

Geometric machine learning informed by ground truth: Recovery of conformational continuum from single-particle cryo-EM data of biomolecules

Evan Seitz,¹ Francisco Acosta-Reyes,² Suvrajit Maji,² Peter Schwander*,³ and Joachim Frank*^{1,2}

¹⁾ Columbia University, Department of Biological Sciences, New York, NY 10027, USA

²⁾ Columbia University Medical Center, Department of Biochemistry and Molecular Biophysics, New York, NY 10032

³⁾ University of Wisconsin-Milwaukee, Department of Physics, 3135 N. Maryland Ave, Milwaukee, WI 53211, USA

(*Electronic mail: pschwan@uwm.edu, jf2192@cumc.columbia.edu)

This work is based on the manifold-embedding approach to study biological molecules exhibiting continuous conformational changes. Previous work established a method capable of reconstructing 3D movies and accompanying energetics of atomic-level structures from single-particle cryo-EM images of macromolecules displaying multiple conformational degrees of freedom. Here, we introduce an unsupervised geometric machine learning approach that is informed by detailed heuristic analysis of manifolds formed by simulated heterogeneous cryo-EM datasets generated from an atomic structure. These simulated data were generated with increasing complexity to account for multiple conformational motions, state occupancies and typical microscope parameters in a wide range of signal-to-noise ratios. Using these datasets as ground-truth, we provide detailed exposition of our findings using several conformational motions while exploring the available parameter space. Guided by these insights, we build a framework to leverage the high-dimensional geometric information obtained towards reconstituting a quasi-continuum of conformational states in the form of a free-energy landscape and respective 3D density maps for all states therein. As shown by a direct comparison of results, this framework offers substantial improvements relative to the previous work.

INTRODUCTION

Molecular machines, consisting of assemblies of proteins or nucleoproteins, take on a range of unique configurations or *conformational states* as they go through their functional cycles¹. These states are typically characterized by different spatial constellations of relatively rigid domains, and can be organized in a *state space* according to the continuous motions of each domain along a unique coordinate. Specific sequences of the states in this space form pathways along which the molecular machine may transform. When energetics of states are known, in terms of the machine's free-energy landscape, a path is singled out along which the machine performs its metabolic function².

A number of recent studies^{1,3,4} were inspired by the realization that it is possible, through the analysis of experimental data, to gain insights into the rules governing a molecular machine's function. In thermal equilibrium, molecular machines are constantly buffeted by the random motions of nearby solvent molecules, which deform them reversibly as they transition via a series of thermally-driven steps. State-of-the-art single-particle cryo-EM^{5–7} is now capable of providing large numbers of two-dimensional snapshots (i.e., projections) of a molecular machine undergoing this process. When the number of snapshots is sufficiently large – typically several hundred thousand – they capture virtually the entire range of conformations accessible in thermodynamic equilibrium. By virtue of the Boltzmann statistics, the relative number of sightings in each of these states can be translated into changes of free energy^{8,9}. Thus, under assumption of thermodynamic equilibrium, the machine's free-energy landscape can be obtained from an experiment. Accurate estimation of the free-energy landscape for molecular ma-

chines and other biological assemblies is of unparalleled importance in modern structural biology.

The way to utilize the data from a single-particle cryo-EM experiment is not easy, however. Ideally, we would wish to compare 3D structures, but only 2D projections are accessible experimentally. After each 2D projection is assigned angles to define its viewing direction on the 2-sphere (S^2), a set of projections in close proximity to one another can be assigned to a unique *projection direction* (PD)[†]. For any $\text{PD} \subset S^2$, the challenge is that the relationship among the N images therein, represented as a P pixel array, require an analysis of the point cloud formed in vector space \mathbb{R}^P . Similarities between molecules captured in the same PD, but slightly different conformations, appear as closeness between corresponding points in this high-dimensional space. Thus, for a given PD, images of molecules captured in random states are arranged—by virtue of their similarities—according to the continuous motions of the molecule's domains.

The geometric structure formed by such an ensemble is an n -dimensional manifold Ω embedded in a high-dimensional Euclidean space \mathbb{R}^P , with an intrinsic dimension n equal to the number of the system's independent molecular degrees of freedom. By choosing a suitable embedding that maps the data points in Ω into a low-dimensional Euclidean space, we create the foundation for the analysis of the molecule's conformational spectrum and free-energy landscape by a machine learning approach. In the following, we use the term *PD-manifold approach* to refer to this strategy. Specifically, it entails the grouping of cryo-EM data into individual PDs,

[†]A tabulated description of symbols and abbreviations used throughout this document is available in the Appendix.

with the subset of images within each PD analyzed via manifold embedding, and the resulting representations combined into a consolidated conformational spectrum. This approach was first introduced by Dashti et al.¹ and is now termed ManifoldEM^{10,11}. Results from previous ManifoldEM studies on biological systems—including the ribosome¹, ryanodine receptor¹⁰ and SARS-CoV-2 spike protein¹²—have proven its viability and its potential to provide new information on the functional dynamics of molecules.

As manifolds are encountered in many domains of mathematics, science and engineering¹³, the aim of dimensionality reduction has been widely pursued and given rise to a number of well-established techniques to analyze large and complex datasets. Representing data points on Ω in terms of leading eigenvalues and eigenvectors gives valuable insights into its intrinsic structure, with these relationships having been well studied in the context of spectral geometry¹⁴. In the analysis of cryo-EM data, both linear^{3,15–21} and nonlinear^{1,10,20,22} dimensionality reduction methods have been applied, primarily principal component analysis²³ (PCA) and diffusion maps^{24,25} (DM), respectively. Both approaches allow an analysis of the data points in Ω as embedded in \mathbb{R}^N , whose entries are the first N eigenvectors of the respective graph, and noting that only a leading subset of these are needed for retrieving the conformational spectrum.

In the PCA approach, eigenvectors are obtained from the covariance matrix, whereas DM approximates the eigenfunctions of the Laplace-Beltrami operator (LBO) on Ω , sampled at the given data points. Some techniques are not so easily classified, however, such as the method of Laplacian spectral volumes⁴, which relies on both linear and nonlinear dimensionality reduction. The application of these methods can further be classified based on their type of data input – generating embeddings from either 2D projections straight from a cryo-EM experiment (i.e., the ManifoldEM approach), or 3D density maps which have been reconstructed from those projections^{4,26–29}. Regardless of the approach, since the intrinsic structure of such manifolds formed by the data is unknown, these competing reconstruction methods cannot be immediately validated or informatively compared using experimental data alone. They instead require careful evaluation using appropriate synthetic ground-truth datasets.

The purpose of the present study is twofold. For our first endeavor, we provide a heuristic investigation of the manifolds obtained from synthetic quasi-continuous ground-truth datasets, with properties endowed as is anticipated from cryo-EM visualization of biological molecules. To this end, we create several state spaces by simulating a molecule with movable domains, each having undergone a series of independent *conformational motions* (CMs), with the number of CMs constructed for each state space defining the intrinsic dimensionality of the dataset. We then determine how these quasi-continuous motions are reflected in the low-dimensional representations of the manifold's spectral ge-

ometry, obtained by linear or nonlinear dimensionality reduction using kernel methods. For either construction, we derive an explicit expression to account for the geometric structures observed, and describe how to interpret this information as it exists on a hypersurface spanned by multiple degrees of freedom. This heuristic analysis is introduced as a clean slate free from assumptions, aiming to further investigate—using ideal data—the feasibility of manifold embedding techniques under realistic experimental conditions, while exposing any intrinsic uncertainties that may arise. Recently, several issues and limitations have been documented¹¹ in the ManifoldEM framework, further amplifying the motivation of our current pursuit.

For our second endeavor, we introduce a novel methodology (which we will term ESPER: “Embedded subspace partitioning and eigenfunction realignment”) for extraction of conformational information from specific subspaces of PD-manifold (Ω_{PD}) embeddings, which we use to generate the molecular machine’s free-energy landscape and corresponding 3D movies depicting its function. Whereas the previous approach^{1,10,11} reconstructs images via nonlinear Laplacian spectral analysis³⁰ (NLSA) in an additionally embedded space spanned by one or more CMs, ESPER instead captures each CM directly from the initial embedding while retaining the original cryo-EM images. In addition, several novel operations and refinements to the existing PD-manifold approach are introduced, including a previously unaccounted-for high-dimensional eigenbasis transformation that we deem essential for correctly recapitulating ground-truth information, as well as identification of the proper 2D subspaces required to adequately capture each CM. We demonstrate that this alternative methodology provides conformational movies of significantly improved quality, further enabling the use of efficient strategies for generating multidimensional free-energy landscapes not accessible via the founding ManifoldEM framework. Ultimately, we will show that, when certain requirements are met in the quality and structure of a dataset, ESPER offers an alternative strategy with many benefits compared to the current ManifoldEM approach.

METHODS

We first introduce a framework for the creation of synthetic ground-truth single-particle cryo-EM datasets in the form of 2D projections of 3D electron density maps arising from a quasi-continuum of atomic structures^{31,32}. (Note that in reality cryo-EM data represent projections of the electrostatic or Coulomb potential distribution, which is distinct from the electron density distribution “seen” by X-rays. However, for the present analysis, this distinction is irrelevant). To begin, a suitable macromolecule is chosen as a foundational model, defined by available structural information in the form of 3D atomic coordinates from the Protein Data Bank³³ (PDB). Us-

ing this initial PDB structure as a seed, a sequence of states is generated by altering the positions of specific domains of the macromolecule's structure. To mimic quasi-continuous conformational motions, we used equispaced rotations of the domains about their hinge-residue axes. The number of these mutually independent conformational motions[†] defines the intrinsic dimensionality n of the system. By exercising these domain motions independently in all combinations, a set of atomic coordinate structures in PDB-format are generated. In sum, this quasi-continuum of states spans the molecular machine's state space.

For this work, the heat shock protein Hsp90 was chosen as a starting structure due to its simple design, exhibiting two arm-like domains (chain A and B, containing 677 residues each) connected together in an overarching V-shape³⁴. *In vivo*, these arms are known to close after binding of the molecule with ATP, with Hsp90 acting as a chaperone to stabilize the structures of surrounding heat-vulnerable proteins. During its work cycle, Hsp90 naturally undergoes large conformational changes, transitioning from its two arms spread open in a full V-shape (inactive state) to both arms bound together along the protein's central line of two-fold symmetry (active state) following ATP binding. We initiated our workflow with the fully closed state via entry PDB 2CG9, whose structure was determined at 3.1 Å by X-ray crystallography³⁵.

Casting Hsp90's biological context aside, liberties were taken in the choice of the synthetic model's leading degrees of freedom. Instead of a single conformational motion (arms open to closed, as *in vivo*), we decided to create three easily-identifiable and fully-decoupled domain motions, which we refer to as CM₁, CM₂ and CM₃. Each CM was designed to cover a unique range of motions, with the cascade of overlaid states making up CM₁ occupying the largest spatial region, followed in magnitude by CM₂ and then by CM₃. Using combinations of these CMs, three synthetic state spaces were generated, with intrinsic dimensionalities of $n = 1, 2, 3$. This was achieved by changing the positions of the first, the first two, or all three regions defined as rigid domains in their given ranges monotonically and (in the latter two cases) independently. Importantly, in view of a later discussion of boundary conditions, there is no steric hindrance between domains within the ranges of their motions.

In the following analysis, these state spaces are termed SS₁, SS₂ and SS₃, and defined by: (1) 20 states exhibiting one degree of freedom (CM₁); (2) 400 states (20×20) with two degrees of freedom (CM₁, CM₂); and (3) 1000 states ($10 \times 10 \times 10$) with three degrees of freedom (CM₁, CM₂, CM₃), respectively. As a specific example of the ranges of motion present in SS₃, the Root-Mean-Square

Deviation (RMSD) was calculated³⁶ for the differences of the atomic coordinates between neighboring states in each CM, yielding the values of 1.8 Å, 1.3 Å and 0.3 Å along CM₁, CM₂ and CM₃, respectively; with the RMSD between the first and last state of each CM (representing its total span) yielding 15.3 Å, 11.3 Å and 2.4 Å. Altogether, the total spans of these synthetically-constructed CMs cover a wide range of motions, as one might observe in experiment. In-depth details for these datasets, such as exact atomic descriptions of each state, are provided in the supplementary material (SM) section SM-I[‡]. This section should also be consulted for its description of the indexing used for ordering images within each state space, which is essential for interpreting color maps in figures of embedded manifolds throughout this paper.

Our presentation showcases results from detailed evaluation of three data types—termed data-type I, II, and III—with each step incorporating image artifacts and ensemble statistics in our state-space models as is anticipated in a cryo-EM experiment. Detailed information pertaining to the construction of each of these three data types is provided in the supplementary material. We first investigate the pristine *data-type I*, which is given no simulated experimental artifacts or occupancy assignments. Within this construction, for five example PDs, we examine the manifolds (Ω_{PD}) corresponding to each of the three state spaces as obtained via the DM framework, followed by a comparison with those obtained via PCA. Using the eigenfunctions of the LBO, we analytically quantify the trajectories of our simulated conformational changes as embodied by the spectral geometry of each Ω_{PD} .

Next, in establishing *data-type II*, we vary the abundance of images per state in each dataset and add noise to the images with varying signal-to-noise ratio (SNR), so as to investigate the influence of statistical coverage on spectral geometry, and to quantify the robustness of this geometry in the presence of noise. Following this analysis, we further increase the presence of experimental artifacts through application of a contrast transfer function (CTF) with realistic microscopy parameters and random defocus variations (within the typical range expected in the experiment), and apply noise to obtain an experimentally-relevant SNR (*data-type III*). Based on the findings of our analysis, we finally introduce an overview of the ESPER method for reconstructing the conformational motions in the form of 2D and 3D movies as obtained from a collection of PDs along a great circle on S^2 , with occupancies assigned and transformed into a free-energy landscape (*final analysis*). All Python scripts for reproducing this workflow, including extensive documentation therein, have been made available online^{32,37}.

[†]There is a wide range of nomenclature used here between fields and, in some instances, works by the same author. The following are interchangeable: *conformational motions*; *conformational coordinates*; *reaction coordinates*; *collective motion coordinates*.

[‡]All supplementary material sections will be referenced throughout this document in form SM- $\{\text{Roman numeral}\}$. The ordering of sections in SM is arranged to form a cohesive narrative, separate from the order each section is introduced in our main text.

RESULTS

I. Diffusion Maps

A. Data-type I in State Space 1

For its illustrative qualities, we first analyze embeddings constructed from the eigenvectors obtained from the DM framework for SS_1 , representing one degree of freedom sampled with 20 states in CM_1 . As the workflow for obtaining the general diffusion map for a given dataset has been described in several publications^{24,25,38}, we summarize these procedures here while providing detailed information in the supplementary material (see section SM-X, where we define characteristic parameters such as the Gaussian bandwidth, ε). Using the DM framework, we ultimately generate a different embedding for each of the five PD manifolds, with each of the resultant point clouds containing 20 points, and each point therein corresponding to an image of a conformational state from CM_1 .

Upon inspection of each embedding (one per PD) for suitable ε within the range discovered, we found that the corresponding eigenvalue spectrum for each Ω_{PD} showed a staggered falloff in the significance of leading eigenvectors, which decayed slowly to zero. Projecting the resultant set of eigenvectors (forming an *eigenbasis*) onto the leading eigenvector (Ψ_1) alone presented a skewed version of the anticipated mapping, with the indices of states appearing in jumbled sequence near the boundaries. When we alternatively projected the data onto the first two eigenvectors $\{\Psi_1, \Psi_2\}$ —forming a 2D eigenvector subspace—the conformational signal followed a parabolic trajectory (as shown in the first subplot of Fig. 1-B), confirmed by the proper ordering of indices of points along this curve.

Following this analysis, we next proceeded to investigate all other unique 2D combinations of eigenvectors. Mathematically, each such mapping to a 2D vector subspace is the restriction to the N -dimensional embedding of the projection of \mathbb{R}^N onto \mathbb{R}^2 ; given by $\{\Psi_1, \Psi_2, \dots, \Psi_N\} \mapsto \{\Psi_i, \Psi_j\}$, where $i < j$. (For expediency, we will use the term *subspace* to specifically refer to a subspace of an embedded manifold). As seen in Fig. 1-B, a subset of the canonical Lissajous curves³⁹ emerged across the 2D subspaces of each Ω_{PD} , with the curves in this set having the form

$$L_{p,q} = \{\cos(p\pi x) \times \cos(q\pi x) \mid 0 \leq x \leq 1; p \neq q \in \mathbb{Z}^+\}$$

where the operator \times denotes the Cartesian product⁴⁰.

The appearance of these $L_{p,q}$ curves—which are the composite of sinusoids—aligns with known attributes of the Laplace-Beltrami operator. Specifically, the functions $\psi_k = \{\cos(k\pi x) \mid 0 \leq x \leq 1; k \in \mathbb{Z}^+\}$ are the canonical eigenfunctions of the LBO on the interval $[0, 1]$ subject to Neumann boundary conditions⁴¹, with a metric tensor g equal to identity (see section SM-XIII and SM-XIV). By relying on our privileged knowledge of

the ground-truth sequential ordering of CM states, we were able to further investigate these underlying sinusoidal forms. 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 20 total states in SS_1), making sure that the ordering of the points in 1D follows the sequence assigned by the ground-truth index of its corresponding image along CM_1 . As seen in Fig. 1-A, when the collection of points in Ψ_k are ordered appropriately, the eigenfunction’s sinusoidal form emerges along the full extent of the degree of freedom present (i.e., $I \in [1, 20] \mapsto x \in [0, 1]$).

Of course, as the points in an experimental dataset naturally arrive in unordered sequence, one would have to properly sort the image indices to recognize these sinusoids; here, for example, there are $20!$ sequences to consider. In the application, even if an approximation of this sequence were obtained, then in the presence of duplicate CM states (which we anticipate in an experiment), each sinusoid would be irregularly stretched along the x -axis where those duplicate states occurred, forming an unwieldy distorted sinusoidal form. However, as the points in each Ψ_k are always scrambled in the same order in all eigenvectors, the composite of any two will always exist in a readily identifiable $L_{p,q}$ form. For these composites, we found that CM information 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; i.e., $(\{\Psi_1, \Psi_2\}, \{\Psi_1, \Psi_3\}, \dots, \{\Psi_1, \Psi_g\})$, where g 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 kind⁴², 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). First, note that for each sinusoidal subplot in Fig. 1-A, points are equispaced along the x -axis while maintaining the proper sinusoidal form on the y -axis, in correspondence with the uniform rotations of the corresponding atomic-coordinate structures. However, due to the Cartesian product, only non-uniform spatial relationships exist between neighboring states in each $L_{p,q}$. Analytically, this relationship is described by a non-isometric mapping, where lengths in the domain X_a are not preserved in the codomain $X = \Pi_{a \in A} X_a$, and naturally arises when taking a set (indexed via A) of Cartesian products (II) operating on cosine functions $X_a = \{\cos(k\pi x) \mid x \in [0, 1]; k \in \mathbb{Z}^+\}$ that are each uniformly occupied with a finite number of datapoints. As shown in Fig. 2-A, the spacing between points in $L_{1,2}$, which is the composite of two such sinusoids, has an intrinsically nonuniform spatial distribution, with the density of points similarly arranged as seen in the corresponding point clouds.

We denote this aspect with the term *nonuniform rates of change*. As a potential remedy, we investigated the

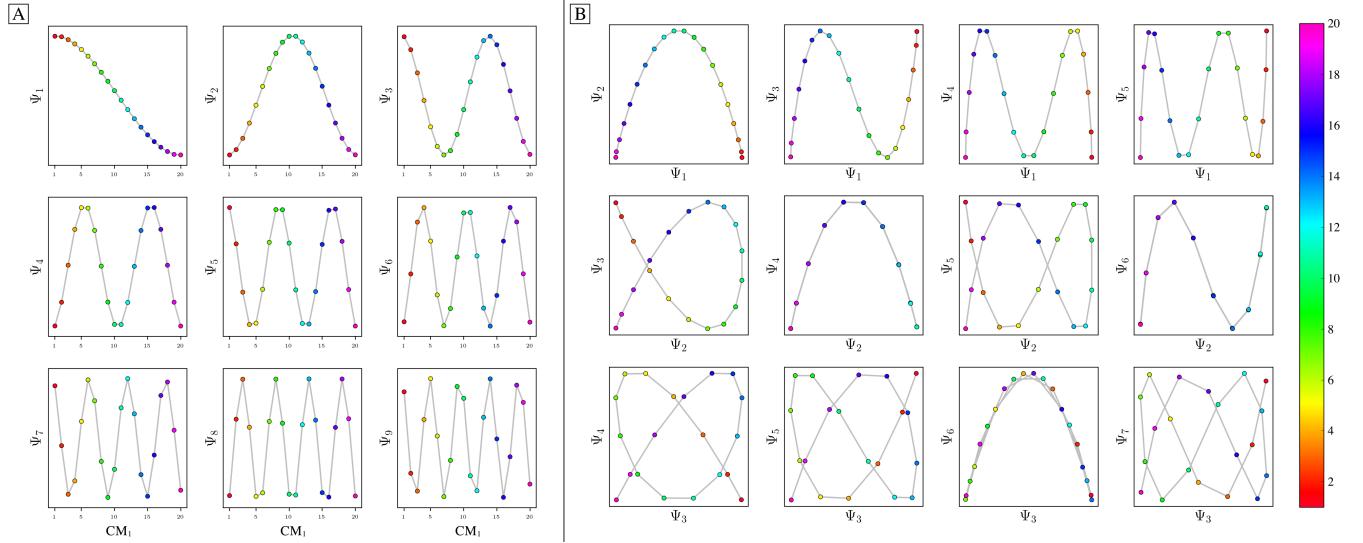


FIG. 1: Analysis of eigenfunctions for PD_1 in SS_1 (i.e., 20 states total making up one degree of freedom). On the left [A] are the sinusoidal forms $\{\cos(k\pi x) \mid k \in \mathbb{Z}^+ \leq N\}$ of each eigenvector Ψ_k that emerge when points (corresponding to images) in each Ψ_k are ordered precisely in the sequence 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 in [B]. In the first row are the Chebyshev polynomials of the first kind, of which the parabola $\{\Psi_1, \Psi_2\}$ is the simplest mapping of the conformational information present. As is explained below, $\{\Psi_2, \Psi_4\}$ and $\{\Psi_3, \Psi_6\}$ represent *parabolic harmonics* of the $\{\Psi_1, \Psi_2\}$ parabola, which obfuscate the CM content. Finally, note the *nonuniform rates of change* along each Lissajous curve – where it can be seen, for example, that points along the $\{\Psi_1, \Psi_2\}$ parabola are most densely packed near the boundaries and vertex.

use of an inverse-cosine mapping on each eigenfunction. Fig. S4 provides the results of this transformation on both (1) the analytically-derived cosine functions $k \in \{1, 2\}$ and (2) the SS_1 eigenfunctions obtained by applying DM on images in PD_1 . The first two subplots in Fig. S4 further highlight the remarkable fidelity of the DM eigenfunctions of the graph Laplacian to the analytical form of the LBO, while the third subplot illustrates the results of inverse-cosine transformation. As can be seen, this mapping presents the coordinates of each eigenfunction in a space with uniform rates of change, consistent with the ground-truth relationships between atomic-coordinate structures. We will leverage this aspect later in our framework, and indicate any eigenvectors Ψ_i under this transformation with the insignia Φ_i .

Next for consideration, as seen in Fig. 1-B, there exist several parabolic trajectories scattered throughout the 2D subspaces of a given Ω_{PD} . As confirmed by the indices of points and the corresponding color map along each curve, only the first of these parabolas describes the full extent of the conformational motion present monotonically, while all other trailing parabolas display a non-monotonic signal. As a specific example, Fig. 2 shows that the first three such parabolas can be generated via $L_{1,2}$, $L_{2,4}$ and $L_{3,6}$ – which repeat the conformational information once, 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)⁴⁰, with all other mappings to higher-order parabolas non-injective surjections. Importantly, since the Cartesian product of continuous functions is continuous and projections from product spaces are also continuous, this bijection further meets the requirements of a *homeomorphism*: a bijective correspondence that preserves the topological structures involved⁴⁰. 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 mapping a given CM; a problem that becomes more challenging in the following subsections as more degrees of freedom are added to the system.

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 each Ω_{PD} that should be expected, 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 along the PD under investigation, ground truth is preserved to different degrees. Going forward, we will refer back to this notion under the label *PD disparity*. This disparity affects all Ω_{PD} characteristics, and will become more relevant as we investigate the embeddings of datasets with multiple degrees of freedom.

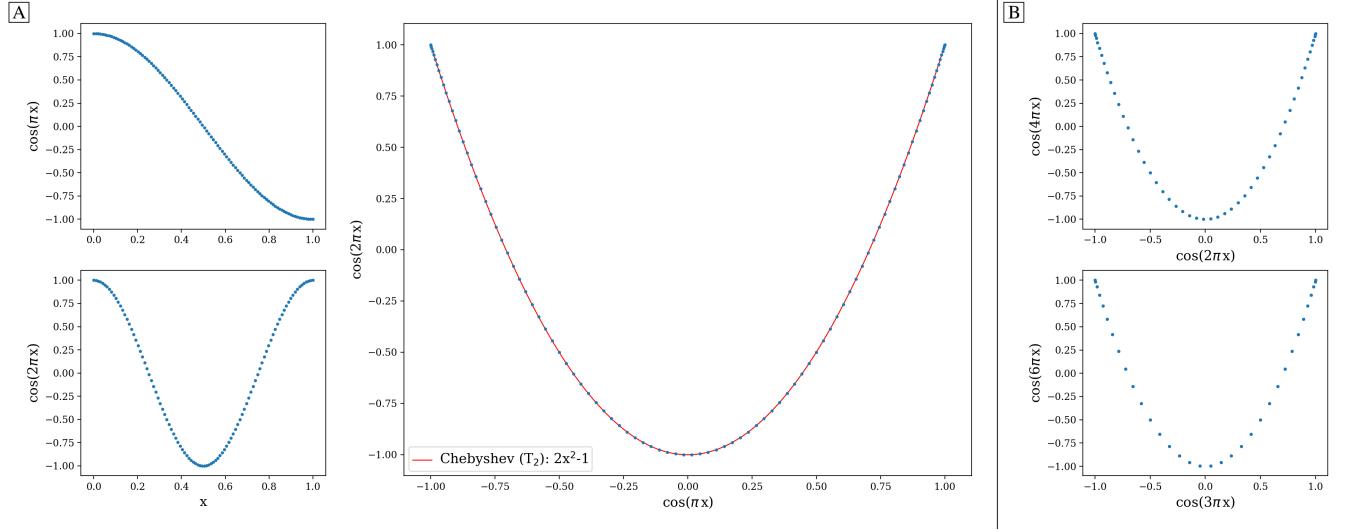


FIG. 2: The analytical generation of the Lissajous curve $L_{1,2} = \{\cos(\pi x) \times \cos(2\pi x) \mid \text{uniform } x \in [0, 1]\}$, where $L_{1,2} \in ([-1, 1] \times [-1, 1])$ is shown in [A]. Note the naturally-induced nonuniform spacing between points near the boundaries and vertex of the 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], the 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 $L_{1,2}$.

B. Data-type I in State Space 2

To further understand these conformational-variation signals with increasing intrinsic dimensionality n , we next investigated the embeddings generated for SS_2 . As seen in Fig. 3-A, by plotting the points in each eigenvector in the specific ground-truth sequence constructed for CM_1 against a uniform index (now for the 400 states in SS_2 ; i.e., $\{1, 2, 3, \dots, 400\}$), a similar but now interspersed pattern of sinusoids appeared. Specifically, the appearance of the sinusoids (with increasing $k \in \mathbb{Z}^+$) only manifested in a subset of all eigenvectors present, while for all other eigenvectors outside of this set, more arcane patterns emerged. Following this observation, we next reordered the indices of points within all eigenvectors to instead correspond with 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. 3-B, which manifested 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—we conclude that for n degrees of freedom in a given dataset, there must be n independent sets of these sinusoids $\{\cos(k\pi x_q) \mid q \in n\}$, with each set interspersed throughout the collection of available eigenvectors $\{\Psi_i \mid i \in N\}$.

Following our previous discovery of a single set of orthogonal Chebyshev polynomials spanning specific 2D subspaces of SS_1 , we next investigated whether similar

patterns existed in SS_2 . In doing so, we found that for every conformational motion present in a given state space, there exists a corresponding set of Lissajous curves interspersed across specific $\{\Psi_i, \Psi_j\}$ projections of the N -dimensional embedding. Specifically, in the case of PD_1 , independently projecting the data for SS_2 onto the planes spanned by its $\{\Psi_1, \Psi_i\}$ and $\{\Psi_2, \Psi_j\}$ combinations (where $i > 1$; $j > 2$) revealed a unique set of Chebyshev polynomials, with the sequence of points along these trajectories corresponding to CM_1 and CM_2 (Fig. 4). With this knowledge in hand, we can now compare the subset of eigenfunctions as obtained in either the reference frame of CM_1 (Fig. 3-A) or CM_2 (Fig. 3-B) with the Chebyshev polynomials in Fig. 4. Indeed, each Chebyshev polynomial mapping CM_1 information in Fig. 4 (visualized with subplots enclosed by blue boxes) corresponds to the subset of sinusoidal eigenfunctions which emerged in the reference frame of CM_1 in Fig. 3-A; with similar relations holding for CM_2 in Fig. 3-B. (For convenience, we will refer to a set of Chebyshev polynomials corresponding to a given CM as the *conformational modes*). Thus, even though the knowledge required to view these CM sinusoids is unavailable outside of ground-truth studies, our analysis confirms that these CM relationships are ever-present, and further, that we can rely on their existence—via the composites of carefully chosen eigenvectors—to elucidate conformational type and order.

