Cryo-EM images are intrinsically low dimensional

Luke Evans[1]* **Octavian-Vlad Murad**[2]* **Lars Dingeldein**[3] **Pilar Cossio**[1]
Robert Covino[3,4] **Marina Meila**[2]

[1]Flatiron Institute [2]University of Washington [3]Frankfurt Institute For Advanced Study
[4]Institute of Computer Science, Goethe University Frankfurt

Abstract

Simulation-based inference has been successful in the analysis of cryo-electron microscopy, particularly for estimating biomolecular conformations from individual images. However, the latent representations learned during training contain more information, which can support or complement the predictions learned in supervised mode. Here, with images of hemagglutinin, we demonstrate that the simulated and experimental data representations can be modeled as a low-dimensional smooth manifold. We identify the (non-linear) directions of variation of the main parameters of interest, and link physical parameter values to the experimental images. By using state-of-the-art manifold learning, we provide accurate visualizations of the data, quantitative supporting evidence to validate the neural predictions with insights into physical properties of the latent representation.

1 Introduction

Cryogenic electron-microscopy (cryo-EM) is a structural biology technique for imaging individual biomolecules at atomic resolution. In a cryo-EM experiment, a biomolecular sample is imaged with a transmission electron microscope, and the resulting data is processed to yield a large dataset of unlabeled 2D images with one molecule per image (particles). Reconstruction algorithms [1] can estimate the 3D structure of the biomolecule from the 2D particles. In many cases, biomolecules coexist in different conformational states in the sample.

Machine learning methods, including diffusion maps [2] and deep-generative models [3, 4, 5], have become central in cryo-EM for reconstructing heterogeneous conformations of biomolecules [6, 7]. These methods project the high-dimensional conformational space on to a low-dimensional latent representation, but these latent spaces lack physical interpretability [8]. Applying physical constraints during training [9] or comparing to ground truth data [10] can help mitigate some of these issues. However, extracting physical information from the featurized images remains challenging due to non-linear feature mapping, low signal-to-noise ratio (SNR) and uncertainty in pose assignment, which can be confused with conformational changes.

Recent simulation-based techniques from integrative structural biology [11] and probabilistic machine learning [12] hold great promise for analyzing cryo-EM data. CryoSBI [13] uses simulation-based inference [14, 15] (SBI) to infer conformations and uncertainties from cryo-EM particles by training a latent representation and normalizing flow with simulated cryo-EM experiments. The trained networks can be quickly evaluated on large experimental particle datasets. Because the training is only done with simulated data, a key feature of cryoSBI is that it enables linking of physical properties of the molecules and the experiment to experimental data. We hypothesize that the representations learned by the neural network are near low dimensional manifolds inside the latent space. The objective of this work is to study the geometry of the data using manifold learning techniques [16, 17, 18, 19]. First, we will seek to ascertain whether the learned representations correspond to well-behaved low-dimensional manifolds, and second, whether these are parameterized

*Equal contribution. Correspondence to: `levans@flatironinstitute.org, ovmurad@uw.edu`

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

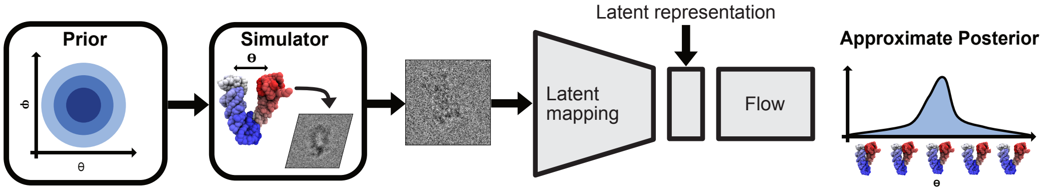


Figure 1: Schematic workflow for learning the surrogate posterior with cryoSBI. Parameter samples are drawn from the prior to simulate synthetic cryo-EM images. These images are then used to approximate the posterior by jointly training a summary network and a normalizing flow.

by generative variables important in predicting the posterior over the conformation. Our analysis quantitatively validates the latent space of cryoSBI and leads to a general computational workflow both for interpreting latent spaces of cryo-EM heterogeneity analysis methods and more broadly for learned summary statistics in simulation-based inference.

2 CryoSBI and Latent Spaces

CryoSBI [13] is a new method to quantify the probability that a given image I depicts a molecular conformation θ. We assume to have a set of structures, e.g. from molecular simulations or Alphafold [20], which we expect to find in the sample. For simplicity, we also assume that θ is a one-dimensional parameter, and we aim to infer the conformation θ of the molecule observed in the image, i.e., compute the Bayesian posterior $p(\theta|I)$. The posterior quantifies how compatible θ is with the observed image I.

To model the image formation process, one must consider experimental details such as microscope aberration, noise, and random orientation of the molecule. To simulate a cryo-EM image, one samples conformations from the prior $\theta_i \sim p(\theta)$, and imaging parameters from $\phi_i \sim p(\phi)$ and then generates a synthetic image $I_i \sim p(I|\theta_i, \phi_i)$ using a forward model of the imaging process (Appendix A.1), accumulating a data set of simulated images and ground truth parameters $\mathcal{D} = \{\theta_i, \phi_i, I_i\}_{i=1}^{N}$. The nuisance parameter vector ϕ_i includes random orientations, a wide range of defocus values, center translations, and SNRs.

Feature Latent Representation and Neural Posterior Estimation. CryoSBI follows the Neural Posterior Estimation framework [21, 22], jointly training a latent representation network $S_\psi(\cdot)$ to extract summary statistics and a normalizing flow $q_\varphi(\cdot)$ as surrogate model of the posterior $q_\varphi(\theta|S_\psi(I)) \approx p(\theta|I)$. This is done by maximizing the average log-likelihood $\mathcal{L}(\varphi, \psi) = \frac{1}{N}\sum_{i=1}^{N} \log q_\varphi(\theta_i|S_\psi(I_i))$ of the posterior probability under the training samples $\mathcal{D}$ (Appendix A.2). In principle, after training S_ψ should *i*) compress images to predict the relevant features and *ii*) enable efficient comparison of simulated images to 'nearby' experimental images. For example, the latent representation should distinguish images due to conformation, SNR and projection direction, as these are the primary experimental factors determining how precisely we can estimate a molecular configuration from a single image. In practice, while the feature representation for Neural Posterior Estimation - and cryoSBI- offers powerful inference capabilities, it is not immediately interpretable, making it challenging to check for model misspecification [23].

