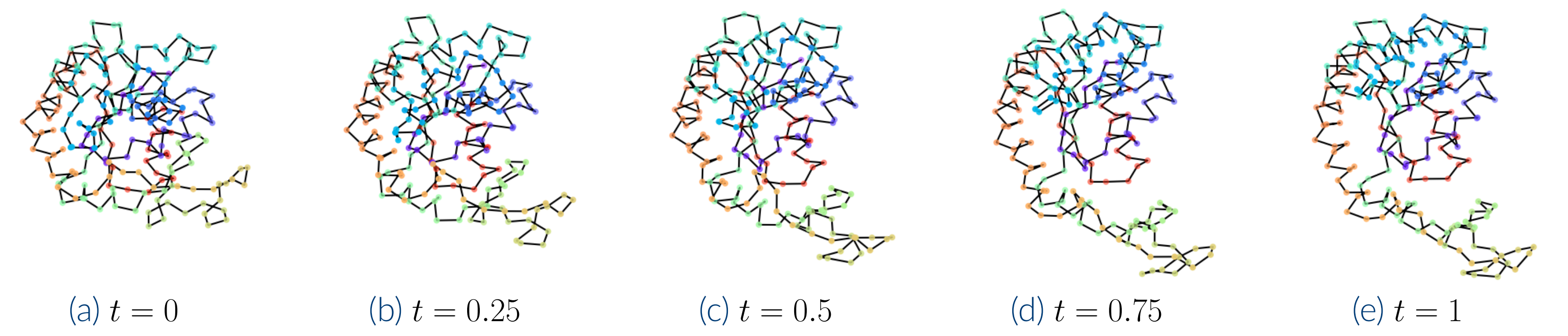
We argue that the framework of **Riemannian geometry** could be a suitable choice for representation learning under the *manifold hypothesis*. Here,

- **interpolation** can be performed over non-linear **geodesics** of the lower dimensional manifold,
- **extrapolation** can be done using the **Riemannian exponential mapping**,
- the **data mean** is naturally generalized to the **Riemannian barycentre**,
- and **low-rank approximation** finds the most important geodesics using the **Riemannian logarithmic mapping**.

Riemannian geometry and **diffeomorphic learning** we can achieve all four by learning a metric preserving diffeomorphism based on a local Euclidean distance approximation of the geodesic.

Data Analysis of an Adenylate Kinase Trajectory

An increasingly common assumption is that protein dynamics data lie on a lower dimensional data manifold $\mathcal{M}$. Here, we consider data of the time-normalized closed-to-open transition of adenylate kinase consisting of $N = 102$ conformations.



(a) $t = 0$ (b) $t = 0.25$ (c) $t = 0.5$ (d) $t = 0.75$ (e) $t = 1$

We approximate the geodesic distance of the data manifold through Isomap [1] and train a diffeomorphism to approximate a metric preserving Euclidean latent space. This leads to three key questions:

- How can we parameterize a metric preserving diffeomorphism?
- What should the objective be to optimize such a metric preserving diffeomorphism?
- Can we use the Riemannian geometric mappings to interpolate and perform low-rank approximations to find better geodesics on the data manifold?

A Diffeomorphic Learning Approach

We propose to approximate a smooth diffeomorphism $\varphi : \mathcal{M} \subset \mathbb{R}^n \to \mathbb{R}^d$, where $\mathcal{M}$ is a d-dimensional Riemannian manifold. In Proposition 2.1 of [2] it has been shown that from this diffeomorphism we get all relevant geometric mappings:

1. Distance: $d^{\varphi}_{\mathbb{R}^d}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\varphi(\boldsymbol{x}_i) - \varphi(\boldsymbol{x}_j)\|_2$
2. Length-minising geodesics: $\gamma^{\varphi}_{\boldsymbol{x}_i,\boldsymbol{x}_j}(t) = \varphi^{-1}(\varphi(\boldsymbol{x}_i)(1-t) + \varphi(\boldsymbol{x}_j)t)$
3. Logarithmic map: $\log^{\varphi}_{\boldsymbol{x}_i}(\boldsymbol{x}_j) = \varphi_*^{-1}(\varphi(\boldsymbol{x}_j) - \varphi(\boldsymbol{x}_i))$
4. Exponential map: $\exp^{\varphi}_{\boldsymbol{x}_i}(\Xi_{\boldsymbol{x}_i}) = \varphi^{-1}(\varphi(\boldsymbol{x}_i) + \varphi_*(\Xi_{\boldsymbol{x}_i,\boldsymbol{x}_j}))$

