

Cite this: *Digital Discovery*, 2023, 2, 702

Beyond ManifoldEM: geometric relationships between manifold embeddings of a continuum of 3D molecular structures and their 2D projections†

Evan Seitz, ^{ab} Joachim Frank^{*ab} and Peter Schwander ^{*c}

ManifoldEM is an established method of geometric machine learning developed to extract information on conformational motions of molecules from their projections obtained by cryogenic electron microscopy (cryo-EM). In a previous work, in-depth analysis of the properties of manifolds obtained for simulated ground-truth data from molecules exhibiting domain motions has led to improvements of this method, as demonstrated in selected applications of single-particle cryo-EM. In the present work this analysis has been extended to investigate the properties of manifolds constructed by embedding data from synthetic models represented by atomic coordinates in motion, or three-dimensional density maps from biophysical experiments other than single-particle cryo-EM, with extensions to cryo-electron tomography and single-particle imaging with a X-ray free-electron laser. Our theoretical analysis revealed interesting relationships between all these manifolds, which can be exploited in future work.

Received 16th November 2022
Accepted 21st March 2023

DOI: 10.1039/d2dd00128d

rsc.li/digitaldiscovery

1 Introduction

Biological molecules and their assemblies, including molecular machines, assume a continuum of conformational states as they go through work cycles required for executing their metabolic function.^{1,2} Single-particle cryo-EM³ of suitable *in vitro* systems affords the ability to collect a large number of projections that originate from an ensemble of 3D structures presenting a continuum of states in thermal equilibrium.^{4–8} This information, however, comes buried among typically hundreds of thousands of unorganized snapshot images, each a 2D projection of a molecular instance, corrupted by aberrations of the microscope and noise that often exceeds the signal by more than an order of magnitude. By virtue of machine learning algorithms, it is possible to determine a low-dimensional representation of the conformational spectrum, with leading coordinates corresponding to each of the system's degrees of freedom.

From the numbers of sightings of the observed states in this low-dimensional space, a free-energy landscape may be obtained following a fundamental relationship of statistical mechanics.^{4,6,9} Together this provides a complete mapping of

the system's state space while articulating its energetics topographically in the form of sprawling hills and valleys.^{10,11} From such a landscape, a minimum-energy path can be derived representing the most probable sequence of transitions taken by the molecular machine between any two states.^{12,13} And along this path, a sequence of 3D structures can be extracted for biophysical analysis, allowing the basis for molecular function to be elucidated. The ability to experimentally determine energy landscapes, in conjunction with 3D conformational movies of a molecular machine along distinct low-energy paths, opens a new horizon in structural biology and, by extension, in molecular medicine.

In general, obtaining this desired information is a difficult task, and numerous methods have been developed over the last decade to tackle this problem.^{5,14–16} The ManifoldEM method^{5,17} we have adopted is based on an unsupervised geometric machine learning approach using manifold embedding to recover the distribution and occupancies‡ of states. Its viability has been demonstrated in its application to four experimental systems—the ribosome,⁵ ryanodine receptor (RyR1),¹⁷ vacuolar ATPase,¹⁸ and SARS-CoV-2 spike protein.¹⁹ In each of these studies, conformational states of the respective macromolecular complex were characterized by different spatial constellations of its relatively rigid domains, and organized in a state space according to the continuous motions of each domain along a unique coordinate.

‡ We use the term “occupancy” throughout the manuscript, which corresponds to the number of states in conformational space. It is related to the conformational density of states $\rho(C)$, so that the number of states in a small interval $[C \pm \Delta C]$ is proportional to $\rho(C) \times \Delta C$.

^aDepartment of Biochemistry and Molecular Biophysics, Columbia University Medical Center, New York, NY 10032, USA. E-mail: ees2192@columbia.edu; jj2192@cumc.columbia.edu

^bDepartment of Biological Sciences, Columbia University, New York, NY 10027, USA

^cDepartment of Physics, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA. E-mail: pschwan@uwm.edu

† Electronic supplementary information (ESI) available: Data and code repository. See DOI: <https://doi.org/10.1039/d2dd00128d>



Recently, we devised a protocol using ensembles of simulated macromolecular complexes to interpret the conformational manifolds learned by two prominent dimensionality-reduction methods: principal component analysis²⁰ (PCA) and diffusion maps²¹ (DM, used in ManifoldEM), with a strong focus on the latter. With the knowledge gained from this analysis, we were able to introduce several advancements to the ManifoldEM approach in a novel method called ESPER (Embedded Subspace Partitioning and Eigenfunction Realignment), which we validated using experimental data of the ribosome and RyR1 obtained *via* cryo-EM.²² Altogether, ESPER constitutes a substantial improvement of the original ManifoldEM framework that advances the ability to accurately and efficiently construct the free-energy landscape of macromolecular complexes from cryo-EM data.²² Our *post hoc* interpretation strategy, however, was only briefly summarized in that work.

The current article brings our interpretation strategy center stage to detail the analysis that proved instrumental in edifying our understanding of these complex manifold geometries. In contrast to the scope of our companion article,²² which deals solely with the properties of 2D projections of macromolecular conformers, we also now focus on their 3D structures represented either by electron density maps or sets of atomic-coordinate structures. Ultimately, we embark on an in-depth analysis of different types of synthetic-continuum datasets and their embeddings, with each “datatype” modeled so as to emulate the output of a different experiment from the same conformations of a molecular machine.

Importantly, the insights gained from our analysis of these datatypes are not limited to the field of cryo-EM, but also apply to other biophysical methods supplying information on conformational continua. In order to put the properties of these datatypes into the context of established theories, we further frame our results in comparison to a simple approximative mathematical model and derive useful explicit relationships. In the process, many striking similarities emerge between the results from various datatypes, and when differences do emerge, the information is revelatory.

Through this analysis, we are able to demonstrate the effect of different choices of continuum datatypes on the performance of geometric machine learning, which has significant implications for the reconstruction of complex information obtained from a range of widely used experimental methods. Given these results, we are able to highlight the power of manifold embedding in elucidating important biophysical properties—provided that minimum requirements are met in the data—while describing several key challenges and limitations that must be considered when analyzing datasets with multiple types of statistical uncertainties. Finally, informed by these implications, we provide a general outlook on the future of these methods within the scope of cryo-EM and beyond.

1.1. Additional background

As described in our companion work,²² we first devised a protocol for generating ensembles of simulated cryo-EM images derived from an atomic-coordinate structure (ACS) of

a known protein. Specifically, we chose the heat-shock protein Hsp90 due to its simple design, exhibiting two arm-like domains (chain A and B) connected in an overarching V-shape which are known to naturally undergo large conformational changes.²³ We altered the 3D atomic coordinates of this molecule in its closed state (PDB 2CG9)²⁴ so as to exercise multiple independent easily-identifiable conformational motions§ (CMs¶) with $n = 1, 2$ or 3 degrees of freedom. As one example, to simulate the motions for $n = 2$, we independently altered the positions of the molecule's two large arm-like domains using equispaced rotations of each domain about its hinge-residue axis, which we refer to as CM_1 and CM_2 , respectively. By exercising these domain motions independently in all combinations, we generated a set of $M_2 = 400$ structures—each represented by its atomic coordinates in PDB format—altogether spanning a 20×20 state space (SS_2 , with intrinsic dimensionality $n = 2$).

These M_n structures were then transformed into simulated 3D electron density|| maps (EDM) with a resolution of 3 Å, where each atom is represented by a 3D Gaussian with radius defined by the corresponding number of electrons.²⁵ Using parallel line integrals along the direction of the electron beam, 2D projections of each EDM were obtained to simulate weak-phase contrast images as generated in a transmission electron microscope (TEM) operated in bright-field mode.³ These projections were generated for evenly-spaced viewing angles across the entire angular space, so that a full set of images depicting M_n conformational states of the molecule was available for each projection direction (PD). Each projection was further modified by application of the contrast transfer function (CTF), duplication to create multiple sightings per conformational state, and addition of noise to simulate realistic images of a cryo-EM experiment.

Selected ensembles of such simulated single-particle cryo-EM images were then independently embedded in a low-dimensional space—forming a manifold of intrinsic dimension n —and the resulting spectral properties²⁶ of each embedding were analyzed and compared. Using first pristine (*i.e.*, noiseless and aberration-free) synthetic data, we observed surprising patterns in the geometries of these embeddings. For instance, each PD-manifold followed a clearly-defined high-dimensional parabolic surface, but also included eigenfunction harmonics and other unexpected aberrations. The same geometric features were present for datasets with additive noise applied to the images, with the fidelity of these features to the pristine case dependent on the signal-to-noise ratio (SNR), the

§ In the literature, there is a wide range of nomenclature used here among fields and, in some instances, works by the same authors. For clarity, the following terms are interchangeable: conformational motions (CMs); conformational coordinates (CCs); reaction coordinates (RCs); collective motion coordinates.

¶ A tabulated description of symbols and abbreviations used throughout this document is available in ESI.†

|| In reality, cryo-EM data (under the regime of weak phase contrast in bright-field transmission electron microscopy) are projections of the Coulomb potential distribution, which is distinct from the electron density distribution determined in X-ray crystallography. For the present analysis, however, this distinction is irrelevant.



number of times τ images were generated for each state M_n (totaling $N_n = \tau M_n$), and introduction of CTF.

But it was only through a much deeper investigation of the manifolds of atomic-coordinate structures and 3D electron density maps of the same conformer—which we aim to present in this article—that a complete understanding of these manifolds was obtained. Using this knowledge, we were able to determine the exact form of the eigenfunctions in each PD-manifold, including identification of each set of harmonics and high-dimensional manifold rotations.

Guided by results of these ground-truth studies on both pristine and noisy data encompassing up to three degrees of freedom, the compendium of our heuristic findings provided new insights into the origin of longstanding ManifoldEM problems, leading to the development of the ESPER method for correcting them. Along with several novel operations and refinements to the preexisting approach, the ESPER method provides a thorough rationale for properly handling the n -dimensional PD-manifold embeddings in the presence of experimentally-relevant noise, CTF and several degrees of freedom to accurately generate the free-energy landscape of a molecular machine as well as 3D movies depicting its function.

2 Relationships between manifolds formed for different representations of continuum data

In this section, we expose our findings on ACS, EDM and PD continuum datasets in detail. As the analysis leading to these findings is rooted in established spectral theory,^{26–28} we encourage readers to peruse the ESI Section A,† where we detail both the PCA and DM approach, and for the latter, introduce the Laplace-Beltrami Operator (LBO) and define key parameters such as the Gaussian bandwidth ϵ . Briefly, we note that the DM eigenvectors Ψ converge to the eigenfunctions ψ of the LBO on a manifold Ω , sampled at the given data points, and carry useful information about the intrinsic geometry of Ω .

As an outline of our exposition, in Section 2.1, we use the DM method to investigate the known eigenfunctions of the LBO on the interval and rectangular domains and compare these results to the manifolds formed by a quasi-continuum of atomic-coordinate structures. Following this analysis, we detail how the structure of manifolds Ω obtained from a conformational state space transforms as the data type is changed successively from atomic-coordinate structures (Ω_{ACS}) in Section 2.2, to 3D density maps (Ω_{EDM}) in Section 2.3, and finally to 2D projections (Ω_{PD}) of those maps in Section 2.4. A schematic of this framework is provided in Fig. 1, showing how the Hsp90 ground-truth data are innately represented by each of these distinct datatypes.