Hemagglutinin Dataset. The CryoSBI latent space we analyze here corresponds to the hemagglutinin dataset considered in ref. [13]; it consists of latent representations of the simulated and experimental images. CryoSBI training was performed as in [13] using cryo-EM simulations by sampling the priors (Appendix A.3). After training, we valuated a simulated dataset $\mathcal{D}_s$ consisting of $N_s = 100,000$ feature vectors with i-th datapoint $x_i = S_\psi(I_i) \subseteq \mathbb{R}^{256}$, nuisance parameters ϕ_i, ground-truth conformation parameter θ_i, posterior mean $\hat{\theta}_i$ and width σ_i of the posterior $q_\varphi(\cdot|x_i)$, so that $\mathcal{D}_s = \{\hat{\theta}_i, \sigma_i, \theta_i, \phi_i, x_i\}_{i=1}^{N_s}$. The experimental dataset $\mathcal{D}_e$ consists of $N_e = 271558$ tuples$\{\tilde{x}_i, \hat{\theta}_i, \sigma_i\}$ with $\tilde{x}_i = S_\psi(\tilde{I}_i)$, for whitened single particle-images $\{\tilde{I}_i\}_{i=1}^{N_e}$ from EMPIAR 10532 [24], where $\hat{\theta}_i, \sigma_i$ are the inferred posterior parameters (note that the experimental images have no ground truth θ or ϕ). We denote the representations learned by S_ψ, $\mathcal{X}_s = \{x_i\}_{i=1}^{N_s}$ and $\mathcal{X} = \{\tilde{x}_i\}_{i=1}^{N_e}$.

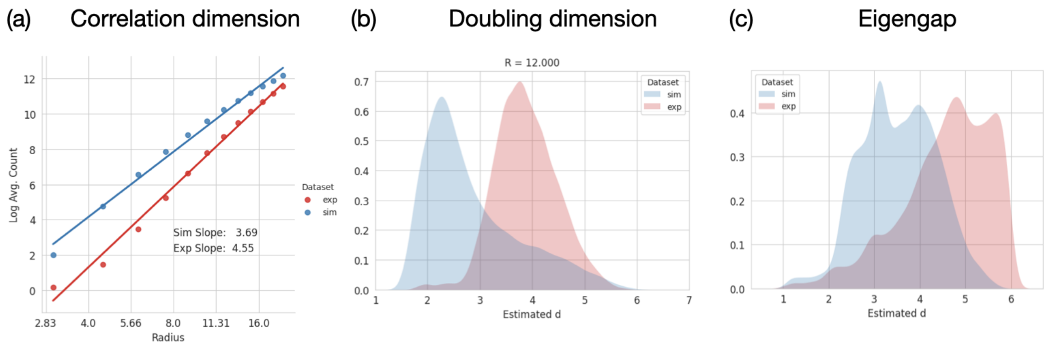


Figure 2: Estimation of the intrinsic dimension d_s (blue) and d_e (red) of the manifolds $\mathcal{M}_s$ and $\mathcal{M}_e$, respectively, using the correlation dimension (**a**), doubling dimension at R=12 (**b**), and Eigengap (**c**) methods. Note that for **(b)** and **(c)**, we plot the distribution of the local estimates of d, while for **(a)** the prediction is global. The results suggest that $2 \leq d_s \leq d_e \leq 6$.

3 Geometric Analysis of the CryoSBI Latent Space

Now, we proceed to study the shape of the data cloud $\mathcal{X}_e$, under the hypothesis that it is low-dimensional, i.e. near a smooth manifold $\mathcal{M}_e$. The simulated data $\mathcal{D}_s$ support the interpretation of $\mathcal{M}_e$, but some aspects of its geometry will also be considered, in particular the low dimensionality hypothesis. In the following, we will determine the intrinsic dimensionality of the datasets, assess how well the simulated data covers the experimental space, and uncover the physical interpretation of the latent representations. The data preprocessing, consisting of removing outliers, and resampling the data to avoid large variation in density is described in Appendix A.4.

Are the data low dimensional? We estimate the intrinsic dimension of the experimental and simulated data (d_e and d_s, respectively). Due to the challenges of reliably estimating dimensions for noisy data, we employ three different methods for greater accuracy. Two of these methods leverage the rate of growth of the volume of a ball of radius R in a manifold with intrinsic dimension d, which is $\sim R^d$. The correlation dimension [25] uses the number of neighbors $N_e(x)$ of radius R of $x \in \mathbb{R}^{256}$, which satisfies $\log N_e(x) \approx d \cdot \log R + \text{const}$, allowing us to estimate d as the slope of a regression line. Similarly, following [26], we use $\frac{N_{2R}(x)}{N_e(x)} \approx 2^d$ as a local statistic to estimate d, called *doubling dimension*[2]. The third method, Eigengap, is that of ref. [27]. This method locally estimates the intrinsic dimension d by finding the largest gap between two consecutive eigenvalues in the local covariance matrix. We implement a variation of this method, by combining it with the neighborhood scale selection of ref. [28]. This method and the doubling dimension give *local* estimates of d around a point x. A global d is then selected by majority vote; we modify this by using smoothed histograms for the former and softmax for the latter. In Figure 2 we present the results of the estimation using these methods. All three methods indicate that the data have *low intrinsic dimension*. This is partly due to the neural network training algorithm that is optimized for predicting a low-dimensional function $p(\theta|I)$. We find a dimension near 2 for the simulated data and slightly higher dimension for the experimental data. The discrepancy, where $2 \leq d_e \leq 6$, with a peak near $d_e = 5$, is likely due to experimental noise and dependencies not captured by the simulated noise model (see also Figure 3, (a)). Based on the estimated intrinsic dimensions, this suggests that the manifold assumption is supported by the data.

Does the simulated data cover the experimental data well? For this, we first estimate the data densities p_e and p_s in $\mathbb{R}^{256}$ by kernel density estimators (KDE) [29] $\hat{p}_e$ and $\hat{p}_s$. The bandwidths $h_e = 0.34$ and $h_s = 0.48$ are obtained by cross-validation. While it is known that KDE is poor in high dimensions, the method is *adaptive*, meaning that it will work when the intrinsic dimension is low, as in this case. We use samples of size 17000 for fitting $\hat{p}_{e,s}$. We do not expect p_e to equal p_s, but we would like to confirm that p_s is predictive of the experimental data. Thus, on two held out datasets $\mathcal{X}_e^{test}$ and $\mathcal{X}_s^{test}$, with $|\mathcal{X}_s^{test}| = |\mathcal{X}_e^{test}| = n^{test} = 3000$, we calculate the negative log-likelihoods (i.e., cross-entropies) $-\frac{1}{n^{test}} \log \hat{p}_{r,s}(\mathcal{X}_{r,s}^{test})$ (in Table 1) and the Kullback-Leibler

[2] Note that this estimate depends on R (Appendix Figure 1).

divergences $D_{KL}(p_e||p_s) = 97.6$, $D_{KL}(p_s||p_e) = 1824.9$. These show that the simulated data can predict the experimental data well; meanwhile, the experimental data does not completely cover the simulated data. For further analysis, we retain in $\mathcal{X}_s$ only the samples that are near the experimental data. The hypothesis that we can infer what generative parameters best describe the experimental data, is so far supported since we can, for most experimental $\tilde{x} \in \mathcal{X}_e$, find enough near-by synthetic $x \in \mathcal{X}_s$ to perform this prediction in a robust manner.