Combining this empirically-obtained knowledge with our *a priori* understanding of the eigenfunctions of the LBO on known domains, we were able to intuit the analytical form of these Ω_{PD} eigenfunctions. A detailed ex-

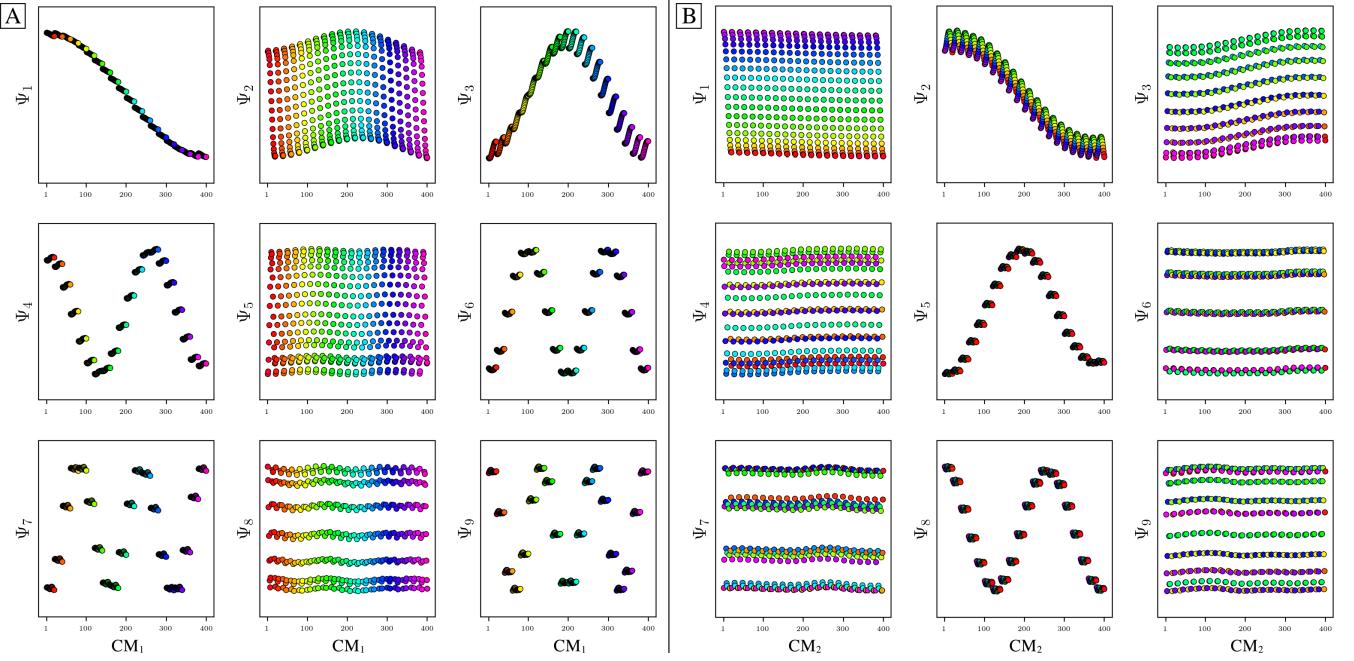


FIG. 3: Visualization of eigenfunctions for PD_1 in SS_2 (i.e., $20 \times 20 = 400$ states total making up two degrees of freedom). On the left [A] are the sinusoidal forms $\{\cos(k\pi x) \mid k \in \mathbb{Z}^+\}$ that emerge for only a specific subset of eigenvectors $\{k = 1, 3, 4, 6, 7, 9, \dots\}$ when points in each Ψ_k are ordered precisely in the sequence of CM_1 (as assigned when the ground-truth images were initially constructed). Likewise, in [B], when points in each Ψ_k are instead ordered precisely in the sequence of CM_2 , a new set of sinusoids emerge $\{k = 2, 5, 8, \dots\}$ precisely for those remaining Ψ_k not in the previous CM_1 subset. Hence, it can be seen in [A] and [B] 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 that degree of freedom. However, as stated for SS_1 , such frames of reference are unavailable *a priori*.

position of this discovery is provided in section SM-XIV, showing how the canonical eigenfunctions on a rectangular domain transform as the data type is translated stepwise from atomic models to 3D density maps to 2D projections. In close approximation, the leading Ω_{PD} eigenfunctions are of the form

$$\Psi_i = \cos(\theta)\cos(v\pi x) + \sin(\theta)\cos(w\pi y) = A\psi_v + B\psi_w$$

such that a given eigenvector Ψ_i may contain some linear combination of n canonical interval-like eigenfunctions $\{\cos(k\pi x_q) \mid k \in \mathbb{Z}^+\}$ corresponding to each degree of freedom $x_q \subset \mathbb{R}^n$. In Fig. S26, we use this explicit form to near-perfectly emulate the heuristic results obtained in Fig. 3 and Fig. 4. We also show that the these coefficients are conserved across pairs of eigenvectors (i.e., $A^2 + B^2 = 1$), such that the base functions $\Psi'_i = \psi_v$ and $\Psi'_j = \psi_w$ can be expressed as a rotation $\Psi = \mathbf{R}^T\Psi'$, having the form

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

From our analytical expression, it is clear that depending on the PD, CM 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 $\{\Psi_3, \Psi_4\}$ of PD_3 (Fig. S5), where the parabolic surface described

by the Chebyshev polynomial is significantly tilted 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. S5. In section SM-XIV-C, we demonstrate that this feature is a result of *PD disparity*. Specifically, we generate an embedding (via DM) of the SS_2 collection of 3D electron density maps (EDMs, from which the PD datasets originate), and demonstrate near-perfect decoupling of all cosine eigenfunctions such that they become independent eigenvectors. Thus, it is clear that the need for eigenfunction realignment is due to the change in interatomic distances dependent on projection direction (Fig. S25). This disparity among PDs is inevitable, and poses a fundamental problem that must be addressed.

As a remedy to this problem, we aim to stitch the CM information of each Ω_{PD} together into one consolidated orthogonal coordinate system. As already shown, since each CM is represented by a set of orthogonal sinusoids (one per degree of freedom), we thus aim to isolate these sinusoids in their complete form within each PD-manifold eigenbasis. As detailed in section SM-XIV, by use of appropriate rotation operators $R_{i,j}$, the summands within each eigenfunction pair can be maximally separated among a set of eigenvectors (e.g., $\Psi_i = \psi_v$ and $\Psi_j = \psi_w$) such that an ideal (i.e., canonical) eigenbasis is recovered. As a result of this decoupling of eigenfunc-

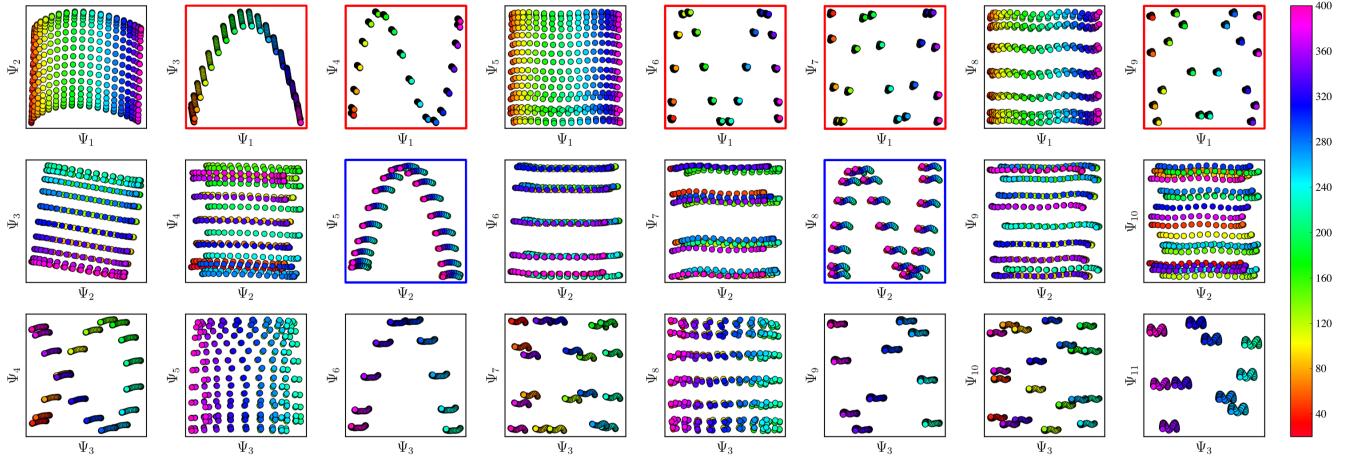


FIG. 4: A subset of the space the 2D subspaces for PD_1 in SS_2 . As demarcated in red and blue boxes, a set of conformational modes exists for both CM_1 (red boxes, $\{\Psi_1, \Psi_i\}$) and CM_2 (blue boxes, $\{\Psi_2, \Psi_j\}$; where $i > 1$ and $j > 2$), interspersed throughout each row. The indices for points in each set of polynomials can be visualized here via the corresponding color mapping, where CM_1 points follow along the full spectrum of colors (i.e., a rainbow with indices 1-400) while CM_2 points are approximately uniform in color map value (i.e., magenta with indices a multiple of 1-20, with all other colors similarly underlaid). Additionally, note the occurrence of the first parabolic harmonic for CM_1 located at $\{\Psi_3, \Psi_6\}$. See Fig. S5 for similar plots obtained for the remaining four PDs.

tions onto a set of appropriate eigenvectors, each corresponding parabolic surface becomes manually aligned within its 2D subspace, such that the projected structure is again that of the 2D Chebyshev parabola carrying information about a single CM along its curve. In this projected view, states differing in coordinates that are orthogonal to the projection plane (and thus describe ulterior CM information embedded on a higher-dimension surface) overlap – a feature we will take full advantage of later when generating 2D conformational movies. Thus, as long as each parabolic trajectory corresponding to a given CM is aligned with the plane of an independent 2D subspace, we can restrict our study to an analysis of only a few essential subspaces.

To provide rationalization for this technique, Fig. 5-A shows the eigenvectors for the highly-misaligned PD_3 eigenbasis, ordered along CM_1 . As seen in the first column of Fig. 5-A, while the sinusoids for $\Psi_1 = \{\cos(\pi x) \mid \text{CM}_1\}$, $\Psi_4 = \{\cos(2\pi x) \mid \text{CM}_2\}$ and $\Psi_5 = \{\cos(3\pi x) \mid \text{CM}_1\}$ are in agreement with expectations, the graphs of $\Psi_2 = \{\cos(2\pi x) \mid \text{CM}_1\}$ and $\Psi_3 = \{\cos(\pi x) \mid \text{CM}_2\}$ appear heavily deformed. As a direct consequence, any Lissajous curve that inherits one of these deformed sinusoids (e.g., any subspace composed in combination with Ψ_2 or Ψ_3) will be misaligned with respect to its ideal form (Fig. 5-B).

Given this insight, we now introduce a method for correcting these misalignments using orthogonal transformations. Specifically, we apply a d -dimensional rotation operator of sufficiently large dimensions to single-handedly reorient all aberrant surfaces in their respective 2D subspaces. The results of this operation on the embedding associated with PD_3 can be seen in Fig. 5-B and Fig. 5-C; before and after applying a 5D rotation matrix, respec-

tively. Mathematically, this d -dimensional rotation is a subgroup of the orthogonal transformation in d dimensions with determinant one. These orthogonal transformations are linear and represented by a $d \times d$ matrix O with the property $O \times O^T = I$, where O^T is the transpose of O and I is the identity matrix. As a consequence, orthogonal transformations leave lengths and angles between vectors unchanged. Each such matrix O can further be represented by $d(d - 1)/2$ rotation sub-matrices $R_{i,j}$, with each sub-matrix parameterized by a unique angle and operating on a specific plane. For the specific case of the 5D rotation matrix used in Fig. 5, there exist 10 rotation sub-matrices in total, with each corresponding to a specific planar rotation on the eigenbasis. Of these 10 matrices, we found that only one had to be altered to achieve the results shown, having general form

$$R_{2,3}(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & \cos(\theta) & -\sin(\theta) & 0 & \dots \\ 0 & \sin(\theta) & \cos(\theta) & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

As this $R_{2,3}(\theta)$ operator corresponds to transformations performed solely on Ψ_2 and Ψ_3 (row 2 and 3, respectively), eigenvectors previously identified as problematic in PD_3 are thus isolated. The result of this transformation on the full set of eigenvectors can be seen in the three columns of Fig. 5-A, which visualize the $R_{2,3}(\theta)$ rotation under 0° , 10° and 20° , respectively (where only Ψ_2 and Ψ_3 undergo change, as expected). Intuitively, the outcome of this operation is equivalent to the traditional notion of a vector rotation (for example, consider $\mathbf{e}_1 = (1, 0) \in \mathbb{R}^2$: just as \mathbf{e}_1 is given some combination of the secondary dimension y with its initial dimension x during a rotation

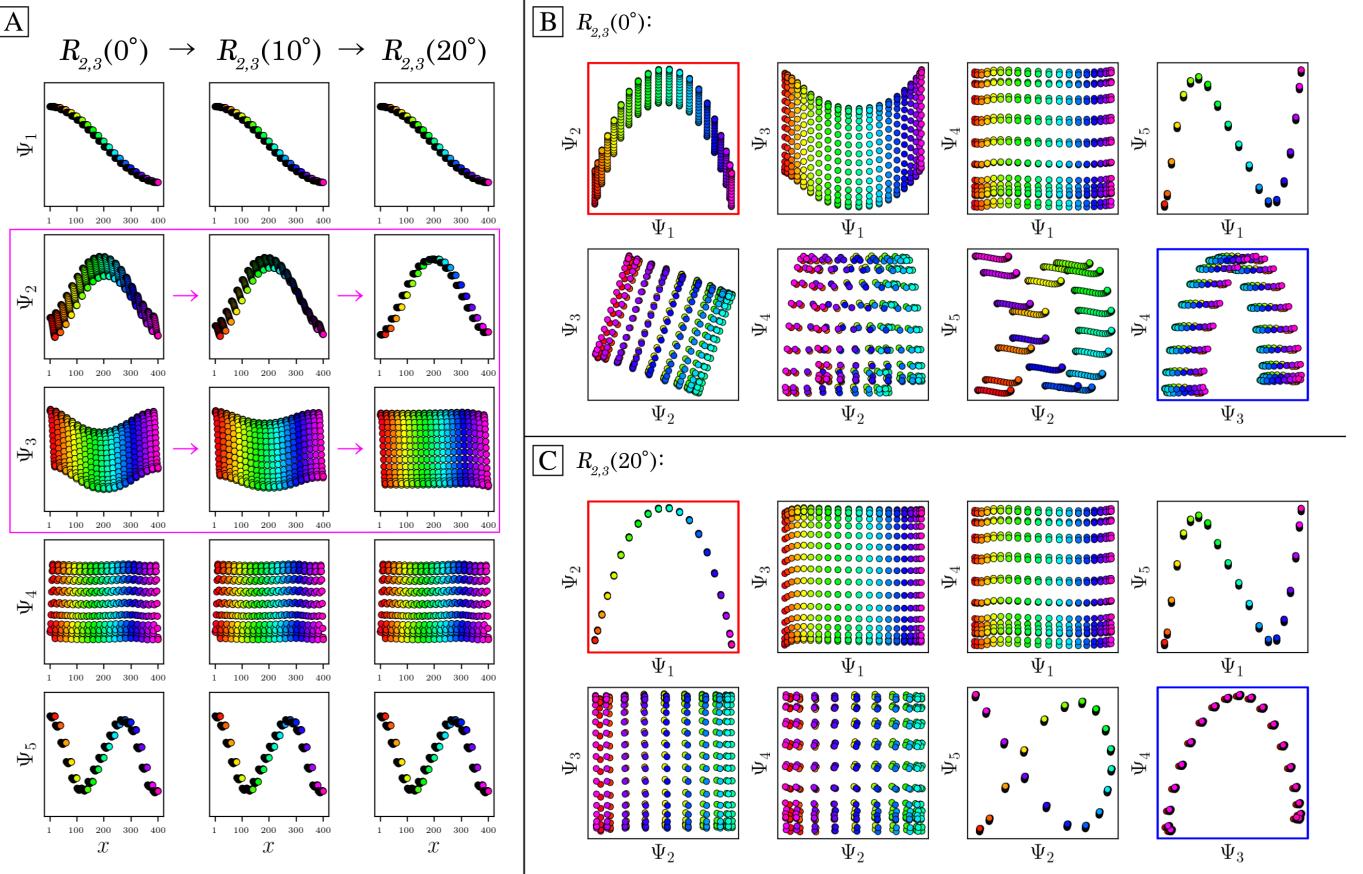


FIG. 5: Application of a 5D rotation matrix $R_{2,3}(\theta)$ on the embedding generated for PD_3 from SS_2 . The three columns in [A] display the individual eigenfunctions (as plotted by indices corresponding to the CM_1 frame of reference) before the rotation is applied, at $R_{2,3}(10^\circ)$, and finally at $R_{2,3}(20^\circ)$, respectively. Note that $R_{2,3}(20^\circ)$ maximally decomposes Ψ_2 and Ψ_3 into unique sinusoids (recalling that the planar distribution in Ψ_3 is in fact a sinusoid when visualized in the CM_2 frame of reference, and vice versa for Ψ_2). The before and after effects of these rotations on the Lissajous curves can likewise be seen in [B] and [C], respectively. Applying $R_{2,3}(20^\circ)$ properly orients both parabolic surfaces corresponding to CM_1 and CM_2 (denoted with red and blue boxes, respectively), such that the eigenvectors are orthogonally aligned with the eigenbasis of the CMs.

(via vector additions and scalar multiplications), so too is Ψ_2 weighted with Ψ_3 , and vice versa. As seen in Fig. 5-A, along the way in reaching $R_{2,3}(20^\circ)$, Ψ_2 and Ψ_3 have effectively transferred between each other an equal share of their initial content via a series of continuous deformations, with each initial eigenvector thus sharing some combination of the other's initial sinusoidal form (as is also analytically demonstrated in Fig. S20-C). After this exchange, the initially overlapping sinusoidal information contained in part between Ψ_2 and Ψ_3 is maximally separated between both eigenvectors, ultimately resulting in the alignment of all corresponding Lissajous surfaces with their 2D subspaces (Fig. 5-C), as desired. Later in our analysis we will return to this topic under the moniker *eigenfunction realignment*, and describe a strategy for automating these corrective actions for noisy datasets.

C. Data-type I in State Space 3

We next investigated the 1000 states making up SS_3 . As before, the eigenvalue spectra were similarly found to be slowly decreasing, but falling off more gradually than in the SS_1 and SS_2 spectra. Additionally, for each conformational motion present in a given PD dataset (this time for CM_1 , CM_2 and CM_3), a set of unique Lissajous curves were again found spanning specific 2D subspaces of the embedded manifold, with the Chebyshev subset explicitly describing the corresponding CM along a 2D trajectory explicitly. Fig. S7 shows the set of 2D subspaces where these modes exist for PD_5 . To note, due to the increased complexity of the SS_3 state space, these patterns were much more interspersed throughout the N -dimensional embedding, but still followed a similarly predictable ordering. In addition, due to the relatively small range of motion exhibited by the third conformational domain (as seen from these PDs and as designed in the ground-truth

structures), all CM_3 modes were found in higher-order eigenvectors; e.g., Ψ_5 and higher for these five PDs. As similar patterns were identified in SS_3 as in previous accounts, for the remainder of our paper, we will hone our focus on mapping datasets generated specifically from SS_2 .

II. Principal Component Analysis

Data-type I in State Space 2

Following our analysis of manifolds using the DM framework, we next performed linear dimensionality reduction on the SS_2 images in PD_1 using PCA. Instead of defining the Gaussian kernel as previously used in the Markov transition matrix, we performed PCA on the array of all pixels, with dimension defined by the number of images and pixels in each image (i.e., on a dataset Z of dimension $P \times N$). Before embedding, we standardized the images in each dataset by removing the mean and scaling to unit variance, and generated eigenfunctions of the resultant $N \times N$ matrix $Z^T Z$. To note an important comparison between PCA and DM, the matrix $Z^T Z$ is symmetric and positive semi-definite (i.e., all eigenvalues are non-negative)⁴³, which is also the case for the Markov transition matrix used in the DM framework.

A set of different projections of this embedding as obtained from selected eigenvectors (i.e., principal components, PC_i) can be seen in the first column of Fig. S17, with results from DM similarly presented for comparison in the fourth column. As demonstrated, the eigenvalue spectra and eigenvectors obtained from performing PCA and DM are almost identical, except for subtle differences in the spacing between states and boundaries for the pristine case (SNR_∞ , i.e., data-type I). These similarities align with our *a priori* knowledge of the existence of quadric surfaces for positive semidefinite matrices, as described in the section SM-XII). The similarity between these manifolds holds for all subspaces explored, and, as will be seen in the next section, the distinction is diminished in the presence of noise. The results of PCA versus DM on SS_1 and SS_3 show similar behavior.

III. Influence of SNR and Statistical Coverage

Data-type II in State Space 1 and 2

As SNR is an important attribute of any experimental dataset, we next sought to understand how the structure of these manifolds change with varying SNR and state space coverage. To this end, we first compared the manifolds from PCA and DM for PD_1 with additive Gaussian noise (generated as described in section SM-IV), such that the images in each dataset had unique $\text{SNR} \in \{1, 0.1\}$ consistently applied to all images in a set. The results of this procedure can be seen in the remaining

columns of Fig. S17. For both dimensionality reduction techniques, the fidelity of the resulting spectral geometry to the state space ordering decayed with increasing noise level. Within low-SNR regimes, the behavior of the embeddings from PCA and DM were highly consistent, with DM generated within its optimal range of Gaussian bandwidth. Overall, the corresponding spectral geometry obtained from each framework became increasingly similar as the SNR was decreased (Fig. S18).

We next investigated the effects of varying state space coverage across several SNR regimes, and its effects on the robustness of the corresponding manifolds produced by PCA. As the choice of PCA or DM proved irrelevant in these low-SNR regimes, PCA was chosen here so as to bypass uncertainties introduced by the need of additional parameters in DM; i.e., the Gaussian bandwidth. For this study, we used the 20 images in PD_1 representing SS_1 (i.e., one full range of conformational motion), and varied both the number of times (τ) these $M = 20$ ground-truth states were duplicated as a group—with each instance having a different realization of additive Gaussian noise—and the SNR of each image therein. Here, Gaussian noise of constant variance was applied for each SNR regime and uniquely added to each of the $\tau M = N$ images independently, as shown in Fig. S8. An excerpt from the results of our analysis is shown in Fig. 6, where a highly structured pattern emerged. Specifically, when increasing levels of noise was added to each image (decreasing SNR), increasingly larger values of τ were required to reestablish coherent structure in the spectral geometry; i.e., the set of Lissajous curves and corresponding Chebyshev polynomials.

To quantify these relations, each member of the set of PCA-embedded manifolds in a τ -series was fitted with a set of leading Chebyshev polynomials, as seen in Fig. S9 for $\text{SNR} = 0.1$. The coefficient of determination (R^2), which can be interpreted as the proportion of variance in one variable accounted for by another⁴⁴, was then computed for each mode therein. The resultant trends across several SNR regimes are plotted in Fig. S10. Our findings show that as τ is increased, the rate at which the geometry of each subspace reaches its most stable regime is dependent on SNR. A critical τ_c value was determined both visually and analytically by assessment of the asymptotes for each SNR regime, beyond which larger values of τ provided no further improvement to the spectral geometry.

Further, across all of these regimes, each subsequently higher-order Chebyshev polynomial required a larger value of τ_c to be properly resolved (Fig. S10-A), which is a consequence of higher-frequency patterns requiring more points to resolve when the number of their points (i.e., images) is held constant. For our purposes, recall that the accurate acquisition of only the parabolic trajectory is relevant. As τ_c fluctuates based on numerous unknowns in the experiment, determination of its value for a given experimental dataset is infeasible. Parameters influencing τ_c include not only unknowns such as the number of ground-truth states M and SNR regime, but

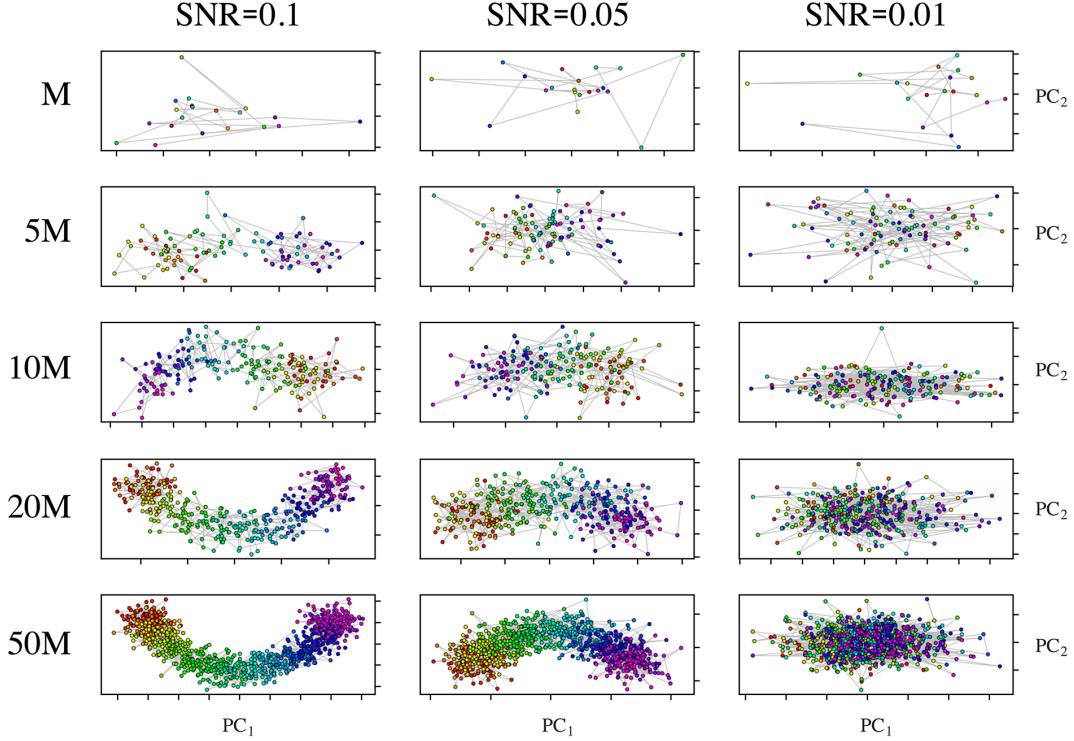


FIG. 6: Set of $\{\text{PC}_1, \text{PC}_2\}$ subspaces produced by PCA from PD_1 images in SS_1 over a range of SNR values and state space coverage. As can be seen in the columns, the fidelity of the point-cloud distribution in each subspace to the parabolic form increasingly deteriorated due to decreasing SNR regimes. However, as the $M = 20$ state space was populated by increasing values of τ in each of these SNR regimes, the intrinsic parabolic structure of the embedding reemerged. To be precise, $5M$ represents five exact copies of the 20 SS_1 images ($\tau M = 100$ images), with unique Gaussian noise added to each image independently as prescribed by its SNR regime. It can be seen here that all values of τ shown (up to 50) in the SNR=0.01 regime are too low for recapitulating the intrinsic parabolic structure of the embeddings, and, as further illustrated by the color mapping of their points, no sensible ordering of snapshots can be ascertained within these subspaces.

also the intrinsic dimensionality of the dataset and the free energy of the system.

We next describe specific characteristics of CM subspaces obtained from datasets generated with these noisy-duplicate images. Specifically, we examine the parabolas generated via PCA from SS_2 PDs with $\text{SNR} = 0.1$ and $\tau = 10$, which will guide several choices made for our framework in the following section. Fig. 7 shows the composite parabolic trajectory and corresponding sinusoidal form of each eigenfunction for CM_1 and CM_2 of PD_1 , as well as a collection of similar CM subspaces from randomly-selected PDs. Each subplot has been assigned a color map matching the ground-truth sequence of states of the CM to which it corresponds with this sequence partitioned into 20 equally-occupied bins (i.e., CM states). As can be seen, while each of the two underlying point clouds corresponding to a unique eigenfunction maintains well-defined structure after introduction of noise, CM state partitioning becomes increasingly disordered in their composite parabolic point cloud.

Additionally, due to PD disparity, the characteristics of each CM-parabola can be seen to vary significantly de-

pending on viewing direction. These variations include average thickness, length, density, trajectory, and spread of data points in each parabolic point cloud, with aberrations occurring most frequently in CM subspaces generated from PDs where the apparent range of the given CM is diminished. As a result, while the CM subspaces for all PD manifolds carry reliable content for recovery of 3D density maps along a conformational trajectory, certain clusters of $\text{PDs} \subset S^2$ offer less reliable geometric structure for accurately estimating occupancies of CM states therein. From these initial observations, it is clear that effectively delimiting states in these highly-variable subspaces will require robust solutions to be subsequently explored. As a final note for this section, all trends described here for PCA were likewise found to exist for embeddings of manifolds obtained using DM (Fig. S18).

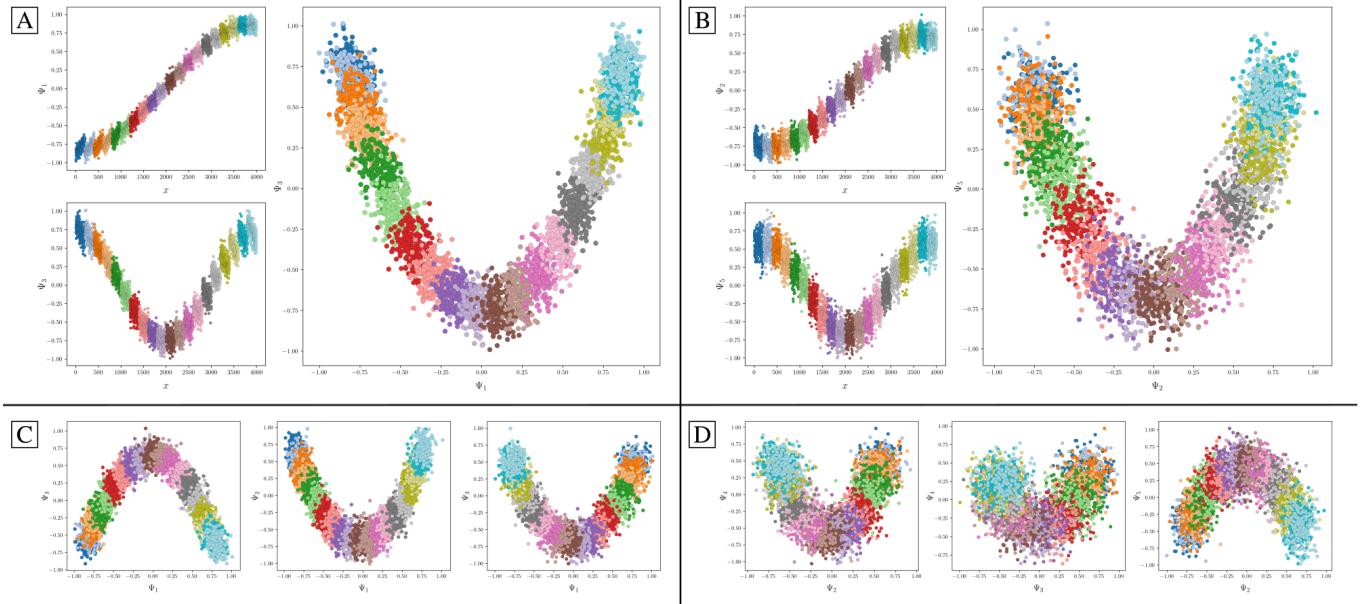


FIG. 7: CM subspaces for a set of five PDs generated with $\text{SNR} = 0.1$ and $\tau = 10$ and embedded via PCA. The coordinates within each point cloud are colored to indicate their ground-truth CM state assignment, such that each point belongs to one of 20 CM bins, and each bin contains 200 points (with the same coloring scheme used regardless of CM). In [A], the parabolic CM₁ subspace of PD₁ is shown along with its two leading cosine eigenfunctions (with each cosine ordered according to its ground-truth sequence). Similarly, in [B], the parabolic CM₂ subspace of PD₁ is shown with its own set of leading cosine eigenfunctions. The remaining subplots show a variety of CM₁ [C] and CM₂ [D] subspaces for three randomly-oriented PDs, so as to emphasize the variability in features prevalent in manifolds embedded from noisy images.