Where φ_* denotes the pushforward of φ and $\Xi_{\boldsymbol{x}_i} \in \mathcal{T}_{\boldsymbol{x}_i}\mathcal{M}$ denotes a tangent vector in $\boldsymbol{x}_i$. We parameterize the diffeomorphism by using an adaptation of Neural ODE. Where following the flow forward in time equates to φ and following the flow backward in time equates to φ^{-1}.

$$\frac{d\boldsymbol{z}(t)}{dt} = f_{\boldsymbol{\theta}}(\boldsymbol{z}(t)) \tag{2}$$

Now assume you have a metric space $(\mathbb{R}^n, (\cdot,\cdot))$, where $d_{i,j}$ is the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ on the manifold $\mathcal{M}$. Then we find the diffeomorphism φ by optimizing the following objective function:

$$\begin{aligned}
\mathcal{L}(\theta) = {} & \alpha_1 \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|d^{\varphi}_{\mathbb{R}^d}(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_{i,j}\|_2^2 && \text{(metric preserving loss)} \\
& + \alpha_2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \|(d^{\varphi}_{\mathbb{R}^d}(\boldsymbol{x}_i, \boldsymbol{x}_k) - d^{\varphi}_{\mathbb{R}^d}(\boldsymbol{x}_j, \boldsymbol{x}_l)) - (d_{i,k} - d_{j,l})\|_2^2 \quad \forall k, l && \text{(graph matching loss)} \\
& + \alpha_3 \frac{1}{N} \sum_{i=1}^{N} \| \begin{bmatrix} I_{n-d} & \emptyset \\ \emptyset & \mathbf{0}_d \end{bmatrix} \varphi_\theta(\boldsymbol{x}_i))\|_2^2 && \text{(low-dimensional loss)} \\
& + \alpha_4 \frac{1}{N} \sum_{i=1}^{N} \int_0^1 \|\boldsymbol{\epsilon}^T \nabla f_{\boldsymbol{\theta}}(\boldsymbol{z}_i(t))\|_1 \, dt, && \text{(stability regularization)}
\end{aligned}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$. [2] also uses **metric preserving loss** and **low-dimensional loss**. We propose to use the **graph matching loss** [3] for training metric preserving diffeomorphisms. This objective enforces an uncorrelated error in the distance matrix ($d^{\varphi}_{\mathbb{R}^d}(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_{i,j}, \ \forall i, j$) of the metric preserving diffeomorphism. We further improve the objective function of [2] by using **stability regularization** [4] to steer the solution to be locally Lipschitz arount the data. This new formulation compared to [2] is metric tensor free and thereby more efficient and scalable for training high-dimensional metric preserving diffeomorphisms.

Contributions

- **Improved parameterization of diffeomorphism**: We improve the expressiveness of the diffeomorphism parameterization by using Neural ODEs compared to [2].
- **Efficient and scalable objective**: We show that we can train stable diffeomorphisms by only having access to a k-NN Euclidean approximation (Isomap) of the distances of the points on the data manifold, avoiding the expensive calculation of the metric tensor.
- **Experimental protein dynamics data**: We show the theory and results from [2] can be applied to experimental data of protein dynamics trajectory of the adenylate kinase protein.

Case study 1: Isometric Embedding

Adenylate kinase consists of 214 aminoacids in 3 dimensions, thus for our experiments $n = 642$ and we assume $d = 1$.