Modeling the low dimensional cryo-EM images manifold. We use a suite of manifold learning techniques [16, 28, 17, 30] to map the neural representations $\mathcal{X}_e \subseteq \mathbb{R}^{256}$ down to much lower dimensional embeddings Φ_e, which we here interpret geometrically and in the following section from the physical point of view, in relation to the simulated data $\mathcal{X}_s$. We use Diffusion Maps [16] with a kernel width parameter ϵ selected by the method of ref. [28] to compute the low-dimensional embedding $\Phi_e \in \mathbb{R}^d$ of $\mathcal{X}_e$; similarly we compute Φ_s for the filtered $\mathcal{X}_s$ data (see Appendix Figure 2). The Diffusion Maps embedding is based on the eigendecomposition of the Laplacian matrix $\mathbf{L}$ [16], and in a first stage we compute it up to the m'th non-zero eigenvalue, for $m = 20$, and denote these coordinates with $\Phi_{1:m} \in \mathbb{R}^n$, with $n = |\mathcal{X}_e|$. The analysis of the principal eigenvalues of $\mathbf{L}$, which are slowly growing and well above 0 (Appendix Figure 3), indicates that the manifold $\mathcal{M}_e$ is connected, that is, there are no isolated clusters and no outliers for the postprocessed data. However, the presence of clusters as high-density regions in the data is not precluded by this analysis, and Figure 3, (c), as well as Appendix Figure 4 map a sample from the original p_e into $\mathcal{M}_e$. Next, we perform IES [17] to select $d = 3$ independent and low-frequency coordinates from $\Phi_{1:20}$. We use these coordinates, denoted Φ_e, to visualize and interpret the experimental data. As shown previously, $d = 3$ is likely close to the true intrinsic dimension of $\mathcal{M}_s$ and $\mathcal{M}_e$, meaning we can expect to capture most of the relevant structure of the experimental data by analysing these Φ_e coordinates. We apply Riemannian Relaxation [30] to push Φ_e closer to being isometric to $\mathcal{X}_e$. The resulting embedding is shown in Figure 3. We perform similar steps with the simulated data $\mathcal{X}_s$ (Appendix A.5).

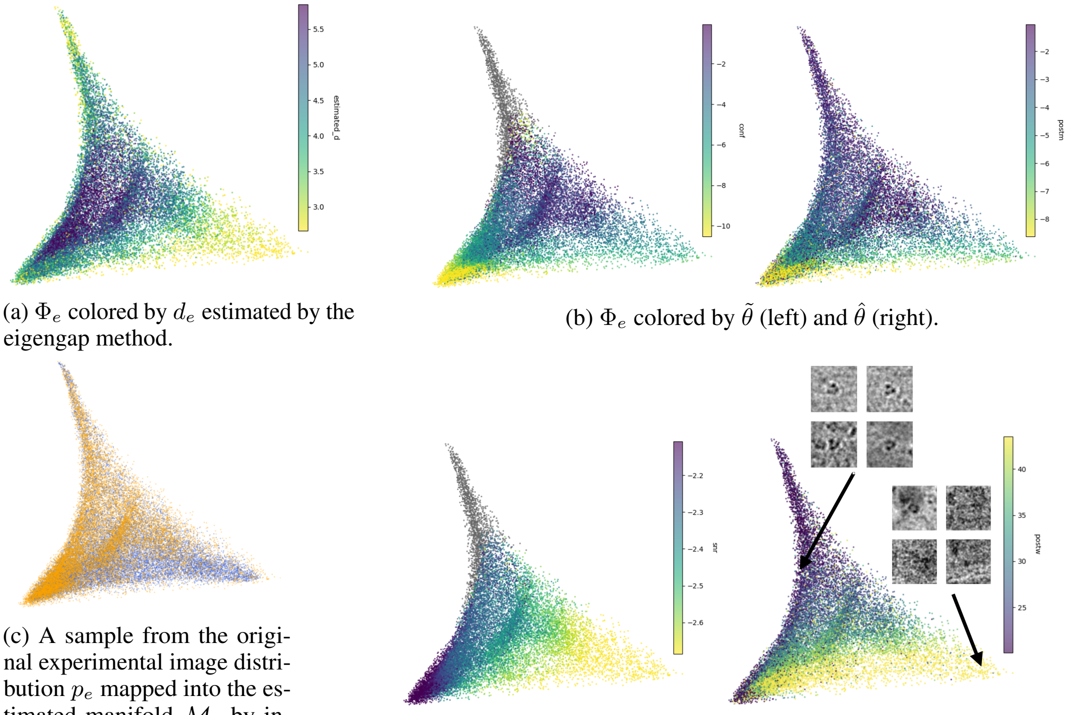
$-\frac{1}{n}\log p_{model}(\mathcal{X}^{test})$		
	$\mathcal{X}_e^{test}$	$\mathcal{X}_s^{test}$
p_e	84.9	2005.7
p_s	182.5	180.8

Table 1: Test data negative log-likelihoods under p_e and p_s.

Physical interpretation of the experimental data manifold In the absence of ground truth generative parameters for the experimental data, we have to find alternative ways to determine whether S_ψ is a good predictor for the true conformational parameter θ, and the noise level, an important nuisance parameter. While this can be done with a manually labeled test set, we focus on indirect geometric methods that don't require scientific labeling. We first use a statistical method, TSLasso [18] to interpret the embedding Φ_e. Afterwards, we support its results and expand the analysis with visualizations. TSLasso searches for the optimal interpretation of an embedding in a *dictionary* $\mathcal{F} = \{f_k : \mathcal{M}_e \to \mathbb{R}, k = 1 : p\}$ of (smooth) potential coordinate functions on $\mathcal{M}_e$. Here, each $f_k \in \mathcal{F}$ represents one of the simulation parameters (the conformation θ or one of the nuisance parameters in ϕ), hence $|\mathcal{F}| = 10 = p$. TSLasso recovers a subset f_S of $\mathcal{F}$ which parametrizes $\mathcal{M}_e$, by selecting d functions whose gradients "most economically" span the tangent spaces of the manifold at a sample of the data. Since the functions f_k are unknown on the experimental data, we infer them by interpolation (Appendix A.6), obtaining $\tilde{\theta}$ and $\tilde{\phi}$ for the experimental data. We also estimate the gradients ∇f_k (Appendix A.7). TSLasso is run 20 times using random subsets of 500 data points. We find that f_S almost always consists of conformation θ, SNR, and one of the rotation coordinates in $\phi/\tilde{\phi}$ (albeit not always the same one). The full results are presented in Appendix A.7. For completeness, we apply the same algorithm to the simultated data. Our results show that this combination of functions parametrizes both $\mathcal{M}_s$ and $\mathcal{M}_e$. We have confirmed statistically, without any visualization, that the two parameters θ and SNR inferred from nearby simulated data, vary smoothly along the experimental data manifold $\mathcal{M}_e$ (as well as along $\mathcal{M}_s$), therefore, supporting the neural network predictions for $\mathcal{X}_e$. The visualizations are shown in Figure 3 (b) and (d).