2.1. Eigenfunctions of the latent space

We first introduce the simplest possible continuum model, which we construct using a spatially delimited set of points X_n in \mathbb{R}^n , forming a metric space with Euclidean metric²⁹ defined by the equation

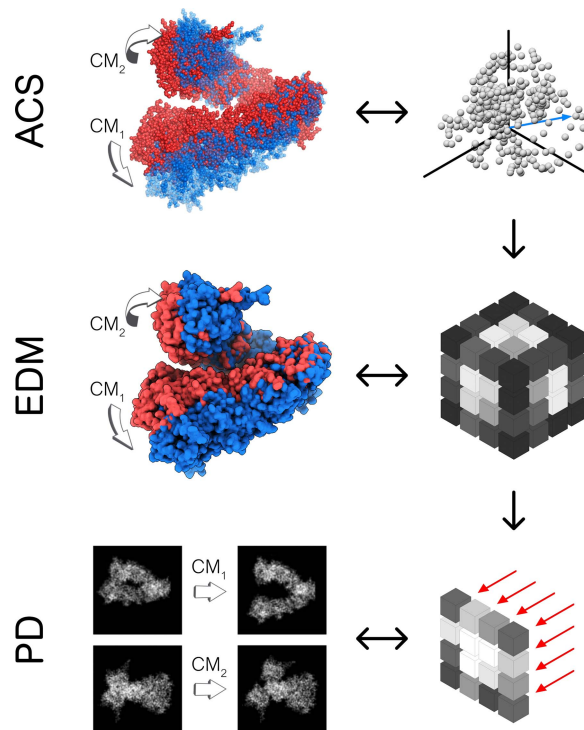


Fig. 1 Outline of our framework for generating continuum data composed in three distinct datatypes: 3D atomic-coordinate structures (top); 3D electron density maps (middle); and 2D projections of 3D electron density maps (bottom). In the left-hand column are the respective appearance of these datatypes for the Hsp90 continuum models, showing the type and general range of motion for two conformational motions (CM_1 and CM_2). On the right is a corresponding schematic, showing how the continuum models are represented by the respective datatype. Specifically, a collection of 3D coordinates (top) defines the position of each atom, while a set of 3D voxels (middle)—or pixels (bottom)—defines the electron density within each cubic—or its projection in a square—region in space (to note, our schematic showing the projection operation along a principal coordinate is strictly for conceptual aid; in application, projections occur along any direction).

$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\| = [(a_1 - b_1)^2 + \dots + (a_n - b_n)^2]^{1/2} \quad (1)$$

which gives the “standard” distance between any two vectors \mathbf{a} and \mathbf{b} in Euclidean n -space. Specifically, we define X_n as a Cartesian product of equispaced points along each linear dimension: for example, a line of points for $n = 1$, or an array of points arranged inside a rectangle for $n = 2$. Altogether, we term these elements and their relations the latent n -space.

At first glance, such a latent space seems to have little relation to the relatively complex synthetic continuum dataset of Hsp90 conformers. However, this abstraction is motivated by the representation of our ground-truth state space of atomic models, where the relationship between equispaced coordinates in the latent space matches the relationship between equiangular molecular-domain rotations. By embedding the data of the atomic-coordinate structures of Hsp90 conformers occupying SS_n and comparing the resulting eigenvectors to those obtained from the embedding of the data of the latent



space X_n , we will show in Section 2.2 that the results are nearly identical. We will then demonstrate in Section 2.3 and 2.4 that for the embeddings of data from the 3D electron density maps as well as their 2D projections, the relation relative to the latent space becomes distorted, which can be explained by a change of the metric in the process of switching from one data type to the other.

2.1.1 State space 1. First, we will show that in the 1D space X_1 , a set of pairwise distances between a collection of equispaced coordinates on a line carries all essential information necessary to model the pairwise distances between a sequence of atomic models in SS_1 with a molecular domain rotated by a constant angular increment. To represent our SS_1 dataset, we uniformly sample $N = 50$ equispaced points from a 1D interval $X_1 \in [0, \ell = 1] \subset \mathbb{R}$, with each of these points representing a unique state of the molecule. Following the DM method, we then calculate the distance matrix for this collection of points and embed the data in a low-dimensional space, spanned by the leading eigenvectors. Furthermore, we will show that two characteristic regimes emerge depending on the choice of

Gaussian bandwidth, which we will denote with ε_\downarrow and ε_\uparrow for the small and large regime, respectively.

For the regime of small Gaussian bandwidths, a cosine series emerged for all eigenvectors (Fig. 2A), in very good agreement with the Laplacian eigenfunctions on a 1D Euclidean interval with Neumann boundary conditions. Specifically, we anticipate and retrieve canonical eigenfunctions²⁸ of the form $\psi_\nu(x) = \{\cos(\nu\pi x/\ell) | \nu \geq 1\}$. As the Gaussian bandwidth was incrementally increased from ε_\downarrow to ε_\uparrow , this cosine series smoothly transformed into a different complete, orthogonal set: the Legendre polynomials³⁰ (Fig. 2B). However, we note that these polynomials only occur for boundary conditions of hyperrectangles, which are n -dimensional Cartesian products of orthogonal intervals.

2.1.2 State space 2. Next, to represent our SS_2 dataset, we uniformly sample $N = 50 \times 50$ points from a 2D interval $X_1 \times X_2 \in [0, \ell_x = 1] \times [0, \ell_y = 1.1] \subset \mathbb{R}^2$, where the operator \times denotes the Cartesian product.²⁹ For ease of illustration, we avoid degeneracy by ensuring that $(\ell_x/\ell_y)^2$ is not a simple ratio.²⁸ Again, we follow the DM method by calculating the

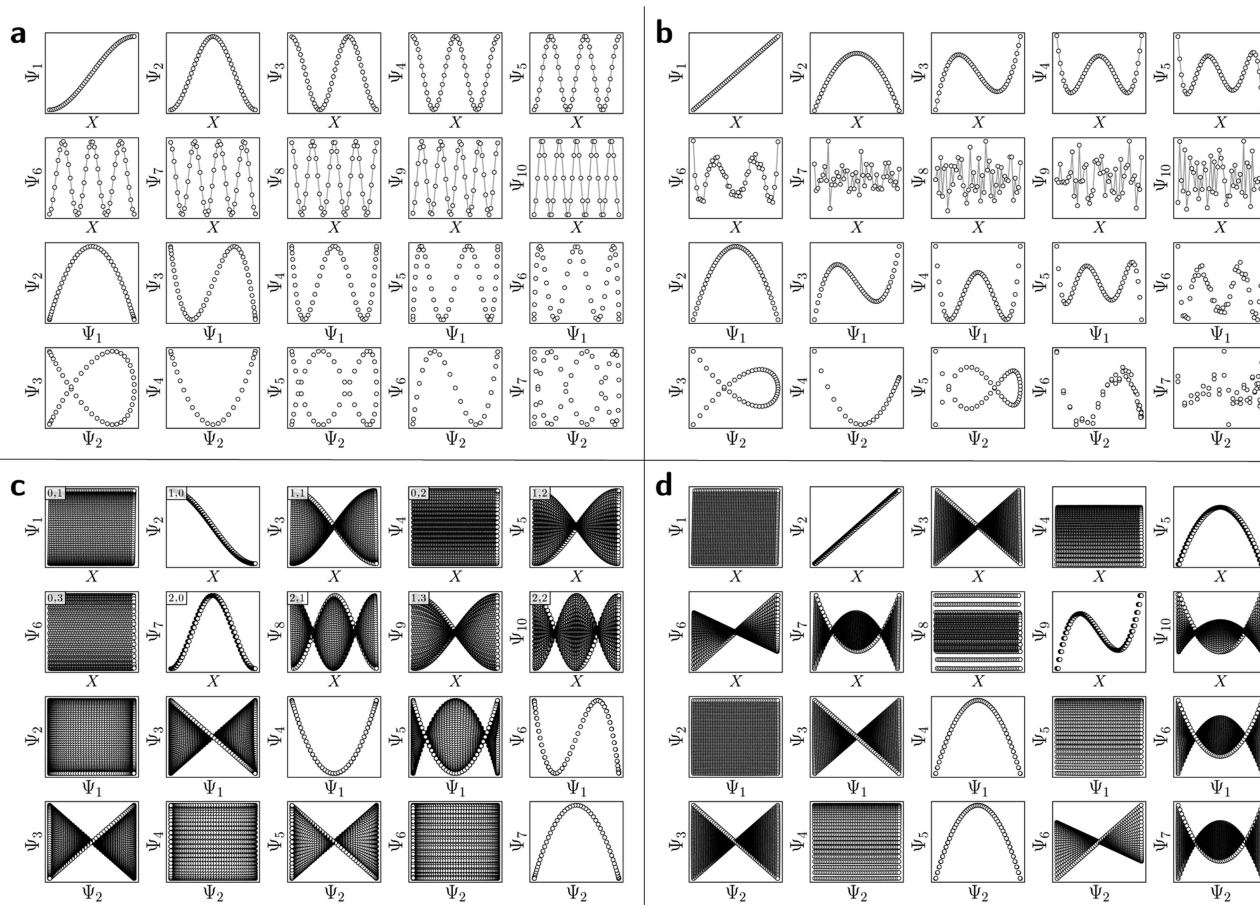


Fig. 2 DM eigenvectors in the nondegenerate latent 1-space and 2-space. The DM eigenvectors of the 1D interval for small ($\varepsilon_\downarrow = 5 \times 10^{-5}$) and large ($\varepsilon_\uparrow = 10$) Gaussian bandwidths are shown in (a) and (b), respectively. Likewise, eigenvectors of the $N = 50 \times 50$ rectangular (nondegenerate) domain for small and large Gaussian bandwidth are shown in (c) and (d), respectively. As will be done throughout this text, eigenfunctions have been independently displayed by indexing each by its ground-truth ordering (here *via* sequential x -coordinates). For (c) and (d), a similar appearance of eigenfunction plots, albeit interchanged, would be seen when indexing instead *via* sequential y -coordinates. In (c), each eigenfunction's corresponding modes (ν, w) have also been provided in the top left-hand corner. For all four subplots, pairwise combinations of eigenfunctions are additionally shown, which can be visualized after an embedding without any ground-truth knowledge.



pairwise distances between these points and embedding the data in a low-dimensional space. As demonstrated in Fig. 2C, the set of eigenvectors obtained in the smaller Gaussian bandwidth regime matched our expectations for the Laplacian eigenfunctions on a rectangular domain with Neumann boundary conditions. These canonical eigenfunctions are

$$\psi_{vw}(x, y) = \{ \cos(v\pi x/\ell_x) \cos(w\pi y/\ell_y) \mid v, w \geq 0 \}, \quad (2)$$

which follow the same pattern for higher-dimensional domains $\Omega_R = [0, \ell_1] \times \dots \times [0, \ell_n] \subset \mathbb{R}^n$ (with $\ell_i > 0$).²⁸ Again, as we incrementally increased the Gaussian bandwidth from ε_1 to ε_\uparrow , this set of complete and orthogonal cosines smoothly transformed into the set of orthogonal Legendre polynomials, which are now functions of both x and y , as expected (Fig. 2D). Importantly, the leading Legendre polynomials provide a direct linear mapping of the input data points, which is a consequence of the linear terms $P_1(x)$ and $P_1(y)$. While these linear relationships are more convenient than the cosine form, we will show in Section 2.3 and 2.4 that they are absent in the embeddings of 3D EDMs and 2D projections.

It is important to take notice that, alongside leading eigenfunctions in the first two rows of each subplot in Fig. 2, the leading composites of these eigenfunctions $\{\Psi_i \times \Psi_j \mid i < j\}$ have also been plotted in the rows that remain, with each composite forming a unique 2D subspace. Mathematically, each such mapping to a 2D subspace is the restriction to the N -dimensional embedding of the projection of \mathbb{R}^N onto \mathbb{R}^2 ; given by $\{\Psi_1 \times \Psi_2 \times \dots \times \Psi_N\} \rightarrow \{\Psi_i \times \Psi_j\}$ (for expediency, we will use the term subspace to specifically refer to a subspace of an embedded manifold). Of interest, among the available subspaces, a leading parabolic trajectory exists for each degree of freedom present; for example, $\{\Psi_1 \times \Psi_4\}$ and $\{\Psi_2 \times \Psi_7\}$ in Fig. 2C, which correspond to the sequence of states along Y_1 and X_1 , respectively. While less significant for the scope of the current section, the study and use of these 2D subspaces will be crucial in the sections to come, especially when dealing with experimental data, which we describe below.

Specifically, as the points in an experimental dataset naturally arrive in unordered sequence, one would have to properly sort the dataset indices to recognize the sinusoids shown in Fig. 2A and C; here, for example, there would be 50! sequences to consider. In the application, even if the ground-truth ordering was obtained, then in the presence of duplicate states (which we anticipate in an experiment), each sinusoid would be irregularly stretched along the horizontal axis where those duplicate states occurred, forming an unwieldy distorted sinusoidal form. However, as the points in each eigenvector are always scrambled in the same order for all eigenvectors, the composite of any two will always appear in a readily identifiable form. For example, as seen in Fig. 2A, a subset of the canonical Lissajous curves³¹ emerges across the 2D subspaces of each Ω_{PD} , with the curves in this set having the form

$$L_{p,q} = \{ \cos(p\pi x) \times \cos(q\pi x) \mid 0 \leq x \leq 1 \} \quad (3)$$

such that $p \neq q \in \mathbb{Z}^+$.