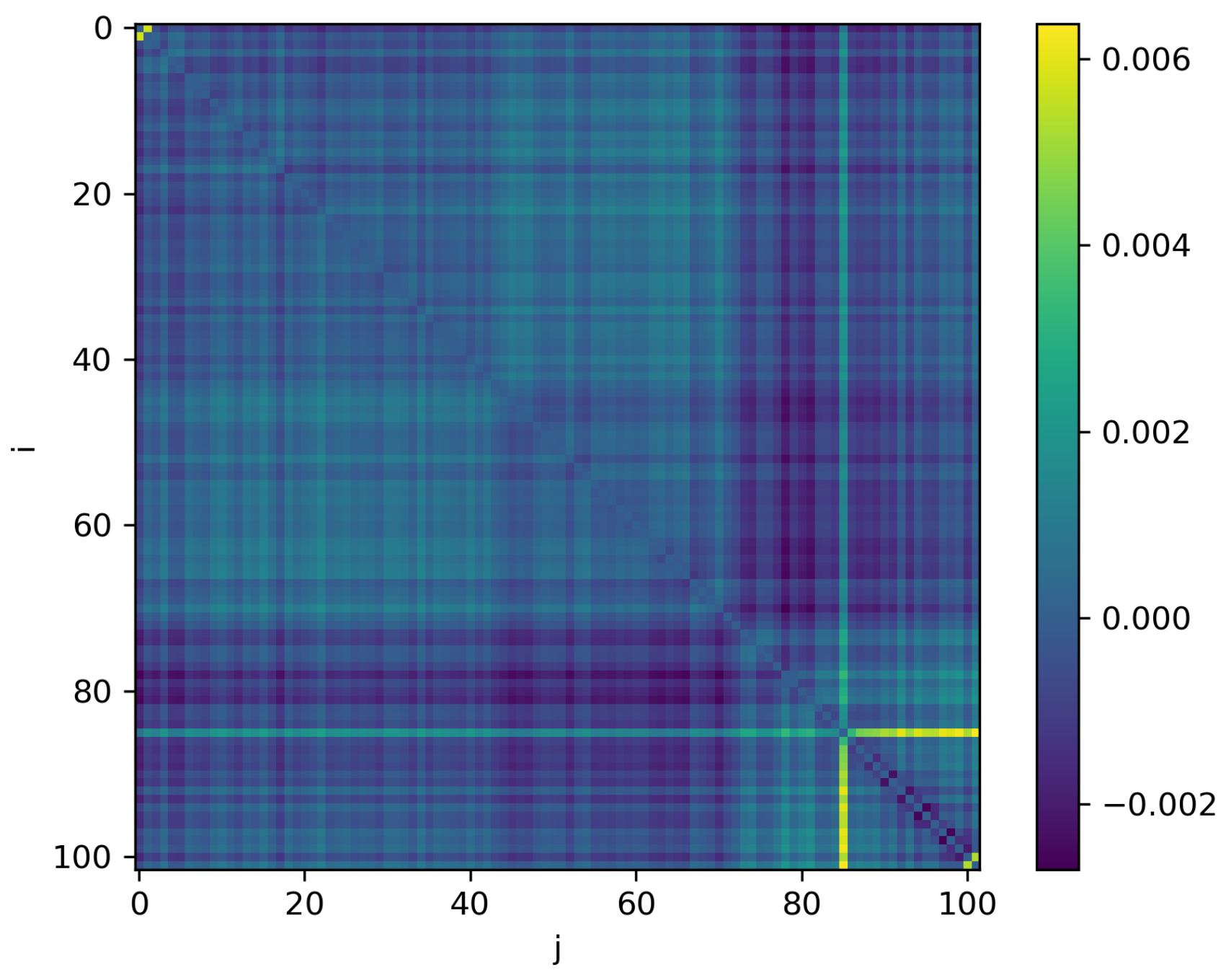


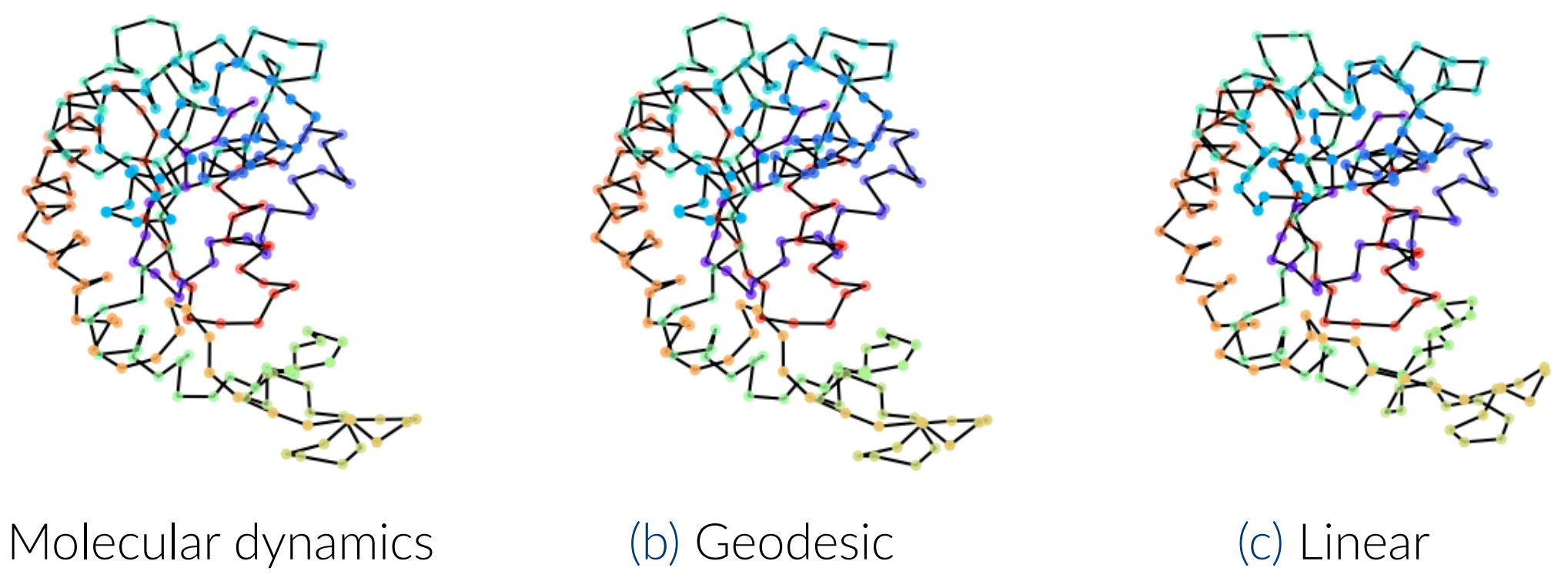
Figure 2. $\frac{d_{i,j} - d^{\varphi}_{\mathbb{R}^d}}{\max_{i,j} d_{i,j}} \quad \forall i, j$

In Figure 2 we see that we successfully find a diffeomorphism that embeds the data $\boldsymbol{x}_i \in \mathcal{M} \subset \mathbb{R}^{214 \times 3}$ for $i = 1, \dots, N$ into a $d = 1$ dimensional metric preserving Euclidean latent space.

For more results, scan me!

Case study 2: Geodesic Interpolation

Next, we compare linear interpolation with linear interpolation of the metric preserving latent space for the longest geodesic in the data (between $\boldsymbol{x}_1$ and $\boldsymbol{x}_{102}$).



(a) Molecular dynamics (b) Geodesic (c) Linear

Figure 3. Comparison of conformations for $t \approx 0.5$ between the molecular dynamics $\boldsymbol{x}_{\mathbf{51}}$, geodesic approximation $\boldsymbol{\gamma}^{\varphi}$ and linear interpolation $\boldsymbol{\gamma}^{\varphi}$.

In Figure 4 we observe that interpolation in the latent space accurately (RMSE ≤ 0.5 Å) approximates the geodesics on the manifold.