IV. Influence of TEM Contrast Transfer Function

Data-type III in State Space 2

Carrying forward our knowledge gained from evaluation of data-types I and II, we next turn to data-type III for analyzing the PD manifolds obtained from image ensembles generated with experimentally-relevant CTFs and SNR as is encountered in a Transmission Electron Microscope (TEM). For these trials, we first generate and apply a CTF to each image as described in section SM-VI. Specifically, using images from PD₂ of SS₂ with $\tau = 10$, we assign to each image a random defocus value from the interval [5000, 15000] Å. Such a wide range is usually chosen to compensate for the zero-crossings of CTFs where no information is transferred (Fig. S12), with similar intervals typically used in modern cryo-EM experiments. Likewise for each image, constant values are used for voltage (300 kV), spherical aberration coefficient (2.7 mm), and amplitude contrast ratio (0.1) to emulate typical TEM conditions. These parameters are jointly used to construct a unique CTF for each image, which is applied via multiplication to the image's Fourier transform. With the collection of images modified by unique CTFs, additive Gaussian noise is next applied such that the SNR of each image in the resultant ensemble is approximately 0.1.

We next set out to measure the extent of interference of the Contrast Transfer Function on the corresponding

manifold for an example PD. However, since the collection of images are now sampled using a range of defocus values, they are no longer directly comparable using a standard distance metric. Instead, an adjustment to the kernel must first be made to account for our introduction of CTF. We show here the results of applying the previously established *double-filtering* kernel, which ensures a zero Euclidean distance between any two images that differ only in defocus¹. In application, during each pairwise Euclidean distance calculation, the Fourier transform of each image is multiplied by the CTF of the image under comparison to compensate for the defocus difference. The corresponding manifold embedding is shown in Fig. 8, which juxtaposes these results with the same dataset generated without CTF and using a standard Gaussian kernel.

As seen on the right-hand side of Fig. 8, there is a noticeable inward-curling at the ends of the CM subspace parabolas generated using the double-filtering kernel. Notwithstanding this artifact, we found that the double-filter kernel was successful in preserving the most important aspects of the manifold. This approach also proved superior to alternative techniques explored, such as embedding using a standard kernel from sets of CTF-corrected images. To note, perfect defocus assignments were used here for CTF correction, when in reality these values would be estimated first using established algorithms^{45–47}.

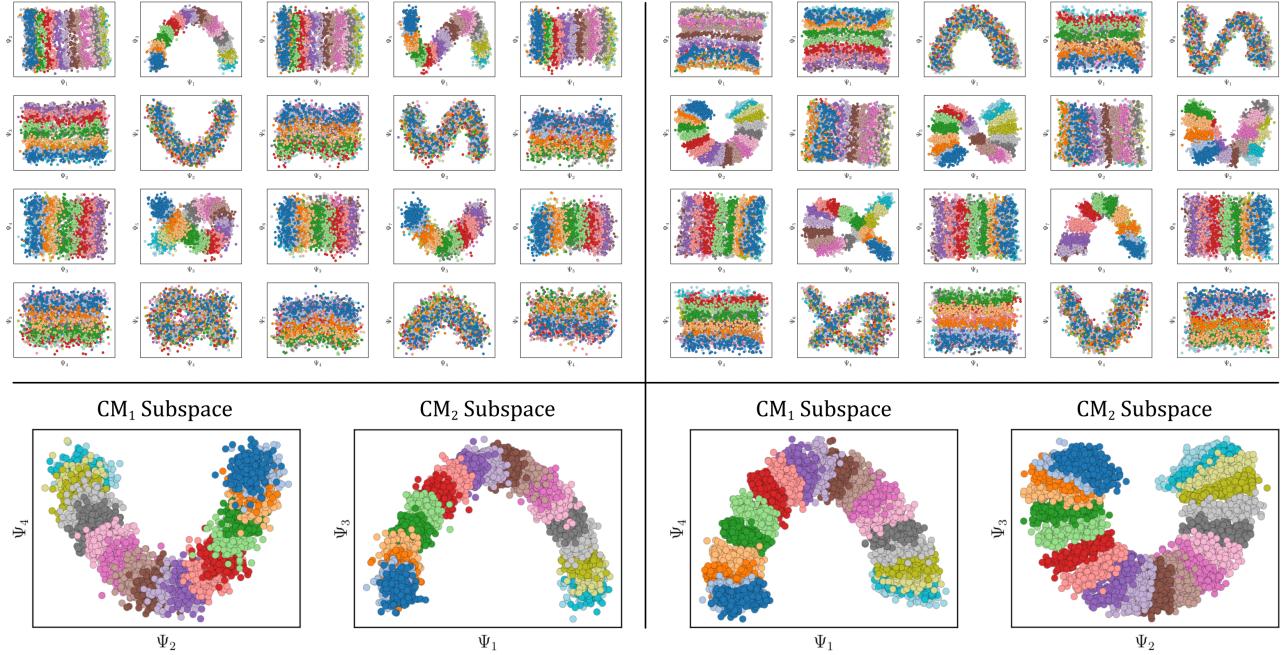


FIG. 8: SS_2 manifold embeddings (PD_2 , $\tau = 10$, $\text{SNR} = 0.1$) obtained with and without CTF applied, as shown on the right and left-hand side, respectively. For the case of the embedding obtained from images without defocus, protocols for synthetic generation follow those established in Fig. S12 (A, B). Likewise, on the right, protocols follow synthetic generation of images with microscopy parameters as shown in Fig. S12 (D, E). The non-CTF manifold embedding was generated via DM with the standard Gaussian kernel, while the CTF-manifold embedding was obtained via DM with the double-filtering kernel. On the top insets, colors displayed represent the ground-truth CM_2 bins, while for the bottom insets, for both sides, colors represent CM_1 bins (left) and CM_2 bins (right).

V. Overview of the ESPER Framework

Having explored all three data-types including CTF and noise, we now lay out the ESPER strategy for recovery of conformational motions in the form of 3D movies and a corresponding free-energy landscape. This methodology requires several steps that will be introduced in turn, with the entire process schematized in Fig. 9.

The general intuition for our approach is as follows. Ideally, for each Ω_{PD} embedding, we first wish to translate the n conformational-variation signals residing along a high-dimensional parabolic surface into a rectilinear n -dimensional state space. To this end, one can imagine forming a coarse n -dimensional grid along this desired hypersurface—with each n -cube (bin) on the grid nonuniformly stretched to occupy an equal volume as required to account for nonuniform rates of change along its complex surface—and accruing the set of points (and thus indices of corresponding images) falling within each bin’s boundary. This procedure should then be repeated for each Ω_{PD} independently. To reconcile the contents of these PD manifolds on S^2 , which may contain conformational information along different coordinates due to PD disparity, the orientation of each n -dimensional grid (and thus ordering of bins therein) must be aligned so as to match across all PD manifolds. Next, the set of images belong-

ing to each compiled bin can be combined to reconstruct a 3D density map of the molecule, with the total image count used to define a state occupancy. As a result of this construction, an n -dimensional occupancy map (and thus free-energy landscape) can be formed, along with a set of corresponding 3D density maps representing every state.

In application, however, there are many complications to this procedure. For one, the desired high-dimension parabolic surface presents difficulties in both discovery and direct mapping. Since there are many such potential subspaces housing parabolic surfaces (even for $n = 2$) within the embedding of a given Ω_{PD} , there exists ambiguity as to which one contains the desired information, further exacerbated in the presence of harmonics and experimental artifacts. In addition, due to the complex nature of these hypersurfaces—which can vary in features ranging from elliptical, to parabolic, to hyperbolic, and each with boundary aberrations—it is much easier to instead fit and partition the set of its orthogonal components; i.e., the parabolas residing in easily identifiable 2D subspaces. Given this route, several operations can next be performed on these parabola-housing subspaces to approximate an idealized, straightened trajectory for each CM, ultimately allowing the formation of a rectilinear coordinate system when this set of straightened CM

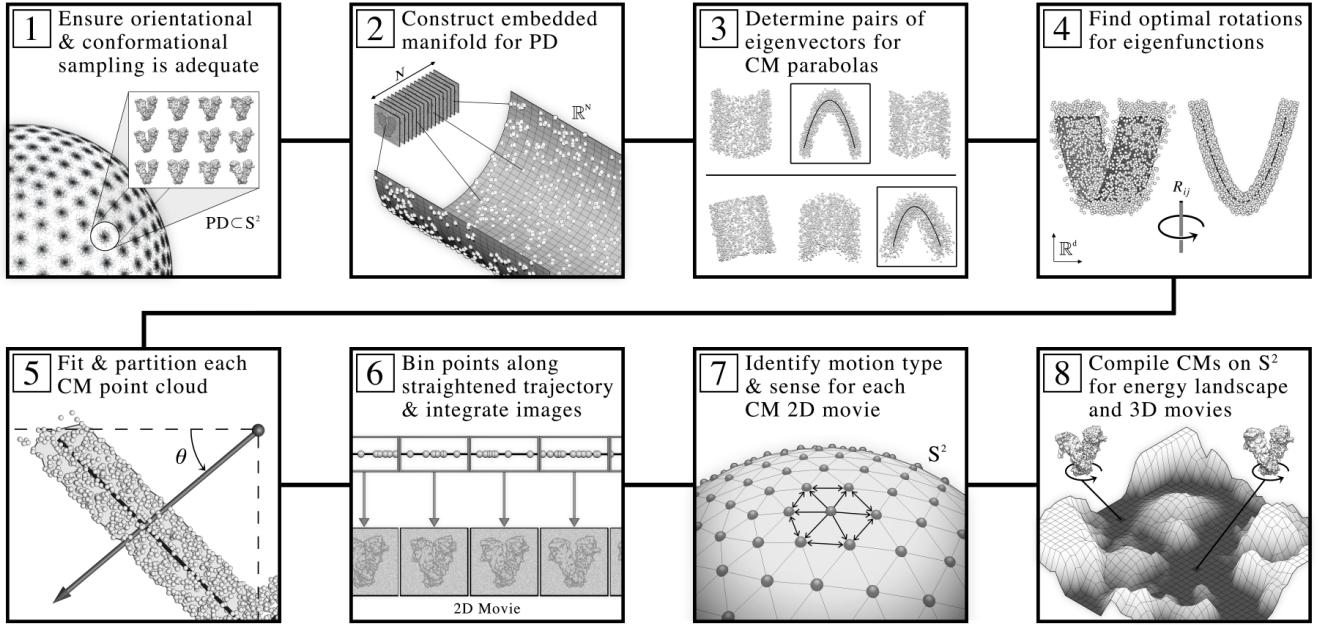


FIG. 9: Schematic detailing the ESPER workflow for recovery of conformational continuum as informed by our heuristic analysis. Through this framework, 3D movies and corresponding free-energy landscapes are obtained for the set of conformational motions in a given dataset. Note that the previous ManifoldEM workflow branches off after completion of step 2 above, and after performing a series of alternative steps required by NLSA, then enters again with our pipeline at step 7, before splitting off again to form final outputs (as we achieve independently via ESPER) in step 8. As will be described fully in our Discussion, given certain requirements are met in the quality and structure of a dataset, our method provides an alternative avenue to NLSA for obtaining conformational outputs within the ManifoldEM framework.

trajectories are recombined. Finally, these rectilinear coordinate systems must be organized in such a way that CM content is matched across all PDs, and compiled.

Following this rationale, the ESPER approach first aims to find the set of parabola-housing subspaces (via least-square conic fits) required for elucidating higher-dimensional CM information, while accounting for eigenvector rotations and carefully eliminating harmonics [steps 3, 4]. Each parabola-housing subspace is next transformed (along with its conic fit) via the inverse-cosine mapping to account for nonlinear rates of change, with points partitioned into contiguous equal-area bins [step 5]. Each bin is then filled with a set of image indices corresponding to all points falling within its geometric bounds. Images belonging to each bin are next integrated to form the frame of a 2D movie [step 6], which is used to identify both the type of CM and its directionality (i.e., *sense*) of its motion [step 7]. As the location of each point (and thus image index) present in a given CM subspace is coupled to its coordinates in all other orthogonal CM subspaces (on the high-dimensional surface), we can reconstruct this joint geometrical relationship using only the intersection of image indices obtained in all pairwise combinations of bins spanning all CMs. By means of this approach, when this information is accumulated across all PD manifolds, the desired occupancy map and index sets required for full recovery of 3D electron density

maps in all bins are thus obtained [step 8].

In the following three subsections, we provide a more detailed description of these steps. For the purposes of this exposition, we will use SS_2 via data-type II for initial demonstrations of eigenfunction realignment and subspace partitioning (subsection A and B, respectively), followed by use of our final analysis data-type (i.e., SS_2 via data-type III with free-energy landscape) in subsection C to furnish final outputs and assess their validity.

A. Eigenfunction Realignment for Data-type II

We describe here our methodology for calculating the rotations required for eigenfunction realignment of each embedded Ω_{PD} in the presence of noise with experimentally-relevant SNR. We consider this calculation to be the first step in the ESPER framework (i.e., [step 3] in Fig. 9). To note, after generation of each embedding from PD-images as previously described [steps 1, 2], our methodology deviates here from the existing ManifoldEM method^{1,10,11}, which would next move on to NLSA without accounting for eigenfunction realignment on Ω_{PD} subspaces. First, recall that depending on the PD, the observed conformational eigenfunctions may be misaligned with respect to the ideal eigenfunctions of the LBO, requiring application of a d -dimensional rota-

tion matrix to align the subspace (Fig. 5). As an example, the effects of applying a 4D rotation to the 4D subspace for data-type II corresponding to PD_2 (containing all parabolic modes of CM_1 and CM_2 in SS_2) can be seen in movie M5, where only one of the six required rotation matrices is altered by 28.65° (with the remaining five unaltered; i.e., 0°) to single-handedly realign both parabolic modes (one per CM) to the plane of their respective 2D subspaces.

Based on this behavior, we have developed a technique to automate the discovery of the rotations required to realign essential eigenfunctions in each Ω_{PD} embedding. This algorithm is informed, first and foremost, by our heuristic findings of the existence of parabolic surfaces in each embedding, which correspond to a specific CM. In the case of noisy data, as each corresponding 2D subspace is rotated, it exhibits a unique profile that can be characterized by a sequence of 2D histograms on that subspace, with one 2D histogram per each rotation angle corresponding to a given R_{ij} rotation operator. When we plot the number of nonzero bins in the corresponding 2D histogram as a function of rotation angle, the minimum in this distribution corresponds to the angle required to properly counter-rotate each 2D subspace by the current operator (movie M6). After careful observation of all PDs across numerous datasets, we have determined that the exact rotational operators R_{ij} required to adequately rotate each 2D subspace are linked to the indices of those eigenvectors housing each CM parabola. As a consequence, we need to first determine the 2D subspaces housing parabolas, and these are identified via the best least-squares fits in each eigenvector row (movie M7). A detailed description for the procurement of this information is available in the section SM-XVI, with an example visualization of this workflow provided in movie M7.

Once these CM subspaces have been isolated, a final 2D in-plane rotation still needs to be applied to orient the parabola into its canonical form. We thus perform a least-squares fit Ψ_{fit} using the implicit equation of a general conic defined by an irreducible polynomial of degree two

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

This general conic form can account for parabolas, ellipses or hyperbolas (discriminant $b^2 - 4ac$ equal to zero; less than zero; or greater than zero, respectively). As will be seen, this flexibility is essential for fitting parabolic-like point clouds with nonzero discriminant, which are encountered for point clouds with boundary aberrations, and especially those obtained from images modified by the CTF. In this form, the xy -term rotates the graph, providing for the possibility of encountering subspace trajectories with an axis of symmetry unaligned from the 2D eigenbasis. This equation can thus be rewritten with a new set of coefficients⁴⁸ to effectively rotate the coordinate axes such that they come to alignment with the axis of symmetry.

B. Subspace Partitioning for Data-type II

Once the required Ω_{PD} eigenfunctions are correctly rotated into a common eigenbasis as defined by the desired CMs, and each 2D subspace housing CM information is identified for each PD [steps 3, 4], we next partition these 2D subspaces into contiguous equal-area bins [step 5] representing a quasi-continuum of conformational states. Here ESPER differs decisively from the preexisting ManifoldEM workflow encompassing NLSA. To help distinguish between these strategies, a brief summary of the NLSA workflow has been provided in section SM-XVIII.

The motivation for the ESPER subspace-partitioning approach stems from our analysis of PD disparity in the presence of noise (as shown in Fig. 7), where it is observed that the ground-truth bins and overall area of each point cloud manifest in a variety of sizes depending on PD viewing angle. These observations inspired an area-based point-cloud fitting approach able to correctly chart spatial discrepancies while remaining unencumbered by changing densities (i.e., occupancies) along each trajectory. Fig. 10 provides an overview of our novel strategy for splitting up each CM subspace into a sequence of equal-area bins, with subplots detailing recovery of CM_1 states and corresponding occupancies for a single PD.

To initiate this procedure, for each CM subspace of a given Ω_{PD} , we first scale both eigenvectors $\{\Psi_i, \Psi_j\}$ between $[-1, 1]$ and apply an inverse-cosine transformation on each: $f : \{\Psi_i, \Psi_j\} \mapsto \{\Phi_i, \Phi_j\}$. As previously shown in Fig. S4, it is expected such a mapping will induce a space with uniform rates of change between states. To mitigate the overall complexity of operations, the axis of symmetry of the conic equation (with $b' = 0$) is then used to split the subspace into two halves, such that each half can be operated on individually. We next apply a ball tree algorithm^{49,50} to temporarily prune outlier points for heightened accuracy during subsequent steps. Specifically, the ball tree approach clusters points in a series of nesting 1-spheres based on the Euclidean metric, from which we select only those clusters having a minimum number of members.

Following this preparation, we define the overall area of each halved subspace with a polygon enclosing a majority of our remaining points (Fig. 10-B and Fig. 10-C). This construction is achieved via the alpha shapes algorithm^{51–53}: a generalization of the convex hull that defines the boundaries of the point cloud by a series of α -discs (1-spheres of radius $1/\alpha$), such that an edge of the alpha-shape (polygon) is drawn between two members of the point cloud whenever there exists an α -disc containing no members of the point cloud and carrying the property that the two points lie on its boundary. A family of alpha-shapes can thus be defined for each halved subspace via the α -parameter, ranging from coarse (a convex hull) to increasingly finer fits around the point cloud. Within this class of polygons, there exists a member providing an optimal level of refinement, in which the alpha-shape area and point-cloud area are equal⁵².

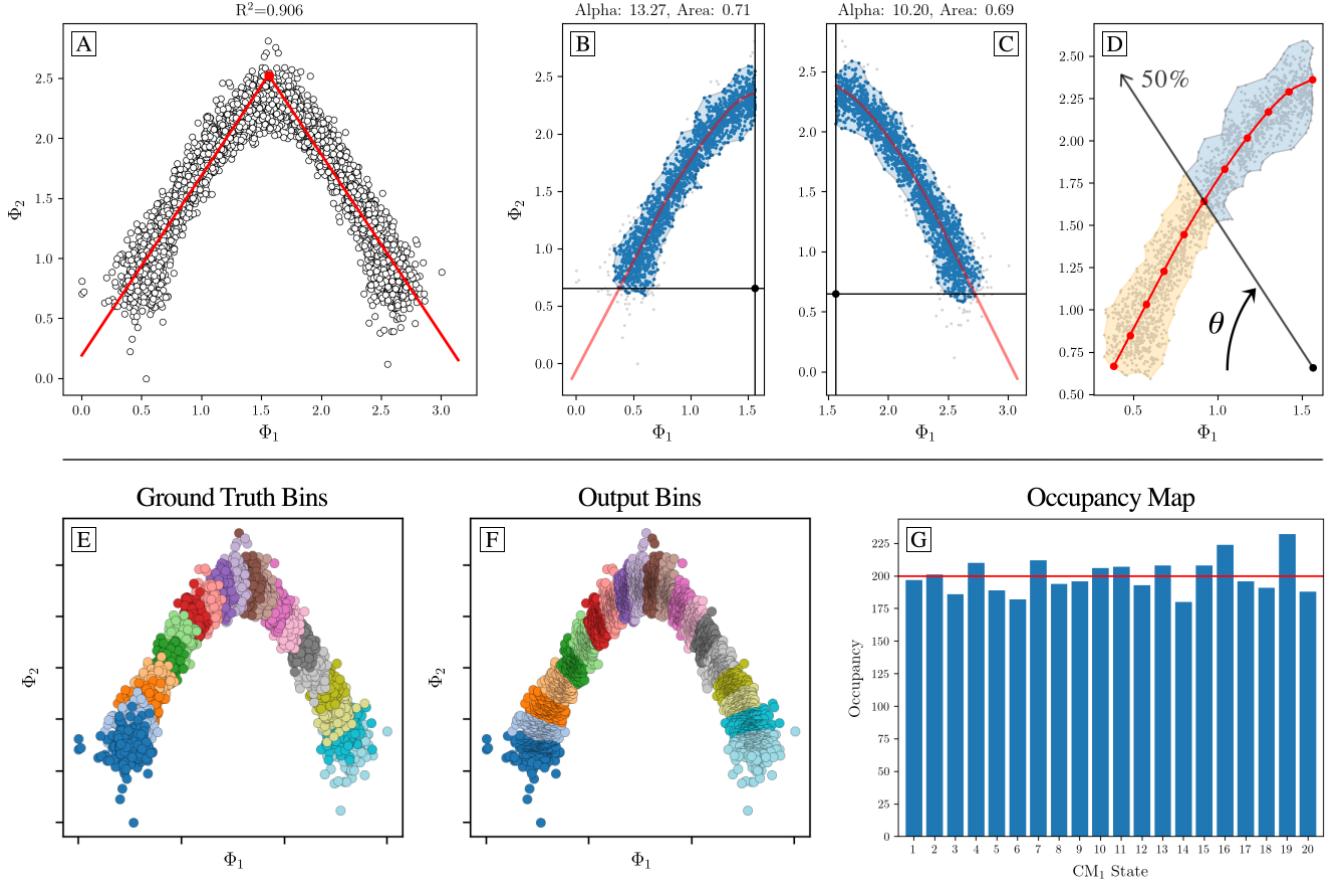


FIG. 10: Overview of area-based method for extracting sequential conformational information from a given CM subspace. Subplots [A] through [G] display our algorithm’s outputs on the CM₁ subspace of an arbitrary PD from data-type II. First, [A] shows the inverse-cosine transformation and corresponding preliminary fit using an absolute value function. Subplots [B] and [C] demonstrate the alpha-shape polygon and Φ_{fit} trajectory defined on each halved subspace, with the anchor-point designated within the central alcove. In [D], a ray is shown passing from the anchor-point through the point cloud. At the current angle θ shown, half of the area of the alpha-shape has been traversed, demarcating the boundary between the 5th and 6th (of 10) CM₁ bins. Subplots [E] and [F] compare the ground-truth bins—as visualized via the known sequence of images in each state—with the final output bins defined via this framework. Finally, the 1D occupancy map is provided in [G], where the horizontal red line (200 images) represents the ground-truth occupancy assignment per CM₁ state.

For our purposes, the determination of a suitable value for this parameter was automated by generating a sequence of alpha-shapes of increasingly finer complexity up until the resulting alpha-shape – previously defining one polygon – collapsed into two polygons. Through this construction, our point cloud is enclosed by a fine polygon representing the key features of its geometric shape.

Next, the general conic fit Ψ_{fit} is transformed by inverse cosine to form the trajectory Φ_{fit} and split along its axis of symmetry for use on each half of the $\{\Phi_i, \Phi_j\}$ subspace. The intersection of Φ_{fit} with the outer boundary of the alpha-shape is used in combination with the position of the initial vertex to form a new anchor-point nested within the central alcove of the point cloud (Fig. 10-B and Fig. 10-C). For each image-point in the point cloud, a ray is next drawn connecting the anchor-point with the image-point, with the intersection of that ray with

Φ_{fit} recorded. As a result of this construction, all image-points are uniquely projected onto the Φ_{fit} trajectory. With each image-point now assigned an index along Φ_{fit} , a method is next employed to partition the trajectory into segments representing CM states, such that each image-point is ultimately assigned to a single state. Here, we define a ray emanating from the anchor-point, initiated with $\theta = 0^\circ$ (Fig. 10-D). For each angle $\theta \in [0^\circ, 90^\circ]$, the area of the lower sub-polygon formed by the intersection of the alpha-shape with the ray is determined⁵⁴ to record the overall ratio of that sub-polygon to the whole. We form 10 bins in this fashion along the Φ_{fit} trajectory (per halved subspace), making up 20 bins in total. Finally, we tally the number of image-points assigned to each of these 20 bins (as visualized in Fig. 10-F) to form the 1D occupancy map for the current CM. Importantly, we store the image indices belonging to each bin along

the given CM for subsequent use in forming an $n > 1$ occupancy map (to be detailed in the following section C).

A comparison of our outputs with ground-truth is provided in Fig. 10-E and Fig. 10-G, showing an overall agreement with expectations. To note, we have programmed these steps to require no intermediate supervision by using a robust automation strategy, with details for specific subtasks available in the comments of our corresponding code³⁷. As a result of this procedure, the points in the embedding corresponding to each conformational motion are independently lined up within a corresponding 2D subspace, such that averaging points together in that subspace only reveals the conformational-variation signal corresponding to the current CM. Hence, images in each bin can next be averaged to generate each frame of the respective CM’s 2D movie. This process is then repeated for the 2D subspace where the second parabolic mode resides (CM_2), and so on for higher degrees of freedom. The results of this procedure can be found in supplementary movie M2, where we showcase 2D movies for both CM_1 and CM_2 as obtained from a subset of these 126 PDs.

C. Final Analysis – Recovery of 3D Conformational Motions

We demonstrate here the efficacy of our entire framework with a comprehensive dataset of PDs occupying states in SS_2 , using ground-truth images modified with experimentally-relevant CTF and noise. We will additionally note slight alterations to the previously-described ESPER methodology [steps 3-6] that are required for handling our final analysis dataset, which now includes introduction of CTF as well as noise. Finally, once 2D movies have been obtained from each Ω_{PD} , we describe here the conclusion of the ESPER framework via an efficient method for compiling CM information from all PDs [step 7] to create a free-energy landscape and corresponding set of 3D movies [step 8].

First, we note that the minimum number of equispaced PDs (PD_{min}) on a great circle required for 3D tomographic reconstruction at a given resolution is defined by the Crowther criterion⁵⁵ $PD_{min} = \pi D/r$. Here D is the particle diameter (120 Å, as measured in state 20_20 of SS_2) and r is the targeted resolution of the reconstructed volume (for our purposes, 3 Å as chosen to match the resolution of our ground-truth maps). According to this criterion, we generated 126 equidistant PDs spaced approximately 1.5° apart along one half of a great circle (labeled as Great Circle 1 in figures that follow), chosen so as to avoid redundant information due to diametric mirroring. Each of the 400 SS_2 states in each of these PDs was then duplicated based on assignments imposed from a fictitious occupancy map (see section SM-VII) resulting in 4000 images per PD, with each particle modified by an individual CTF having randomly-assigned defocus [5000, 15000] Å and the same microscopy parameters

as previously described. Finally, additive Gaussian noise ($SNR = 0.1$) was applied to each image such that 504,000 unique images were created in total. Euclidean distances among the 4000 images within each PD were then calculated using the aforementioned defocus-tolerant kernel with matching CTF assignments. Finally, following the DM framework, a Markov transition matrix was produced for each distance matrix and diagonalized for subsequent Ω_{PD} analysis.

As was discovered for data-type I and II, we found that a substantial number of these 126 PD-manifold embeddings had misaligned eigenfunctions from our preferred, common coordinate system, with the magnitude of counter-rotations required varying significantly from one PD to the other. For all 126 PDs, our previously-described rotation-automation strategy (see ESPER subsection A) correctly isolated CM_1 and CM_2 and counter-rotated each CM trajectory into the plane of its 2D subspace. As one small adjustment, we substituted the use of the general conic fit for constrained parabolic fit at the beginning of our eigenfunction-rotation algorithm, which proved essential for evaluating initial subspaces housing inward curling parabolas due to the presence of CTF (Fig. S14). A histogram of the magnitudes of rotations used across all rotation operators—which varied nontrivially among the PDs—is provided in Fig. S28.

With the required CM eigenfunctions in these 126 PD-manifold embeddings correctly rotated into a common eigenbasis and each 2D CM subspace identified, we next proceed to partition each 2D subspace into bins representing different conformational states. This procedure follows our previous description (see ESPER section B), with only slight modifications required for the case of images modified with CTF. First, to combat the effects of inwards curling (Fig. 8) that now encumber ray projections, the location of the anchor-point was moved closer towards the central alcove of the Φ_{fit} trajectory, such that it comes to reside halfway between the center of the $\{\Phi_i, \Phi_j\}$ subspace and its previous position on the y -axis. Additionally, the emergence of a 20-PD “blind spot” on our 126-PD great circle—where images of CM_2 were highly obfuscated while CM_1 remained pronounced—inspired the creation of an alternative branch in our partitioning procedure. In the presence of CTF, we found that while CM_1 subspaces remained highly parabolic for these 20 PDs, the parabolas corresponding to obfuscated CM_2 signals were much less appreciable in structure. These instances could be easily predicted by simple evaluation of each subspace’s coefficient of determination (R^2), with a low score indicating that the conic fits and corresponding anchor-points had proven suboptimal. Using an R^2 cutoff (0.6) as a criterion, we devised an alternative strategy for dealing with these aberrant subspaces, where, instead of a general conic fit, the subspaces were fit using an absolute-value function (as is shown, for demonstration only, on a more well-behaved subspace in Fig. 10-A), and anchor point re-assigned such that nearly vertical projections were taken across the point cloud. In effect,

of the two eigenvectors making up the subspace coordinates, the influence of the leading eigenvector was made more prominent, such that the contiguous bins were delimited with near-vertically aligned borders. This added flexibility made our partitioning algorithm robust in the presence of less-structured distributions, with results validated via examination of ground-truth bins.