For these composites, we found that information pertaining to the given degree of freedom is portrayed most simply (without overlap) along a specific subset of L , here as seen across the set of 2D subspaces defined in pairwise combination with the leading eigenvector; e.g., $\{(\Psi_1 \times \Psi_2), (\Psi_1 \times \Psi_3), \dots, (\Psi_1 \times \Psi_z)\}$, where z is the index of the smallest non-zero eigenvalue. Specifically, this subset $T_k \in L$ corresponds to the known Chebyshev polynomials of the first kind,³⁰ of which we observed that the parabolic form is the lowest-order member present in each Ω_{PD} embedding.

Given their significance, these 2D subspaces have several important properties worth highlighting for their eventual use (or avoidance) during interpretation. First, note that for each sinusoidal subplot in Fig. 2A, points are equispaced along the horizontal axis in correspondence with the equidistance between points in the corresponding latent space. However, as a result of taking the Cartesian product of sinusoidal eigenfunctions, only non-uniform spatial relationships exist between neighbouring states in each $L_{p,q}$; a relationship described by a non-isometric mapping²⁹ where distances in a domain are not preserved in its codomain. As shown in Fig. 3, the spacing between points in $L_{1,2}$, which is the composite of two such sinusoids, has an intrinsically nonuniform distribution. The density of points is similarly arranged as seen in the

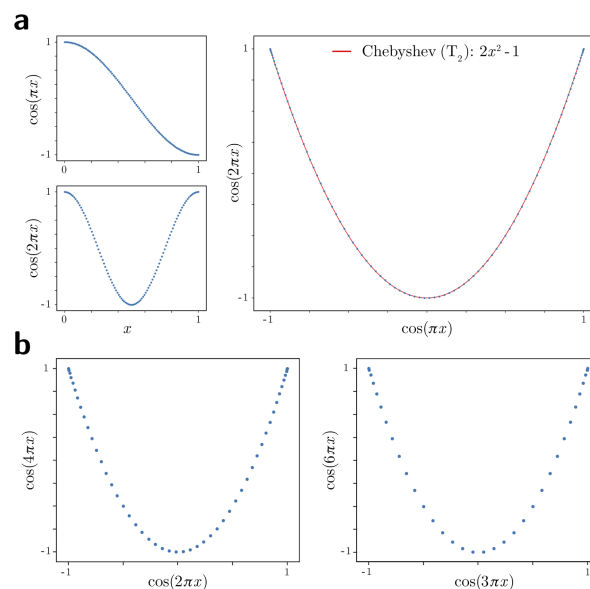


Fig. 3 Analytical generation and analysis of Lissajous curves. The analytical generation of the Lissajous curve $L_{1,2} = \{\cos(\pi x) \times \cos(2\pi x) \mid \text{uniform } x \in [0, 1]\}$ is shown in (a). Note the naturally-induced nonuniform spacing between points near the boundaries and vertex of the parabola. As a simple demonstration, we also fit this curve with the Chebyshev T_2 polynomial, which is a subset of the Lissajous curves; however, T_2 does not share the same nonuniformity in spacing as $L_{1,2}$. In (b), parabolic harmonics are likewise generated for $L_{2,4}$ and $L_{3,6}$. While the same x -coordinates were used to generate all underlying cosines for parabolas in both (a) and (b), more than one point in the domain ends up mapping to each coordinate of these parabolic harmonics. As such, these harmonics obfuscate the true conformational information, which is intact on $L_{1,2}$.



corresponding point clouds. For general reference, we denote this aspect with the term nonuniform rates of change.

Next for consideration, as seen for example in Fig. 2A, there exist several parabolic trajectories scattered throughout the 2D subspaces. However, only the first of these parabolas (here, $\{\Psi_1 \times \Psi_2\}$) describes the full extent of the variational information present monotonically, while all other trailing parabolas (such as in the $\{\Psi_2 \times \Psi_4\}$ subspace shown) display a non-monotonic signal. As an analytical demonstration, Fig. 3 shows that the first three such parabolas can be generated *via* $L_{1,2}$, $L_{2,4}$ and $L_{3,6}$. The latter two repeat the conformational information twice and three times, respectively, within one span of the parabolic trajectory.

As a consequence, only the mapping from the sinusoids to the first parabola in this set is bijective (injective and surjective),²⁹ while all other mappings to higher-order parabolas are non-injective surjections. Importantly, since the Cartesian product of continuous functions is continuous, and so are projections from product spaces, this bijection further meets the requirements of a homeomorphism: a bijective correspondence that

preserves the topological structures involved.²⁹ We denote the higher-order parabolas (formed *via* the non-injective surjections) as parabolic harmonics, which do not preserve topological structure and must be avoided when determining a given degree of freedom; a problem that becomes more challenging as more degrees of freedom are added to the system.

In general, the presence of these patterns extends into latent n -spaces with $n > 1$. For every degree of freedom present in a state space SS_n , there exists a corresponding set of Lissajous curves interspersed across specific $\{\Psi_i \times \Psi_j\}$ projections of the corresponding embedding. Specifically, in the case of Fig. 2C, independently projecting the data for SS_2 onto the planes spanned by its $\{\Psi_1 \times \Psi_j\}$ and $\{\Psi_2 \times \Psi_j\}$ combinations (where $i > 1; j > 2$) reveal a unique set of Chebyshev polynomials, with the sequence of points along these trajectories corresponding to CM_1 and CM_2 . For convenience, a set of Chebyshev polynomials³⁰ corresponding to a given CM will be referred to as its conformational modes). Thus, even though the knowledge required to view these sinusoids is unavailable outside of ground-truth studies, one can rely on their existence—*via* the

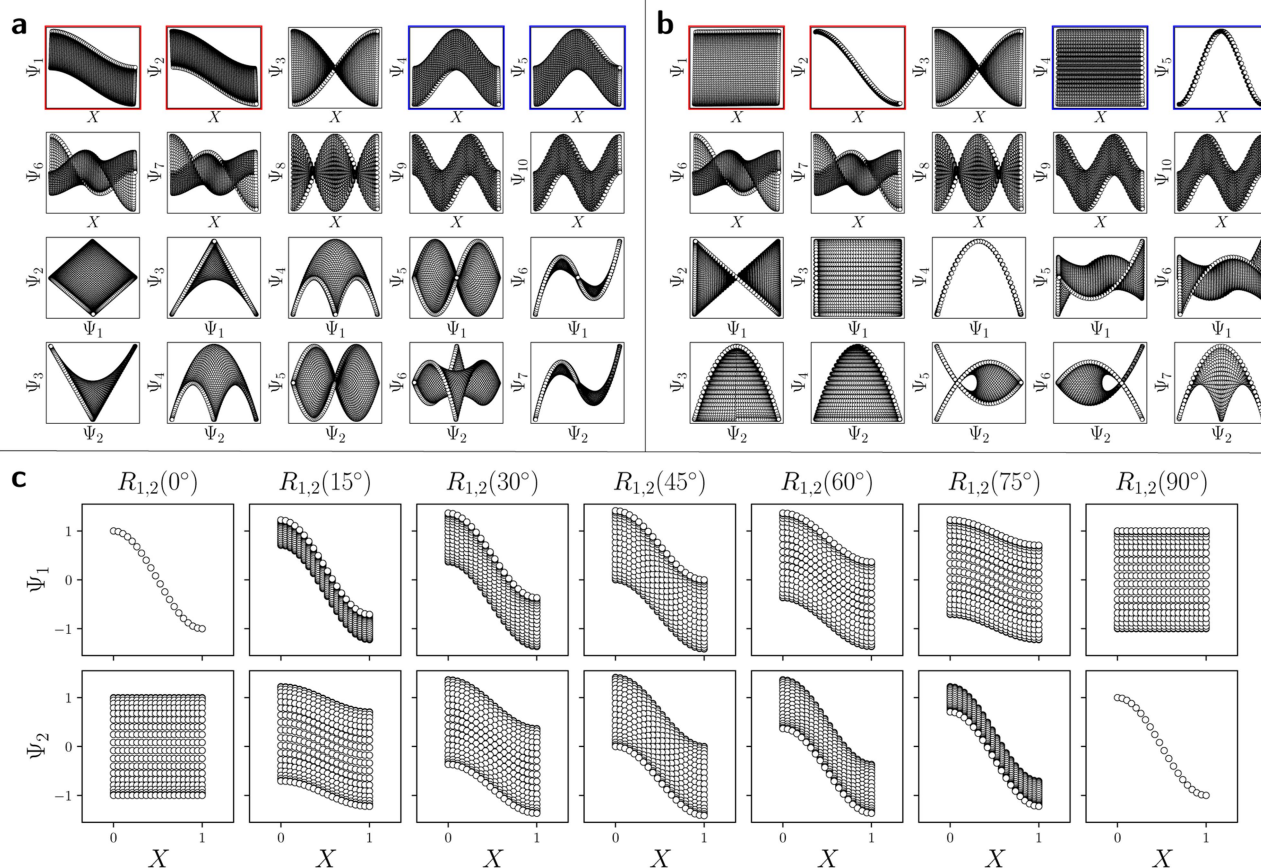


Fig. 4 DM eigenvectors of the degenerate latent 2-space. DM eigenvectors of the $N = 50 \times 50$ square domain for small Gaussian bandwidth ($\epsilon_1 = 5 \times 10^{-5}$) are shown in (a) and (b) before and after high-dimensional rotations, respectively. It can be seen here that pairs of eigenfunctions exist that contain relationships aberrant to the canonical eigenfunction form seen in Fig. 2C. Two such pairs have been highlighted in red and blue, respectively, with the members of each pair always rotated 90° apart. To note, as any rotation can happen in the presence of degeneracy, this initial rotation is an arbitrary one. We demonstrate this property *via* the schematic in (c), which shows the angular relationship between two analytically-generated functions ($\cos(\pi x)$ and $\cos(\pi y)$), each displayed in the reference frame of states in X_1 as they are jointly rotated 90° . By applying rotation operators $R_{1,2}(\theta) = 45^\circ$ and $R_{4,5}(\theta) = 45^\circ$ independently to two such aberrant pairs in (a), the canonical eigenfunction form begins to recover in (b), and more so as additional operators are appropriately applied.



composites of carefully chosen eigenvectors—to elucidate conformational type and order.