4 Discussion

In summary, our study of the latent embedding representations of hemagglutinin cryo-EM data from cryoSBI, has revealed that these live near a well-behaved low dimensional manifold in $\mathbb{R}^{256}$ space where the simulated images cover (almost entirely) the experimental ones. Therefore, we can use the simulated data (on which we have full control) to interpret the experimental data in the latent



(a) Φ_e colored by d_e estimated by the eigengap method.

(b) Φ_e colored by $\tilde{\theta}$ (left) and $\hat{\theta}$ (right).

(c) A sample from the original experimental image distribution p_e mapped into the estimated manifold $\mathcal{M}_e$ by interpolation (blue) vs Φ_e (orange).

(d) Φ_e colored by $\tilde{\mathrm{SNR}}$ (left) and posterior width σ (right), with example particles corresponding to indicated regions.

Figure 3: Diffusion Maps embedding Φ_e in $d = 3$ dimensions. **(a)** Φ_e colored by local d_e. The highest intrinsic dimension is in regions with medium SNR, while high SNR regions have $d_e \in [3, 4]$. **(b)** Φ_e colored by the predicted conformation from manifold interpolation $\tilde{\theta}$ and the conformation estimated posterior mean $\hat{\theta}$. **(c)** Difference in density between the sample from p_e (blue) and the sample used to compute Φ_e (orange). p_e is much denser in the low SNR regions. **(d)** Φ_e colored by the interpolated SNR and posterior width σ.

space. Furthermore, we have identified the physical and geometrical features that explain the different directions in the latent space.

We presented visualizations (e.g., by postprocessed Diffusion Maps embedding) that accurately display the data shape by being almost isometric. We are also excited by the possibilities of replacing visual analysis with quantitative measures, and principled algorithms in creating and validating low dimensional models of cryo-EM data. Examples of such tasks include detecting the intrinsic dimensionality, interpreting the manifold by physical coordinates, measuring the smoothness of functions over the data manifold (not included here, but straightforward via the Laplacian operator), detecting if clusters exist, and measuring local distortion [31].

From the methodological point of view, we present a pipeline for analyzing, exploring and visualizing high dimensional data presumably living near a smooth manifold. The pipeline components integrate state of the art geometric algorithms and theoretical results. However, we note that we do not propose to replace the trained neural network predictor with (a variant of) the methods presented here. Typically, dimension reduction methods do not outperform a neural network trained in supervised mode. What our method offers is interpretability of the latent representations and a connection of the experimental data to the physical simulator.

At the same time, we acknowledge that the data might not align perfectly with the manifold hypothesis. Our current understanding does not yet enable us to predict, comprehend, or control how finer-scale data structures— e.g., what we consider "noise"—affect geometric algorithms, which should be a matter of further investigation.

References

[1] Eva Nogales and Sjors HW Scheres. Cryo-em: a unique tool for the visualization of macromolecular complexity. *Molecular cell*, 58(4):677–689, 2015.

[2] Ali Dashti, Peter Schwander, Robert Langlois, Russell Fung, Wen Li, Ahmad Hosseinizadeh, Hstau Y. Liao, Jesper Pallesen, Gyanesh Sharma, Vera A. Stupina, Anne E. Simon, Jonathan D. Dinman, Joachim Frank, and Abbas Ourmazd. Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl. Acad. Sci. U. S. A.*, 111:17492–17497, 2014.

[3] Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods*, 18(2):176–185, 2021.

[4] Muyuan Chen and Steven J Ludtke. Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. *Nature Methods*, 18(8):930–936, 2021.

[5] Ali Punjani and David J Fleet. 3D flexible refinement: structure and motion of flexible proteins from cryo-EM. *BioRxiv*, 2021.

[6] Claire Donnat, Axel Levy, Frederic Poitevin, Ellen D Zhong, and Nina Miolane. Deep generative modeling for volume reconstruction in cryo-electron microscopy. *Journal of structural biology*, 214(4):107920, 2022.

[7] Wai Shing Tang, Ellen D Zhong, Sonya M Hanson, Erik H Thiede, and Pilar Cossio. Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy. *Current Opinion in Structural Biology*, 81:102626, 2023.

[8] Roy R Lederman and Bogdan Toader. On manifold learning in plato's cave: Remarks on manifold learning and physical phenomena. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–7. IEEE, 2023.

[9] David A Klindt, Aapo Hyvärinen, Axel Levy, Nina Miolane, and Frédéric Poitevin. Towards interpretable cryo-em: disentangling latent spaces of molecular conformations. *Frontiers in Molecular Biosciences*, 11:1393564, 2024.

[10] Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, Ryan Feathers, Alkin Kaz, Sonya M Hanson, Pilar Cossio, and Ellen D Zhong. Cryobench: Diverse and challenging datasets for the heterogeneity problem in cryo-em. *arXiv preprint arXiv:2408.05526*, 2024.

[11] Michael P Rout and Andrej Sali. Principles for integrative structural biology studies. *Cell*, 177(6):1384–1403, 2019.

[12] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[13] Lars Dingeldein, David Silva-Sánchez, Luke Evans, Edoardo D'Imprima, Nikolaus Grigorieff, Roberto Covino, and Pilar Cossio. Amortized template-matching of molecular conformations from cryo-electron microscopy images using simulation-based inference. *bioRxiv*, 2024.

[14] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[15] George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

[16] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[17] Yu-Chia Chen and Marina Meila. Selecting the independent coordinates of manifolds with large aspect ratios. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Samson J Koelle, Hanyu Zhang, Octavian-Vlad Murad, and Marina Meila. Consistency of dictionary-based manifold learning. In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4356. PMLR, 2024.

[19] Marina Meilă and Hanyu Zhang. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11, 2024.

[20] Gabriel Monteiro da Silva, Jennifer Y Cui, David C Dalgarno, George P Lisi, and Brenda M Rubenstein. High-throughput prediction of protein conformational distributions with subsampled alphafold2. *nature communications*, 15(1):2464, 2024.

[21] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.

[22] Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. Neural methods for amortised parameter inference. *arXiv preprint arXiv:2404.12484*, 2024.

[23] Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Detecting model misspecification in amortized bayesian inference with neural networks: An extended investigation. *arXiv preprint arXiv:2406.03154*, 2024.

[24] Yong Zi Tan and John L Rubinstein. Through-grid wicking enables high-speed cryoem specimen preparation. *Acta Crystallographica Section D: Structural Biology*, 76(11):1092–1103, 2020.

[25] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1):189–208, 1983.

[26] Patrice Assouad. Plongements lipschitziens dans $\setminus^n$. *Bulletin de la Société Mathématique de France*, 111:429–448, 1983.

[27] Guangliang Chen, Anna V. Little, and Mauro Maggioni. Multi-resolution geometric analysis for data in high dimensions, 2013.

[28] Dominique Perrault-Joncas and Marina Meila. Improved graph laplacian via geometric self-consistency, 2014.

[29] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[30] James McQueen, Marina Meila, and Dominique Joncas. Nearly isometric embedding by relaxation. 29, 2016.