Finally, we note that when CTF-modifications are present, we first individually CTF-correct and Wiener-filter⁵ each image before integration within each CM bin. A subset of the final 2D movies produced through this framework are available to view in supplementary movie M3. We have also provided a comparison of these final ESPER outputs to NLSA for one degree of freedom in the section SM-XIX. Through analysis shown in several movies, we found that ESPER generates 2D movies of noticeably higher quality than NLSA, and with significantly more accuracy in occupancy assignments for corresponding states. We further highlight problems that can emerge in the alternative NLSA pipeline, such as the inability to capture certain conformational motions captured properly by ESPER, and the possibility of delivering nonsensical (i.e., physically impossible) results. In addition, we found that the overall computation time of NLSA far exceeds the one needed by ESPER.

After generating all 2D movies (one per CM for each PD), both the type of conformational motion present in the 2D movie (e.g., CM₁ or CM₂) as well as its *sense* must be determined individually for each PD. As to the definition of the *sense* of each movie, it cannot be said *a priori* in what direction (i.e., the sequential ordering of states) a CM trajectory is following along any path⁵⁶. For example, for CM₁, the parabolic mode could either be charting the trajectory from states 1 to 20 or from 20 to 1. This uncertainty is due to arbitrary eigenfunction polarity which naturally arises via eigendecomposition⁵⁷. Although a comprehensive method has been developed to solve this problem for datasets with large numbers of sufficiently occupied PDs using optical-flow and belief-propagation algorithms⁵⁶, considering we only had 126 PDs to decipher, we opted to instead determine the type of CM and its sense with perfect accuracy by visual inspection of the 2D movies. With CM types and senses assigned, the 2D movie of a given CM—housing indices of all images within its frames—can next be compiled together with all other 2D movies (and corresponding 1D occupancies) of that same CM across all PDs. If we desired only one degree of freedom as output, our task would next be complete after reconstructing 3D density maps from the images accumulated in each frame of the given 2D movie (and similarly, as it applies, for the 1D NLSA approach).

For presentation of the intermediate 1D occupancy results, we performed this compilation on both CM₁ and CM₂ independently, with corresponding occupancy statistics accumulated for each state therein, and compared with ground-truth knowledge (as shown on the left in Fig. 11). For our comprehensive dataset, we found that

the 1D occupancy map distributions were in strong agreement with ground-truth knowledge, with states on average monotonically captured along each subspace trajectory and having a relatively small spread of uncertainty for each bin. Still, noticeable disagreement emerged for both CMs near the boundaries of these distributions, where inward curling of the parabolic point cloud due to CTF is most prominent. As a result, 1D occupancy assignments are slightly skewed (overestimated) near the boundaries in comparison to ground-truth expectations (Fig. S13).

As an aside, in order to further investigate these trends, we repeated this analysis for data-type II independently along three orthogonal great circle 126 PD-trajectories. The results of this analysis are provided in Fig. S11, where we plot 1D occupancy maps for CM₁ and CM₂ along with corresponding histograms detailing the R^2 values of all 126 respective CM subspaces. First we note that occupancy statistics become aberrant near the boundaries to different extents in all occupancy maps observed, a problem seemingly unavoidable based on our observations of corresponding embedded geometries. As this problem exists, to some lesser extent, even without CTF modification of the data, it can likely be attributed to Neumann boundary conditions^{24,57} (i.e., vanishing normal derivatives). It is also clear that the overall geometric quality of the collection of CM subspaces present (described here via R^2) is a defining factor affecting the fidelity of the corresponding occupancy assignments to ground truth. Artifacts from PD disparity aside, all occupancy maps remained in excellent sequential agreement, with the significance of occupancy surface fluctuations highly dependent on the quality of CM subspaces available. With this understanding, we now return to our final data-type analysis to conclude with the remaining steps of the ESPER framework.

As previously pointed out, these two CM coordinates are intrinsically linked by the independent occurrence of image indices from the same PD image stack. This fact is used by ESPER to generate a 2D occupancy map, and likewise for any number of degrees of freedom present in a dataset. Specifically, this operation was performed by taking the intersection of image indices (overlap) corresponding to each pairwise combination of bins in the CM₁ and CM₂ trajectories: effectively reconstructing the hypersurface on which they jointly reside. For more information on this procedure, see section SM-XVII. The simplicity and efficiency of this operation is based on our method's continued use of the raw cryo-EM images and initial manifold embedding. In contrast, the original ManifoldEM^{1,10} workflow requires a radially dense set of 1D profiles derived via NLSA (180 in total for $n = 2$, each obtained from an independent NLSA analysis) from the n -dimensional subspace formed by the previously-selected set of n eigenvectors. Naturally, this operation rapidly increases in computational time as n is increased. An inverse Radon transform is next applied to these NLSA profiles to reconstruct an n -dimensional

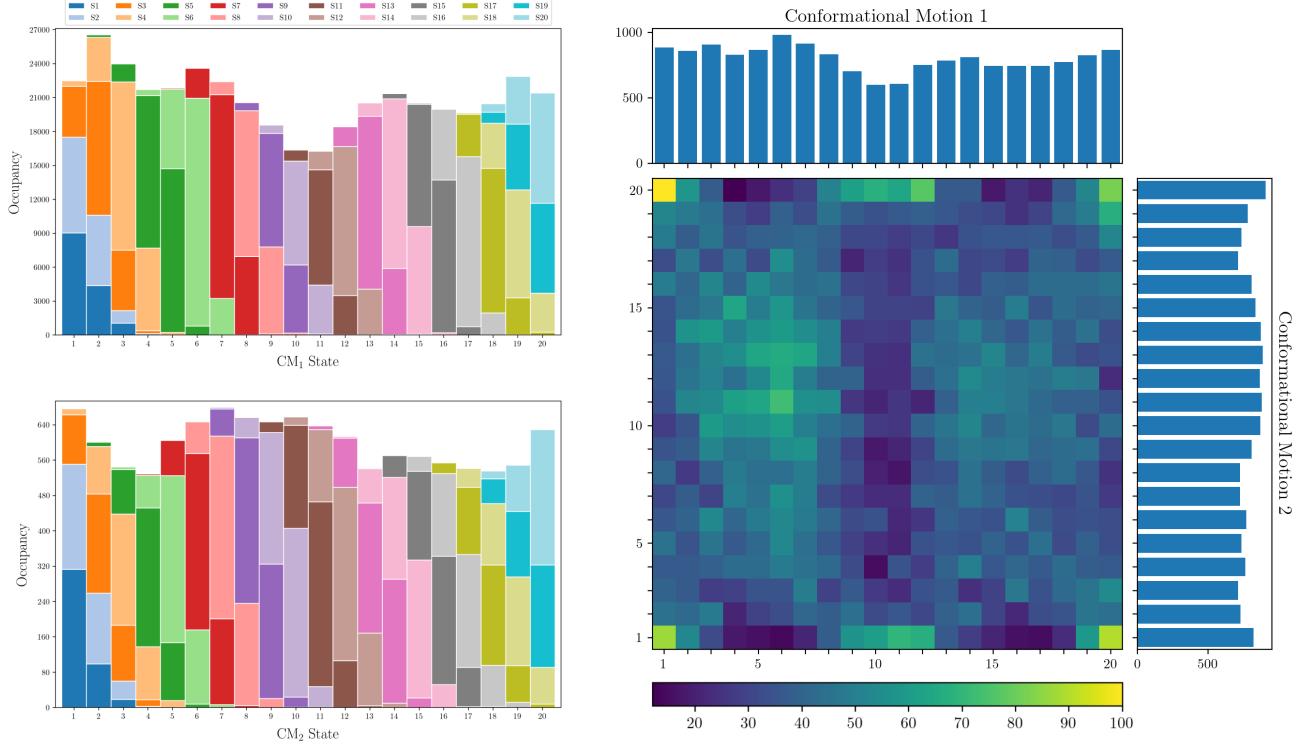


FIG. 11: On the left, the final occupancy maps for the 20 states in CM₁ (top) are shown alongside an equivalent representation for the 20 states in CM₂ (bottom). Each plot was obtained by integration of the corresponding 20 bins (corrected for sense) in each of the 126 PDs. The total number of images as assigned to each state via our subspace fitting procedure is shown by the height of the 20 bars. Within each bar, the different colors represent how many of the assignments therein belonged to which ground-truth states (as seen in the legend), allowing an assessment of the True Positive rate. On the right, the final 2D occupancy map for the 400 states formed by CM₁ and CM₂ is shown; obtained via the intersection of image indices in all pairwise combinations of CM₁ and CM₂ bins (corrected for sense) in each of the 126 PDs. Refer to Fig. S13 for a direct comparison with ground truth. Finally, to circumvent issues stemming from inclusion of CM subspaces with poor geometric structure, we note that while all images are used for subsequent 3D reconstructions, only those occupancy assignments for CM subspaces above an R^2 -threshold value (0.7) were integrated during this analysis. We additionally note that all results shown are the product of our robust automation strategy involving no case-by-case intervention; assuredly, these deviations could be further mitigated by enforcing key parameter choices with supervision.

occupancy map. The result of our ESPER operation is shown on the right in Fig. 11, which encompasses occupancies for all 400 (20×20) bins in a 2D state space. As each of these 400 bins contains the indices of images sharing a given bin coordinate in the $\{\text{CM}_1, \text{CM}_2\}$ plane, 3D density maps can next be produced for each state in the state space.

Following this assessment, image stacks, one for each state, were generated and paired with an alignment file that carried the input alignment and microscopy information for each image therein, as initially defined for each PD along the 126-PD great circle trajectory. This file was then used as input for the `relion_reconstruct` module⁴⁶ to create a 3D reconstruction for each of the 400 states. These 3D density maps were loaded in sequence to create 3D movies seen from different views using Chimera⁵⁸. As shown in Fig. 11 and qualitatively expressed in movie M4, these 3D density maps uphold the spatial relationships in the ground-truth CMs with

striking accuracy. Of most importance, and presenting a key distinction from NLSA results, is the fact that the mobile domains in all states are as well resolved as the immobile domains.

As a quantitative validation, we calculated the Fourier shell correlations⁵ (FSC) curves between several ground-truth 3D density maps and their corresponding ESPER recovered 3D density maps. FSC curves are a routinely used tool for map validation in cryo-EM and suited to provide a global indicator of agreement between two 3D density maps. The FSC curves confirmed a good recovery of the different states up to a resolution near 3 Å, the ground-truth value. Additionally, Q-scores⁵⁹ were used, not to estimate the resolution, but as a local quantitative validation of the structural fidelity of these outputs. Using this approach on the ground-truth atomic-coordinate structures and their corresponding ESPER recovered 3D density maps, we found highly favorable agreement across all residues for each state. Results for

an example state are available in section SM-XX.

DISCUSSION

Through our analysis, we have identified the way sets of images originating from a varying atomic structure are represented in the manifold embeddings obtained by DM or PCA dimensionality-reduction techniques, and how this information can be used to retrieve the original, ground-truth conformational motions. Our findings on synthetic noisy datasets provide a number of new insights, and emphasize the need for a refined workflow when analyzing the eigenvectors from embeddings of single-particle cryo-EM datasets of molecules exhibiting conformational motions. Several of the operations introduced in this study offer straightforward improvements on the founding PD-manifold approach ManifoldEM^{1,10,11}, such as our informed subspace fitting procedure using specific combinations of eigenfunctions, exclusion of parabolic harmonics, and a novel direct retrieval of each CM using the raw cryo-EM snapshots as arranged within the initial embedding. In the last case, the use of the raw images improves both the accuracy of occupancies and final resolution of 3D structures, while providing a vastly simplified workflow for generation of multidimensional free-energy landscapes. Further, we found the corrections for previously unaccounted d -dimensional rotations to be essential; the absence of awareness of which can lead to serious systematic errors downstream (see section SM-XIX).

All of the findings within this study are based on heuristic information obtained from ideal, controlled datasets, analyzed so as to maximize the fidelity of our final outputs with ground truth, while uncovering key limitations and uncertainties that could potentially emerge within this unsupervised machine-learning framework. It is important to be aware that results from synthetic data will always be superior to experimentally-obtained data, since even the most sophisticated simulations will be unable to capture all complexities existent in an experimental system. These complexities can be considered as introducing higher-order terms in our parameter space, which has been designed to emulate all lower-order terms up to a threshold deemed satisfactory. Any limitations or uncertainties that do emerge using synthetic data should be anticipated to arise in real-world data, and potentially in exacerbated form.

Contextually, this heuristic analysis is focused on data models originating from molecules undergoing collective rigid-body motions, which we believe are sufficient for most molecular machines, but may fall short of addressing instances involving more complex motions. This is especially the case for those machines entailing the concerted binding and release of ligands, which naturally require a separate state space for each possible combination of the machine with each of its binding partners. For such a situation, a study of ligands has been per-

formed with the founding ManifoldEM approach¹⁰ using two state spaces, which could similarly be explored in an extension of the ESPER framework.

It is also worth noting that while the results of our heuristic analysis are most relevant to machine-learning methods dealing with projection data (i.e., requiring the PD-manifold approach), several portions of our analysis can be extended to other experimental techniques dealing with alternative manifold inputs, such as the use of atomic models in molecular dynamics and 3D density maps in cryo-electron tomography. Specifically, in section SM-XIV, we detail how the structure of manifolds obtained from a conformational state space transforms as the data type is translated stepwise from atomic models to 3D density maps to 2D projections. Broadly, we believe that there is a potential for the application of our methodology to a wide range of experimental datasets, and particularly those obtained from systems exercising multiple, continuous degrees of freedom.

With that said, we return to the flowchart presented in Fig. 9, to discuss each step of the analysis in detail, focusing on both implementation and existing limitations.

1. Ensure that orientational and conformational sampling is adequate

Before initiating this workflow, it is first necessary to ensure that adequate coverage has been obtained via cryo-EM imaging of the heterogeneous ensemble of molecules. We have identified two main categories of coverage: (i) *coverage of S^2* : overcoming the effects of missing orientations with sufficient sampling of viewing directions; and (ii) *coverage of states*: imaging all available conformational states present and in sufficient abundance for the given SNR to obtain robust statistical coverage in the corresponding manifold. For datasets with low image counts in many PDs, the corresponding final 3D reconstructed volumes will ultimately suffer a loss in resolution. This loss is manifested when the CMs in each Ω_{PD} are stitched together, as not enough 2D CM information is present to properly depict its corresponding aspect of the 3D structure. For assessing coverage of S^2 , we have confirmed that within our framework the availability of data over a great circle is sufficient for recapitulating ground-truth conformational information (movie M4).

As for state space coverage, we found that the combination of two parameters—image SNR and state-space occupancy (e.g., $\tau M = N$ for uniform distributions)—together served as a strong indicator for the ability of each Ω_{PD} embedding to render coherent conformational-variation signals. Specifically, for sets of images in SNR regimes of approximately 0.05 to 0.1, as encountered in low-exposure cryo-EM^{60,61}, multiple noisy duplicates (as defined by τ) of the fundamental state space (having M states) were required to recapitulate ground-truth structure within the spectral geometry (Fig. S9 and Fig. S10). To note, for experimental conditions where each state

occurs with a given frequency as dictated by its underlying free energy, this condition must be met for the least abundant states in the dataset. This extended coverage effectively serves to drown out the experimental noise up to the point where only the conformational-variation signal remains. Without this additional coverage, in the worst-case scenarios, the ordering of the ground-truth points within the embedding will be jumbled in an uninterpretable form, with the distribution of these points closely resembling a Gaussian distribution (e.g., as seen in the last column of Fig. 6).

However, even embeddings of manifolds with adequate state space coverage can still appear globular in structure while retaining proper arrangement of states, with this trend decreasing as coverage is increased (as shown in the second and third columns of Fig. 6). We found that as the τ -value is increased, there exists a lower τ -threshold (τ_c) such that the arrangement of points in the embedded manifold is in highest achievable consistency with its ground-truth state space. In other words, there is a fixed amount of coverage that is sufficient. Unfortunately in an experiment, since the number of ground-truth states M for a given molecular machine is unknown, so too is this threshold particle count (τ_c), which can be affected by the sample characteristics as well as the quality of the collected data. This threshold was additionally found to vary depending on both the intrinsic dimensionality and energetics of the ground-truth data – with knowledge of either of these parameters not immediately clear for a given dataset. In practice, then, the insight gained about threshold particle count is of no help in guiding the experimental design, and hence the decision on adequacy of state space coverage must be made by trial and error.

One would first aim to collect as many particles as possible during the experiment, and, after generation and embedding of manifolds, make a decision on the adequacy of that collection based on the presence of robust parabola-housing subspaces. Such subspaces were directly observed and showcased, for example, in the conformational embedding of the ribosome¹. As one potential scheme, for each Ω_{PD} individually, after 2D subspaces have been fit and R^2 -values computed (Fig. S14), an R^2 -threshold can be used across all subspaces to assess geometric conditions and demarcate use of either NLSA or ESPER on that manifold. As seen in movie M8, these two frameworks share considerable overlap in their 2D movie outputs, and may need only their final PD-outputs subsequently combined for production of sensible 3D movies. Additionally to consider, while it has been shown that ESPER is by far the better choice for manifolds with geometrically-structured subspaces, only NLSA can be applied in the regime of manifold embeddings completely lacking discernible form, while still potentially incurring its known limitations and uncertainties¹¹.

As a final note for this section, experimental data suffer from a wider range of nuisances than we have accounted for in our simulation, including the occurrence of aberrant particles (such as ice shards or foreign bod-

ies); uncertainty in CTF estimation; and uncertainty in angular alignments, which are more pronounced in heterogeneous data. While numerous preprocessing algorithms exist to handle each of these instances^{27,46,47}, it is obvious that such artifacts and errors, if left unaccounted for, can have detrimental consequences for the fidelity of the corresponding manifolds to the rules we uncovered using the simulation.

2. Construct embedded manifold for PD

The choice of method for obtaining eigenvectors—through either linear or nonlinear dimensionality reduction frameworks—proved to have relatively minor consequences compared with many other choices in our workflow. As previously noted, both PCA and DM approaches aim to achieve a description of the dataset’s most fundamental form, defined by the multidimensional relationship among all images in a PD. Surprisingly, PCA and DM produced almost identical eigenvectors for all data examined in this work, particularly in the presence of noise (as seen in Fig. S17 and Fig. S18). The high similarity in the performance of these methods did not conform to previous expectations, as the superiority of nonlinear dimensionality reduction frameworks to linear ones is a belief often cited in the field^{1,4,24,38,62}. In contrast, however, a comparative review of twelve prominent dimensionality frameworks¹³ has also shown that, despite their ability to learn the structure of complex manifolds, most nonlinear techniques are unable to outperform PCA on experimentally-obtained datasets. In the presence of noise, our discoveries indicate that biological objects (such as macromolecules) undergoing combinations of rigid body conformational motions fall in the latter camp and can be effectively studied using either linear or nonlinear techniques.

Given such a choice in embedding strategies, it would still make sense to give preference to DM, as it leaves room for the possibility of cases yet to be encountered, such as those where PCA may have difficulty in untangling certain types of complex conformational relationships. The DM framework additionally offers a reduction in computational load of the eigendecomposition due to its sparse matrices, which becomes increasingly relevant as the number of snapshots increases. However, certain caveats must still be considered. For one, discretion is introduced in all nonlinear frameworks through the fitting of additional parameters required for their optimal performance, failure of which can lead to systematic errors¹³. Using the right strategies, these uncertainties can be minimized. In the case of DM, while incorrect choice of the Gaussian bandwidth had drastic consequences, this parameter was consistently put into the correct ballpark using bandwidth estimation plots. It is also necessary to point out that while DM and PCA examined in detail here are two of the most prominent dimensionality reduction techniques, it might be well worth investigating

the performance of other approaches¹³ using simulated datasets such as ours.

Finally for this section, we discuss the use of the defocus-tolerant kernel¹ for calculating Euclidean distances between images on which a CTF was imposed. We found that its success dwarfed that of two other techniques explored. For comparison, we first observed that manifold embeddings obtained using the standard kernel without any CTF correction were completely incoherent in structure, as expected. Better behaved but still flawed were embeddings obtained from sets of CTF-corrected images (Fig. S12-G) using the standard kernel, which ultimately provided structure suboptimal to those obtained using the double-filtering kernel. We consider this study the first demonstration that the previous double-filter kernel introduced¹ is more effective than CTF-correction of individual images. However, as previously noted, this kernel is not free of artifacts.

Most significantly, CM subspaces typically had a strong proclivity for curling inwards near the outer edge, with states clumped more densely in these regions (Fig. 8). This trend was apparent enough to require the introduction of general conic fitting strategies for elucidating CM subspaces (Fig. S14). Future studies could further examine the effects of this defocus-tolerant kernel on a wide range of defocus-value intervals and relative magnitudes. As a note, when altering the defocus range, one must carefully choose the particle box size. If the box size is too small to fit the broadening due to the point-spread function (i.e., the Fourier transform of the CTF), the presence of inward curling on each CM subspace is greatly exacerbated. There are additional higher-order aberrations of the CTF relevant for high-resolution cryo-EM⁶³ (such as astigmatism and beam tilt) which can be easily incorporated in the CTF of the double filter. While not explored here, these may further affect the spectral geometry if not accounted for properly.

3. Determine pairs of eigenvectors for CM parabolas

Specific techniques must next be applied to discover the set of conformational modes corresponding to each CM within each Ω_{PD} embedding, with the number of these sets defined by the dataset's intrinsic dimensionality. However, as described in section SM-XI, there exists an initial uncertainty about how to best determine the intrinsic dimensionality of any given dataset^{64,65} (i.e., the number of CMs present to search for), for which both an evaluation of the eigenvalue spectrum (for both PCA and DM) and bandwidth estimation strategy (for DM) have proven unsatisfactory for our purposes. To circumvent this uncertainty, we have introduced an elimination procedure to locate subspaces where conformational information most likely resides, and based on information gleaned from those findings, eliminate unsuitable subspaces from further study.

Our analysis demonstrated that the minimum informa-

tion required during this discovery was the attainment of the lowest-order Chebyshev polynomial (T_2) for each degree of freedom, with all other features in the embedding irrelevant to our needs. For any number of CMs defined by the dataset's intrinsic dimensionality, the point-cloud resemblance to these Chebyshev polynomials can be found spanning specific 2D subspaces. However, care must still be taken to prevent overfitting, since blindly performing the subsequent steps in this framework on subspaces that do not unequivocally represent one of the dataset's CMs (note that there are principally many such subspaces, such as those housing harmonics) can produce nonsensical conformational movies. For instance, we have provided several 2D NLSA movie outputs corresponding to independent degrees of freedom in movie M8 showcasing physically impossible motions (further descriptions are provided in section SM-XIX). Likewise, we have demonstrated how sets of parabolic harmonics for each CM naturally exist in all Ω_{PD} embeddings and, due to the overlapping nature of their point clouds, are unavoidable for mapping conformational information.

In contrast, the ESPER algorithm automatically fits each 2D subspace with parabolas, and signals the best-fit subspace for each eigenvector row for use in subsequent steps. Further, the eigenvector indices of these best-fit subspaces can be used to procedurally eliminate the use of CM harmonics, as is described in section SM-XVI. The availability and use of this information acts to circumvent the potential for contextual confusion in outputs, and is a direct remedy to one of the most prominent uncertainties¹¹ in the founding ManifoldEM framework (see section SM-XIX). However, in practice, the ability to strategically avoid harmonics is only applicable up to the number of CMs present having pronounced geometric structure viable for reliable parabolic fits. It is likely that more advanced strategies could be applied to optimize our routine to additionally eliminate a larger subset of higher-order harmonics. Finally, it is important to carefully examine the 2D movies produced for each potential CM by eye. The emergence of nonsensical (i.e., physically impossible) patterns should be seen as a strong indicator for overfitting of the Ω_{PD} embedding. In a similar vein, the possibility of underfitting (analyzing too few CM subspaces) must also be taken into account, though it carries significantly less risks.

4. Find optimal rotations for the eigenfunctions

As we have seen to some extent in every embedding explored, depending on the choice of PD, the conformational modes previously discovered can appear misaligned with the plane of their 2D subspace. As shown in section SM-XIV, this effect arises from the introduction of altered, foreshortened distances in each 2D projection (i.e., PD disparity), and could be countered via rotation of specific eigenvectors to ease retrieval of the CM. Given this knowledge, we have presented an automated

procedure for the discovery of these optimal rotation angles for each PD. Despite its remarkable performance on our four 126-PD datasets (Fig. 11 and Fig. S11), we believe there is still room for this technique to be further developed in future works so as to account for rare events such as complex boundary conditions (e.g., due to steric hindrance between domains, as discussed in section SM-XV) as well as datasets with $n > 2$ degrees of freedom (see section SM-XVI). In the prior case, the use of additional rotation operators may be required, creating a more complex collection of decisions, possibly well-suited for a maximum-likelihood approach. Furthermore, for noisier, less-structured embeddings, we have observed that the 2D histogram method may provide close but not perfectly-aligned counter-rotations. It is possible that this method can be further elaborated by using additional eigenfunctions, which would allow the fitting with parabolic sheets instead of—or perhaps alongside with—histogram distributions.

These findings have directly established a new set of requirements in the types of datasets most viable for PD-manifold studies of conformational continuum. If no geometric form can be deciphered in all or even a significant number of Ω_{PD} embeddings, we claim it is effectively impossible to find the proper counter-rotation such that all PD eigenbases are displayed in a common coordinate system defined by CMs. We have further observed that the inability to align eigenfunctions onto this common eigenbasis can to varying extents subvert the quality of 2D movies and occupancy maps (see PD₃₃ in movie M8), and thus ultimately the quality of the final 3D reconstructions. This fact explains several of the limitations¹¹ recently documented in the founding PD-manifold approach^{1,10}, since the previous ManifoldEM approach does not account for the tendency of Ω_{PD} embeddings to be unpredictably unaligned from a common coordinate system (e.g., Fig. S28). For a more detailed analysis of limitations and uncertainties observed, see section SM-XIX. Thus, in order to maximize fidelity of final outputs with ground truth, a dataset must first fulfill a minimum set of requirements that ultimately provide a well-structured spectral geometry when embedded. The performance of ESPER hinges on the presence of this geometric information, and as we have shown and will continue to discuss in the sections that follow, a great number of benefits emerge when it is available.

5. Fit and partition each CM point cloud

Once the set of conformational modes are identified and counter-rotated for each Ω_{PD} embedding, a further obstacle is encountered in the heightened variability of these point clouds (Fig. 7) – which vary in average thickness, density, length and type of trajectory, as well as spread of data points – depending on CM and PD. These complications were addressed through use of a robust automation strategy, including least-squares fitting and area-based

partitioning of the essential subspaces. Overall, the area-based procedures employed—including the ball tree and alpha shape frameworks—are strongly affected by large changes in parameters and thus may require initial supervision on externally-obtained datasets. For the purpose of this study, we investigated over 500 subspaces, with each of our parameters broadly tuned for robust, high-quality performance on the SS₂ CM₁ and CM₂ subspaces from data-type II and III. Detailed notes on these procedures have been provided in our published repository³⁷, which includes comments describing less-significant decisions not explicitly noted in our main text. As a final note, we analyzed the outputs of these procedures on both the parabolic $\{\Psi_i, \Psi_j\}$ and transformed $\{\Phi_i, \Phi_j\}$ CM subspaces, and found consistently better agreement with ground-truth occupancies in the latter kind.

6. Bin points along straightened trajectory and integrate images

After achieving these prior conditions, points (one corresponding to each image) projected onto their respective least-squares fit must be binned, with bins created via the aforementioned area-based approach. Each of these bins represents one of the system’s unique states along the CM corresponding to the current 2D subspace. While we used a bin size as informed by our ground-truth knowledge to enable a direct comparison between inputs and outputs, it is another issue entirely to decide on the proper bin size for real-world data. As detailed in section SM-XVII, note that the bin size effectively controls the precision to which we can locate each point in the higher-dimensional surface, and it influences the range of images falling within each state that we group together as virtually identical for means of our final outputs. Naturally, the use of this optimal value should maximize the amount of information ascertainable in our system. In theory, we desire a minimum number of snapshots in the lowest-occupancy bin, such that every frame of the subsequent movie has significant content; e.g., as possibly defined via both the average SNR of each image and the number of images in the lowest-occupancy bin.

For each CM in a given dataset (regardless of intrinsic dimensionality: 1, 2, and 3 were tried), we were able to project the set of our images onto curves determined by least-squares fits, organize them into bins, and produce near-perfect 2D movies displaying each conformational motion independently of all others (movies M2 and M3). The corresponding occupancy maps showed a favorable agreement with our ground-truth knowledge (Fig. 10-G, Fig. 11, Fig. S11), and were significantly more accurate than those produced by NLSA for all datasets explored (movie M8). Overall, the procurement of accurate occupancy maps for each PD was by far the most trying endeavor, requiring a robust workflow able to account for large variations in a wide range of manifold characteristics. While it is very easy to split the CM subspace

crudely into only two or a handful of states, as the number of states requested is increased, the degree of sophistication in mapping and segmenting each point cloud must also increase in turn. As previously noted, when dealing with experimental data, these occupancies contain vital information for the energetics of the ensemble, and are directly linked to the free-energy landscape spanned by the conformational degrees of freedom. We will return to this topic in the next section when discussing the final occupancy maps generated via integration of CM content across all PDs on S^2 .

7. Identify motion type and sense for each CM 2D movie

With the acquisition of each isolated conformational motion for each PD, lastly in such a framework, these CMs must be stitched together^{1,56} using equivalent CM information from each PD across S^2 . For our workflow, this procedure required manually solving two subproblems: (i) identifying each set of CMs across S^2 (e.g., such that CM₁ in a given PD is matched with CM₁ in all other PDs, etc.) and (ii) identifying the sense of each CM in each PD. We draw attention again to the external, comprehensive method developed to solve this problem with heightened accuracy, which uses optical flow and belief propagation algorithms⁵⁶, and strongly advise against making such assignments arduously by hand.

8. Compile CMs on S^2 for free-energy landscape and 3D movies

After CMs have been properly identified and matched among all PDs, the indices of points (designating images) and related statistics can be organized to produce an n -dimensional occupancy map, with the images assigned to each n -dimensional bin therein accumulated to form a corresponding 3D density map. Once these previous steps were performed for our dataset, we compared the state assignments of all images from S^2 to their ground-truth indices. We found that the majority of images in each bin (state) were correctly assigned, with each bin also encompassing a small set of images that were actually ground-truth members of neighboring bins (Fig. 11 and Fig. S11).

Our final 2D occupancy map (also seen in Fig. 11), formed by the intersection of snapshots in both 1D occupancy distributions, showed similar consonance with ground-truth expectations (Fig. S13). As a result of these accurate occupancy assignments, we ultimately observed a remarkable fidelity between each of the 400 reconstructed 3D density maps obtained by ESPER (movie M4) and their respective ground-truth 3D density maps. In addition, the resolution of all volumes were close to 3 Å, matching the resolution of the ground-truth structures. Overall, the small spread of states into neighboring bins appeared to have only a marginal effect on the qual-

ity of final 3D density maps. Through our analysis, it is apparent that the majority of aberrant information in each bin (corresponding to neighboring states) was effectively averaged out, such that only the conformational-variation signals corresponding to the current state dominated in each subsequently-generated 3D reconstruction. These findings were further supported quantitatively by an assessment of the structural fidelity of ESPER outputs with ground-truth atomic-coordinate structures and 3D electron density maps using Q-score calculations on residues and FSC curves, respectively (see section SM-XX).