With these details set aside, we next return to the smaller of the two Gaussian bandwidth regimes, in order to compare the previous nondegenerate rectangular results to those from a degenerate square domain, with $N = 50 \times 50$ points equispaced identically along X and Y (Fig. 4A). Due to the presence of degenerate eigenvalues, which can arise for domains with a rational ratio $(\ell_x/\ell_y)^2$, we encounter pairs of eigenfunctions that appear different from the nondegenerate case of the rectangle.²⁸ This can be seen, for example, by eigenvector pairs $\{\Psi_1, \Psi_2\}$ and $\{\Psi_4, \Psi_5\}$ in Fig. 4A. In Fig. 4C, we illustrate that these eigenfunctions are just rotated within their degenerate space, exactly as expected. We note that an eigenfunction associated with a degenerate eigenvalue is a linear combination of the degenerate eigenfunctions,²⁸ where the normalization of the eigenfunctions restricts this linear transformation to a rotation and reflection (*i.e.*, the group of orthogonal transformations). For example, the $\{\Psi_1, \Psi_2\}$ eigenvector pair is of form $\Psi' = R^T\Psi$ such that

$$\begin{bmatrix} \Psi'_1(\theta) \\ \Psi'_2(\theta) \end{bmatrix} = \begin{bmatrix} \cos(\theta)\cos(\pi x) + \sin(\theta)\cos(\pi y) \\ -\sin(\theta)\cos(\pi y) + \cos(\theta)\cos(\pi x) \end{bmatrix} \quad (4)$$

As seen for eigenvector Ψ_6 in Fig. 4A, these summands can also have the form of two products: $\Psi_6 = \beta_1 \cos(\pi x) \cos(2\pi y) + \beta_2 \cos(2\pi x) \cos(\pi y)$, with any β_1 and β_2 such that $\beta_1^2 + \beta_2^2 \neq 0$. Hence, it can be seen that these aberrant eigenfunction pairs are defined by an admixture of cosines in a higher-dimensional space, with form

$$\Psi_i = \beta_1 \cos(v\pi x) \cos(w\pi y) + \beta_2 \cos(w\pi x) \cos(v\pi y) = \beta_1 \psi_{vw} + \beta_2 \psi_{wv} \quad (5)$$

By using an appropriate rotation operator $R_{i,j}$, the summands within each eigenfunction pair can be maximally separated between two members $\Psi_i = \psi_{vw}$ and $\Psi_j = \psi_{wv}$, such that the canonical eigenbasis is recovered (Fig. 4B). As demonstrated using analytical expressions $\psi_{1,0} = \cos(\pi x)$ and $\psi_{0,1} = \cos(\pi y)$ in Fig. 4C, this separation occurs multiples of $\theta = 90^\circ$ apart. In the Fig. 4C example, at $R_{1,2}(45^\circ)$, these eigenfunctions have the form

$$\begin{bmatrix} \Psi'_1(\theta = 45^\circ) \\ \Psi'_2(\theta = 45^\circ) \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 \cos(\pi x) + \sqrt{2}/2 \cos(\pi y) \\ -\sqrt{2}/2 \cos(\pi y) + \sqrt{2}/2 \cos(\pi x) \end{bmatrix} \quad (6)$$

which decouples back into two distinct modes (*i.e.*, $\cos(\pi y)$ and $\cos(\pi x)$ for Ψ_1 and Ψ_2 , respectively) at $R_{1,2}(90^\circ)$. A similar result is obtained by applying this operation on the appropriate eigenvectors obtained *via* DM, with each initially assuming a random rotation angle (Fig. 4A) requiring a specific correction $R_{i,j}(\theta)$, as seen in Fig. 4B.

While degeneracy is a coincidence which can be directly identified from the eigenvalue spectrum, a similarly rotated appearance (*i.e.*, eigenfunction misalignment) will later turn up during our investigation of PD manifolds. Pairs of misaligned eigenfunctions, at least approximately, can also appear when

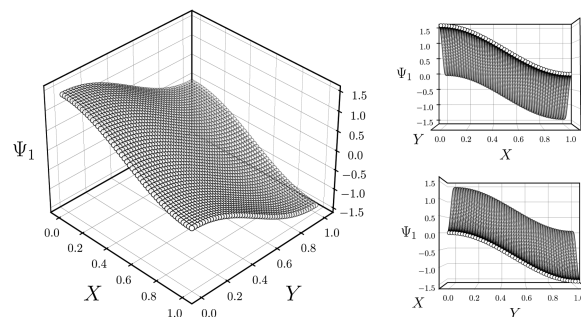


Fig. 5 Intuition for sequential ordering of eigenfunctions based on ground truth. Subplots are analytically generated so as to match the appearance of Ψ_1 in Fig. 4A. For this presentation, the equation $\Psi_1 = \cos(\theta) \cos(\pi x) + \sin(\theta) \cos(\pi y)$ was used with $\theta = 45^\circ$. As seen in the largest subplot, the eigenfunction exists in an n -dimensional space defined by the n degrees of freedom of the system. By displaying points in sequence corresponding to a known degree of freedom, here X or Y , we are effectively viewing each eigenfunction on a projected plane in its n -dimensional space.

domains have undergone certain elementary geometric transformations. For example, by performing an affine transformation on a rectangle Ω_R to form a parallelogram Ω_P , we observed a rotation of the first two eigenvectors, as similarly seen in Fig. 4A. Recall that an affine mapping preserves collinearity and ratios of distances, but in general not distances and angles.³² In Section 2.3, we will explore the possibility of other classes of transformations.

As a final point in this section, we illustrate our method for retrieving the canonical eigenfunctions buried within an embedding, which has been used in Fig. 2 and 4, and extensively throughout the remainder of this work. Fig. 5 provides a schematic using the known analytical eigenfunctions chosen so as to match the results from DM on the square (degenerate) Ω_R . As shown, a given sequence corresponding to the ground-truth arrangement of points along each degree of freedom (here X_1 or Y_1) captures the eigenfunction on a projected plane in the n -dimensional space where it resides.

2.2. Eigenfunctions of the 3D atomic-coordinate structures

We next investigate the manifolds obtained from the state spaces formed from a quasi-continuum of atomic-coordinate structures, each represented by a set of m 3D atomic-coordinates (*e.g.*, as visualized in Fig. 1). Importantly, the set of these 3D atomic-coordinate structures in $\mathcal{Q}_{ACS} \subset \mathbb{R}^{3m}$ represents the structural identity of each state, from which the cryo-EM experiment could only obtain two-dimensional information in the form of images.

2.2.1 State space 2. Following the DM approach, we first calculated the distance matrix for SS_2 , which we obtained by the root-mean-square deviation (RMSD) for each pair of the $M_2 = 400$ atomic-coordinate structures (as given by their PDB files). The RMSD between two atomic-coordinate structures $A = (a_1, a_2, \dots, a_m)$ and $B = (b_1, b_2, \dots, b_m)$, each composed of m atoms, is defined as



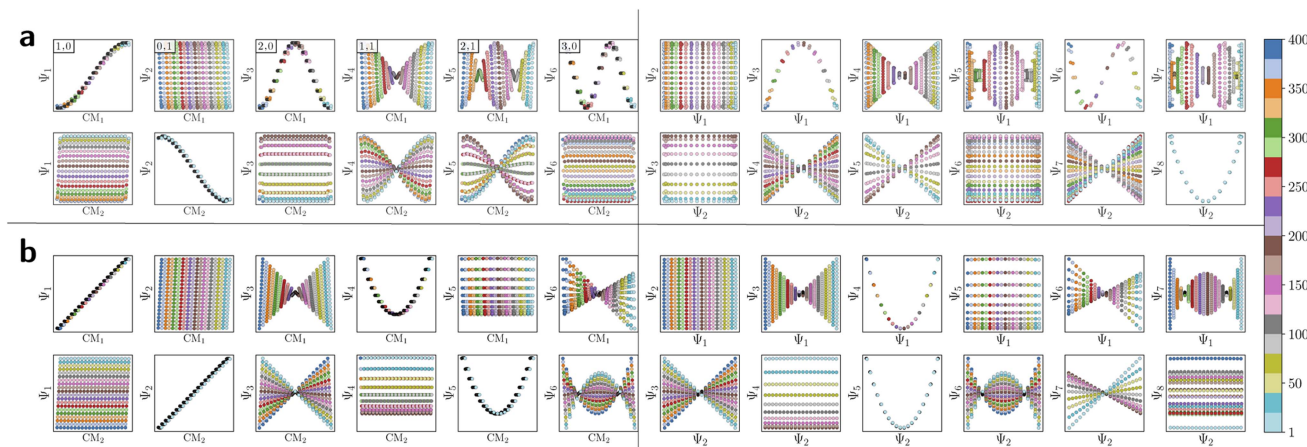


Fig. 6 DM eigenvectors for the quasi-continuum of atomic-coordinate structures. Eigenvectors obtained for $20 \times 20 = 400$ atomic models occupying SS_2 for small ($\varepsilon_{\dagger} = 0.1$) and large ($\varepsilon_{\dagger} = 1000$) Gaussian bandwidths are shown in (a) and (b), respectively. Leading eigenvectors are displayed in the first and second rows *via* sequential indexing along the ground-truth $CM_1 \in [1, 400]$ coordinates (*i.e.*, equispaced rotations of chain A) and $CM_2 \in [1, 400]$ coordinates, respectively. The color mapping illustrates the position of ground-truth states, where CM_1 points can be seen following along the full spectrum of colors (indices 1–400) while CM_2 points are approximately uniform in color map value (*i.e.*, light blue for points in front, with all other colors similarly underlaid with indices a multiple of 1–20). In (a), the modes $\{v, w\}$ corresponding to each eigenvector are provided in the top left-hand corner, showing exceptional agreement with the LBO eigenfunctions on a rectangular domain with Neumann boundary conditions. We additionally note the absence of any significant eigenfunction misalignments.

$$\begin{aligned} \text{RMSD}(A, B) &= \sqrt{\frac{1}{m} \sum_{i=1}^m \|a_i - b_i\|^2} \\ &= \sqrt{\frac{1}{m} \sum_{i=1}^m \left((a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2 + (a_{iz} - b_{iz})^2 \right)} \end{aligned} \quad (7)$$

which is up to a trivial constant factor of $m^{-1/2}$ equal to the Euclidean distance eqn (1).

The resulting DM embeddings for the small and large Gaussian bandwidth regimes, which are shown in Fig. 6A and 6B, respectively, share a strong resemblance with those found for the latent space (Fig. 2). Again, we note the presence of cosine eigenfunctions for the small Gaussian bandwidth regime, and a nearly-perfect linear form (*via* leading Legendre polynomials) in the large Gaussian bandwidth regime; *i.e.*, $\{\Psi_1, \Psi_2\}$ in Fig. 6B. For the latter, we will show that such a linear form cannot be obtained in the other data types, 3D EDMs and 2D PDS, to be explored. In the small Gaussian bandwidth regime, we can identify both CM_1 and CM_2 parabolas residing in the subspaces $\{\Psi_1 \times \Psi_3\}$ and $\{\Psi_2 \times \Psi_8\}$, respectively. Similar results—albeit for different dimensions—were found for the Ω_{ACS} embeddings from SS_1 and SS_3 .

Briefly, we note that the results of PCA on this same dataset most closely resembles the geometries observed for DM in the large Gaussian bandwidth regime (in this instance, Legendre-like), with a slightly less uniform distribution. This suggests that DM better approximates intrinsic relationships in the data, as expected. A general similarity between the geometries observed using PCA and DM in the ε_{\dagger} regime will continue to manifest throughout the remaining sections of this article.

The striking similarity between the eigenfunctions of the latent space and the eigenfunctions of the atomic models can

be rationalized as follows. Provided that the range of a single-body rotation is moderate ($\leq 30^\circ$), the distance D_{ij} between any two states i and j within this range is to a good approximation $D_{ij} = \Theta_{ij} \left(\sum_{k=1}^m r_k^2 \right)^{1/2}$, where r_k is the distance of atom k away from the rotation axis, m the number of atoms of the body, and Θ_{ij} the angular difference between the states. Therefore, D_{ij} is to a good approximation directly proportional to Θ_{ij} . If there are multiple independent body rotations (*i.e.*, CMs) present, the individual distances add in quadrature as in a Euclidean space. While not further discussed in this article, the linearity also holds for rigid body translations, where the distance is precisely proportional to the magnitude of the translation. Thus, the agreement between the eigenfunctions of the latent space and the ones of the atomic models is a direct consequence of the approximately linear relationship between distance and extents of multi-body motions (*i.e.*, moderate rotations and translations of any magnitude).

2.3. Eigenfunctions of the 3D density maps

We next investigated how the conformational relationships between states are changed when representation of structures by atomic coordinates is transformed into one by 3D electron density maps (EDMs; as visualized in Fig. 1). To this end, we generated the EDMs for each of the 3D atomic-coordinate structures for all previously-defined state spaces. We next calculated the pairwise Euclidean distances between these EDMs in $\mathcal{Q}_{EDM} \subset \mathbb{R}^V$, with V the number of voxels, and performed an embedding *via* the DM method.