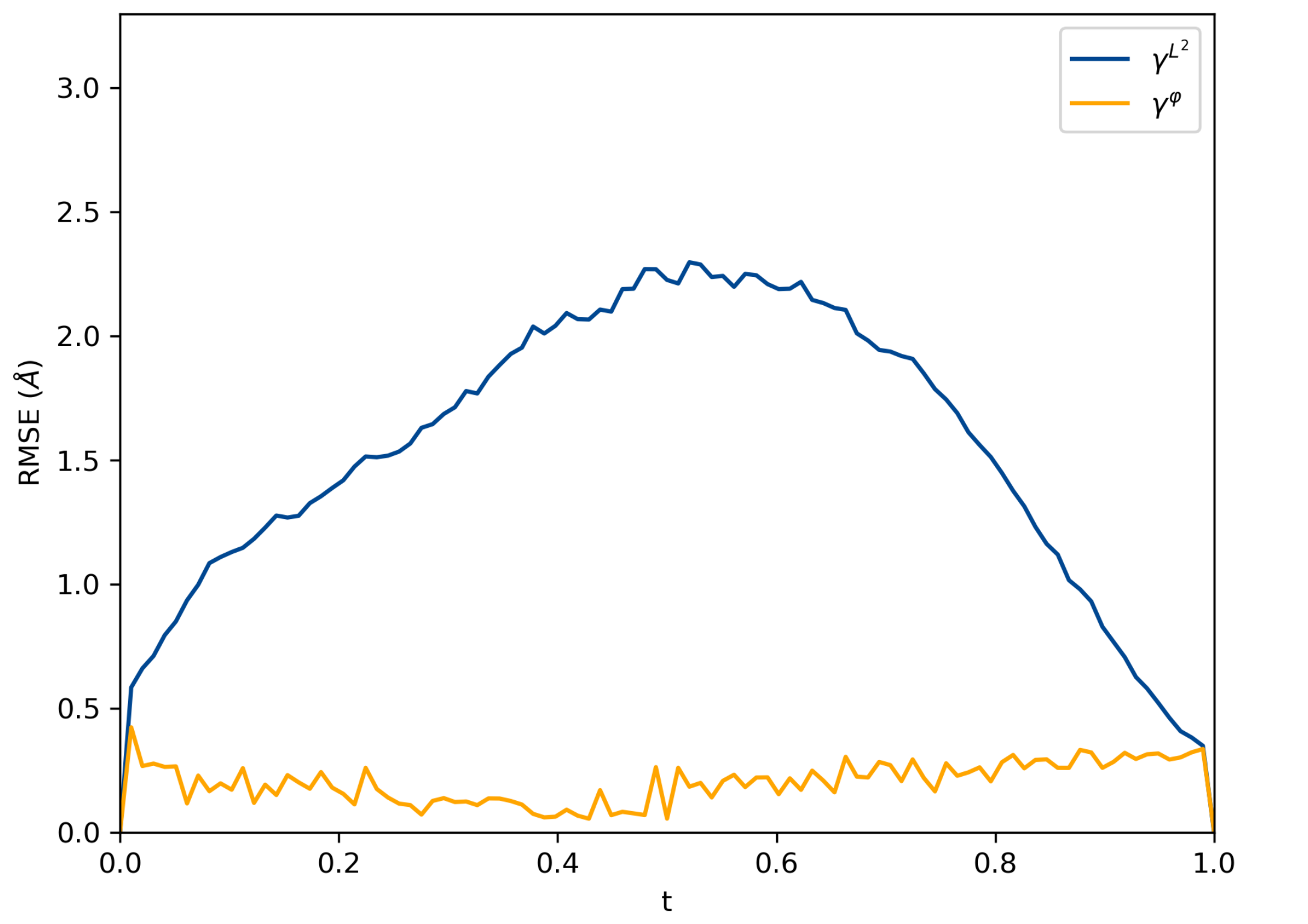


Figure 4. RMSE (Å) for interpolation between $\boldsymbol{x}_1$ and $\boldsymbol{x}_{102}$ for $\boldsymbol{\gamma}^{\boldsymbol{L^2}}$ and $\boldsymbol{\gamma}^{\varphi}$.

Case study 3: Low-Rank Approximation

We use the logarithmic maps from the Riemannian barycentre to perform low-rank approximation of the data manifold ($\boldsymbol{\varphi_{lr\_approx}}$) and compare that to PCA ($\boldsymbol{L^2_{lr\_approx}}$).