As a final note, since our method retains the original image content for each image index assigned to a given bin, it is possible to further improve these image assignments after they have been generated by our subspace fitting routines. Our ability to leverage the final image content to further improve 3D density maps and corresponding occupancy distributions stands in contrast to the founding PD-manifold approach¹, which relies on histogram equalization (technically *histogram matching*) to match the occupancy distributions across PDs. Although not pursued here, one possible way is a maximum-likelihood approach aimed at comparing images within each bin and reassigning erroneously-assigned subpopulations into the neighboring bin in which they most likely belong. To note, a maximum-likelihood approach does already exist that aims to extract such granular conformational heterogeneity⁶⁶, as does a method based on neural networks⁶⁷.

CONCLUSIONS

An ensemble of synthetic cryo-EM projections of the Hsp90 protein undergoing quasi-continuous conformational changes has been generated as an exemplary ground-truth model to determine how macromolecular motions appear in low-dimensional representations of their respective dataset using the PD-manifold approach. Based on knowledge obtained by this analysis, we have introduced a novel, unsupervised workflow with several enhancements that substantially improve the ManifoldDEM framework and its ability to recover continuous conformations from single-particle cryo-EM data. These include essential eigenvector rotations to consistently align the spectral geometry of all Ω_{PD} embeddings, use of the original cryo-EM images to form high-quality 2D and 3D movies, and efficient strategies for building multidimensional free-energy landscapes. Along with the introduction of this workflow, we have also pointed out challenges, fundamental limitations and uncertainties that emerge in geometric machine learning of heterogeneous cryo-EM data. Finally, we hope that the insights gained from these machine-learning heuristics will be useful not only in cryo-EM, but also in the development of other techniques aimed at untangling complex systems exercising multiple, continuous degrees of freedom.

ACKNOWLEDGMENTS

We would like to express our gratitude to Abbas Ourmazd, Ali Dashti and Ghoncheh Mashayekhi for their insights and expertise, providing a number of stimulating conversations throughout this development. This research was supported by the National Institutes of Health Grants GM29169 and GM55440 (to J.F.) and by the U.S. National Science Foundation Award No. STC1231306 (to P.S.).

CONTRIBUTIONS

Author contributions listed alphabetically below. All authors reviewed the final manuscript.

- Conceptualization: ES, JF
- Formal analysis: ES
- Methodology and software: ES, FAR, PS, SM
- Validation: ES, FAR, PS
- Direction and project administration: JF
- Manuscript draft: ES
- Manuscript review and editing: ES, FAR, JF, PS, SM

Detailed contribution notes for the ESPER algorithm suite have been supplied in the header of all scripts in our online repository³⁷.

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The Python software repository³⁷ is available at the following link:

https://github.com/evanseitz/cryoEM_ESPER

This repository includes several pristine image stacks and algorithms to modify them with different occupancy assignments, CTF and SNR. In addition, a collection of pre-generated manifolds are supplied for immediate use in computations employing the eigenfunction realignment and subspace partitioning algorithms presented here.

Appendix: Description of symbols and abbreviations

Name	Description
CM	Conformational motion
n	Intrinsic dimensionality of the state space (SS_n): one degree of freedom per CM
Ω	Compact Riemannian n -dimensional manifold
N	Number of images in each PD manifold (Ω_{PD})
P	Number of pixels per image
S^2	2-sphere: set of angles for 3D viewing orientations
PD	Projection direction $\subset S^2$; for any PD, typically $N \ll P$
ψ_k	Eigenfunction of Laplace-Beltrami operator
Ψ_i	Eigenvector of N -dimensional manifold; $i \in \{1, 2, \dots, N\}$
λ_i	Eigenvalue of N -dimensional manifold; $i \in \{1, 2, \dots, N\}$
Φ_i	Inverse cosine of Ψ_i
L	Lissajous curves; $L_{p,q} \subset L$
T_k	Chebyshev polynomial of the first kind; e.g., $T_2(x) = 2x^2 - 1$ for the parabola
d	Dimension of orthogonal matrix O (and R_{ij}) applied on embedded space in \mathbb{R}^d
R_{ij}	Rotation sub-matrix operating on $\{\Psi_i, \Psi_j\}$, of which there are $d(d-1)/2 \in O$
ε	Gaussian bandwidth used in DM Gaussian kernel
M	Number of unique ground-truth states
τ	Number of times a given state space is uniformly duplicated, such that $\tau M = N$
R^2	Coefficient of determination
m	Number of atoms in an atomic-coordinate structure

REFERENCES

- ¹A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseiniadeh, H. Y. Liao, J. Pallesen, G. Sharma, V. A. Stupina, A. E. Simon, J. D. Dimman, J. Frank, and A. Ourmazd, "Trajectories of the ribosome as a brownian nanomachine," *Proceedings of the National Academy of Sciences* **111**, 17492–17497 (2014), <https://www.pnas.org/content/111/49/17492.full.pdf>.
- ²E. Seitz and J. Frank, "Polaris: Path of least action analysis on energy landscapes," *Journal of Chemical Information and Modeling* **60**, 2581–2590 (2020), pMID: 31999117, <https://doi.org/10.1021/acs.jcim.9b01108>.
- ³P. A. Penczek, M. Kimmel, and C. M. Spahn, "Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images," *Structure* **19**, 1582–1590 (2011).
- ⁴A. Moscovitch, A. Halevi, J. Andén, and A. Singer, "Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes," *Inverse Probl* **36** (2020).
- ⁵J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State* (Oxford University Press, Oxford, New York, 2006).
- ⁶J. Frank, "Generalized single-particle cryo-EM—a historical perspective," *Microscopy (Oxf)* **65**, 3–8 (2016).
- ⁷J. Frank, "Single-particle reconstruction - story in a sample." Nobel Lecture (2017).
- ⁸N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, and H. Stark, "Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy," *Nature* **466**, 329–333 (2010).
- ⁹P. 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, *Dynamic views of ribosome function: Energy landscapes and ensembles* (Springer, Vienna, 2011).
- ¹⁰A. Dashti, G. Mashayekhi, M. Shekhar, D. Ben Hail, S. Salah, P. Schwander, A. des Georges, A. Singhary, J. Frank, and A. Ourmazd, "Retrieving functional pathways of biomolecules from single-particle snapshots," *Nat Commun* **11**, 4734 (2020).
- ¹¹G. Mashayekhi, "Manifoldem matlab repository," https://github.com/GMashayekhi/ManifoldEM_Matlab (2020).
- ¹²T. 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, "A glycan gate controls opening of the sars-cov-2 spike protein," *Nature Chemistry* (2021), 10.1038/s41557-021-00758-3.
- ¹³L. V. D. Maaten, E. Postma, and J. Herik, "Dimensionality reduction: A comparative review," (2009).
- ¹⁴T. R. M. Craioveanu, M. Puta, *Old and New Aspects in Spectral Geometry* (Springer Science, 2001).
- ¹⁵W. Liu, N. Boisset, and J. Frank, "Estimation of variance distribution in three-dimensional reconstruction. II. Applications," *J Opt Soc Am A Opt Image Sci Vis* **12**, 2628–2635 (1995).
- ¹⁶P. Penczek, "Variance in three-dimensional reconstructions from projections," in *Proceedings IEEE International Symposium on Biomedical Imaging* (2002) pp. 749–752.
- ¹⁷P. A. Penczek, C. Yang, J. Frank, and C. M. Spahn, "Estimation of variance in single-particle reconstruction using the bootstrap technique," *J Struct Biol* **154**, 168–183 (2006).
- ¹⁸P. A. Penczek, J. Frank, and C. M. Spahn, "A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation," *J Struct Biol* **154**, 184–194 (2006).
- ¹⁹H. Y. Liao and J. Frank, "Classification by bootstrapping in single particle methods," *Proc IEEE Int Symp Biomed Imaging* **2010**, 169–172 (2010).
- ²⁰P. Schwander, R. Fung, and A. Ourmazd, "Conformations of macromolecules and their complexes from heterogeneous datasets," *Philos Trans R Soc Lond B Biol Sci* **369**, 20130567 (2014).
- ²¹A. Punjani and D. J. Fleet, "3d variability analysis: Directly resolving continuous flexibility and discrete heterogeneity from single particle cryo-em images," *bioRxiv* (2020), 10.1101/2020.04.08.032466.
- ²²J. Frank and A. Ourmazd, "Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM," *Methods* **100**, 61–67 (2016).
- ²³K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine* **2**, 559–572 (1901).
- ²⁴R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps," *Proc Natl Acad Sci U S A* **102**, 7426–7431 (2005).
- ²⁵R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis* **21**, 5–30 (2006), special Issue: Diffusion Maps and Wavelets.
- ²⁶S. H. Scheres, "Processing of Structurally Heterogeneous Cryo-EM Data in RELION," *Methods Enzymol* **579**, 125–157 (2016).
- ²⁷A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination," *Nat Methods* **14**, 290–296 (2017).
- ²⁸T. Nakane, D. Kimanius, E. Lindahl, and S. H. Scheres, "Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION," *Elife* **7** (2018).
- ²⁹J. Zivanov, T. Nakane, B. O. Forsberg, D. Kimanius, W. J. Hagen, E. Lindahl, and S. H. Scheres, "New tools for automated high-resolution cryo-EM structure determination in RELION-3," *Elife* **7** (2018).
- ³⁰D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability," *Proc Natl Acad Sci U S A* **109**, 2222–2227 (2012).
- ³¹E. Seitz, F. Acosta-Reyes, P. Schwander, and J. Frank, "Simulation of cryo-em ensembles from atomic models of molecules exhibiting continuous conformations," *bioRxiv* (2019), 10.1101/864116.
- ³²E. Seitz, "Repository: Cryo-em synthetic continua (initial release)," Zenodo , 10.5281/zenodo.3561105 (2019).
- ³³H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide protein data bank," *Nature Structural Biology* (2003).
- ³⁴D. Goodsell, "Pdb-101 molecule of the month: Hsp90," (2008).
- ³⁵M. M. Ali, S. M. Roe, C. K. Vaughan, P. Meyer, B. Panaretou, P. W. Piper, C. Prodromou, and L. H. Pearl, "Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex," *Nature* **440**, 1013–1017 (2006).
- ³⁶L. Schrödinger, "The pymol molecular graphics system," (2015).
- ³⁷E. Seitz, "ManifoldEM: ESPER," Zenodo , 10.5281/zenodo.3884160 (2021).
- ³⁸A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, "Systematic determination of order parameters for chain dynamics using diffusion maps," *Proc Natl Acad Sci U S A* **107**, 13597–13602 (2010).
- ³⁹H. Cundy and A. Rollet, *Mathematical Models*, 3rd ed. (Tarquin Pub., 1989) pp. 242–244.
- ⁴⁰J. R. Munkres, *Topology*, 2nd ed. (Prentice Hall, Incorporated, 2000).
- ⁴¹D. S. Grebenkov and B.-T. Nguyen, "Geometrical structure of laplacian eigenfunctions," *SIAM Review* **55**, 601–667 (2013), <https://doi.org/10.1137/120880173>.
- ⁴²M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. (Dover, 1972) pp. 771–802.
- ⁴³D. Lay, S. Lay, and J. McDonald, *Linear algebra and its applications*, 5th ed. (Pearson, 2016).
- ⁴⁴P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *Anesth Analg* **126**, 1763–1768 (2018).
- ⁴⁵G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "EMAN2: an extensible image processing suite

- for electron microscopy,” *J Struct Biol* **157**, 38–46 (2007).
- ⁴⁶S. H. Scheres, “RELION: implementation of a Bayesian approach to cryo-EM structure determination,” *J Struct Biol* **180**, 519–530 (2012).
- ⁴⁷A. Rohou and N. Grigorieff, “CTFFIND4: Fast and accurate defocus estimation from electron micrographs,” *J Struct Biol* **192**, 216–221 (2015).
- ⁴⁸J. Stewart, *Calculus*, 6th ed. (Brooks/Cole, 2008).
- ⁴⁹F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- ⁵⁰S. Omohundro, “Five balltree construction algorithms,” (2009).
- ⁵¹H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, “On the shape of a set of points in the plane,” *IEEE Transactions on Information Theory* **29**, 551–559 (1983).
- ⁵²J. D. Gardiner, J. Behnsen, and C. A. Brassey, “Alpha shapes: determining 3D shape complexity across morphologically diverse structures,” *BMC Evol Biol* **18**, 184 (2018).
- ⁵³K. Bellock, N. Godber, and P. Khan, “Python alpha shape repository,” (2021).
- ⁵⁴S. Gillies *et al.*, “Shapely: manipulation and analysis of geometric objects,” (2007–).
- ⁵⁵P. F. Gilbert, “The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. II. Direct methods,” *Proc R Soc Lond B Biol Sci* **182**, 89–102 (1972).
- ⁵⁶S. Maji, H. Liao, A. Dashti, G. Mashayekhi, A. Ourmazd, and J. Frank, “Propagation of Conformational Coordinates Across Angular Space in Mapping the Continuum of States from Cryo-EM Data by Manifold Embedding,” *J Chem Inf Model* **60**, 2484–2491 (2020).
- ⁵⁷R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, “Graph Laplacian tomography from unknown random projections,” *IEEE Trans Image Process* **17**, 1891–1899 (2008).
- ⁵⁸E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, “UCSF Chimera—a visualization system for exploratory research and analysis,” *J Comput Chem* **25**, 1605–1612 (2004).
- ⁵⁹G. Pintilie, K. Zhang, Z. Su, S. Li, M. F. Schmid, and W. Chiu, “Measurement of atom resolvability in cryo-EM maps with Q-scores,” *Nat Methods* **17**, 328–334 (2020).
- ⁶⁰W. T. Baxter, R. A. Grassucci, H. Gao, and J. Frank, “Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules,” *J Struct Biol* **166**, 126–132 (2009).
- ⁶¹S. H. Scheres, R. Núñez-Ramírez, Y. Gómez-Llorente, C. San Martín, P. P. Eggermont, and J. M. Carazo, “Modeling experimental image formation for likelihood-based classification of electron microscopy data,” *Structure* **15**, 1167–1177 (2007).
- ⁶²C. O. S. Sorzano, A. Jiménez, J. Mota, J. L. Vilas, D. Maluenda, M. Martínez, E. Ramírez-Aportela, T. Majtner, J. Segura, R. Sánchez-García, Y. Rancel, L. Del Caño, P. Conesa, R. Melero, S. Jonic, J. Vargas, F. Cazals, Z. Freyberg, J. Krieger, I. Bahar, R. Marabini, and J. M. Carazo, “Survey of the analysis of continuous conformational variability of biological macromolecules by electron microscopy,” *Acta Crystallogr F Struct Biol Commun* **75**, 19–32 (2019).
- ⁶³X. Zhang and Z. H. Zhou, “Limiting factors in atomic resolution cryo electron microscopy: no simple tricks,” *J Struct Biol* **175**, 253–263 (2011).
- ⁶⁴M. Hein and J.-Y. Audibert, “Intrinsic dimensionality estimation of submanifolds in rd,” in *Proceedings of the 22nd International Conference on Machine Learning*, ICML ’05 (Association for Computing Machinery, New York, NY, USA, 2005) p. 289–296.
- ⁶⁵F. Camastrà and A. Staiano, “Intrinsic dimension estimation: Advances and open problems,” *Information Sciences* **328**, 26–41 (2016).
- ⁶⁶J. Giraldo-Barreto, S. Ortiz, E. H. Thiede, K. Palacio-Rodriguez, B. Carpenter, A. H. Barnett, and P. Cossio, “A bayesian approach for extracting free energy profiles from cryo-electron microscopy experiments,” *Scientific Reports* **11** (2021).
- ⁶⁷E. Zhong, T. Bepler, B. Berger, and J. Davis, “Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks,” *Nature Methods* (2021), 10.1038/s41592-020-01049-4.
- ⁶⁸J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith, “SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields,” *J Struct Biol* **116**, 190–199 (1996).
- ⁶⁹E. D. Zhong, T. Bepler, B. Berger, and J. H. Davis, “CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks,” *Nat Methods* **18**, 176–185 (2021).
- ⁷⁰P. Buser, *Geometry and Spectra of Compact Riemann Surfaces* (Springer, 1992).

Geometric machine learning informed by ground truth: Recovery of conformational continuum from single-particle cryo-EM data of biomolecules

Evan Seitz,¹ Francisco Acosta-Reyes,² Suvrajit Maji,² Peter Schwander*,³ and Joachim Frank*^{1,2}

¹⁾ Columbia University, Department of Biological Sciences, New York, NY 10032, USA

²⁾ Columbia University Medical Center, Department of Biochemistry and Molecular Biophysics, New York, NY 10032

³⁾ University of Wisconsin-Milwaukee, Department of Physics, 3135 N. Maryland Ave, Milwaukee, WI 53211, USA

SM-I. SIMULATION OF CRYO-EM ENSEMBLES FROM ATOMIC MODELS

This section provides additional information on the construction of our synthetic data, with more detailed documentation¹ and a repository of code for generating custom datasets available online². In the time since its conception, this synthetic framework^{1,2} has already been used as a benchmark in one external cryo-EM study on free-energy landscapes³. Here we repeat much of the previous presentation of our methodology to provide a description of the atomic-coordinate displacements between states for the specific case of our Hsp90 model with $n = 2$ exercising two conformational motions (CM₁ and CM₂). These CMs were designed so as to be fully decoupled from each other, such that no overlap occurs between the two sets of distinct atomic displacements (Fig. S1).

Atomic manipulations of the original PDB⁴ coordinate file were done using PyMOL⁵. For CM₁, the chain A arm was rotated outwards (directly away from chain B) from

its central hinge at residue 677, in 1° increments until a series of 20 rotational states were defined (Fig. S2-A). For each of these 20 CM₁ states, all residues above chain B's elbow (residues 12-442) were then rotated perpendicularly to the CM₁ motions in 2° increments until a series of 20 rotational states were defined for CM₂ (Fig. S2-B). These operations resulted in the creation of a total of 400 unique conformational states. The ensemble of these states can be organized in a 20×20 state space, where each entry represents one of the possible combinations of CM₁ and CM₂. This state space represents our synthetic model's complete ensemble of physically allowable conformations (i.e., the quasi-continuum). The specific size of the state space (400 states) was chosen based on the relative scale of the protein and range of its motions, providing approximately 3 Å and 2 Å gaps between states over a total arc length of 60 Å and 40 Å for CM₁ and CM₂, respectively (as visualized in Fig. S3). To note, geometry correction and energy minimization of generated states were skipped to avoid introduction of unintentional coupling of CMs.

These 400 structures represented by PDB files were then each transformed into 3D electron density maps (EDMs, formatted as MRC volume files) with a sampling rate of 1 Å per voxel and simulated resolution of 3 Å using the EMAN2⁷ module e2pdb2mrc, where each atom is represented by a Gaussian with radius defined by the appropriate number of electrons (Fig. S3). Projections of each 3D density map can then be obtained via standard parallel line integrals along selected directions so as to simulate images generated in the transmission electron microscope (TEM) operated in the bright-field mode⁸. This core framework is used as a basis for the creation of images that are noisy duplicates or, finally, incorporate a CTF in their simulation:

- **Data-type I:** one copy of each state with pure signal and no CTF
- **Data-type II:** uniformly duplicated states with experimentally-relevant SNR
- **Data-type III:** uniformly duplicated states with experimentally-relevant SNR and CTF

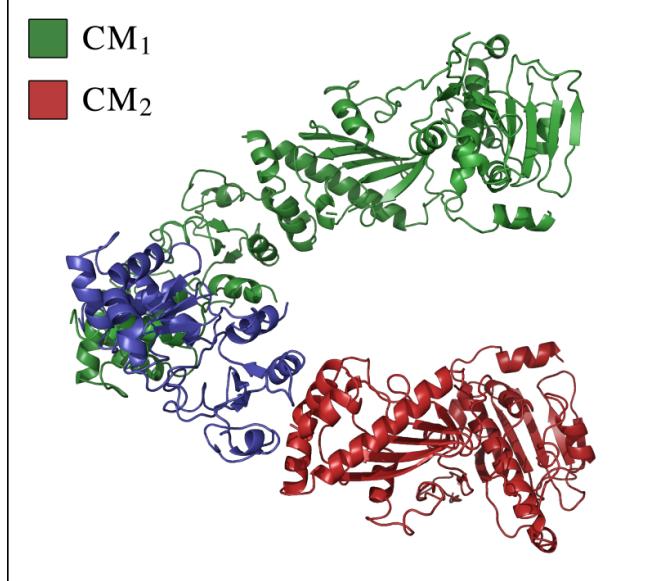


FIG. S1: Cartoon representation of atomic coordinates for state 20_01 with residues demarcated for CM₁ (green) and CM₂ (red). The CM₁ central hinge can be found near the intersection of the blue and green regions at residue 674, while the CM₂ hinge is found near the intersection of the blue and red regions at residue 443. To note, for SS₂, the residues making up the blue region are immobile throughout the entire state space.

An additional alteration on Data-type III is applied for use in our *final analysis*, namely the introduction of a nonuniform occupancy map.

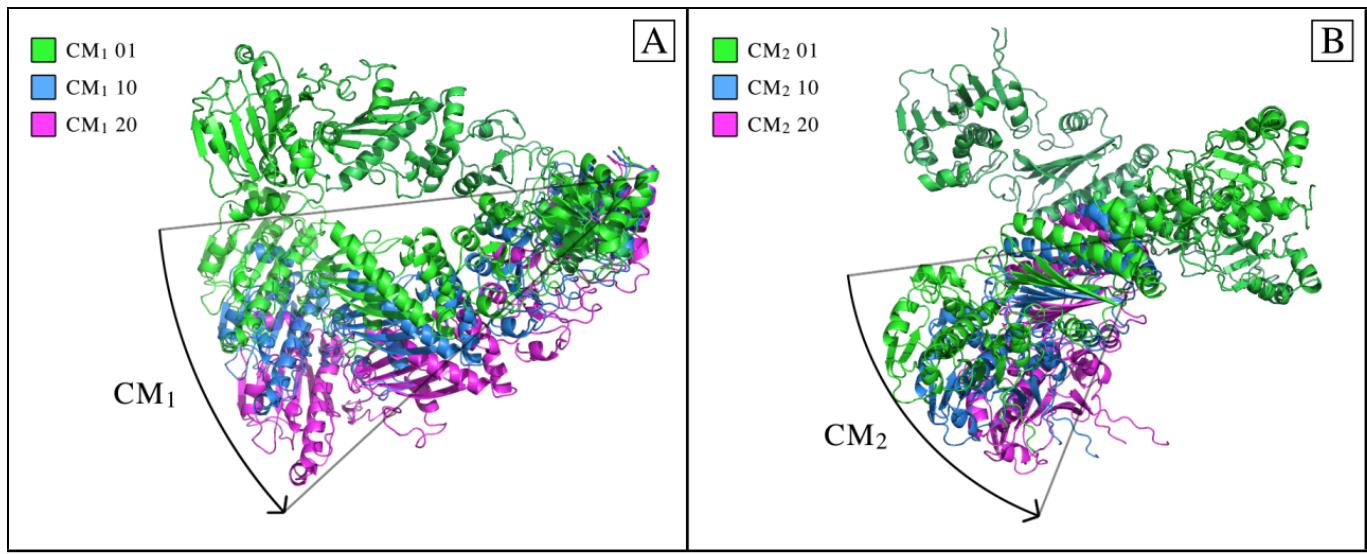


FIG. S2: In [A], a comparison of the CM₁ first, middle and final state is shown. Likewise, in [B], a comparison of the CM₂ first, middle and final state is shown. The CM₂ rotation was performed perpendicularly to the CM₁ rotational hinge where only those atoms above chain B's elbow-region were displaced. As a note, to remove the potential for overlapping residues within certain states, residues 1-11 (making up a loose tail) were removed from both chain A and B.

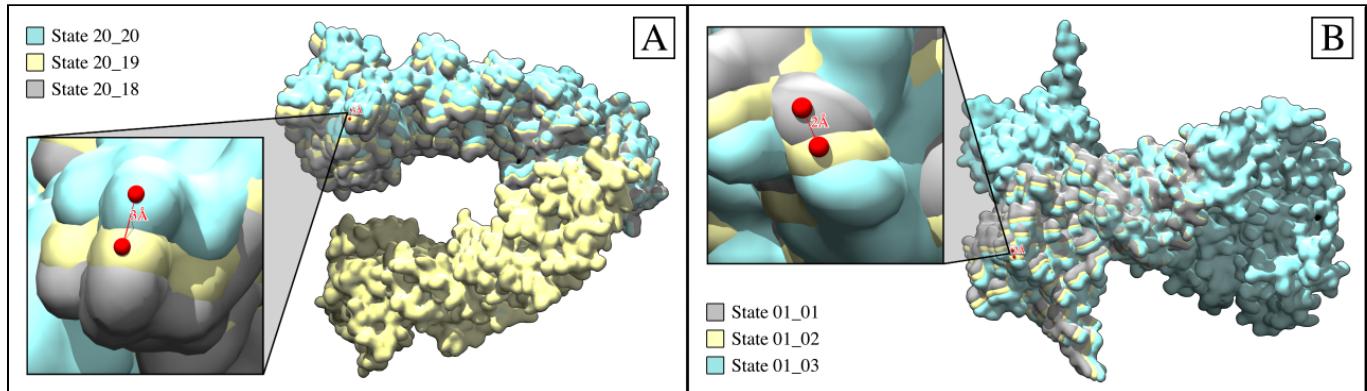


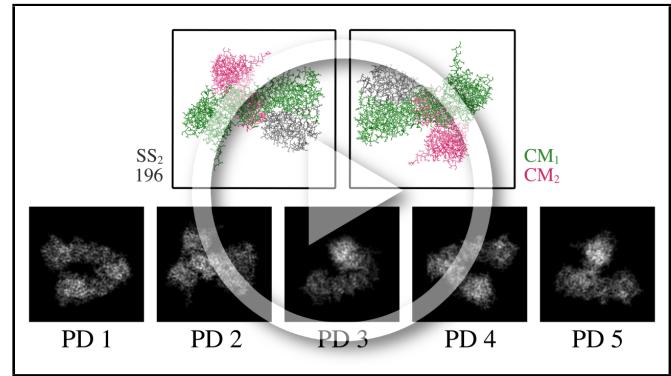
FIG. S3: In [A], a volumetric overlay of the first three rotational states of CM₁ (state 01_01, state 02_01, state 03_01) is presented, visualized as electron density maps (MRC format) via Chimera⁶. As can be seen, only the upper arm (chain A) has been rotationally displaced along CM₁, with 3 Å gaps measured between each consecutive state at the peripheral ends of this rotated region. In [B], a volumetric overlay of the last three rotational states of CM₂ (state 20_18, state 20_19, state 20_20) is shown. Only the upper region of chain B (above the elbow) has been rotationally displaced along CM₂, with 2 Å gaps measured between each consecutive state at the peripheral ends of this rotated region. As a note, the EMAN2 module e2pdb2mrc excludes calculations such as atomic form factors and molecular orbitals. While more accurate maps can be constructed using other methods, this does not affect the results of our heuristic analysis.

SM-II. SIMULATING CRYO-EM ENSEMBLES FOR DATA-TYPE I

In the pristine scenario, for each electron density map within each state space (SS_1 , SS_2 and SS_3), sets of images in five chosen projection directions (PDs) are first obtained. The first of these five PDs was chosen to be normal to the plane of the CM_1 rotation, such that all CM_1 motions from that perspective only underwent changes in the plane of the projection. A similar choice was made for PD_2 , which was projected into the plane of CM_2 motions, with the remaining three PDs chosen at arbitrary positions in angular space. All projections were generated using the EMAN2⁹ module `e2project3d` with no CTF applied, so as to maximally conserve the information contained in each projected EDM. (See movie M1 for a conformational animation of the molecule as viewed in each of these five PDs). For each state space, the set of these images as generated from one of these five PDs forms a high-dimensional manifold Ω_{PD} . In all, this initial procedure resulted in the creation of 400×5 unique 2D images for each of our 400 3D density maps.

As pointed out in the main text, it is important to describe the order in which the states for each SS are indexed, with this ordering repeatedly visualized by color maps throughout this work and ultimately used to locate each state's coordinates within the embedded manifold. The ordering of SS_1 is trivial, with states following a sequence showing CM_1 transition from closed to open ($CM_1\{1\} \rightarrow CM_1\{20\}$; as described by the “arm” motion in Fig. S1-S3), while the ordering of SS_2 and SS_3 follow a clockwork pattern. For SS_2 , the ordering of the states of CM_2 progresses like the second hand of a clock, with each progression from $CM_2\{1\} \rightarrow CM_2\{20\}$ iterating CM_1 (akin to the minute hand) forward once.

Thus, of the 400 states in SS_2 , the first index corresponds to state $CM_1\{1\}_CM_2\{1\}$, the second to state $CM_1\{1\}_CM_2\{2\}$, the 21st to state $CM_1\{2\}_CM_2\{1\}$, and the 400th to state $CM_1\{20\}_CM_2\{20\}$. A similar pattern holds for SS_3 , which now additionally includes an hour hand in this analogy (CM_1), followed by a minute hand for CM_2 and a second hand for CM_3 . An animation of the molecule transforming through these sequences has been provided in movie M1. Of course, the resulting manifold embedding is independent of the ordering of the data points¹⁰; we have merely chosen one such ordering that heightens our awareness of trends in the subsequent outputs.



MOV. M1: An animation cycling through the 400 SS_2 states as represented by the Hsp90 atomic structures (top) and corresponding PDs (bottom), as analyzed in the SS_2 section of Results. The ordering of these states chronologically follows the indices described using our second hand and minute hand analogy. https://www.dropbox.com/s/2qzhrzu0jzc8bbu/M1_SS2_GroundTruth.mp4?dl=0

SM-III. ANALYSIS OF MANIFOLD SUBSPACES FOR DATA-TYPE I

The following figures display several subspaces obtained from embeddings of pristine (noise-free) datasets. First, in Fig. S4, we provide an example of the inverse-cosine mapping $\{\Psi_i, \Psi_j\} \mapsto \{\Phi_i, \Phi_j\}$, as described in the main text. Next, we show the location of each set of Chebyshev polynomials within several exemplary subspaces, and demonstrate how each set corresponds to a unique conformational motion.