[31] Dominique Perraul-Joncas and Marina Meila. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery. *arXiv preprint arXiv:1305.7255*, 2013.

[32] Evan Seitz, Francisco Acosta-Reyes, Peter Schwander, and Joachim Frank. Simulation of cryo-em ensembles from atomic models of molecules exhibiting continuous conformations. *BioRxiv*, page 864116, 2019.

[33] Julian Giraldo-Barreto, Sebastian Ortiz, Erik H. Thiede, Karen Palacio-Rodriguez, Bob Carpenter, Alex H. Barnett, and Pilar Cossio. A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci. Rep.*, 11(1):13657, December 2021.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural Spline Flows. June 2019. arXiv: 1906.04032.

[36] Andrew J Hanson. Visualizing quaternions. In *ACM SIGGRAPH 2005 Courses*, pages 1–es. 2005.

A Appendix

A.1 Cryo-EM image formation forward model

We simulate cryo-EM particles from 3D molecular structures with the forward model of [32, 33]. The electron density $\rho(X)$ of a given structure X is approximated as a Gaussian mixture model with centers on the positions of the C_α atoms, and standard deviations γ. Then, we apply a rotation R_q with quaternion q and projection P_z onto the $z-$axis to $\rho(X)$, then convolve with a point-spread function (PSF), which incorporates the microscope defocus and aberration. The PSF is more straightforward to apply in Fourier space, where the convolution becomes a point-wise multiplication with the Fourier transform of the point-spread function, known as the Contrast Transfer Function (CTF). The CTF is defined as $\mathrm{CTF}_{A,b,\Delta z}(s) = e^{-bs^2/2}\left[A\cos(\pi\Delta z\lambda_e s^2) - \sqrt{1-A^2}\sin(\pi\Delta z\lambda_e s^2)\right]$, with reciprocal radius component $s = 2\pi/\sqrt{x^2+y^2}$, amplitude A, b-factor b, defocus Δz and electron wavelength λ_e. After applying the point-spread function, we translate the image by τ and add Gaussian noise with variance $\sigma^2_{\text{noise}} = \sigma^2_{\text{signal}}/\text{SNR}$, where σ^2_{signal} is the variance of the signal and SNR is the signal-to-noise ratio. The variance of the signal σ^2_{signal} is computed by applying a circular mask with a predefined radius on the noiseless image and then calculating the mean squared intensity. The image formation forward model is then

$$I(x,y|\phi,\rho) = \mathrm{PSF}_{A,b,\Delta z} * (P_z R_q \rho(X) + \boldsymbol{\tau}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0,\sigma^2_{\text{noise}})\,, \tag{1}$$

where $*$ denotes convolution. The imaging parameters utilized for simulating cryo-EM images in CryoSBI are the Gaussian mixture width γ, quaternion q, translation τ, noise level σ_{noise}, and PSF parameters $A, b, \Delta z$, with $\phi = \{\gamma, q, \tau, A, b, \Delta z, \sigma_{noise}\}$.

A.2 CryoSBI feature latent network and conditional density estimation

The latent network S_Ψ follows a ResNet-18 architecture [34] as implemented in ref. [13], with modifications for grascale image input and 256-dimensional feature vector output. For the density estimator q_φ, e implement a Neural Spline Flow (NSF) [35] with the same architecture and training as utilized in ref. [13], and likewise generating each batch of synthetic images on demand in training.

A.3 CryoSBI priors for hemagglutinin

All data processing and SBI procedures for Hemagglutinin data were carried out as in ref. [13], with experimental hemagglutinin images obtained from EMPIAR 10532 [24]. The conformations from hemagglutinin were obtained from a normal mode analysis on atomic structure built from a 3Å reconstruction (PDB id: 6wxb), resulting in 20 conformations indexed by RMSD displacement $\theta_i, i = 1, \ldots, 20$. The conformation prior $p(\theta)$ was taken as a uniform distribution over the possible conformational displacements $\{\theta_i\}$, and the logarithm of the SNR was sampled from a uniform distribution values between $\log 10^{-1}$ and $\log 10^{-3}$. The prior on the quaternions q was chosen so that rotations R_q were sampled uniformly in SO(3) [36]. The other imaging parameters were sampled from uniform distributions in each parameter within bounds chosen in ref. [13]. All nuisance parameters comprising ϕ were assumed independent and sampled independently from their respective priors.

A.4 Data Pre-processing

We begin by randomly sampling $N_e = N_s = 50000$ data points from $\mathcal{D}_e$ and $\mathcal{D}_s$. This is not a requirement and it was done to reduce the computational load. Our next step is to compute, for each $x_i \in \mathcal{X}_s$ and $\tilde{x}_i \in \mathcal{X}_e$, the number of neighbors $N_s(x_i), N_e(\tilde{x}_i)$ within various radii R. We pick a radius R that gives an approximately uniform distribution over the number of neighbors in both datasets. We use $R = 7.5$ and $R = 9.0$ for the experimental and simulation data, respectively.

We remove points with low-connectivity, very likely outliers, by removing all entries from $\mathcal{D}_e$ and $\mathcal{D}_s$ that have $N_e(x_i), N_e(\tilde{x}_i) < 8$. This leaves us with $N_e = 40846$ and $N_s = 36783$. These are the datasets that are used for the coverage analysis between the simulated and the experimental data. Because our objective is to analyze $\mathcal{M}_e$, we also remove all entries in $\mathcal{X}_s$ that are not within $R = 7.5$ of some experimental data point. After this step, $N_s = 26051$ entries remain in $\mathcal{X}_s$.

As shown in [16] , one can remove the biases due to non-uniform sampling density when estimating the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$. However, this result is asymptotic, and assumes that the sampling density does not vary too much. For real data, it is recommended to avoid large variations in data density, for instance by resampling as we do. There is another practical reason to remove large density variations: this allows one to do reliable manifold estimation with a single kernel width ϵ. Empirically we found support for this practical advice;we obtain better results when we subsample the data in such a way that we encourage the sample to be as uniform as possible over $\mathcal{M}_e$. In order to do this, we take the remaining data entries in $\mathcal{D}_e$ and $\mathcal{D}_s$, and sample 20000 data points from each using a distribution over the data entries proportional to $1/N_{7.5}(\tilde{x}_i)$ and $1/N_{9.0}(x_i)$, respectively for the experimental and simulated sets. As shown in Appendix Figure 4, this has the secondary effect of sampling less from the noisy, low SNR, and likely uninformative regions of the manifold. Thus, the embeddings obtained from these samples are encouraged to capture the true geometries of $\mathcal{M}_e$ and $\mathcal{M}_s$, while reducing potential side-effects due to density variations over the manifolds.