2.3.1 State space 2. Over a wide range of Gaussian bandwidths, the structure of the resulting eigenfunctions is very similar to the structure of eigenfunctions retrieved for the



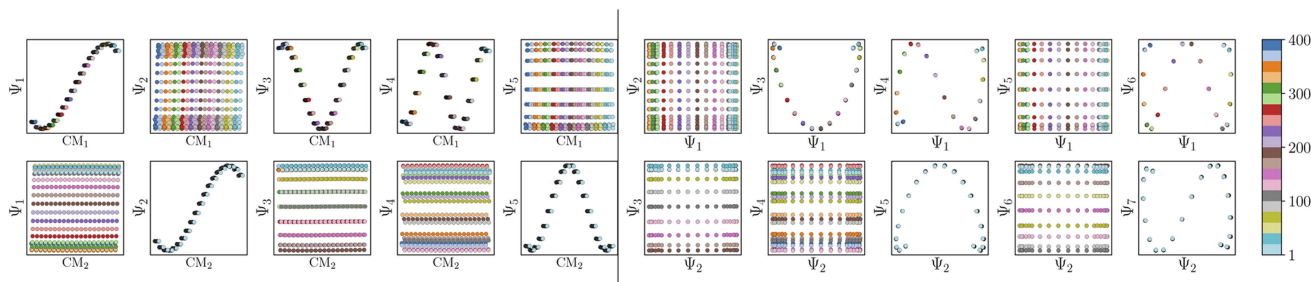


Fig. 7 DM eigenvectors for the quasi-continuum of electron density maps. The results of the DM embedding of pristine EDMs from SS_2 are shown. Leading eigenvectors as indexed by $\text{CM}_1 \in [1, 400]$ and $\text{CM}_2 \in [1, 400]$ are displayed in the first two rows on the left, followed by their composites on the right. Overall, there is near-perfect alignment of these eigenvectors with the canonical eigenfunctions, such that no eigenvector rotations are required. As an aside, the pronounced inward curling at the boundaries of certain subspaces (e.g., $\{\Psi_1, \Psi_3\}$) is due to insufficient sampling.

atomic models in the small Gaussian bandwidth regime. Importantly, as in Section 2.2, these eigenfunctions are still of the form ψ_{vw} , with subspaces having no significant appearance of eigenfunction misalignments.

However, there are a few attributes to consider that distinguish the manifolds obtained for EDMs from those retrieved for the previous data types. First, the difference between small and large Gaussian bandwidth regimes is much less drastic, such that only the cosine eigenfunctions appeared in both regimes. For small Gaussian bandwidth regimes (*i.e.*, a few orders of magnitude below the optimal value ε , determined by the bandwidth estimation method), we found that the leading CM_2 eigenfunctions are buried deeply in low-ranking eigenvectors (e.g., Ψ_8 and higher), with numerous CM_1 eigenfunctions occupying the eigenvectors in between. In addition, eigenvectors with cross terms $\Psi_i = \{\psi_{vw} \mid v, w \neq 0\}$ were scattered mostly in mid-range positions (e.g., Ψ_{12} and higher).

In contrast, for larger Gaussian bandwidth regimes (*i.e.*, near and significantly above ε), eigenvectors with cross-terms are buried in much deeper subspaces (e.g., Ψ_{34} and deeper), with the majority of leading eigenvectors housing content exclusively for either CM_1 ($w = 0$) or CM_2 ($v = 0$). These CM eigenfunctions also have a near-perfect distribution of points, whereas for the ε_1 regime, the distribution of points has noticeably less precision to the ideal form. Notably, the embeddings obtained above and below these regimes are incoherent in form.

We conclude that the eigenfunctions obtained from the larger Gaussian bandwidth regime would be preferred for several reasons. First, the desired CM_1 and CM_2 parabolas occupy leading subspaces and are thus easily identifiable. The paucity of leading cross-term eigenfunctions is also convenient, since they provide no useful information for our analysis while also obfuscating our search for desired subspaces. Additionally, the geometric structure of all subspaces obtained *via* ε_{\uparrow} consistently appears much closer to the canonical form. In Fig. 7, we display the DM eigenfunctions obtained from this regime for the $20 \times 20 = 400$ EDMs occupying SS_2 . Subspaces indexed in the CM_1 reference frame (row one on the left in Fig. 7) and the CM_2 reference frame (row two, left) are displayed, as well as a set of leading composites of these eigenfunctions—forming 2D subspaces—on the right-hand side of the figure.

Importantly, as there was no Gaussian bandwidth value that could ‘recover’ the preferred Legendre-like form, it appears that this feature is lost upon transformation from atomic models to EDMs, caused by a change in the metric (to note, the curved geometry formed by cosines was also in close agreement with the results of applying PCA on this same dataset). As a main agent for this distinction, the distance measure pertaining to EDMs is fundamentally different from the one for the 3D atomic coordinates. Instead of the 3D coordinate points that stand for the positions of atoms in each structure, the data for each EDM is represented by a 3D array of values, one at each voxel (Fig. 1). A key difference, then, is that in the latter case the displacement of atoms is no longer accounted for individually. Instead, every voxel in the 3D map of one state is now compared to a voxel at the same location in another state, with only changes in the value at each voxel entering the distance measure.

Hence, while the eigenfunctions are similar, the relationship between states in these two data types is fundamentally different. To demonstrate this change, Fig. 8 shows a comparison of the pairwise distances between states as calculated for the rectangular latent space, atomic-coordinate structures, and EDMs. As noted in the caption, by assessment of the close similarity between the distances from the latent space and atomic models, we can infer that these two data types are both confined to the rectangular manifold Ω_R (albeit of different sizes). More precisely, the metric space corresponding to ACS conformers in SS_n and to elements in X_n are very similar even though the two datatypes are very different. Since only the metric enters the LBO, their eigenfunctions are nearly identical.

In contrast, we see that the distances from the EDMs are starkly different from the rectangular pattern, where neighboring states are spatially arranged *via* an asymptotic-like trend. From these findings, we must infer that the corresponding data ‘live’ in an altogether different manifold. Although the explicit geometric form of Ω_{EDM} is unknown, we have shown that the leading eigenfunctions of the Laplacian on Ω_{EDM} are essentially preserved *via* the mapping from the latent space. While detailed knowledge of Ω_{EDM} is certainly of interest, it is inconsequential here since our analysis only requires an understanding of the leading eigenfunctions of a manifold.



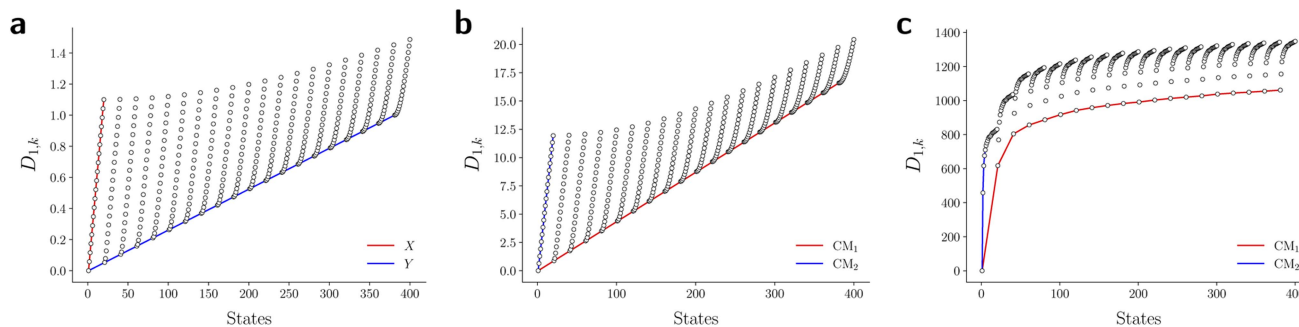


Fig. 8 Comparison of metric for the three data types. The first row of the distance matrix D is plotted for the rectangular Euclidean space (a), the 3D atomic-coordinate structures in SS_2 (b), and the EDMs in SS_2 (c). Given our ordering of states, the first row $D_{1,k}$ corresponds to the pairwise distance calculated between state 01_01 and all 400 states. For the pattern in (a), which was calculated on a rectangular domain $\Omega_R \in [0, 1] \times [0, 1.1]$, one can identify the distance of the first state ($x_1 = 0, y_1 = 0$) to all other coordinates, such that the red line depicts the left-hand side of the rectangle (with maximum distance $D((x_1, y_1), (x_{20} = 1.1, y_{20} = 0)) = 1.1$), and the blue line depicts the rectangle's base (with maximum distance 1). In (b), a similar rectangular pattern arises for the RMSD values calculated between atomic models. The pattern in (c), however, is starkly different from (a) and (b), such that no rectangular (or rectangle-like) domain could be drawn to reproduce this trend.

2.4. Eigenfunctions of the 2D projections

Before providing a detailed description of the eigenfunctions of the LBO on $\Omega_{PD} \in \mathbb{R}^P$, with P being the number of pixels, we first provide a general analysis of their relationship with those from previously-established datatypes (*i.e.*, latent n -space, ACS and EDMs). For similarities, as was observed for the EDMs, we found that eigenfunction characteristics could be broadly grouped into two classes *via* either the small or large Gaussian bandwidth regime. In either regime, the eigenfunctions of the PD manifolds are again of the form ψ_{vw} , such that only cosines emerge. The lack of the Legendre-like form and a similar asymptote-like appearance of distances between images suggests that the PD states in Ω_{PD} reside on a manifold similar to Ω_{EDM} .

The overall difference between eigenfunctions obtained *via* ε_{\downarrow} and ε_{\uparrow} is also much more relevant for PDs than for the EDMs. In the small Gaussian bandwidth regime, CM_2 subspaces exhibit a severely suboptimal point distribution, such that in some PDs, identification of the CM_2 parabola-housing 2D subspace is completely obstructed. These CM_2 subspaces are also buried in trailing eigenvectors and interspersed among those with cross-terms. We also note that the value determined by the bandwidth estimation method (ε_{\cdot}) provides similar spectral properties, making it a suboptimal choice for pristine data. In contrast, the large Gaussian bandwidth regime (*i.e.*, one order of magnitude larger than ε_{\cdot} and spanning numerous orders of magnitude above it) proves superior in every sense, with CM_1 and CM_2 eigenfunctions having ideal point distributions and corresponding subspaces occupying leading eigenvectors. And again, the cross-term eigenfunctions are present only in far trailing eigenvectors (*e.g.*, Ψ_{31} and higher), and would thus not be obstructive during the analysis.**

Regardless, for either Gaussian bandwidth regime, we found that significant eigenfunction misalignments can emerge—

with varying magnitude—depending on the direction of projection. The eigenfunctions of an example PD, determined with an appropriate choice of Gaussian bandwidth, are shown in Fig. 9. By plotting the points in each eigenvector in the specific ground-truth sequence constructed for CM_1 against a uniform index (for the M_2 states in SS_2 indexed *via* $\{1, 2, 3, \dots, 400\}$), a similar display of sinusoids and grid-like patterns emerges as seen in Fig. 7, albeit now misaligned with the eigenvector basis (to note, PCA returned similar spectral features for all PD datasets explored, as demonstrated extensively in our companion article).²²

Since we have previously shown that no such property is apparent in the manifold embeddings generated from the 3D EDMs from which these PDs originate, this suggests that the emergence of these eigenfunction misalignments must be tied to a changed metric in the PD manifolds. Specifically, as different 2D projections are taken of the EDMs *via* $p: \Omega_{EDM} \rightarrow \Omega_{PD}$, the geometry of Ω_{EDM} can become contorted due to the change of pairwise interatomic distances resulting from foreshortening in projection (Fig. 13 in ESI†), such that the apparent span of one CM to another depends on PD. We next describe these abnormalities in greater detail, before quantifying them exactly.

First, it is unclear what effect these eigenfunction misalignments may have on interpretability, since the frames of reference used so far (*e.g.*, as demonstrated in Fig. 9) are unavailable without *a priori* knowledge. To understand the effect of these eigenfunction misalignments on the PD manifolds, a collection of Ω_{PD} embeddings are next analyzed in greater detail and introduced in sequence of increasing intrinsic dimensionality. Notably, the explicit expressions derived in the previous sections will be used to account for the geometric structures observed in each PD embedding, which generally describe perturbations of a hypersurface spanned by the quasi-continuous conformations.

For the following analysis, sets of pristine images in five chosen PDs are obtained. The first PD is chosen to be normal to

** Upon introduction of noise (SNR of 0.1) and $\tau = 5$ noisy duplicates of each state, ε_{\cdot} becomes the most suitable choice (along with numerous orders of magnitude above it), with anything below this range completely inadequate.