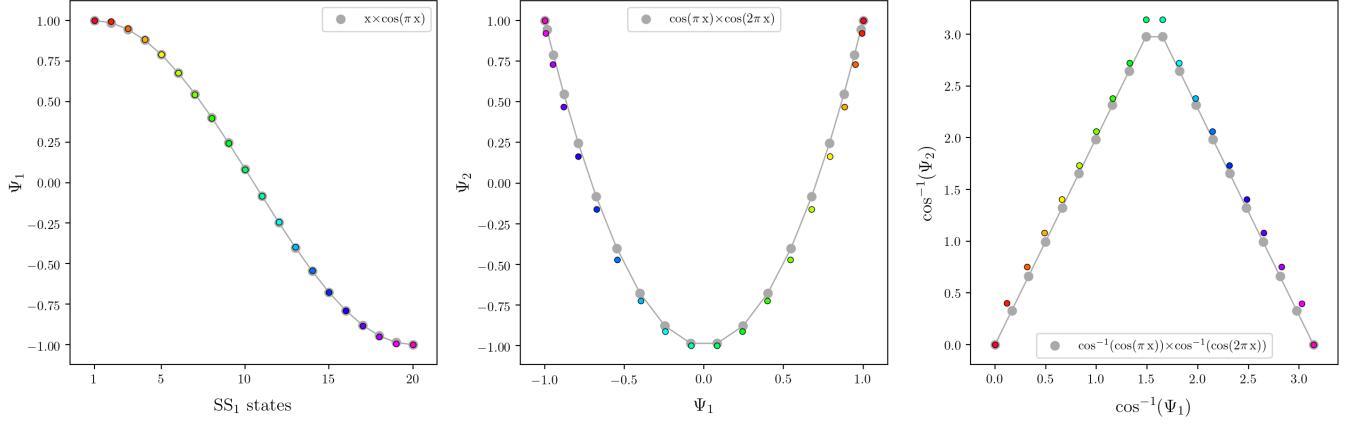


FIG. S4: For each of the three subplots, (1) analytical cosine functions are colored in gray and overlaid with (2) the coordinates of corresponding PD₁ eigenvectors, shown in different vibrant colors denoting the SS₁ sequence of states. To compare these two representations within the same coordinate system, the DM eigenvectors have been scaled to match the range of the analytical cosines. Specifically, for $x \in [0,1]$, 20 x -values were used to generate each $\cos(k\pi x)$ function and subsequently scaled to match the number of SS₁ states; i.e., $x' \in [1,20]$. Similarly, each eigenvector generated by DM was scaled to match the range of each cosine, such that $\Psi_k \in [-1,1]$.

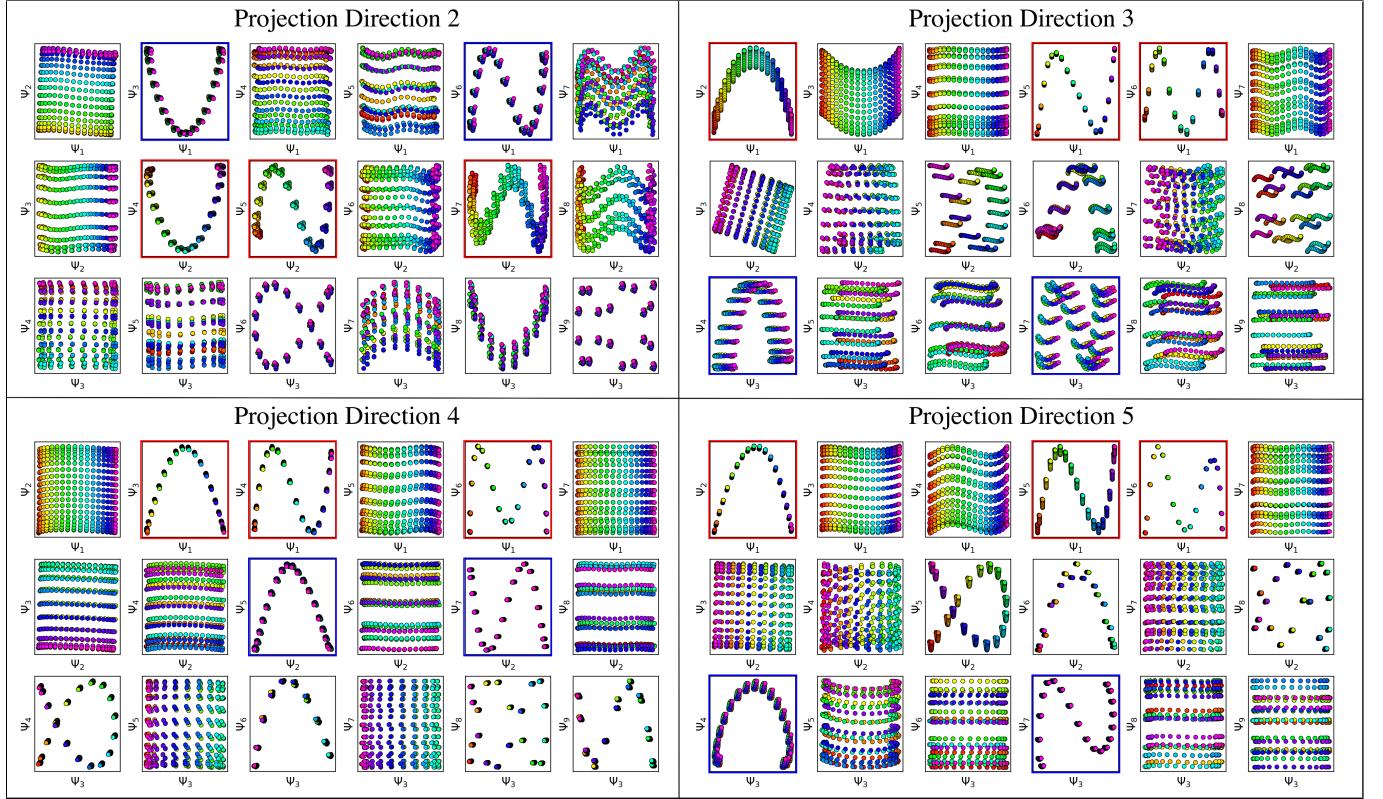


FIG. S5: A similar presentation as is shown in Fig. 4 for the remaining four PDs. Again, for each PD, red and blue boxes indicate the unique set of modes for CM₁ (red) and CM₂ (blue), which are interspersed throughout each row in specific 2D subspaces obtained from their respective N -dimensional embeddings. Subspaces requiring eigenvector rotations (e.g., both parabolas in PD₃) and housing subtle boundary problems (e.g., the curling inwards of the point-cloud trajectory in $\{\Psi_3, \Psi_4\}$ of PD₅) can also be seen in certain 2D subspaces. Note that in PD₂, the hierarchy of CM information is actually reversed from those seen in the other 4 PDs, with the CM₂ Chebyshev polynomials instead present along $\{\Psi_1, \Psi_i\}$ combinations (in the first row) and CM₁ Chebyshev polynomials instead along $\{\Psi_2, \Psi_j\}$ combinations (in the second row). The ordering of sinusoidal sets in each PD is related to the magnitude of the CMs present as seen from the given viewing direction, and not necessarily defined by the CM undergoing the largest change in the ground-truth atomic structures. Therefore, since CM₂ is projected in the plane of PD₂, the magnitude of its apparent motion is visually greater than CM₁, hence the reversed ranking.

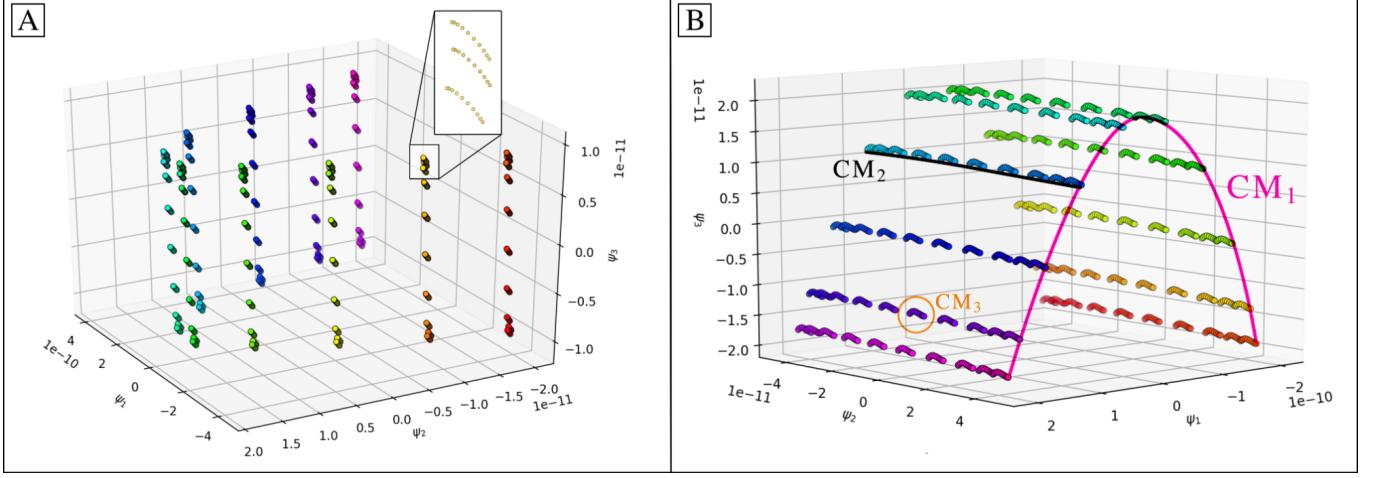


FIG. S6: Comparison of 3D subspaces obtained for images in PD5 [A] and PD4 [B] from SS₃ (1000 states arranged in a $10 \times 10 \times 10$ state space). The inset in [A] illustrates the fractal pattern that emerges when more than two conformational motions are present, where CM₃ can be seen forming a series of mini-parabolas about each of the points in the larger 10×10 parabolic surface. In [B], the hierarchy of these structures is more pronounced, with exemplary CM trajectories depicted by the plotted lines. In section SM-XIV, we demonstrate that these patterns arise due to the additions of orthogonal cosines.

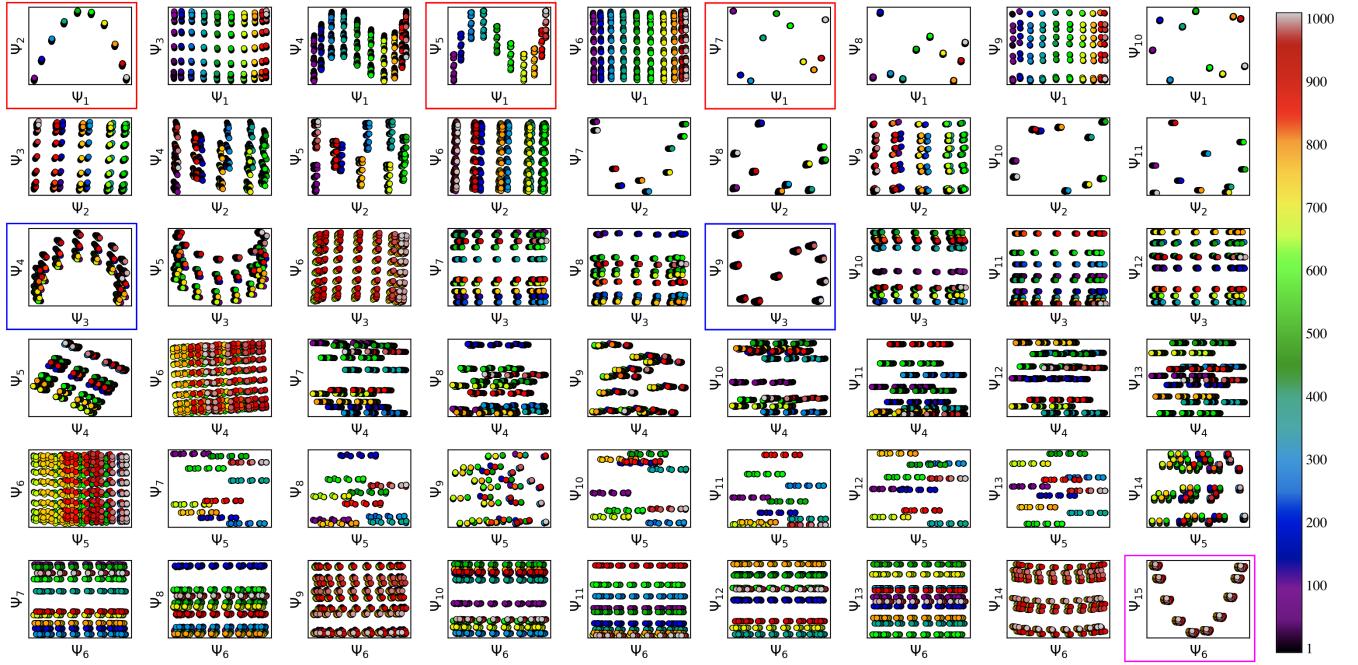


FIG. S7: A set of 2D subspaces projected from the N -dimensional embedding obtained from PD₅ in SS₃. The set of conformational modes corresponding to CM₁ is demarcated by the red boxes around interspersed Ψ_1 plots, and occupy specific $\{\Psi_1, \Psi_i\}$ combinations (where $i > 1$). Likewise, CM₂ and CM₃ are both separately represented by a set of their own conformational modes; demarcated by blue boxes around interspersed Ψ_3 plots and magenta boxes around interspersed Ψ_6 plots (of which only the first is displayed), respectively. Specifically, CM₂ modes span a set of $\{\Psi_3, \Psi_j\}$ combinations (where $j > 3$) while CM₃ modes span a set of $\{\Psi_6, \Psi_k\}$ combinations (where $k > 6$); with the eigenvector depth specified by the apparent span of each conformational motion as seen from the given PD. As expected, points on the trajectories defined by CM₁ modes follow along the full spectrum of colors (i.e., indices 1-1000), while CM₂ and CM₃ points cover a span of 100 and 10 colors, respectively. Note the presence of parabolic harmonics in neighboring rows (for each CM), which are characterized by the presence of α -shaped trajectories (e.g., as seen in $\{\Psi_2, \Psi_5\}$), albeit slightly rotated here from the plane of its 2D subspace).

SM-IV. SIMULATING CRYO-EM ENSEMBLES IN DATA-TYPE II

Data-type II uses images generated for data-type I as a base, from which images are uniformly duplicated τ times. Additive Gaussian noise is next applied to each image individually to grant it a specific SNR that is the same for all images in the set. We define the SNR by the ratio of each image's signal variance (σ_{signal}^2) to its noise variance (σ_{noise}^2)⁸. Here, signal represents the 2D region of pixels corresponding to the average area occupied by the macromolecule. This region is obtained by masking out all pixels within one standard deviation of each image's mean intensity value; in effect, excluding the approximately uniform-intensity background. This process was thus performed by first finding the mean pixel intensity (μ_{signal}) and variance (σ_{signal}^2) of the *signal*, and then calculating $\sigma_{\text{noise}}^2 = \sigma_{\text{signal}}^2 / \text{SNR}$. Using this parameter, we then apply additive Gaussian noise to each image in order to obtain an output image having the desired SNR. In this process, a sample from the Gaussian distribution is added to each pixel's intensity. Each resulting image was then normalized such that the average pixel intensity and standard deviation of pixel intensities was approximately 0 and 1, respectively.

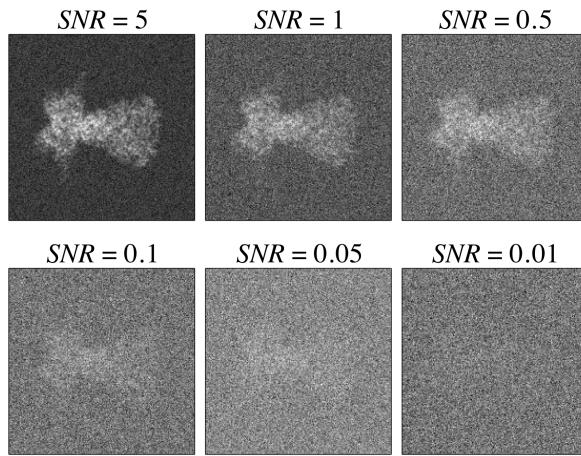
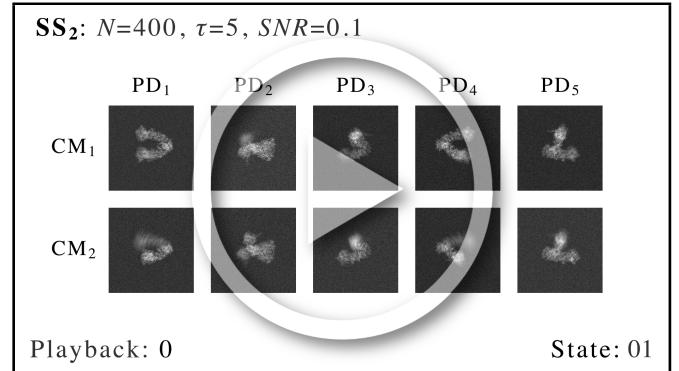


FIG. S8: First image in SS₁ from PD₁ with different values of SNR via additive Gaussian noise. As a note on experimental relevance, SNR = 0.1 has been previously established as a suitable choice for experimental SNR in images obtained by cryo-EM, and its low value can be attributed to the low contrast between macromolecules and their surrounding ice, as well as the limited electron dose required to avoid radiation damage¹¹.

SM-V. ANALYSIS OF SUBSPACE FITTING AND PARTITIONING FOR DATA-TYPE II

In this section, we first present a set of 2D movies captured from SS₂, as seen in movie M2. We then analyze the robustness of recovering several conformational modes while varying τ values and SNR regimes. In Fig. S9, we investigate the robustness for the leading conformational modes (i.e., Chebyshev parabolas and higher-frequency oscillations occupying specific 2D subspaces of PCA embeddings) by fitting across a range of increasing τ values. The coefficient of determination (R^2) is calculated for each mode to provide a measure for the quality of its corresponding fit.

Fig. S10-A shows the R^2 trend for each mode as τ is incrementally increased. In Fig. S10-B, only the R^2 values for the first mode are similarly calculated across several SNR regimes. The asymptotic behavior of each plot is expected, demarcating regions along this trajectory where the intrinsic geometric structure of each mode is optimally reinforced against the background noise. By visual assessment of consecutive τ -defined embeddings, for all asymptotic plots (those in both Fig. S10-A and Fig. S10-B), we found that a suitable value of τ_c could be estimated at approximately half the corresponding plot's maximum value of R^2 . For example, in Fig. S10-A, the value of τ_c for the first mode is approximately 15; in agreement with the first emergence of a robust parabola seen in Fig. 6.



MOV. M2: Set of 2D movies captured along SS₂ subspaces, with each corresponding manifold generated from images with SNR of 0.1 and $\tau = 5$. As seen in both the first and second row, as a result of the integration procedure on each corresponding CM parabola, there is a significant difference in resolution between the desired CM and the CM orthogonal to it. For example, in the first row, the arm motion appears crystal clear in each PD, while the (orthogonal) elbow motion appears blurry; and vice versa for the second row. When 3D movies are eventually constructed from the full set of these 2D movies on S^2 , pairwise information from both CMs is incorporated into each volume such that these anomalies resolve. https://www.dropbox.com/s/z27lp14orpli1xt/M2_2D_Movies_DT2.mp4?dl=0

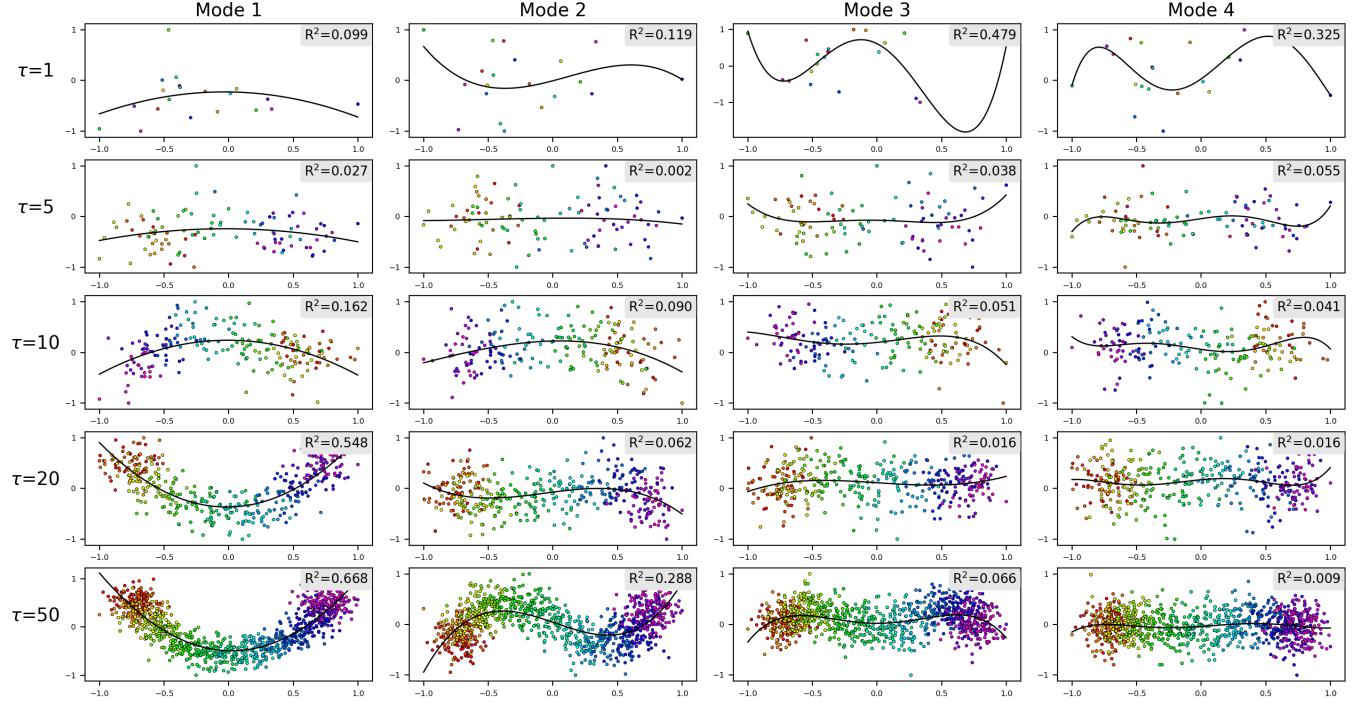


FIG. S9: Each of the above rows show a set of 2D subspaces from a τM -dimensional embedding (in SS_1 where $M = 20$, as obtained from an ensemble of images created with that row's given τ -value and $SNR = 0.1$). Each 2D subspace within each row displays one of the CM's leading Chebyshev modes m via $\{PC_1, PC_{m+1}\}$ (e.g., all 2D subspaces in the first column have x -axis and y -axis defined by PC_1 and PC_2 , respectively). Note that each of these principal components has been scaled to have matching bounds (i.e., $[-1, 1]$). Lines of best fit are then computed with R^2 values recorded, as displayed in the corner of each subplot. As an aside, note the inconsistent orientation of each mode (e.g., upwards or downwards concavity for the parabolic mode), which is due to arbitrary eigenfunction signs which naturally arise during eigendecomposition¹⁰.

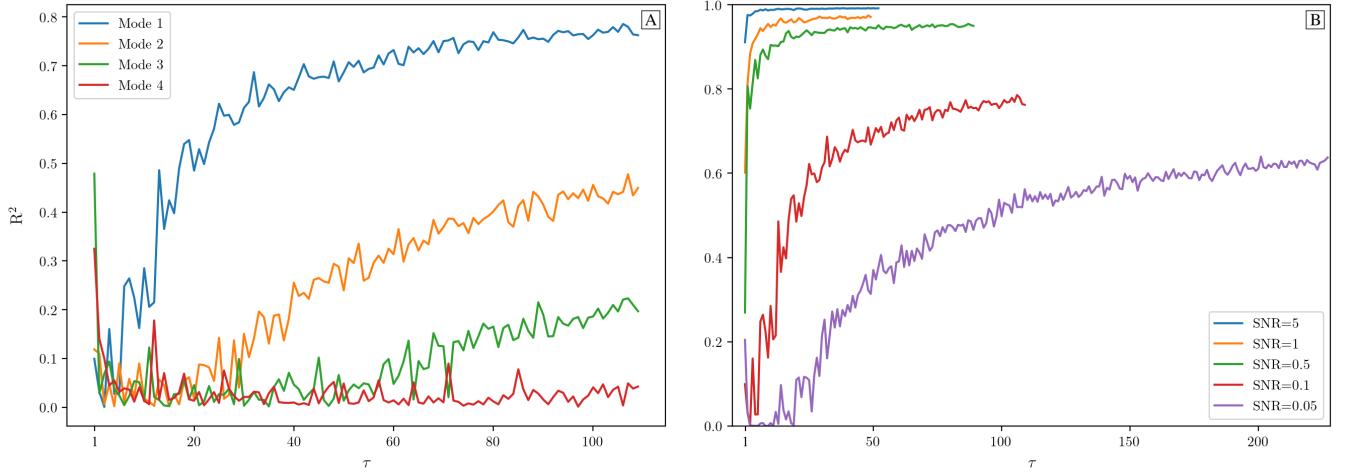


FIG. S10: In [A], the R^2 values are plotted for leading modes with constant SNR. The first mode corresponds to the parabolic Chebyshev mode defined via the projection $\{PC_1, PC_2\}$, with the second Chebyshev mode defined via $\{PC_1, PC_3\}$, et cetera (with exemplary best fits for all modes illustrated in Fig. S9). In [B], only the R^2 values for the first mode are shown while SNR is altered.

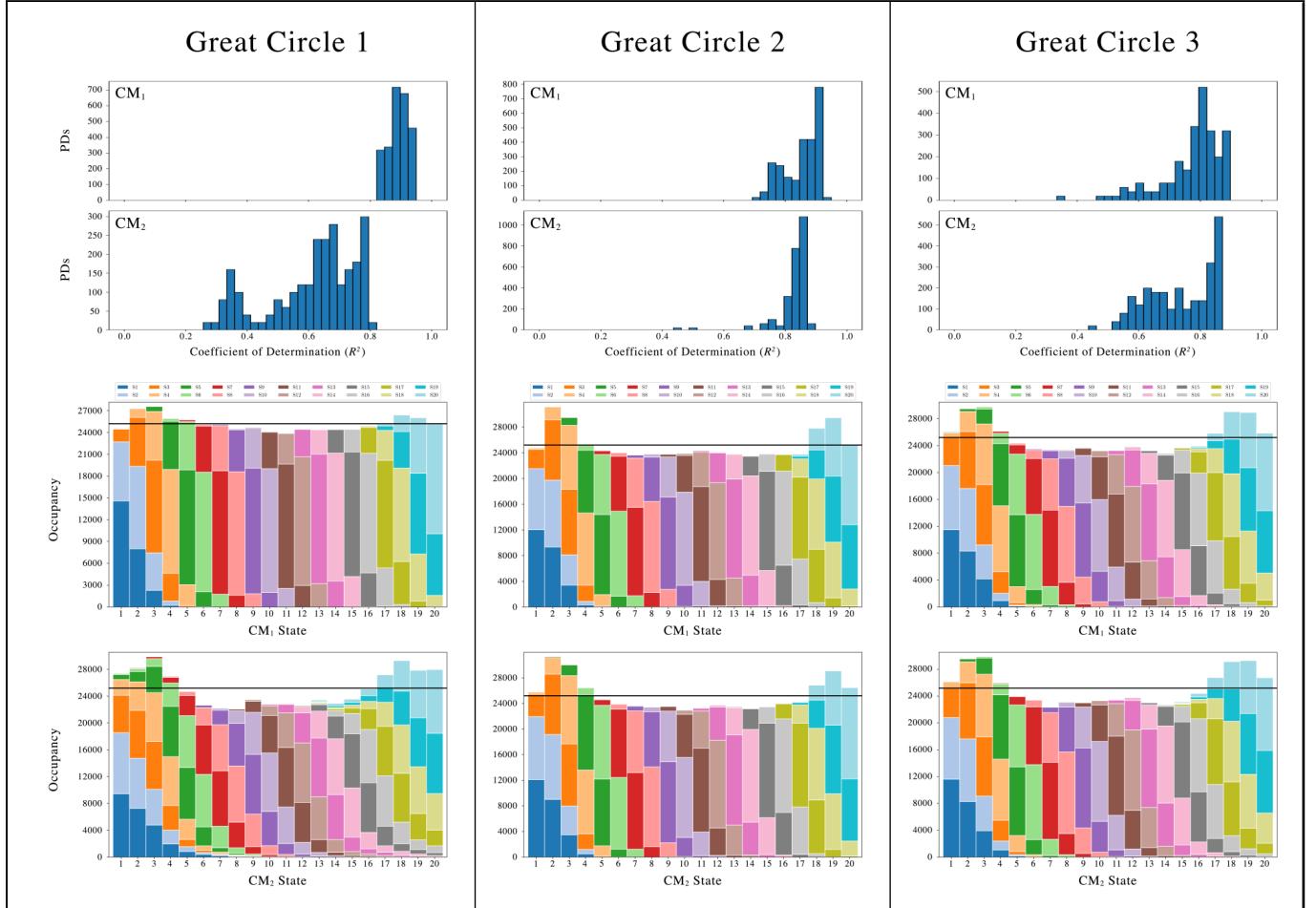


FIG. S11: Each column shows the final outputs of our workflow for one of three equispaced 126-PD great circle trajectories, with each trajectory computed in isolation from all others. To note, each of the great circles used are orthogonal to the other two. In every column, total occupancy maps for CM₁ and CM₂ from SS₂ with $\tau = 10$ are shown, as obtained by integration of the corresponding 20 bins for each CM (corrected for *sense*) in each of the 126 PDs. The total number of images as assigned to each state via our subspace fitting procedure is shown by the height of the 20 bars. Within each bar, the different colors represent how many of the assignments therein belonged to which ground-truth states (as seen in the legend). Note that the black, horizontal line demarcates the expected value of each bin for the ground-truth (flat) distribution; i.e., states \times PDs = $20 \times 10 \times 126 = 25,200$ images. The subplots above each of the great-circle occupancy maps show the corresponding histogram of R^2 scores for each CM as observed in the respective 126 PD-manifold subspaces.

SM-VI. SIMULATING CRYO-EM ENSEMBLES FOR DATA-TYPE III

Data-type III again uses images generated for data-type I as a base. Similar to data-type II, $\tau = 10$ duplicates per state are first generated. However, instead of applying additive Gaussian noise to each image, as was done for data-type II, each image is first filtered by the electron microscope's contrast transfer function (CTF) in an experimentally relevant range. For this task, CTF is generated via the form $\text{CTF} = \sin(\chi) - A\cos(\chi)$, where $\chi = (-\pi\Delta z\lambda k^2 + \frac{1}{2}\pi C_s \lambda^3 k^4)$, Δz is a defocus value randomly assigned in the interval $[5000, 15000]$ Å (positive is underfocus); λ is the wavelength of the electron (calculated via known microscope voltage); k is the spatial frequency; C_s is the spherical aberration; and A denotes the fraction of amplitude to phase contrast. Once the CTF is generated with a choice of relevant microscopy parameters, it is next applied through scalar multiplication with the Fourier transform of the image, followed by an inverse Fourier transform of the product. After this procedure, additive Gaussian noise is uniquely applied to each image ($\text{SNR} = 0.1$) following previous protocol laid out for data-type II. Fig. S12 on the right provides a demonstration of this workflow, which further includes results of CTF correction. It should be noted that we use exactly the same CTF for initially modifying each image as we use later in our pipeline for CTF correction, making no allowance for experimental inaccuracy of CTF estimation.