We use the method of [28] to estimate kernel width parameters ϵ_e, ϵ_s and cutoff radii R_e, R_s that maximize the Laplacian Matrix's $\mathbf{L}$ ability to preserve the geometry of the data. We use $\frac{R_e}{\epsilon_e} = \frac{R_e}{\epsilon_s} = 3$ and find the optimal radii to be $R_e = 17.0$ and $R_s = 15.0$. We remove all entries from $\mathcal{D}_e$ and $\mathcal{D}_s$ whose degrees in the kernel matrices, computed with the widths and radii above, are in the bottom 5-th percentile. This is meant to improve the stability of the eigen-decomposition performed by the Diffusion Maps algorithm. The remaining 19000 data points will be used for computing the Diffusion Maps embeddings Φ_s and Φ_e. We re-estimate $\epsilon_e, \epsilon_s, R_e, R_s$ on these final datasets and obtain $R_e = 16.5$ and $R_s = 13.5$ which will be used for computing $\mathbf{L}$.

A.5 Diffusion Maps Embedding Details

We compute the Diffusion Maps embeddings [16], denoted Φ_s and Φ_e, using the neural representations learned by S_ψ, $\mathcal{X}_s = \{x_i\}_{i=1}^{N_s}$ and $\mathcal{X}_e = \{\tilde{x}_i\}_{i=1}^{N_e}$. Here, and for the remainder of the appendix, $N_s = N_e = 19000$, and $\mathcal{X}_e, \mathcal{X}_s$(and associated $\mathcal{D}_e, \mathcal{D}_s$) are those obtained after the pre-processing steps in Appendix A.4.

Diffusion Maps is based on the eigen-decomposition of the Laplacian matrix $\mathbf{L}$ from which we keep the first m non-zero eigenvectors in increasing order of their eigenvalues, $\Phi_s \in \mathbb{R}^{n \times m}$ and $\Phi_e \in \mathbb{R}^{n \times m}$. We use $m = 20$ in our experiments. For both Φ_s and Φ_e we find that only the first eigenvalue is 0 and that the spectrum increases slowly. This indicates that both $\mathcal{M}_s$ and $\mathcal{M}_e$ are smooth and connected. In Appendix Figure 3, we display the non-zero eigenvalues of the two decompositions. In conjunction with the results from the Main Text and Appendix Figure 4, this provides strong evidence that the neural representations learned by S_ψ are well-behaved low dimensional manifolds.

Next, we perform IES [17] to select three independent and low frequency coordinates from Φ_s and Φ_e. Briefly, IES(Independent Eigencoordinate Selection) selects a subset S of the m coordinates of a smooth embedding $\Phi(\mathcal{M})$ such that $\Phi_S(\mathcal{M})$ is also a smooth embedding striking a balance between having low frequency and having rank consistently close to d, the intrinsic dimension of the manifold $\mathcal{M}$. In our experiments we use $|S| = 6$. Since we don't know the intrinsic dimension d, but we estimate it to be between 2 and 6, we perform IES for all $3 \leq d \leq 6$ and select the coordinates which appear most often across all runs for different d's. We obtain coordinates $S_e = \{0, 1, 3\}$ for the experimental data and $S_s = \{0, 1, 5\}$ for the simulated data. We use these coordinates to visually analyze the embeddings. Since $\mathcal{M}_s$ and $\mathcal{M}_e$ are low-dimensional we fully expect to capture most of the geometric structure by only analyzing these three coordinates. In Figure 3 we display the IES selected coordinates for Φ_e, while in Appendix Figure 2 we display those for Φ_s.

Finally, we apply Riemannian Relaxation [30] to push the embeddings closer to being isometric to their respective neural representations. To do this, Riemannian Relaxation starts from the initial embeddings Φ_e and Φ_s, and iteratively modifies them via gradient descent with respect to a loss function which penalizes local distortions in the estimated pull-back metric at points in the embedding space. In Appendix Figure 5, we display "relaxed" versus "unrelaxed" versions of Φ_e and Φ_s. We note that Riemannian Relaxation is an optional step in our framework that can aid the visual interpretation of the data. In our experiments we use $d = 3, \epsilon_{orth} = 0.5$, and run Riemannian Relaxation until convergence.

A.6 Estimating the parameters of the experimental data by interpolation

In this section, we explain how we infer the generative parameters $\tilde{\theta}$ and $\tilde{\phi}$ for the experimental data and how we embed a new sample from $\mathcal{X}_e$ into the embedding space Φ_e as in Appendix Figure 4. This is done via Nadaraya-Watson Kernel Regression [12] in the neural embedding space. More specifically, for every $\tilde{x}_i \in \mathcal{X}_e$, we estimate the conformation $\tilde{\theta}_i = \frac{\sum_{x_j \in \mathcal{X}_s} K(\tilde{x}_i, x_j)\theta_j}{\sum_{x_j \in \mathcal{X}_s} K(\tilde{x}_i, x_j)}$. Similarly, we obtain estimated nuisance parameters $\tilde{\phi}_i$. To embed a new point $\hat{x}_i \in \mathcal{X}_e$ in the embedding space Φ_e, we compute the c-th coordinate of $\Phi_e(\hat{x}_i)$ as $\Phi_e(\hat{x}_i)_c = \frac{\sum_{\tilde{x}_j \in \mathcal{X}_e} K(\hat{x}_i, \tilde{x}_j)\Phi_e(\tilde{x}_j)_c}{\sum_{\tilde{x}_j \in \mathcal{X}_s} K(\hat{x}_i, \tilde{x}_j)}$.

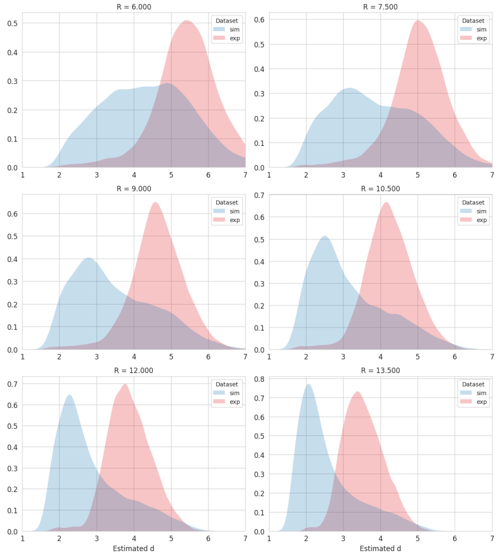
A.7 TSLasso Details

TSLasso [18] is an algorithm which recovers a subset f_S of $\mathcal{F} = \{f_k : \mathcal{M} \to \mathbb{R}, k = 1 : p\}$, where each $f_k \in \mathcal{F}$ represents a potential smooth coordinate function of a manifold $\mathcal{M}$. It does so by finding the subset $f_S \subseteq \mathcal{F}$ whose gradients, which must be either estimated or analytically computable, "most economically" span the tangent spaces of the manifold. More specifically, using a sample of points $x \in \mathcal{M}$, TSLasso first estimates the tangent spaces $T_x\mathcal{M}$, then it projects the gradients $\nabla f_k(x)$ onto these estimated tangent spaces, and finally it attempts to reconstruct a basis of $T_x\mathcal{M}$ using a linear combination of the projected gradients. To force a sparse representation of the tangent spaces over the whole sample, TSLasso regularizes the magnitudes of the linear coefficients B_k with the penalty being applied separately for each $k = 1 : p$. To select $f_S \subseteq \mathcal{F}$ with $|S| = d$, a series of Group Lasso problems is solved for different regularization strengths λ until exactly d linear coefficients B_k are non-zero.