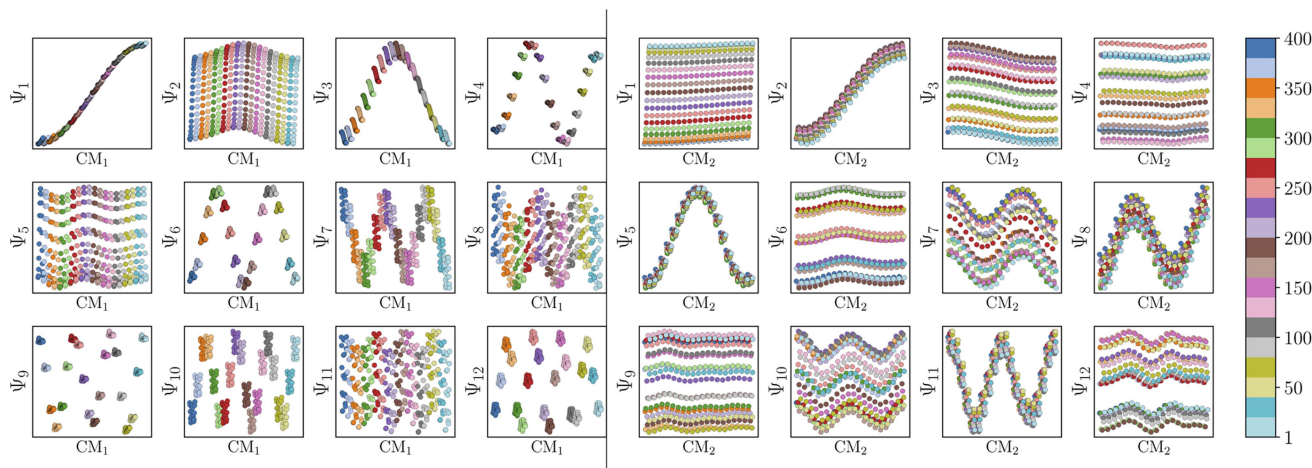


Fig. 9 Analysis of eigenfunctions for an example PD in SS₂. On the left are the sinusoidal forms $\cos(k\pi x)$ that emerge for only a specific subset of eigenvectors ($k = 1, 3, 4, 6, 7, 9, \dots$) when points in each Ψ_k are ordered precisely in the sequence of $CM_1 \in [1, 400]$ as assigned when the ground-truth images were initially constructed. Likewise, on the right, when points in each Ψ_k are instead ordered in the sequence of $CM_2 \in [1, 400]$, a new set of sinusoids emerge ($k' = 2, 5, 8, \dots$) specifically for those remaining Ψ_k not in the previous CM_1 subset. Hence, it can be seen that by systematically ordering the points in each eigenvector in sequence along each degree of freedom present, the corresponding set of sinusoids emerge in the frame of reference of the given degree of freedom.

the plane of the CM_1 rotation, such that all CM_1 motions from that perspective only produce changes in the plane of the projection. A similar choice is made for PD₂, which is projected onto the plane of CM_2 motions (CM_2 only exercised for $n > 1$), with the remaining three PDs chosen along arbitrary directions. For each state space, a set of these M_n images is generated from one of these five PDs to form an n -manifold Ω_{PD} existing in a high-dimensional space \mathbb{R}^P (as a specific example, for SS₁, there are five PD manifolds with 20 images each, while for SS₂, a separate set of five PD manifolds are formed with 400 images each). We next embed data *via* the DM method with Euclidean metric and, using the eigenfunctions of the LBO, analytically

quantify the trajectories of our simulated conformational changes as embodied by the spectral geometry of each Ω_{PD} .

2.4.1 State space 1. We first generated a different DM embedding using a suitable ε value within the range discovered for each of the five PD manifolds in SS₁. Each of the resultant point clouds contains 20 points, with each point corresponding to an image of a conformational state from CM_1 . We next ordered the eigenvector coordinates in each Ω_{PD} embedding to correspond to the ground-truth sequence of CM_1 states; *i.e.*, as understood *via* Fig. 5. As anticipated given our previous analysis, we observed the canonical eigenfunctions of the LBO on the interval $[0, 1]$ subject to Neumann boundary conditions

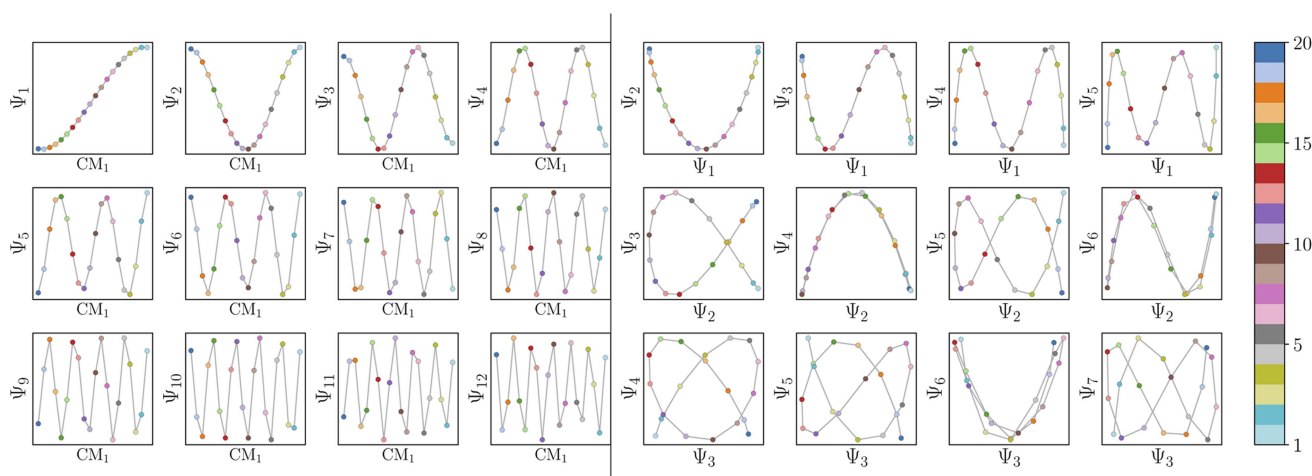


Fig. 10 Analysis of eigenfunctions for PD₁ in SS₁. On the left are the sinusoidal forms ψ_k that emerge when points—corresponding to images—in each eigenvector are ordered precisely in the sequence $CM_1 \in [1, 20]$ in which their ground-truth images were constructed. Regardless of any knowledge of such a sequence, the composites of these eigenvectors will always form well-defined geometries *via* the Lissajous curves, as shown on the right. In the first row on the right are the Chebyshev polynomials of the first kind, of which the parabola $\{\Psi_1 \times \Psi_2\}$ is the simplest mapping of the conformational information present. Figure source: Institute of Electrical and Electronics Engineers (IEEE) article²² (with cosmetic alterations) under licence CC BY 4.0.



$$\psi_k = \{\cos(k\pi x) \mid 0 \leq x \leq 1; k \in \mathbb{Z}^+ \leq N\}, \quad (8)$$

where x is the conformational coordinate represented by a number in the interval $[0, 1]$. For demonstration, we plot each of the 1D points in a given eigenvector as a function of a uniform index $I \in [1, 20]$ (for the total of 20 states in SS_1), making sure that the ordering of points in 1D follows the sequence assigned by the ground-truth index of its corresponding image along CM_1 . As seen in Fig. 10, when the collection of points in each eigenvector are ordered appropriately, the eigenfunction's sinusoidal form emerges along the full extent of the degree of freedom present (*i.e.*, mapping $I \mapsto x \in [0, 1]$).

We next compared these sets of 2D subspaces among the five PDs, and found only subtle differences in the distribution of their point clouds. It is important to underscore here the natural discrepancies between different PD manifolds that should be expected—due to what we term PD disparity—which will continue to manifest in several significant forms throughout this analysis. Naturally, as each 2D projection provides an incomplete representation of the underlying 3D density map, depending on the type of motion and its component on the PD under investigation, ground truth is preserved to different degrees. This disparity affects all Ω_{PD} characteristics and will become more relevant as we investigate the embeddings of datasets generated from structures with multiple degrees of freedom.

2.4.2 State space 2. To further understand these trends for embeddings formed with increasing intrinsic dimensionality, we next investigated the embeddings generated for SS_2 . As previously shown in Fig. 9A, by plotting the points in each eigenvector in the specific ground-truth sequence constructed for CM_1 against a uniform index, a similar but now interspersed

pattern of sinusoids appears. Specifically, the appearance of the sinusoids (with increasing $k \in \mathbb{Z}^+$) only manifest in a subset of all eigenvectors present, while for all other eigenvectors grid-like patterns emerge. These findings concur with our analysis throughout Section 2, where it was shown that such patterns arise as a consequence of viewing each eigenfunction on a projected plane in n -dimensional space (Fig. 5).

As a demonstration of this property, we next reordered the indices of points within all eigenvectors to instead correspond to the specific ground-truth sequence constructed for CM_2 (*i.e.*, $\{1, 21, 41, \dots, 381\}, \dots, \{20, 40, 60, \dots, 400\}$). The output of this operation can be seen in Fig. 9 on the right, which manifests a new subset of interspersed sinusoids, with increasing $k' \in \mathbb{Z}^+$ independent from the previous subset; and inhabiting only those eigenvectors in the complement of the CM_1 subset. By induction—based on these observations and those in Section 2.1—we conclude that for n degrees of freedom in a given Ω_{PD} , there are n independent sets of sinusoids ψ_k . Each set $\{\psi_k^\gamma \mid \gamma \in \mathbb{Z}^+ \leq n\}$, denoted by an index γ per degree of freedom, is interspersed throughout the collection of available eigenvectors $\{\Psi_i \mid i \in N\}$.

Also, in agreement with Section 2.1, we found that for every CM present in the state space, there exists a corresponding set of Chebyshev polynomials. Specifically, in the case of PD_1 , independently projecting the data for SS_2 onto the planes spanned by its $\{\Psi_1 \times \Psi_j\}$ and $\{\Psi_2 \times \Psi_j\}$ combinations (where $i > 1; j > 2$) revealed a unique set of these polynomials, with the sequence of points along these trajectories corresponding to CM_1 and CM_2 (Fig. 11). With this knowledge in hand, we next compared the subset of eigenfunctions as obtained in either the reference frame of CM_1 or CM_2 (Fig. 9, left or right-hand side, respectively) with the Chebyshev polynomials in Fig. 11.

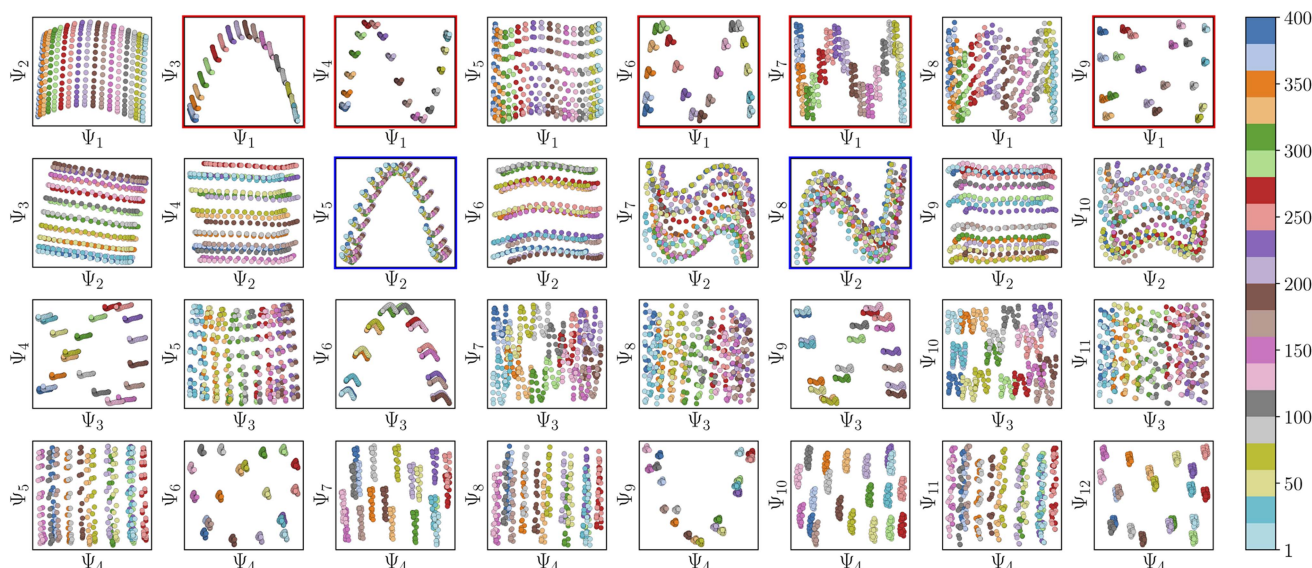


Fig. 11 A subset of the space of 2D subspaces for PD_1 in SS_2 . As demarcated in red and blue boxes, a set of conformational modes exists for both $CM_1 \in [1, 400]$ (red boxes, $\{\Psi_1 \times \Psi_j\}$) and $CM_2 \in [1, 400]$ (blue boxes, $\{\Psi_2 \times \Psi_j\}$); where $i > 1$ and $j > 2$), interspersed throughout each row. Additionally, note the occurrence of the first parabolic harmonic for CM_2 located at $\{\Psi_3 \times \Psi_6\}$. Similar plots for the remaining four PDs are provided in (Fig. 14 in ESI†). Figure source: Institute of Electrical and Electronics Engineers (IEEE) article²² (with cosmetic alterations) under licence CC BY 4.0.