Upon first performing these operations on the pristine states in SS_2 , we found that, in the presence of CTF for certain PDs, gaps emerged between clusters of points, with points within each of these clusters representing an identical ground-truth state. For this experimentally most relevant dataset, it was clear that we needed better sampling of the quasi-continuum conditions. Thus, we shortened the total span of each CM while keeping the number of states constant, effectively increasing the density of states for data-type III. Specifically, we lowered the rotational distance between neighboring states in each CM until the corresponding point-cloud distributions appeared virtually continuous (i.e., without clusters or discernible gaps). The resulting RMSD between consecutive CM_1 states (e.g., state 01_01 and state 02_01) and CM_2 states (e.g., state 01_01 and state 01_02) was 0.4 Å each. In comparison to previous atomic relations (see caption of Fig. S3), the distance between the peripheral atoms of adjacent CM_1 states decreased to 1 Å and for CM_2 states to 1.5 Å.

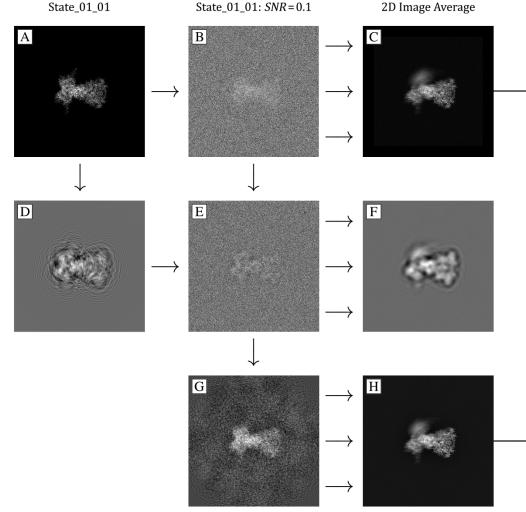


FIG. S12: In the first column, image [A] is a 2D projection of state 01_01 of PD_2 taken without CTF parameters using the EMAN²⁹ module `e2project3d`. Image [D] is that same 2D projection but generated with CTF signal using the following parameter values: defocus 13,500 Å; amplitude contrast ratio 0.1; spherical aberration 2.7 mm; and voltage 300 kV. For both the first and second row, there are 399 other images each, created with respective attributes corresponding to the remaining SS_2 states. All of the 399 images generated similar to [D] were likewise given a random defocus value in the interval [5000, 15000] Å. Image [B] and [E] are the results of adding Gaussian noise to each image [A] and [D], respectively, such that the SNR becomes 0.1. As depicted with three cascaded arrows in the diagram, image [C] is the 2D average of all states generated similar to image [B], and likewise for the relationship of image [F] to image [E]. The image size is 320×320 pixels, with this decision guided by [D] such that the broadening of the point-response function stays within the image bounds. Image [G] is the CTF-corrected version of image [E], as depicted in previous work⁸ using a Wiener filter with $\text{SNR} = 0.1$ and exact assignment of known CTF parameters. Finally, image [H] is the 2D average of all CTF-corrected images similar to image [G]. Note the similarity of [H] with [C], which is a result of filling the missing CTF zero-crossings during the averaging of all CTF-corrected images.

SM-VII. OCCUPANCY ASSIGNMENTS FOR FINAL ANALYSIS

For our *final analysis* dataset, a nonuniform occupancy distribution was created for the $\{\text{CM}_1, \text{CM}_2\}$ map, instead of constant $\tau = 10$ (Fig. S13). As noted earlier, in thermal equilibrium, an occupancy map can be transformed into a corresponding free-energy landscape via the Boltzmann factor¹²: $\Delta G/k_B T = -\ln(n/n_0)$, where n is the occupancy of the current state and n_0 is the occupancy of the maximum occupancy state in the state space. The lowest allowable occupancy for our map was chosen by taking the analysis in section SM-V in consideration.

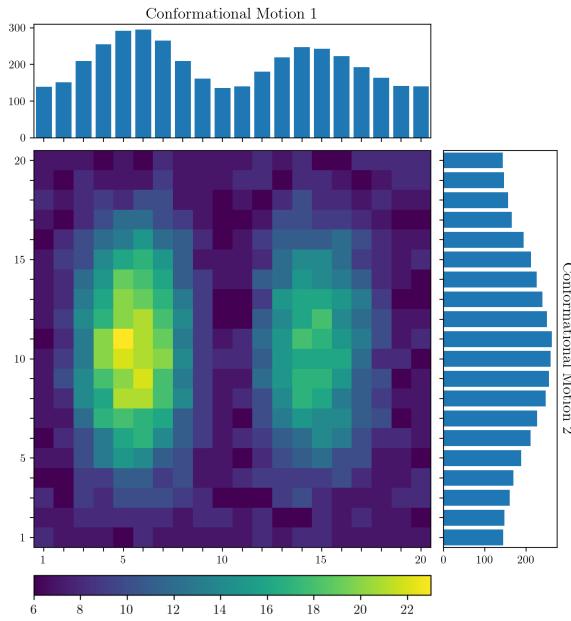
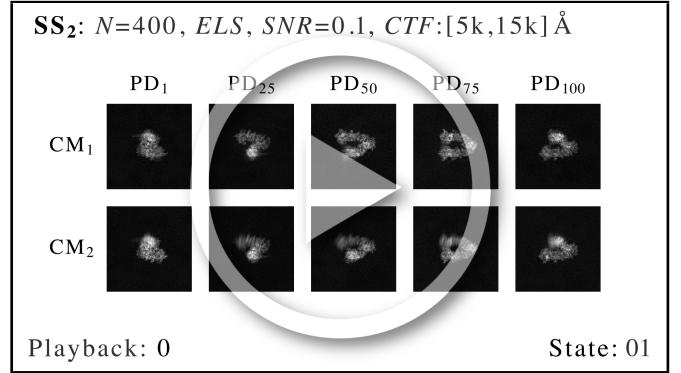


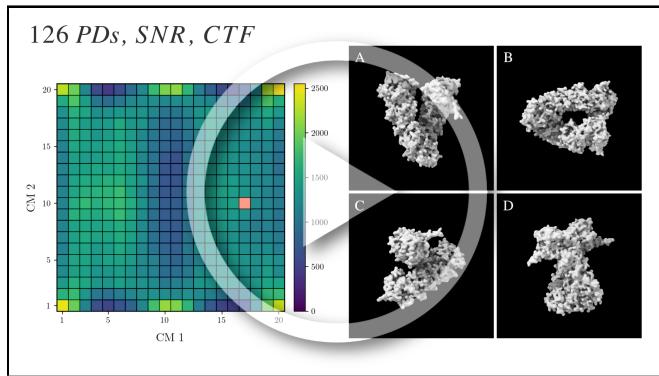
FIG. S13: 2D distribution of occupancies for all 400 states in SS_2 , assigned equally for each PD. The net occupancy of this state space is 4000, such that the total number of images in the complete dataset is the product of 4000 and the number of PDs. The characteristics of this occupancy map were chosen to provide easily distinguishable features along both 1D and 2D conformational motions: specifically, bimodal and unimodal distributions for CM_1 and CM_2 , respectively.

SM-VIII. ANALYSIS OF 2D AND 3D MOVIES FOR FINAL ANALYSIS

The following two movies M3 and M4 show a selection of ESPER 2D and 3D sequences obtained from our final analysis.



MOV. M3: Set of five 2D movies from SS_2 PD-manifold subspaces equispaced on a great circle. Each manifold is generated according to data-type III, with the exception that images are sampled from a nonuniform 2D occupancy map (Fig. S13). The set of CTF-corrected and Wiener-filtered snapshots within each bin are integrated, as opposed to the raw images themselves as previously shown in movie M2. Again, there is a significant difference in resolution between the desired CM and orthogonal CM, with the resolution of the desired CM superb across all states. See movie M4 for 3D movies from the full set of these 2D movies on S^2 , where pairwise information from both CMs is incorporated to fully resolve all molecular domains in each volume. https://www.dropbox.com/s/351288brw25si85/M3_2D_Movies_DT3_ELS.mp4?dl=0



MOV. M4: Output volumes from our final-analysis dataset: 126 PDs, SNR = 0.1 and CTF with microscopy parameters as previously described. A sequence of 69 3D density maps is shown as seen from four orthographic views [A-D] animated along a chosen trajectory in the retrieved 2D state space. Here the 2D occupancy map before R^2 -thresholding has been supplied (in contrast to Fig. 10), with all volumes reconstructed using RELION¹³, without removal of any images in the original ensemble. Post-processing steps for display of each of these volumes included removal of dust (via Chimera's `hideDust` command with size 10) and application of a Gaussian filter (via Chimera, using 1 Å standard deviations of the 3D isotropic Gaussian function).
https://www.dropbox.com/s/kfsddid347zb0j9/M4_3D_Movies_DT3_ELS.mp4?dl=0

SM-IX. ANALYSIS OF SUBSPACE FITTING FOR FINAL ANALYSIS

In the following figure, we provide a collection of 2D subspaces for two PDs (left and right) for the final-analysis data type (i.e., data-type III with nonuniform occupancy assignments). As denoted in the blue boxes, the parabolic CM subspaces tend to curl inwards substantially near their boundaries. This inward-curving effect varies depending on the CM subspace and thus on the type of motion as visualized from the corresponding PD. For all PDs explored, our use of the general conic least-squares fit proved highly robust to these changes.

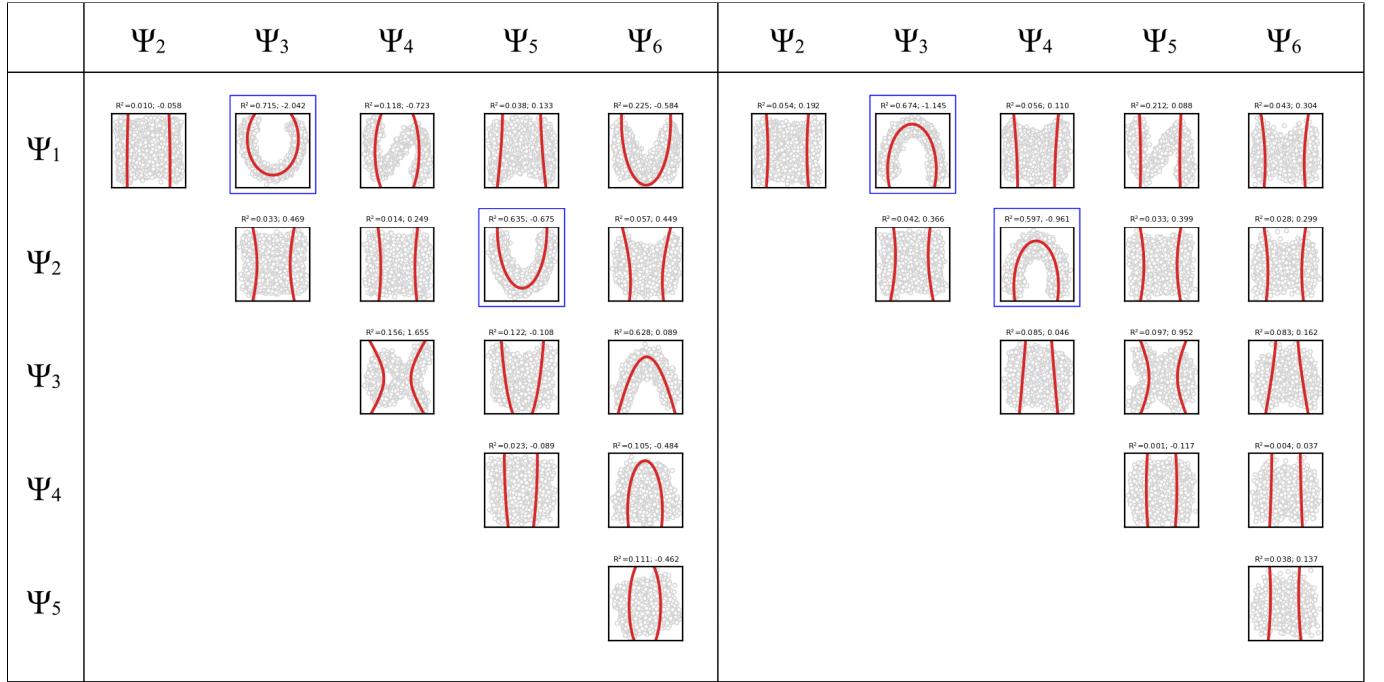


FIG. S14: Here we show the (1) coefficient of determination (R^2); and (2) discriminant of the implicit conic equation above each subplot. The R^2 -value obtained for each 2D subspace is initially used before d -dim rotations to establish the location of parabolic modes. As can be seen above, this score is significantly higher for these parabola-housing subspaces. Across all 126 PDs, non-harmonic parabola-housing subspaces were predominantly defined by a negative (elliptic) discriminant, while harmonics were often associated with a positive (hyperbolic) value. A comparison of these plots with data-type II can be found in movie M7.

SM-X. OVERVIEW OF DIFFUSION MAPS

Given a set of N images, the diffusion map (DM) approach seeks to generate an optimal embedding of the data in a low-dimensional space, so as to preserve all relevant information. Below, we outline the DM framework with considerations taken for the synthetic data explored in our main text. As a preliminary step, we normalize each of the images by removing the mean and scaling to unit variance. As the kernel required for the DM framework is formed using pairwise distances, we first create the distance matrix \mathbf{D} (for each PD independently) by calculating the Euclidean distance for every pairwise combination of its N images. The Euclidean distance between two images $X = (x_1, x_2, \dots, x_P)$ and $Y = (y_1, y_2, \dots, y_P)$, where x_i and y_i denote the intensities at pixel i in images X and Y (each having P pixels), is defined¹⁴ by

$$D_{X,Y} = \left(\sum_{i=1}^P (x_i - y_i)^2 \right)^{1/2}$$

The pairwise distances form a symmetric $N \times N$ square matrix, where a single row represents the distance of the corresponding row-indexed image to each column-indexed image. Next, an isotropic Gaussian kernel is applied to these distances to create a real, symmetric similarity matrix

$$A_{ij} = \exp\left(-\frac{D_{ij}^2}{2\varepsilon}\right)$$

The similarity matrix \mathbf{A} , calculated using a suitable ε value, is then divided by a diagonal matrix of its row sums to construct a symmetric, positive semidefinite stochastic Markov transition matrix \mathbf{M} , representing the relative pairwise affinity between all images¹⁵. Eigendecomposition of the matrix \mathbf{M} is then performed to retrieve an ordered set of N eigenvalues and corresponding eigenvectors (leaving out the steady-state $\{\lambda_0, \Psi_0\}$ eigenvalue-eigenvector pair, where $\lambda_0 = 1$ due to the graph being fully connected¹⁶), which define a nonlinear spectral embedding of the data¹⁰.

The Gaussian bandwidth (ε) in the above expression has a strong influence on the definition of similarity between our simulated images. At small Gaussian bandwidths, the system takes on a relatively fine-grained definition of similarity (i.e., data points only see their direct neighbors). Increasing ε transforms this relationship into a more coarse-grained notion of similarity. These notions of similarity govern the behavior of all subsequent steps, and ultimately impact the geometric structure of the resultant manifold embedding, and thus the DM eigenfunctions. Particularly, in the limit $\varepsilon \mapsto 0$ and $N \mapsto \infty$, and with an appropriate normalization of the similarity matrix¹⁶, the diffusion map eigenvectors converge to the eigenfunctions of the LBO¹⁰.

SM-XI. ESTIMATION OF GAUSSIAN BANDWIDTH AND INTRINSIC DIMENSIONALITY

A detailed analysis of the preferred Gaussian bandwidth regime for both pristine and noisy PD datasets is available in section SM-XIV-D. In summary, for all PD datasets explored, the choice of appropriate Gaussian bandwidth $\varepsilon \in [a, b]$ proved highly flexible, presenting embeddings with virtually identical cosine eigenfunctions across several orders of magnitude. For values slightly below this range, suboptimal yet structured properties emerged. Even farther below this range, the corresponding eigendecompositions did not converge, such that the structures of these embeddings became jumbled in nonsensical patterns. Similar disruptions occurred for all embeddings generated with $\varepsilon > b$, with such anomalies likely arising due to arithmetic underflow encountered during computation of the Gaussian kernel. For noisy datasets, the range $\varepsilon \in [a, b]$ coincided with the prescribed range of ε values defined via a prominent routine for automating this decision (Fig. S15), which uses the correlation dimension as a measure of fractal dimensionality^{10,15,17} (henceforth referred to as the bandwidth estimation method). We have provided an automation strategy¹⁸ in our online repository¹⁹, which uses the inflection point of a fitted hyperbolic tangent to select this value. This automated value was often just shy of retrieving optimal results on pristine datasets, while excelling for cases involving experimentally relevant SNR and τ .

While proponents of the bandwidth estimation method also claim to predict the intrinsic dimensionality n of the dataset¹⁵, we found this recipe to be inconsistent with ground truth (as detailed in Fig. S15 and Fig. S16). The following two figures show the results of using the bandwidth estimation plots as procured following the fractal dimensionality method^{10,15,17}, from which it is expected that (1) the linear region of each plot delimits the range of optimal ε values, and (2) twice its slope defines the intrinsic dimensionality of the dataset. While this method's first prong proved accurate in discovering a suitable ε value for all datasets observed, the second prong always proved highly inaccurate for PD datasets. For example, in Fig. S15, the slopes of each PD within each SS grouping (SS₂ grouped at top and SS₁ at bottom) differ significantly (e.g., within SS₁, the slopes for PD₁ through PD₅ signified a dimensionality range of 1.55 to 1.95). As the ground-truth dimensionality of SS₁ defined via the rotation of a domain of an atomic structure is one, these trends demonstrate that the bandwidth estimation method—while sufficient for measuring the proper ε value—is insufficient for correctly determining the dataset's intrinsic dimensionality. Fig. S16 provides further investigation of these results in the presence of noise, where the intrinsic dimensionality was still significantly overestimated.

Given these results, there remains uncertainty on how to best determine the intrinsic dimensionality of any given dataset (i.e., the number of CMs present to search

for). The estimation of the intrinsic dimensionality of a manifold is a longstanding mathematical problem^{20,21}, which we consider an open issue. As it stands, this information must instead be intuited via a careful analysis of final outputs (i.e., 2D and 3D movies) generated from our framework.

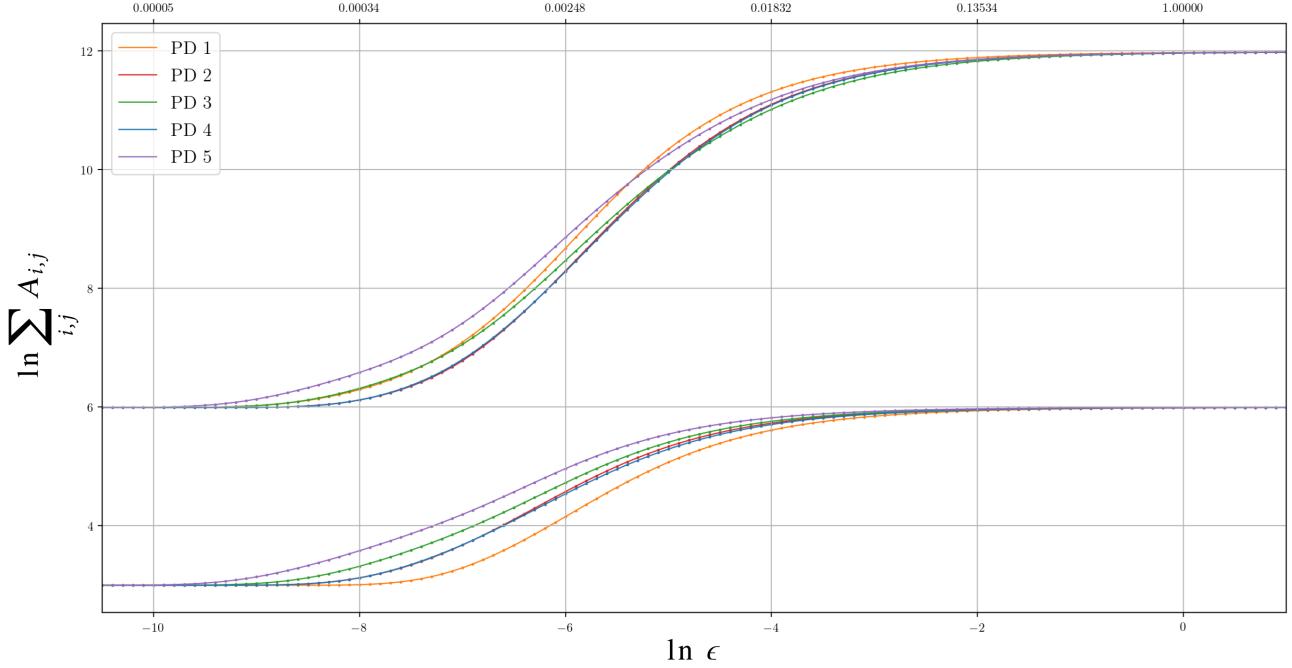


FIG. S15: bandwidth estimation plots as obtained via the fractal dimensionality method. For all PDs, the optimal ε values were manually checked through trial and error, with the optimal range of ε values approximately $[10^{-4}, 10]$. It can be seen that the linear portions of all bandwidth plots for each PD cover a similar ε range, regardless of SS₁ or SS₂ (bottom group and top group, respectively). Instead, PDs from either state space differ mainly by their vertical position in the plot.

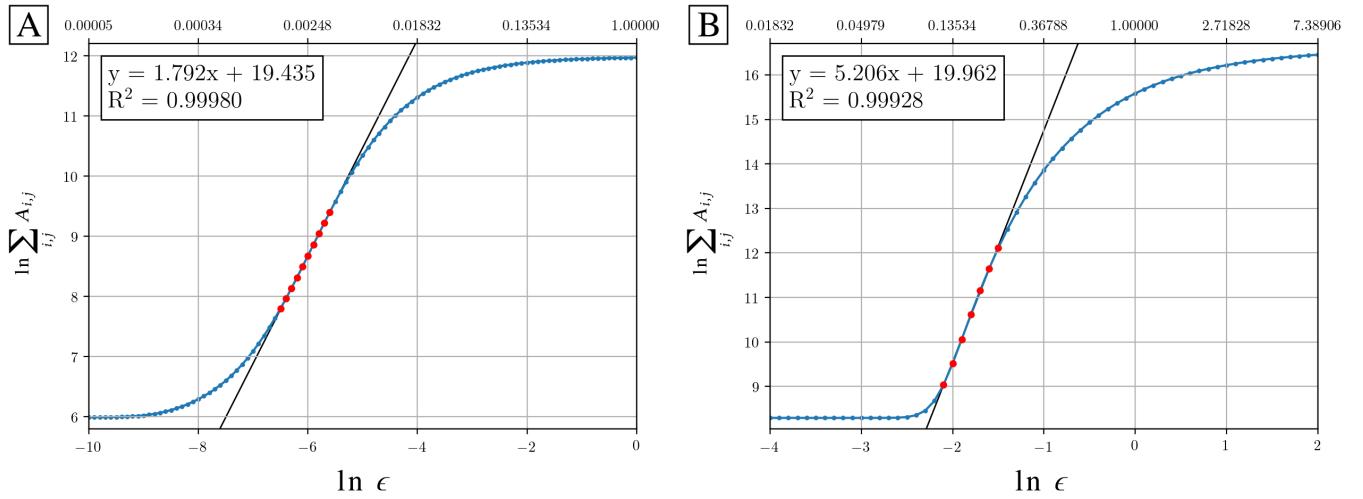


FIG. S16: Bandwidth estimation plots are shown for [A] PD₂ in SS₂ (SNR_{∞} , as previously shown in Fig. S15) and [B] PD₂ in SS₂ ($\text{SNR} = 0.1$ and $\tau = 10$). As can be seen via twice the slope of the fit for each linear region, the ground-truth intrinsic dimensionality of these datasets ($n = 2$) is erroneously defined via this method, and increasingly so with the introduction of noise. However, even in the presence of noise, the criterion for the optimal ε range still proved suitable, with the corresponding eigendecomposition converging for $\varepsilon \in [10^{-11}, 1]$. Note also the asymmetry of each curve and the magnitude of their curvature, which leads to uncertainties in fitting these regions with the prescribed linear model.

SM-XII. COMPARISON OF PCA AND DM

In this section, we complete our comparison of diffusion maps with principal component analysis. First, we briefly readdress our *a priori* expectations as informed by spectral theory²². Both DM and PCA are kernel methods which entail the use of a symmetric matrix (i.e., a Markov transition matrix and a covariance matrix, respectively). Symmetric matrices have many convenient properties, and, particularly for our interest, are diagonalizable with mutually orthogonal eigenvectors, and corresponding real, non-negative eigenvalues. As shown via the principal axes theorem, this diagonalization is determined by the matrix's eigenvectors, which are used to align the innate principal axes of the graph into standard position. When organized along these principal axes, distinct classes of geometries associated with the quadratic form of the symmetric matrix (called “quadric hypersurfaces”) clearly emerge, with the characteristics of each surface specifically defined by its corresponding eigenvalue²².

When a symmetric matrix has only positive or null eigenvalues, the matrix is “positive semidefinite”. Low-dimensional quadrics generated by positive semidefinite matrices include elliptic cylinders, parabolic cylinders, hyperbolic cylinders and cones²³). However, our case is of substantial complexity, as we will be investigating quadric hypersurfaces of graphs generated with inclusion of multiple degrees of freedom and noise. Indeed, as the symmetric matrix of both PCA and DM is positive semidefinite, we anticipate to recognize one of these forms within subspaces of each subsequent embedding (e.g., parabolic cylinders in Fig. S6). With these fundamental similarities in mind, we next display figures (Fig. S17-18) referenced in our main text, which show striking similarity between results obtained by these two techniques. This is further supported by the fact that the eigenvectors of DM generated with very large ε -values converge to the eigenvectors of PCA.

SM-XIII. THE LAPLACE-BELTRAMI OPERATOR ON A RIEMANNIAN MANIFOLD

The Laplace-Beltrami operator (LBO) acting on a scalar function f on a compact (closed and bounded)²⁴ Riemannian manifold is given by²⁵⁻²⁷

$$\nabla^2 f = g^{-1/2} \partial_i (g^{1/2} g^{ij} \partial_j f)$$

where $g = \det(g^{ij})$ and g^{ij} are the components of the metric tensor. Specifically, we are interested in the eigenfunctions of the LBO, $\nabla^2 f = \lambda f$, noting these form a complete basis in the functional space $L_2(\Omega)$ of measurable and square-integrable functions on the manifold Ω ²⁸. For a bounded manifold, the eigenfunctions must further satisfy boundary conditions; for example, DM requires the Neumann boundary conditions¹⁶, such that

the derivatives on the boundaries vanish. Therefore, the eigenfunctions depend also on the boundary of Ω .

If g is constant on Ω , the expression simplifies to $\nabla^2 f = g^{ij} \partial_i \partial_j f$, where g^{ij} are constant values. For example, in two dimensions

$$\nabla^2 f = (g^{11}(\partial_1)^2 + g^{22}(\partial_2)^2 + 2g^{12}\partial_1\partial_2)f$$

Moreover, if $g^{11} = g^{22} = 1$, we recognize the common Laplacian in Euclidean coordinates, which is the sum of pure second derivatives²⁵

$$\nabla^2 = \sum_{i=1}^d \partial_i^2$$

It is well understood that the eigenfunctions of the LBO on a manifold Ω carry useful information about its intrinsic geometry, and are thus important for understanding many systems. For compact manifolds with a boundary, as an example, these eigenfunctions are the modes of vibration of a string (1D) or a membrane (2D). For compact manifolds without a boundary (closed manifolds), the well-known spherical harmonics are eigenfunctions on the surface of the 2-sphere. In the field of structural biology, the eigenfunctions of the LBO on $SO(3)$, which are the Wigner-D functions, have been used for retrieving the unknown orientations of single-particle X-ray and cryo-EM snapshots²⁹. In general, the eigenfunctions of the LBO on different manifolds are fundamental to mathematics and science, and describe a wide diversity of seemingly disparate phenomena—reflecting the so-called “underlying unity of nature”—from quantum mechanics to gravitational fields³⁰.

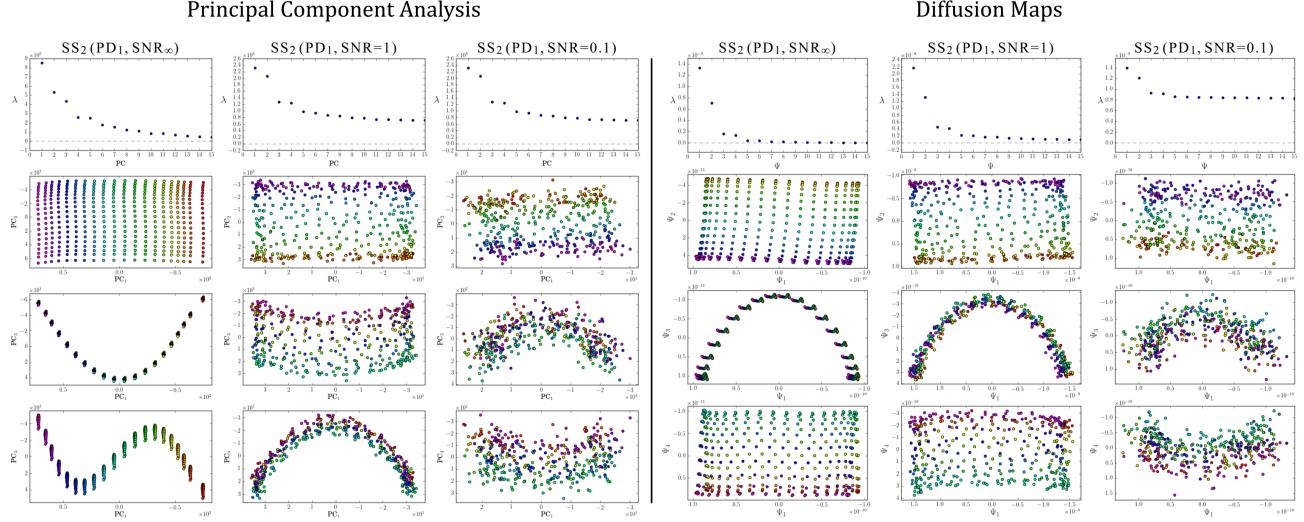


FIG. S17: Comparisons of 2D subspaces and eigenvalue spectra obtained via PCA (left) and DM (right) for PD₁ in SS₂ across three SNR regimes (one SNR regime per column; with uniform occupancy $\tau = 1$). As can be seen in both linear and nonlinear dimensionality reduction frameworks, the well-defined structure of these subspaces deteriorates rapidly as increasing amounts of additive Gaussian noise is introduced on each image. Overall, the outputs of PCA on these datasets revealed a striking resemblance to those produced by DM. Importantly, the parabolic mode is conserved for both frameworks even within experimental regimes (SNR = 0.1).