In our experiments, each $f_k \in \mathcal{F}$ will represent one of the simulation parameters(the conformation θ or one of the nuisance parameters in ϕ), giving us $p = |\mathcal{F}| = 10$. We use $|S| = d = 4$ For the experimental data, we infer these value as in Appendix A.6. We run TSLasso 20 times using samples of size 500. Each run samples points in $\mathcal{X}_s$(or $\mathcal{X}_e$) which have SNRs(or inferred SNRs for the experimental data) in the top q-th percentile over all points. We perform the experiment for $q \in \{0, 5, \dots, 90, 95\}$. We find that f_S almost always consists of conformation θ(or $\tilde{\theta}$), SNR(or inferred SNR), and at least one of the quaternion rotation parameters in ϕ(or $\tilde{\phi}$). The full results and regularization paths are presented in Appendix Figure 6. Our results show that this combination of functions parametrizes both $\mathcal{M}_s$ and $\mathcal{M}_e$.

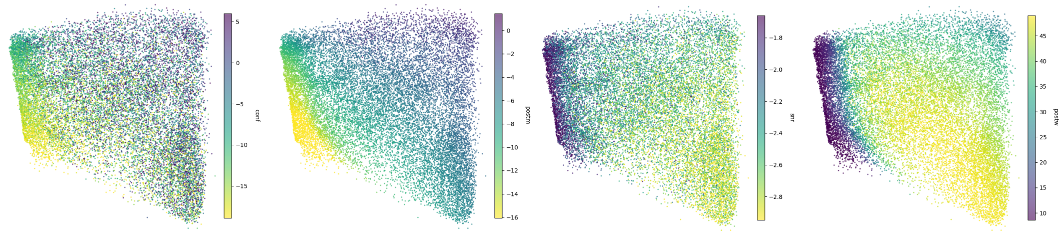
We use a simple procedure to estimate the gradients ∇f_k. We describe the procedure for the simulated data and note that for the experimental data we use the same procedure but the inferred values of the f_k's instead. For each point $x \in \mathcal{X}_s$, we perform weighted local PCA using the same kernel matrix used for Diffusion Maps. We select a local basis around x, $U(x) \in \mathbb{R}^{256 \times d'}$, consisting of the eigenvectors corresponding to the largest d' eigenvalues obtained during PCA. Let $\mathcal{N}_x$ be the set of neighbors of x in the kernel matrix and let $w(x')$ represent, for each $x' \in \mathcal{N}_x$, the entry $K(x, x')$ in the kernel matrix. We create a matrix $\Delta_x(x) \in \mathbb{R}^{256 \times |\mathcal{N}_x|}$, where each column corresponds to $w(x')(x' - x)$. We also create a vector $\Delta_{f_k}(x) \in \mathbb{R}^{|\mathcal{N}_x|}$ where each entry corresponds to $w(x')(f_k(x') - f_k(x))$. Then we solve for $y \in \mathbb{R}^{d'}$ as the weighted least squares solution in $\Delta_{f_k}(x) = [\Delta_x(x)^T U(x)]y$. Here y represents an estimation of the gradient $\nabla f_k(x)$ in the local coordinates $U(x)$. Then, we obtain our estimation as $\nabla f_k(x) = U(x)y$. In our experiments we use $d' = 10$.

B Appendix Figures



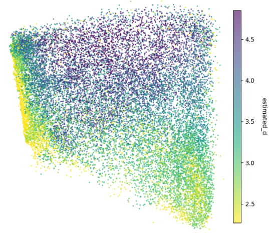
(a) Doubling Dimension

Appendix Figure 1: Estimation of the intrinsic dimension d_s (blue) and d_e (red) of the manifolds $\mathcal{M}_s$ and $\mathcal{M}_s$, respectively, using the doubling dimension for different radii R.



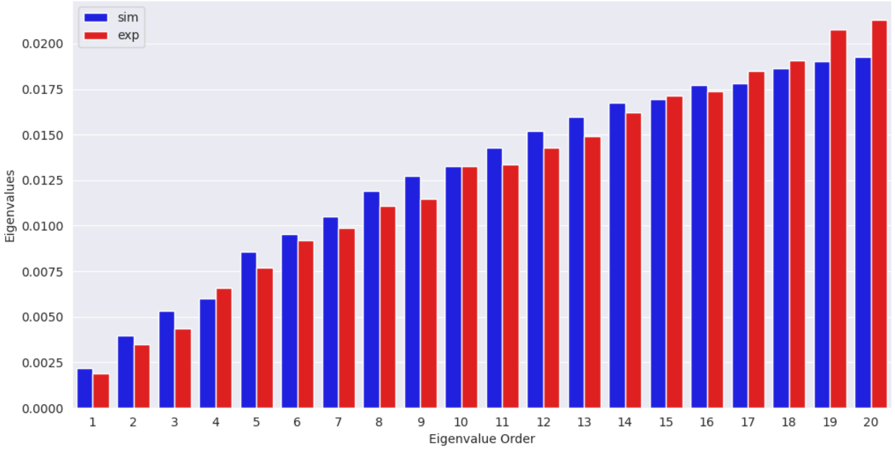
(a) Φ_s colored by θ (left) and $\hat{\theta}$ (right).

(b) Φ_s colored by SNR (left) and σ (right).

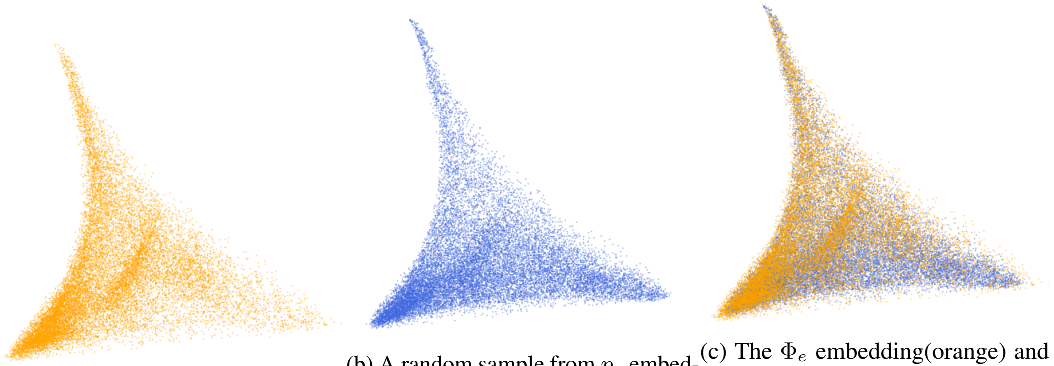


(c) Φ_s colored by d_s estimated by the Eigengap method.