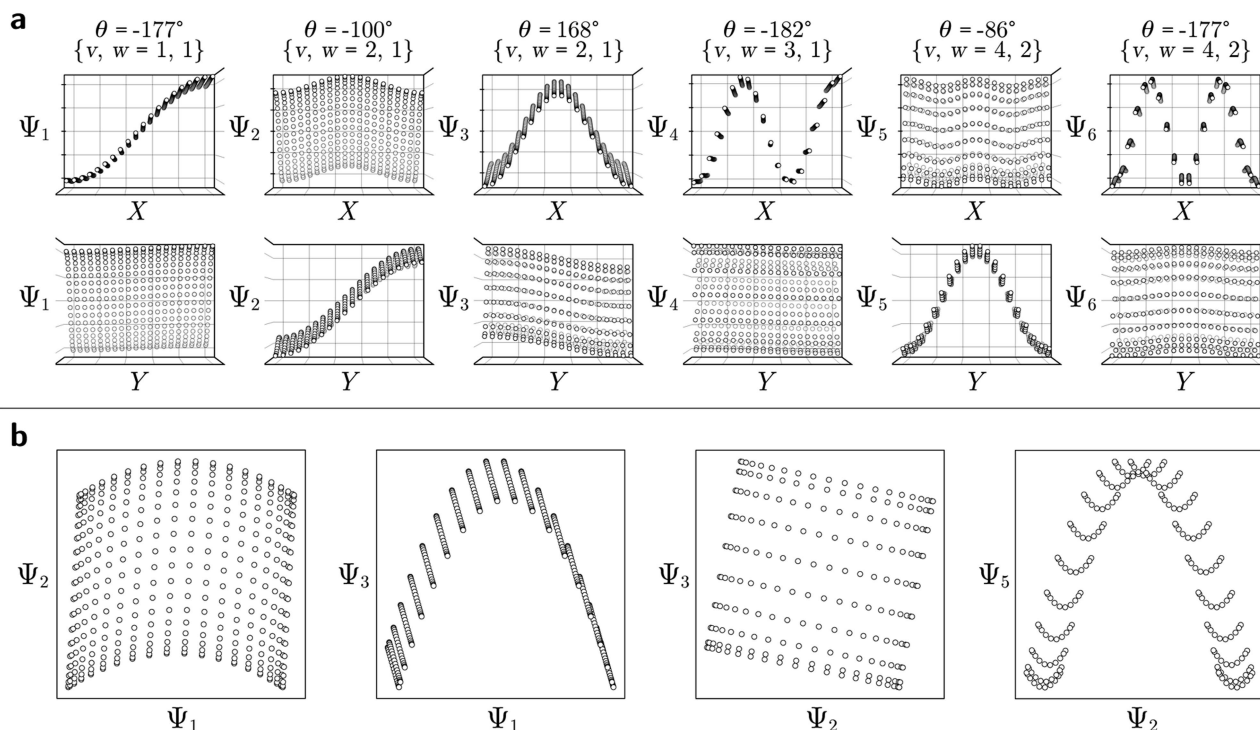


Fig. 12 Comparison of analytically-generated functions with heuristic results previously obtained for PD₁. For each pair of subplots, values for θ were approximated by eye. Our approximations in (a) and (b) share a remarkable similarity with heuristic results shown in Fig. 9 and 11, respectively, and are able to account for geometric minutiae previously unaccounted for, as well as larger-scale rotations seen in the composite of eigenfunctions. For example, discrepancies can be seen in the slightly tilted appearance of the geometry in Ψ_3 of Fig. 9, and corresponding curling at the edges of $\{\Psi_2, \Psi_3\}$ in Fig. 11. These differences can be understood as additional, small-scale perturbations which are currently unaccounted for in our general expression.

Indeed, each of the Chebyshev polynomials mapping CM₁ information in Fig. 11 (visualized with subplots enclosed by blue boxes) corresponds to the subset of sinusoidal eigenfunctions that emerged in the reference frame of CM₁ in Fig. 9A; with similar relations holding for CM₂ in Fig. 9B. Combining this empirically-obtained knowledge with our *a priori* understanding of the eigenfunctions of the LBO on known domains as understood throughout Section 2, we are able to express the analytical form of these Ω_{PD} eigenfunctions. In close approximation, we found that the leading Ω_{PD} eigenfunctions appear in the form

$$\Psi_i = \cos(\theta) \cos(v\pi x) + \sin(\theta) \cos(w\pi y) = \beta_1 \psi_v + \beta_2 \psi_w, \quad (9)$$

such that a given eigenvector Ψ_i may contain some linear combination of the n canonical eigenfunctions $\{\cos(k\pi x_q) \mid k \in \mathbb{Z}^+\}$ corresponding to the n degrees of freedom $x_q \in \mathbb{R}^n$. As shown in Fig. 12, we were able to use this explicit expression to near-perfectly emulate the heuristic results obtained in Fig. 9 and 11.

As also demonstrated in our analysis of eigenfunction misalignments in Section 2.1, the sum of these squared coefficients is conserved across pairs of eigenvectors, such that the base functions $\Psi'_i = \psi_v$ and $\Psi'_j = \psi_w$ can be expressed as a rotation $\Psi = R^T \Psi'$, having the form

$$\begin{bmatrix} \Psi_i(\theta) \\ \Psi_j(\theta) \end{bmatrix} = \begin{bmatrix} \cos(\theta)\psi_v + \sin(\theta)\psi_w \\ -\sin(\theta)\psi_v + \cos(\theta)\psi_w \end{bmatrix} \quad (10)$$

This analytical expression suggests that conformational information—pertaining to each of the system's degrees of freedom—will lie on some linear combination of the embedded manifold's orthogonal eigenvectors. This feature is seen most strikingly in $\{\Psi_3 \times \Psi_4\}$ of Fig. 14B,† where the parabolic surface described by the Chebyshev polynomial is significantly out of alignment with the plane of the 2D subspace containing it. Similar instances, albeit in more subtle form, also arise for surfaces in the remaining three PDs of Fig. 14† (to note, we have also included a brief analysis of state space 3 with similar findings; see Fig. 15 in ESI†). In Section 2, we demonstrated that the need for eigenfunction realignment is due to the change in apparent interatomic distances dependent on projection direction, as illustrated *via* Fig. 13 in ESI.† This disparity among PDs is inevitable and poses a fundamental challenge, which we have addressed using ESPER in our companion article.²²

3 Analysis of complex physical constraints

Finally for consideration, we have so far only dealt here with molecular machines that specifically exhibit each of their



domain motions along an independent and mutually unrestricted sequence of quasi-continuous states. All n -wise combinations of these bounded intervals (one for each conformational motion) produce an n -dimensional shape with a rectangular boundary. In our analysis, we have shown that the corresponding Laplacian eigenfunctions are well defined for this domain.

However, one can still imagine all sorts of other situations, such as a system where one domain blocks—*via* steric hindrance—another domain from its full range of motion in a specific region of the state space. Additionally to consider are state spaces with “holes” (*i.e.*, interior boundaries),²⁸ where the occurrence of certain states is forbidden due to energetic restraints. Indeed, the presence of complex physical constraints can drastically change the boundary shape of the manifold's domain, and consequently its eigenfunctions. Less fundamentally restricted, similar aberrations can occur due to lack of coverage—as obtained by experimental measurements—of the state space. For example, obtaining poorly sampled data from a rectangular domain would allow any number of arbitrary shapes to emerge.

To better understand the effects of these boundary challenges within the scope of our heuristic analysis, we have created a 2D state space with an octagonal domain; noting that, in this case, analytical solutions of the Laplacian eigenfunctions are not available. The octagonal boundaries were created by eliminating states at the four corners of our standard rectangular domain, as shown in Fig. 16A in ESI.† To circumvent potential occurrences of eigenfunction misalignments due to PD disparity—which may complicate the interpretation of the boundary influence—we opted to embed the 3D electron density maps instead of 2D projections for this analysis. The corresponding manifold embedding obtained from this octagonal state space is shown in Fig. 16C,† which features a number of deviations from the canonical rectangular eigenbasis (Fig. 7). Manually, we attempted to find a transformation from the octagonally-derived eigenbasis (Fig. 16C†) to the rectangular form (Fig. 7) by assuming a collection of suitable rotation operators.†† Indeed, we are able to show that such a transformation is possible, up to some level of uncertainty (Fig. 16D†).

4 Discussion

Through our analysis, we have shown how the embeddings of manifolds of experimental data originating from structures undergoing the same conformational changes vary depending on datatype. Specifically, we have identified how the spectral geometry transforms as the datatype is changed stepwise from 3D atomic-coordinate structures to 3D density maps to sets of their 2D projections. In our companion work,²² these findings for 2D projections have already been applied for developing an

improved method, ESPER, to retrieve conformational variability of molecules and their free-energy landscape from ensembles of single-particle cryo-EM images. Similarly, this analysis could be used for the recovery of energy landscapes based on data from other measurement or simulation techniques, including ensembles of atomic models obtained by molecular dynamics (MD) simulations³³ and 3D electron density|| maps obtained by cryo-electron tomography³⁴ (cryo-ET).

During this exposition, we have laid out a framework for generating and analyzing ground-truth data across unique datatypes with detailed guidelines on how to extend our current scope to different experimental approaches. For example, these practices could be applied with minimal alterations to studies of conformational continuum tailored to single-particle imaging with an X-ray free-electron laser (XFEL),³⁵ nuclear magnetic resonance spectroscopy (NMR),³⁶ and time-resolved (TR) cryo-EM.³⁷ Generally, we believe that there is a potential for the application of our methodology to a wide area of research, particularly where biological systems exercise multiple degrees of freedom in a continuous manner. Our companion article provides one such direct application of these findings to the study of single-particle cryo-EM data, where we describe a method for navigating the spectral geometry of 2D projection data in the presence of experimentally-relevant SNR and CTF to recover a quasi-continuum of 3D structures and corresponding free-energy landscape.²² Beyond the potential for additional applications, a comprehensive understanding of how the eigenfunction basis of a manifold changes dependent solely on datatype is an area of research with wide implications for machine learning across a diversity of fields.

More technically, we next summarize key assumptions of this framework that pertain to the relevancy and breadth of our heuristic analysis of structural heterogeneity. Of first consideration, we have focused here on situations where structures undergo collective rigid-body motions, which we believe are sufficient for the description of most molecular machines, but may fall short of addressing instances involving more complex situations. Examples are studies of machines entailing the concerted binding and release of ligands, which naturally require a separate state space for each possible combination of the machine with its binding partners. For such a situation, a similar heuristic analysis could be conducted using synthetic models combining two or more state spaces.