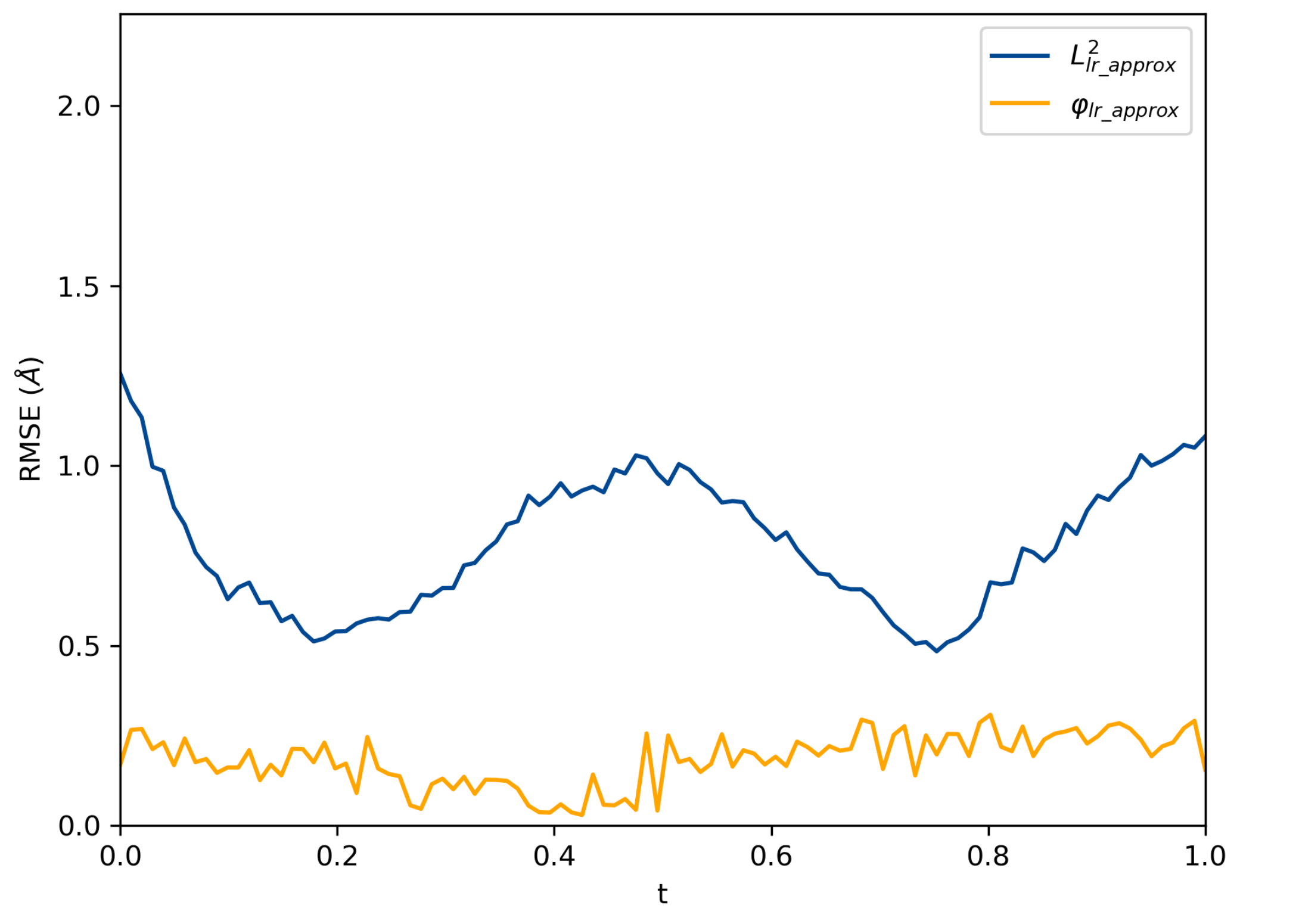


Figure 5. RMSE (Å) for low-rank approximation of molecular dynamics trajectory through PCA ($\boldsymbol{L^2}_{\text{lr\_approx}}$) and PCA on the latent space φ ($\boldsymbol{\varphi_{lr\_approx}}$).

References

[1] Joshua B Tenenbaum, Vin de Silva, and John C Langford.
A global geometric framework for nonlinear dimensionality reduction.
science, 290(5500):2319–2323, 2000.

[2] Willem Diepeveen.
Pulling back symmetric riemannian geometry for data analysis.
arXiv preprint arXiv:2403.06612, 2024.

[3] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen.
Matrix-similarity based loss function and feature selection for alzheimer's disease diagnosis.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3096, 2014.

[4] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman.
How to train your neural ode: the world of jacobian and kinetic regularization.
In *International conference on machine learning*, pages 3154–3164. PMLR, 2020.

frisodk@kth.se