Appendix Figure 2: Diffusion Maps embeddings Φ_s in $d = 3$ dimensions; the plots are rotated to best display the embedding. The three coordinates we display are selected by IES. In **(a)**, for data points with high SNR (the leftmost points), the conformation and posterior mean agree over the embedded points and vary smoothly across the y-axis. In **(b)**, the SNR and posterior width agree over the embedded points and vary smoothly across the x-axis. In **(c)**, the highest intrinsic dimension is in regions with medium SNR. For data with high SNR (the left most points), the intrinsic dimension d_s drops due to the lack of noise; for the noisiest data (lower right of embedding), d_s drops again, as noisy images become more similar to each other.



Appendix Figure 3: The spectrums of the experimental (red) and simulated (blue) eigendecompositions of the Laplacian matrix $\mathbf{L}$ obtained during Diffusion Maps. The smoothness of the spectrum and having only one 0 eigenvalue (not displayed) indicates that both $\mathcal{M}_s$ and $\mathcal{M}_e$ are smooth connected manifolds.

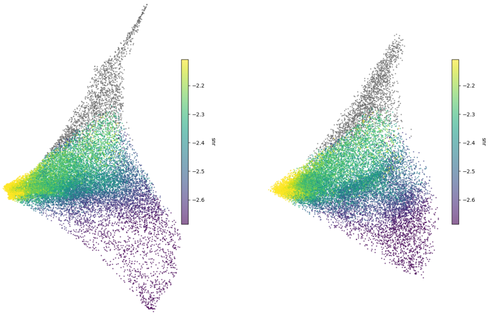


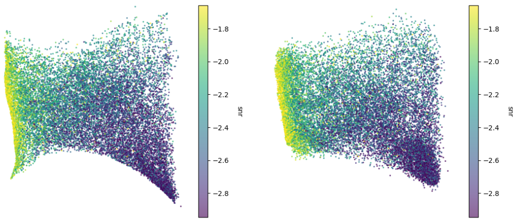
(a) The Φ_e embedding.

(b) A random sample from p_e embedded into the Φ_e space using kernel interpolation.

(c) The Φ_e embedding(orange) and a random sample from p_e(blue) embedded into the Φ_e space using kernel interpolation.

Appendix Figure 4: In **(b)**, we display a random sample from p_e, the density on $\mathcal{M}_e$, which we embed into the Φ_e space **(a)** using the kernel interpolation method presented in Appendix A.6. We observe that this sample has no gaps and no clusters. In **(c)**, we display the difference in density between the sample from p_e (blue) and the sample used to compute Φ_e (orange). This is due to the resampling method described in Appendix A.4 that aims to mimic a uniform distribution over $\mathcal{M}_e$. We note that p_e is much denser in the low SNR regions (see Figure 3). By sampling less from this noisy and uninformative region, we encourage Φ_e(orange) to better capture the geometry of $\mathcal{M}_e$.

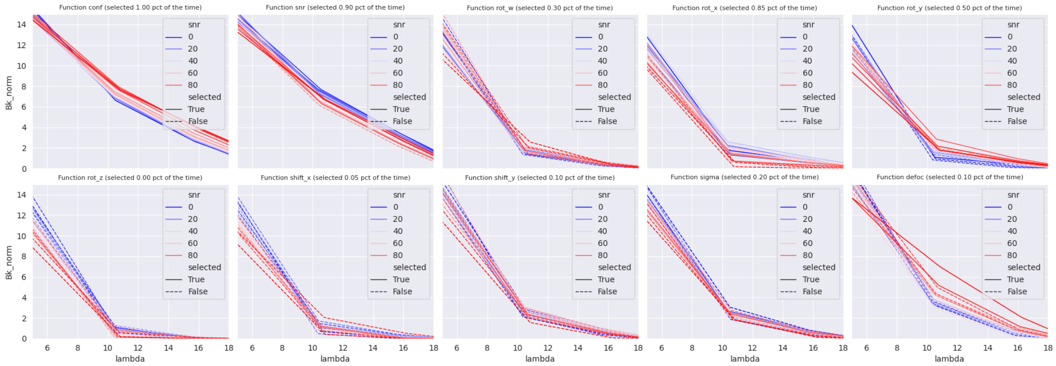




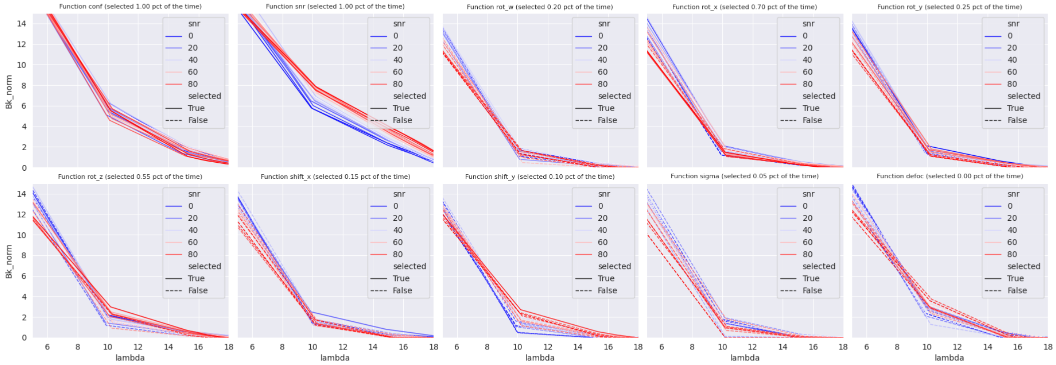
(b) Φ_s before(left) and after(right) Riemannian Relaxation.

(a) Φ_e before(left) and after(right) Riemannian Relaxation.

Appendix Figure 5: Diffusion Maps embeddings Φ_e **(a)** and Φ_s **(b)** before and after Riemannian Relaxation. The embeddings have been slightly rotated to emphasize the effect of the relaxation. Riemannian Relaxation tends to produce smoother embeddings with less curvature and more uniformly distributed points.



(a) TSLasso results for experimental data.



(b) TSLasso results for simulated data.

Appendix Figure 6: The regularization paths of each $f_k \in \mathcal{F}$ obtained over 20 runs of TSLasso for the experimental **(a)** and simulated **(b)** data. Each subplot corresponds to one function $f_k \in \mathcal{F}$, with the name and the selection rate in f_S being indicated in the sub-title. The x-axis represents the value of λ, the strength of the sparsity regularization, while the y-axis represents the average magnitude of B_k, the linear coefficents. Each run consists only of points in the top q-th percentile over all points in terms of SNR. We perform the experiment for $q \in \{0, 5, \ldots, 90, 95\}$ with the lines going from blue to red as q increases. A continuous (dotted) line indicates that f_k was selected (not selected, respectively) in that run. We find that f_S almost always consists of conformation θ (or $\tilde{\theta}$), SNR (or inferred SNR), and at least one of the quaternion rotation parameters in ϕ (or $\tilde{\phi}$).

.