In our previous work,²² we further tested the ability of PCA and DM to correctly embed PD manifolds formed from models exercising more complex domain motions using the so-called “mouth-wing” toy model. Notably, the spheres making up the “mouth” domain in this toy model were uniquely positioned (keyframed) for each state so as to gradually clump together (ultimately presenting a higher density towards the fully-open state), while the “wings” were programmed to open in unison at constant angle increment about their hinge and curl inwards at the same time. Compared to the synthetic framework used to generate the Hsp90 dataset, the construction of mouth-wing data provides a radically different situation, and accounts for complex interactions of spheres within each domain motion, while still obeying conservation of mass. The embedding of the

†† We note that both the indices and number of rotation operators required for this transformation deviated from our PD-manifold findings on eigenfunction realignment performed on rectangular state spaces,²² creating a more complex tree of decisions.



mouth-wing images still manifested all essential spectral characteristics detailed for Hsp90. When random positional noise was applied to each sphere across all mouth-wing models, changes to the spectral geometry were effectively no different than from applying additive Gaussian noise on the rendered image of each model. Although, with the addition of the mouth-wings model, we fall short of an exhaustive coverage of possible data models, we believe the correspondence between its outputs and those of the Hsp90 dataset created with entirely different design principles suggests the generality of our discoveries.

Finally, we will discuss the challenges of analyzing systems with complex physical constraints. For the majority of this study, we have been interested in casting a wide enough net so as to capture the dynamics of a large portion of conformer systems, which we surmise operate within rectangular boundaries of an n -dimensional latent space of relatively-rigid multi-body motions. However, as described in Section 4, one can still imagine other situations where the boundaries could admit any arbitrary shape. As an example, we demonstrated how the eigenfunctions change when our ground-truth data is constrained by octagonal boundary conditions (Fig. 16†). Furthermore, there exists a potential for complex relationships that render the manifold noncompact.²⁹

In general, analytically solving the Laplacian for any arbitrary boundary is impossible, since analytical solutions are only available for certain elementary shapes such as rectangles, discs, ellipses and special triangles.^{28,38} On the other hand, spectral methods such as diffusion maps can efficiently obtain numerical solutions, in principle for any boundary. However, this could present additional challenges, since numerical solutions may be harder to incorporate in the process than analytical forms. Depending on the system under study, the prevalence of non-trivial boundaries inducing eigenfunction dissimilarity may vary, with a more detailed analysis required in each case.

Still, when studying systems described by a compact manifold, it is normally expected that the hypercube will continue to 'fill in' as more data are acquired from the experiment (or simulation in the case of MD). As previously addressed in Section 4, we contend that in the case of conformers of molecular machines, non-trivial boundaries are the exception rather than the rule. When available, it is most lucrative then for experimentalists to aspire towards acquiring this most complete information. In this way, they will maximize the ability to track conformational changes in their analysis.

5 Conclusions

The power of geometric machine learning stems from its potential to detect and follow the intrinsic geometry of a manifold of continuous conformations hidden in a large noisy dataset. In our application of geometric machine learning to experimental cryo-EM single-particle images of large ensembles of biomolecules, we have explored the feasibility of this approach by using simulated ground-truth data of a molecule exercising conformational motions with multiple degrees of freedom. Already, insights gained from our analysis have been used to substantially improve the ManifoldEM framework and

its ability to retrieve continuous conformations and functional pathways from single-particle cryo-EM data. Beyond ManifoldEM, we have shown that our machine learning heuristics have the potential for application to data collected from other experimental methods, particularly from cryo-electron tomography (cryo-ET), X-ray free-electron lasers (XFEL), and molecular dynamics (MD) simulations. More generally, we demonstrate the importance of systematically analyzing simulated ground-truth datasets to inform on key features of different datatypes in geometric machine learning, which has significant implications for the reconstruction of complex information obtained from a wide range of experimental designs.

Graph tools

All figures were made using some combination of Matplotlib,⁴³ Adobe Creative Suite,⁴⁴ Cinema4D,⁴⁵ PyMOL,⁴⁶ and ChimeraX.⁴⁷

Data availability

The Python data and code repository for reproducing this work is available on GitHub.³⁹ As well, the software repository for the ESPER method is available on GitHub⁴⁰ and on the IEEE DataPort.⁴¹ The IEEE repository is a mirror of the GitHub repository, while additionally containing a more extensive catalog of simulated data and results, including the atomic-coordinate structures and 3D density maps used to generate 2D projections of the Hsp90 continuum dataset. Synthetic datasets can likewise be generated from scratch using our synthetic data generation protocol.⁴²

Author contributions

Author contributions are listed alphabetically below. All authors reviewed the final manuscript. Conceptualization: ES, PS, JF. Formal analysis: ES. Investigation: ES. Methodology: ES, PS. Validation: ES, PS. Manuscript draft: ES. Manuscript review and editing: ES, PS, JF. Supervision and project administration: JF.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We are grateful to Francisco Acosta-Reyes and Suvrajit Maji for early discussions and their efforts on the companion article. The work of Peter Schwander was supported in part by the U.S. National Science Foundation under Award STC1231306 and the work of Joachim Frank was supported in part by the National Institutes of Health under Grants R01 GM29169, R01 GM55440, and R35 GM139453.

Notes and references

- 1 J. Frank, *Molecular Machines in Biology: Workshop of the Cell*, Cambridge University Press, Cambridge, 2011.



- 2 B. Roux, *Molecular Machines*, World Scientific, 2011.
- 3 J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*, Oxford Academic, 2006.
- 4 X. Agirrezabala, H. Y. Liao, E. Schreiner, J. Fu, R. F. Ortiz-Meoz, K. Schulten, R. Green and J. Frank, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 6094–6099.
- 5 A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseinizadeh, H. Y. Liao, J. Pallesen, G. Sharma, V. A. Stupina, A. E. Simon, J. D. Dinman, J. Frank and A. Ourmazd, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 17492–17497.
- 6 N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina and H. Stark, *Nature*, 2010, **466**, 329–333.
- 7 P. Schwander, R. Fung, G. N. Phillips and A. Ourmazd, *New J. Phys.*, 2010, **12**, 035007.
- 8 P. Schwander, R. Fung and A. Ourmazd, *Philos. Trans. R. Soc., B*, 2014, **369**, 20130567.
- 9 P. C. Whitford, R. B. Altman, P. Geggier, D. S. Terry, J. B. Munro, J. N. Onuchic, C. M. T. Spahn, K. Y. Sanbonmatsu and S. C. Blanchard, in *Ribosomes: Structure, Function, and Dynamics*, ed. M. V. Rodnina, W. Wintermeyer and R. Green, Springer, Vienna, 2011, pp. 303–319.
- 10 J. B. Munro, K. Y. Sanbonmatsu, C. M. T. Spahn and S. C. Blanchard, *Trends Biochem. Sci.*, 2009, **34**, 390–400.
- 11 D. Wales, *Energy Landscapes*, Cambridge University Press, University of Cambridge, 2004.
- 12 E. Seitz and J. Frank, *J. Chem. Inf. Model.*, 2020, **60**, 2581–2590.
- 13 I. Marcos-Alcalde, J. Setoain, J. I. Mendieta-Moreno, J. Mendieta and P. Gómez-Puertas, *Bioinformatics*, 2015, **31**, 3853–3855.
- 14 D. Haselbach, I. Komarov, D. E. Agafonov, K. Hartmuth, B. Graf, O. Dybkov, H. Urlaub, B. Kastner, R. Lührmann and H. Stark, *Cell*, 2018, **172**, 454–464.
- 15 P. A. Penczek, M. Kimmel and C. M. T. Spahn, *Structure*, 2011, **19**, 1582–1590.
- 16 A. Punjani and D. J. Fleet, *J. Struct. Biol.*, 2021, **213**, 107702.
- 17 A. Dashti, G. Mashayekhi, M. Shekhar, D. Ben Hail, S. Salah, P. Schwander, A. des Georges, A. Singharoy, J. Frank and A. Ourmazd, *Nat. Commun.*, 2020, **11**, 4734.
- 18 E. Seitz, *Doctoral thesis*, Columbia University, 2022, DOI: [10.7916/4n0v-wa24](https://doi.org/10.7916/4n0v-wa24).
- 19 T. Sztain, S.-H. Ahn, A. T. Bogetti, L. Casalino, J. A. Goldsmith, E. Seitz, R. S. McCool, F. L. Kearns, F. Acosta-Reyes, S. Maji, G. Mashayekhi, J. A. McCammon, A. Ourmazd, J. Frank, J. S. McLellan, L. T. Chong and R. E. Amaro, *Nat. Chem.*, 2021, **13**, 963–968.
- 20 K. Pearson, *Philos. Mag.*, 1901, **2**, 559–572.
- 21 R. R. Coifman and S. Lafon, *Appl. Comput. Harmon. Anal.*, 2006, **21**, 5–30.
- 22 E. Seitz, F. Acosta-Reyes, S. Maji, P. Schwander and J. Frank, *IEEE Trans. Comput. Imaging*, 2022, **8**, 462–478.
- 23 F. H. Schopf, M. M. Biebl and J. Buchner, *Nat. Rev. Mol. Cell Biol.*, 2017, **18**, 345–360.
- 24 M. M. U. Ali, S. M. Roe, C. K. Vaughan, P. Meyer, B. Panaretou, P. W. Piper, C. Prodromou and L. H. Pearl, *Nature*, 2006, **440**, 1013–1017.
- 25 G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees and S. J. Ludtke, *J. Struct. Biol.*, 2007, **157**, 38–46.
- 26 M.-E. Craioveanu, M. Puta and T. M. Rassias, *Old and New Aspects in Spectral Geometry*, Springer Netherlands, 2010.
- 27 R. R. Coifman, Y. Shkolnisky, F. J. Sigworth and A. Singer, *IEEE Trans. Image Process.*, 2008, **17**, 1891–1899.
- 28 D. S. Grebenkov and B.-T. Nguyen, *SIAM Rev.*, 2013, **55**, 601–667.
- 29 J. R. Munkres, *Topology*, Prentice Hall Incorporated, Hoboken, 2000.
- 30 M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, <https://digital.library.unt.edu/ark:/67531/metadc40302/>, accessed 16 February 2023.
- 31 H. M. Cundy, A. P. Rollett and H. S. M. Coxeter, *Math. Gaz.*, 1962, **46**, 331.
- 32 M. Berger, *Geometry I*, Springer Berlin, Heidelberg, 1st edn, 1987.
- 33 A. Hospital, J. R. Goñi, M. Orozco and J. L. Gelpi, *Adv. Appl. Bioinform. Chem.*, 2015, **8**, 37–47.
- 34 J. Dubochet, M. Adrian, J.-J. Chang, J.-C. Homo, J. Lepault, A. W. McDowell and P. Schultz, *Q. Rev. Biophys.*, 1988, **21**, 129–228.
- 35 A. Hosseinizadeh, G. Mashayekhi, J. Copperman, P. Schwander, A. Dashti, R. Sepehr, R. Fung, M. Schmidt, C. H. Yoon, B. G. Hogue, G. J. Williams, A. Aquila and A. Ourmazd, *Nat. Methods*, 2017, **14**, 877–881.
- 36 C. Anklin, *Spectrosc. Eur.*, 2018, **30**, 1–4.
- 37 J. Frank, *J. Struct. Biol.*, 2017, **200**, 303–306.
- 38 B. McCartin, *Appl. Math. Sci.*, 2008, **2**, 2891–2901.
- 39 E. Seitz, *Zenodo*, 2023, DOI: [10.5281/zenodo.7657567](https://doi.org/10.5281/zenodo.7657567).
- 40 E. Seitz, *Zenodo*, 2021, DOI: [10.5281/zenodo.5362645](https://doi.org/10.5281/zenodo.5362645).
- 41 E. Seitz, *IEEE DataPort*, 2022, DOI: [10.21227/wmn8-k054](https://doi.org/10.21227/wmn8-k054).
- 42 E. Seitz, *Zenodo*, 2019, DOI: [10.5281/zenodo.3554523](https://doi.org/10.5281/zenodo.3554523).
- 43 J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90–95.
- 44 Adobe Inc., *Adobe Creative Suite (version 5.9.1.377)*, 2022, <https://adobe.com/products>.
- 45 Maxon Computer GmbH, *Cinema4D (version 2023.1.3)*, 2023, <https://www.maxon.net/en/cinema-4d>.
- 46 L. L. C. Schrödinger and DeL. Warren, *PyMOL (version 2.4.0)*, 2020, <http://www.pymol.org/pymol>.
- 47 E. Pettersen, T. Goddard, C. Huang, E. Meng, G. Couch, T. Croll, J. Morris and T. Ferrin, *Protein Sci.*, 2021, **30**, 70–82.