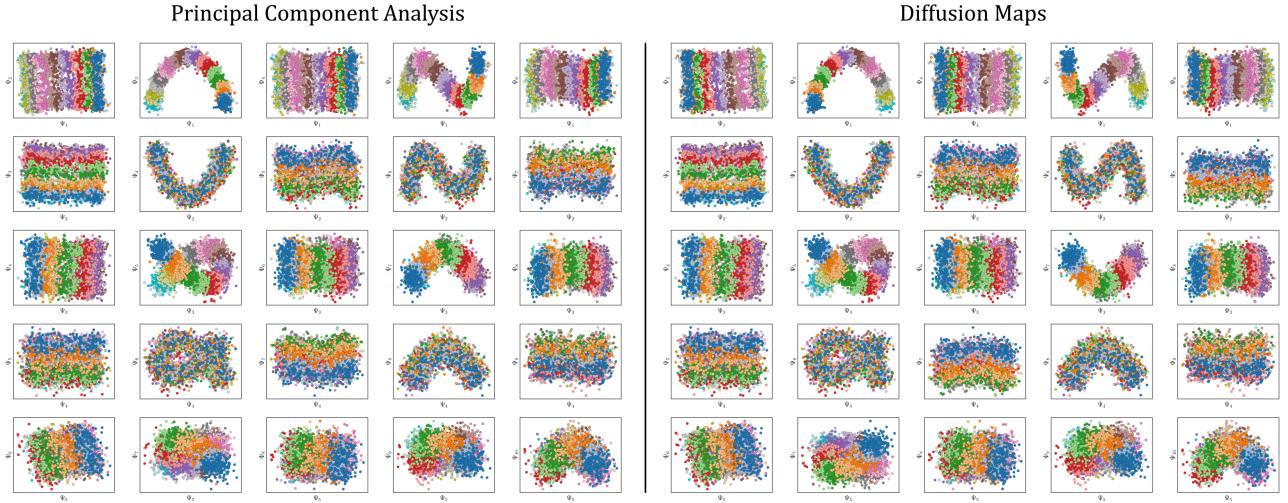


FIG. S18: Comparison of 2D subspaces obtained via PCA (left) and DM (right) for PD₂ in SS₂, with image sets generated with SNR = 0.1 and $\tau = 10$. Colors have been assigned to data points so as to match the ground-truth indices of states covering CM₂, which is the most visually pronounced motion as viewed from this PD. Here, the CM₂ subspaces can be seen in the first row ($\{\Psi_1, \Psi_i \mid i > 1\}$), the CM₁ subspaces in the second row ($\{\Psi_2, \Psi_j \mid j > 2\}$), CM₂ first-harmonics in the third row ($\{\Psi_3, \Psi_k \mid k > 3\}$), CM₁ first-harmonics in the fourth row, and CM₂ second-harmonics in the fifth row. Overall, the similarity in outputs between these two frameworks is undeniable, with the only visual difference appearing in the arbitrary directionality (*sense*) of each coordinate; which is to be expected.

SM-XIV. ANALYSIS OF THE LBO EIGENFUNCTIONS

In the following subsections (A-D), we perform an analysis of the eigenfunctions of the LBO on a set of distinct manifolds Ω . First, in subsection A, we use the DM framework to investigate the known eigenfunctions of the LBO on the interval and rectangular domains, which are cosines. We expect this investigation to inform our heuristic discoveries for the PD embeddings (Ω_{PD}), where similar cosine eigenfunctions were observed for each degree of freedom. As well, we will use these ideal manifolds in A to build intuition for rotations needed to realign essential eigenfunctions, which were observed in the Ω_{PD} embeddings. Following this analysis, we ultimately detail how the structure of manifolds obtained from a conformational state space transforms as the data type is translated stepwise from (B) atomic models Ω_{ACS} , to (C) 3D electron density maps Ω_{EDM} , and finally, to (D) 2D projections Ω_{PD} . In the last case, recall that 2D projections are the only form of data readily accessible in a cryo-EM experiment.

A. Eigenfunctions of the latent space

To get insight into the characteristics of the Ω_{PD} eigenfunctions, we abstract the manifold of the PD-dataset as a Euclidean space with rectangular boundaries. This is motivated by the most simple representation of our ground-truth state space of atomic models, where the relationship between equispaced coordinates in the prior matches the relationship between equiangular molecular domain rotations in the latter. By separately embedding the collection of states in each of these two data types and comparing their resulting eigenfunctions, we will show that these two spaces are nearly identical. In effect, the rectangular domain can be viewed as the conformational *latent space* to which our collection of more advanced state spaces is compared. We will additionally show that for the embeddings of 3D electron density maps and 2D projections, the mapping relative to the latent space becomes distorted. This effect can be explained by a change of the metric induced in the process.

In the 1D space, 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 with molecular domain rotated by a constant angular increment. To represent our SS_1 PD dataset, we uniformly sample $N = 50$ equispaced points from a 1D interval $X \in [0, \ell = 1] \subset \mathbb{R}$, with each of these 50 points representing a unique state of the molecule. Following the DM framework, we then calculate the distance matrix for this collection of points and embed the data in a low-dimensional space. As a note, for all embeddings that follow, we will show that two characteristic regimes emerge depending on choice of Gaussian bandwidth, which we will denote with ε_{\downarrow} and ε_{\uparrow} for the small and large regime, respec-

tively. For the smaller Gaussian bandwidth, a cosine series emerged for all eigenfunctions (Fig. S19-A), in very good agreement with the Laplacian on a 1D Euclidean interval with Neumann boundary conditions. Specifically, we anticipate (and retrieve) canonical²⁸ eigenfunctions of the form $\psi_v(x) = \{\cos(v\pi x/\ell) \mid v \geq 1\}$. 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 polynomials (Fig. S19-B). These Legendre polynomials, however, only occur for hyperrectangles, which are n -dimensional Cartesian products of orthogonal intervals

Next, to represent our SS_2 PD dataset, we uniformly sample $N = 50 \times 50$ points from a 2D interval $X \times Y \in [0, \ell_x = 1.1] \times [0, \ell_y = 1] \subset \mathbb{R}^2$. For its illustrative properties, we avoid degeneracy by ensuring $(\ell_x/\ell_y)^2$ is not a simple ratio²⁸. Again, we follow the DM framework by calculating the pairwise distances between these points and embedding the data in a low-dimensional space. As demonstrated in Fig. S19-C, the set of eigenfunctions obtained in the smaller Gaussian bandwidth regime matched our *a priori* expectations for the Laplacian on a rectangular domain with Neumann boundary conditions. These canonical eigenfunctions are²⁸

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

following the same pattern for higher-dimensional domains $\Omega_R = [0, \ell_1] \times \cdots \times [0, \ell_n] \subset \mathbb{R}^n$ (with $\ell_i > 0$). Again, as we incrementally increased the Gaussian bandwidth from ε_{\downarrow} to ε_{\uparrow} , this set of complete and orthogonal cosines smoothly transformed into the orthogonal Legendre polynomials set, which are now a function of both x and y , as expected (Fig. S19-D). Importantly, the leading Legendre polynomials provide a direct, linear map of the input data points, a consequence of the linear terms $P_1(x)$ and $P_1(y)$. This is preferred over the non-linear map achieved by the cosine functions, as it is unencumbered by nuisances such as non-uniform rates of change and parabolic harmonics. In subsections C and D, however, we will show that this preferred behavior cannot be obtained for 3D electron density maps and 2D projections.

Returning to the smaller of the two Gaussian bandwidth regimes, we next compare the previous non-degenerate rectangular results to those from a degenerate square domain, with $N = 50 \times 50$ points equispaced identically along X and Y (Fig. S20-A). Due to the presence of degenerate eigenvalues, which can arise for domains with a rational ratio²⁸ $(\ell_x/\ell_y)^2$, we encounter pairs of eigenfunctions that appear different from the non-degenerate case of the rectangle²⁸ (as seen, for example, by pairs $\{\Psi_1, \Psi_2\}$ and $\{\Psi_4, \Psi_5\}$ in Fig. S20-A). In Fig. S20-C, 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 eigenfunctions²⁸, where the normalization of the eigenfunctions restricts this linear transformation to a rotation and reflection (i.e., the group of orthogonal

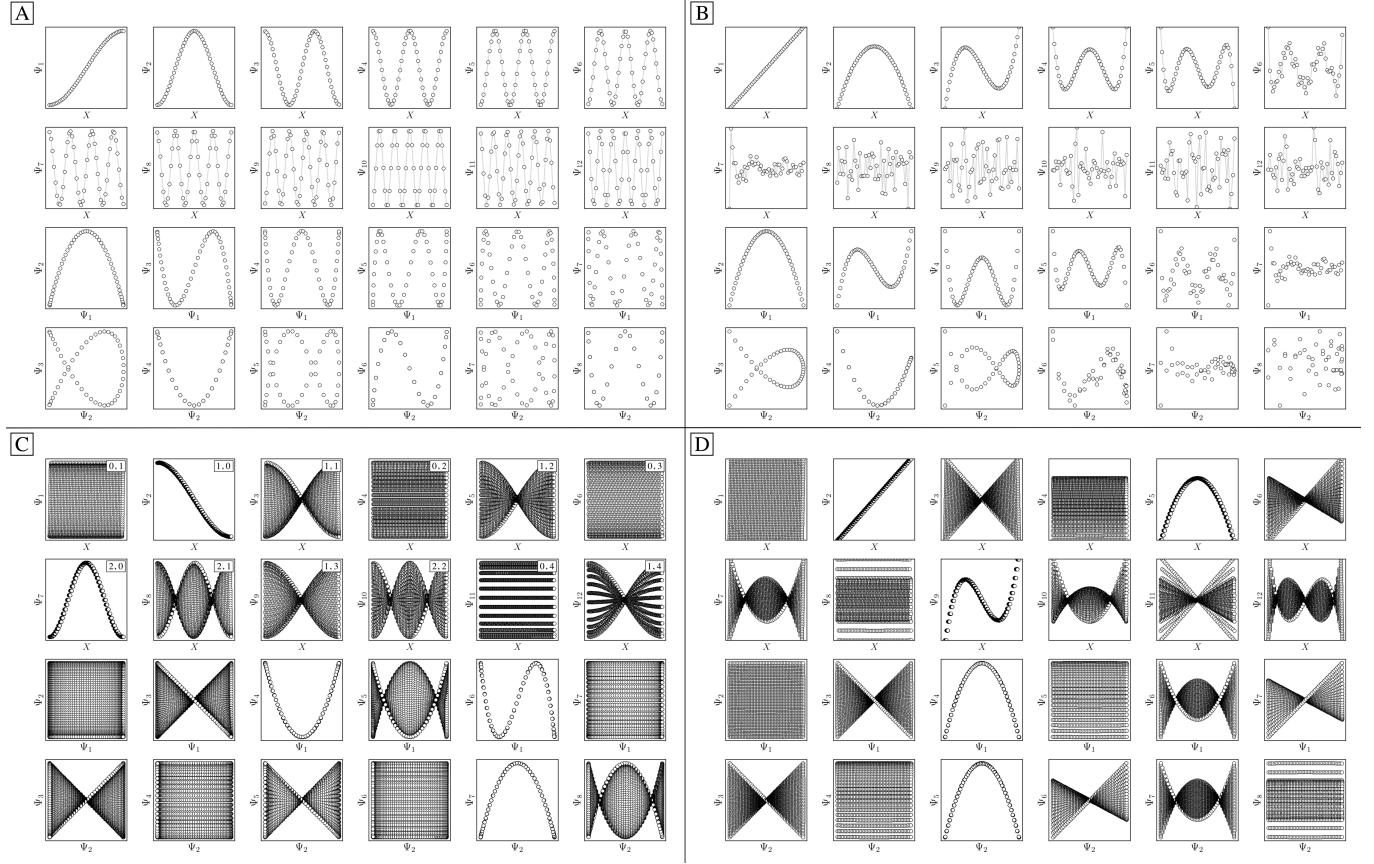


FIG. S19: DM eigenfunctions of the 1D interval for small ($\epsilon_{\downarrow} = 5 \times 10^{-5}$) and large ($\epsilon_{\uparrow} = 10$) Gaussian bandwidths are shown in [A] and [B], respectively. Likewise, eigenfunctions of the $N = 50 \times 50$ rectangular (nondegenerate) domain for small and large Gaussian bandwidth are shown in [C] and [D], respectively. As was done in the main text, eigenfunctions have been independently displayed by indexing each by its ground-truth ordering 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 $\{v, w\}$ have also been provided in the top right-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.

transformations). For example, the $\{\Psi_1, \Psi_2\}$ pair is of form $\Psi' = \mathbf{R}^T \Psi$ such that

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

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

As seen for Ψ_6 in Fig. S20-A, these summands can also have the form of two products $\Psi_6 = A\cos(\pi x)\cos(2\pi y) + B\cos(2\pi x)\cos(\pi y)$, with any A and B such that $A^2 + B^2 \neq 0$. Hence, it can be seen that these aberrant eigenfunction pairs are defined by an admixture of cosines in a higher-dimensional space, with form

$$\begin{aligned} \Psi_i &= A\cos(v\pi x)\cos(w\pi y) + B\cos(w\pi x)\cos(v\pi y) \\ &= A\psi_{vw} + B\psi_{uw} \end{aligned}$$

By using an appropriate rotation operator $R_{i,j}$, the summands within each eigenfunction pair can be

maximally separated among both members $\Psi_i = \psi_{vvw}$ and $\Psi_j = \psi_{uwv}$, such that the canonical eigenbasis is recovered (Fig. S20-B). As demonstrated using analytical expressions $\psi_{1,0} = \cos(\pi x)$ and $\psi_{0,1} = \cos(\pi y)$ in Fig. S20-C, this separation occurs multiples of $\theta = 90^\circ$ apart. In the Fig. S20-C example, at $R_{1,2}(45^\circ)$, these eigenfunctions have form

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

which decouples back into two distinct modes ($\cos(\pi y)$ 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 eigenfunctions obtained via DM, with each initially assuming a random rotation angle (Fig. S20-A) requiring a specific correction $R_{i,j}(\theta)$ (Fig. S20-B).

While degeneracy suggests a possible cause for the appearance of eigenfunction misalignments in Ω_{PD} , we note that it is a rather rare event in our data sets and that it

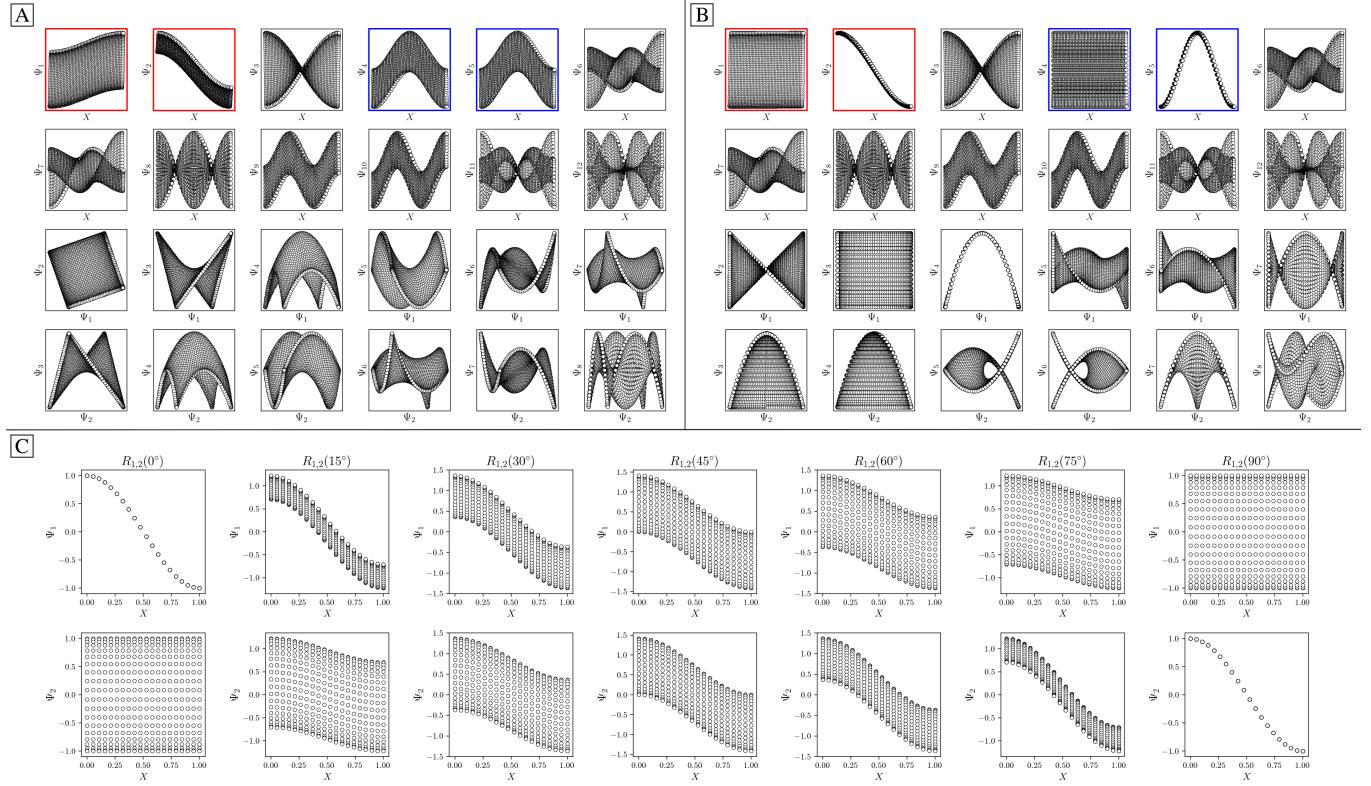


FIG. S20: DM eigenfunctions of the $N = 50 \times 50$ square domain for small Gaussian bandwidth ($\varepsilon_\downarrow = 5 \times 10^{-5}$) are shown in [A] and [B] before and after high-dimensional rotations, respectively. It can be seen here that pairs of eigenfunctions exist that contain relationships aberrant to the canonical eigenfunction form (Fig. S19-C). 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. This property is demonstrated 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 X) as they are jointly rotated 90° . By applying rotation operators $R_{1,2}(\theta) = -19^\circ$ 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 intelligently applied.

can be identified from the eigenvalue spectrum. Pairs of rotated eigenfunctions, at least approximately, can also be mimicked when 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. S20-A. Recall that an affine mapping preserves collinearity and ratios of distances, but in general not distances and angles. In subsection C, we will explore the possibility of other classes of geometric transformations on Ω_R .

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. S19 and Fig. S20, and extensively throughout the main text. Fig. S21 provides a schematic using the known analytical eigenfunctions (Fig. S21-B and Fig. S21-C) chosen so as to match the results from DM on the square, degenerate Ω_R . In Fig. S21-A, we display the eigenfunctions from Fig. S20-A in two different reference frames corresponding to our ground-truth knowledge. Specifically, we

plot the points in each eigenvector in a sequence corresponding to their initial ground-truth arrangement along each degree of freedom (for the rectangular Euclidean space, along either X or Y), which is shown in the first and second row of Fig. S21-A, respectively. As shown in Fig. S21-B and Fig. S21-C, a given reference frame captures the eigenfunction on a projected plane in the n -dimensional space where it resides.

B. Eigenfunctions of the atomic models

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 3D atomic-coordinates $3m$ (e.g., Fig. S2). We generate these structures as described in the first step of our synthetic-generation protocol (section SM-I), which are subsequently used there to produce a corresponding set of 3D electron density maps and 2D projections. Importantly, the set of these 3D atomic-coordinate structures in

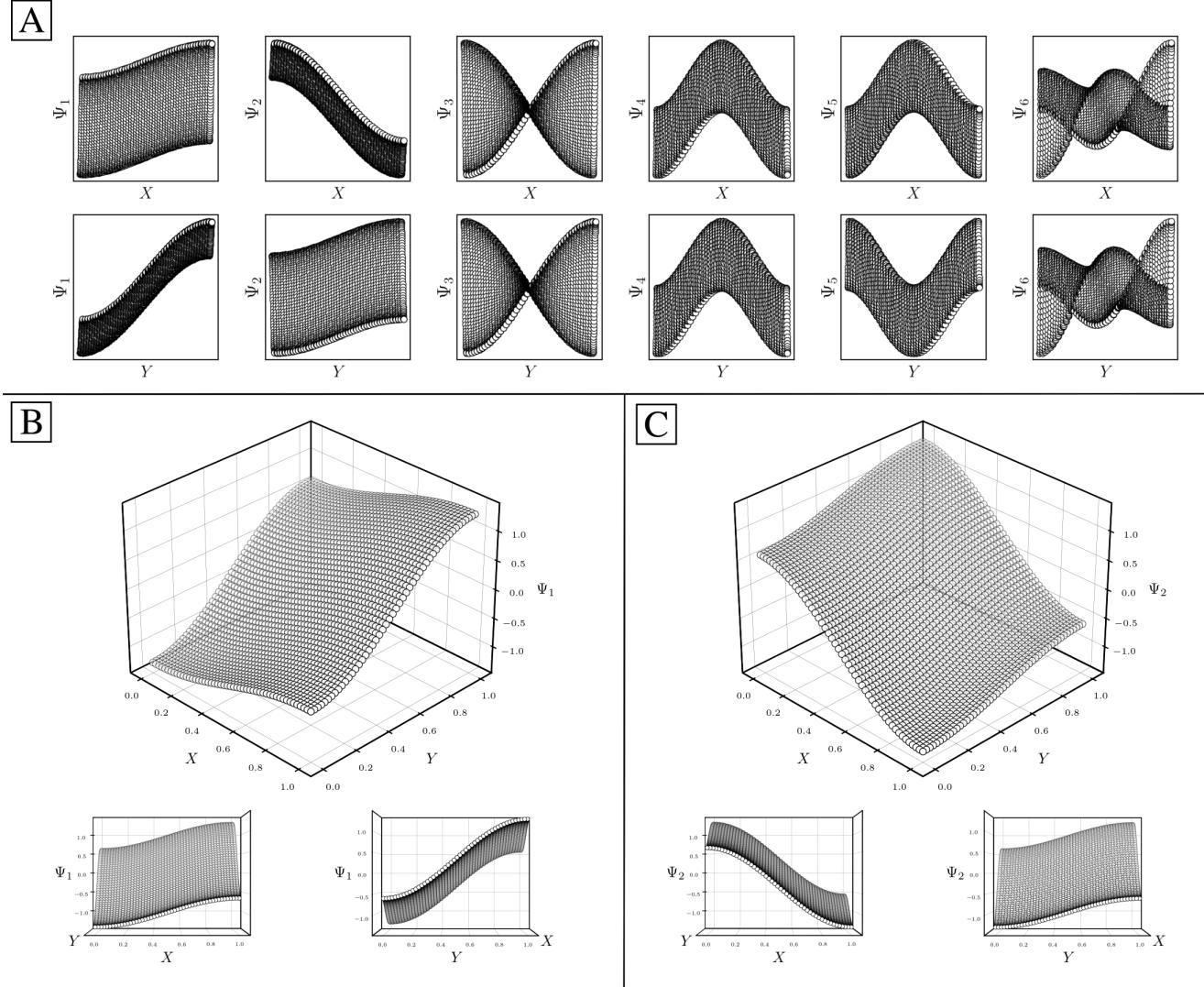


FIG. S21: DM eigenfunctions from Fig. S20-A are shown again in the first row of [A], which were displayed by ordering the points of the embedding in a sequence based on the ground-truth x -coordinates. The second row of [A] displays these eigenfunctions instead via the sequential ordering of ground-truth y -coordinates. Subplots [B] and [C] are analytically generated so as to match the appearance of Ψ_1 and Ψ_2 in the first and second row of [A], respectively. For this presentation, the equations $\Psi_1 = \cos(\theta)\cos(\pi x) + \sin(\theta)\cos(\pi y)$ and $\Psi_2 = -\sin(\theta)\cos(\pi x) + \cos(\theta)\cos(\pi y)$ were used, with $\theta = 250^\circ$. As can be seen in [B] and [C], each eigenfunction exists in an n -dimensional space defined by the n degrees of freedom of the system. Thus, by displaying a point from the manifold along a sequence in the embedding corresponding to a known degree of freedom, we are effectively viewing each eigenfunction on a projected plane in its n -dimensional space.

$\Omega_{\text{ACS}} \subset \mathbb{R}^{3m}$ represent the fundamental biophysical identity of each state, from which the cryo-EM experiment could only obtain two-dimensional information in the form of images. Following the DM framework, we first calculated the distance matrix for SS_2 , which we obtained by the root-mean-square deviation (RMSD) for each pair of its 400 3D atomic-coordinates structures (PDB files). The RMSD between two sets of atomic-coordinate structures $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_m)$ each composed of m atoms is defined as

$$\text{RMSD}(X, Y) = \sqrt{\frac{1}{m} \sum_{i=1}^m \|(x_i - y_i)\|^2}$$

which is up to an irrelevant factor of $m^{-1/2}$ equal to the Euclidean distance $D(X, Y)$.

The resulting DM embeddings for the small and large Gaussian bandwidth regimes are shown in Fig. S22-A and Fig. S22-B, respectively, and share a strong resemblance to those found for the latent rectangular domain (Fig. S19). Again we note the presence of cosine eigen-

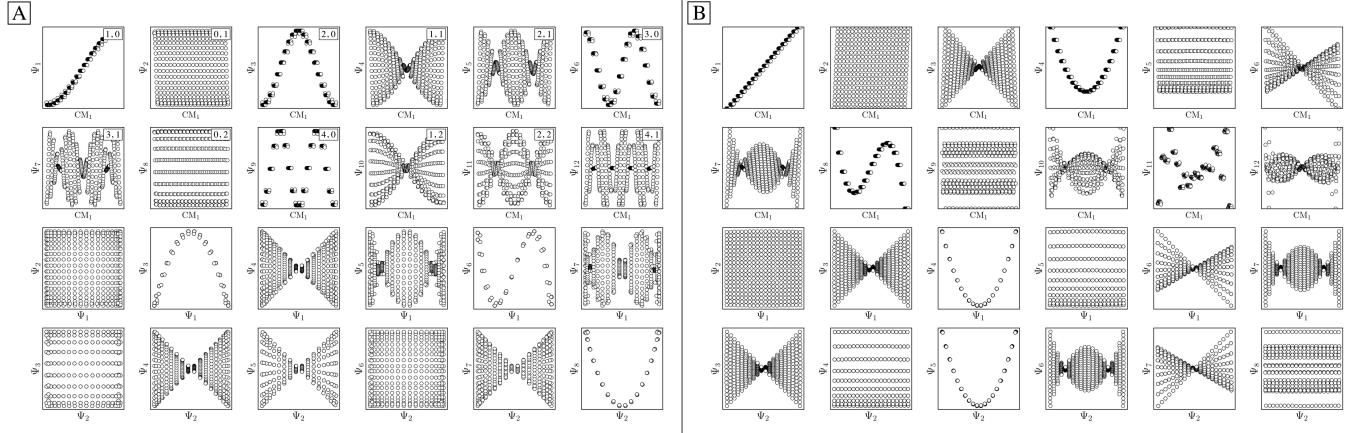


FIG. S22: DM eigenfunctions obtained for $20 \times 20 = 400$ atomic models occupying SS_2 for small ($\varepsilon_\downarrow = 0.1$) and large ($\varepsilon_\uparrow = 1000$) Gaussian bandwidths are shown in [A] and [B], respectively. Leading eigenfunctions are displayed in the first two rows via sequential indexing along the ground-truth CM_1 coordinates (i.e., equispaced rotations of chain A). In [A], each eigenfunction's corresponding modes $\{v, w\}$ are provided in the top right-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.

functions for the small Gaussian bandwidth regime, and a near-perfect linear form (via leading Legendre polynomials) in the large Gaussian bandwidth regime ($\{\Psi_1, \Psi_2\}$ in Fig. S22-B). For the latter, we will show this feature is a luxury not obtained in the other data types explored. In the small Gaussian bandwidth regime, we can identify both our CM_1 and CM_2 parabolas residing in the subspaces $\{\Psi_1, \Psi_3\}$ and $\{\Psi_2, \Psi_8\}$, respectively. Similar results were found for the manifold embeddings generated using 3D atomic-coordinate structures organized in SS_1 and SS_3 .

The striking similarity between the eigenfunctions of the latent space and the eigenfunctions of the atomic models can be rationalized as follows. If the range of a single body rotation is moderate ($\lesssim 30^\circ$), the distance D_{ij} between any two states i and j within this range is to high accuracy[†] $D_{ij} = \Theta_{ij}(\sum_{k=1}^m r_k^2)^{1/2}$, 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 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. Also not investigated in this paper, the linearity holds for body translations as well, where the distance is directly 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 consequence of the linear relationship between distance and the multi-body motions, rotations and translations.

C. Eigenfunctions of the 3D electron density maps

We next demonstrate that the properties of manifolds (as seen in the previous subsection for the 3D atomic-coordinate structures) significantly change when the data representation of their underlying states is altered. Specifically, we investigate how the conformational relationships between states are changed when representation by atomic coordinates is transformed into one by 3D electron density maps (EDMs; e.g., Fig. S3). To this end, we generated the EDMs for each of the 3D atomic-coordinate structures for all previously defined state spaces, as is described in our synthetic-generation protocol (section SM-I). We calculated the pairwise Euclidean distances between these EDMs in $\Omega_{\text{EDM}} \subset \mathbb{R}^V$, with V the number of voxels in an EDM, and performed an embedding via the DM framework. Overall, over a wide range of Gaussian bandwidths, the structure of the resulting eigenfunctions was very similar to the structure of eigenfunctions retrieved for the atomic models in the small Gaussian bandwidth regime. Importantly, these eigenfunctions were still of the form ψ_{vw} (see subsection B), 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 was much less drastic, such that the cosine eigenfunctions appeared in both regimes. For small Gaussian bandwidth regimes (i.e., a few orders of magnitude below the value determined by the bandwidth estimation method ε_* , as defined in section SM-XI), we found that the CM_2 parabola-housing subspace was buried deeply in low-ranking eigenvectors (e.g., Ψ_8 and higher), with

[†] This approximation is justified because the rotation matrix has only linear terms in rotation angle, provided the rotation angle is small.

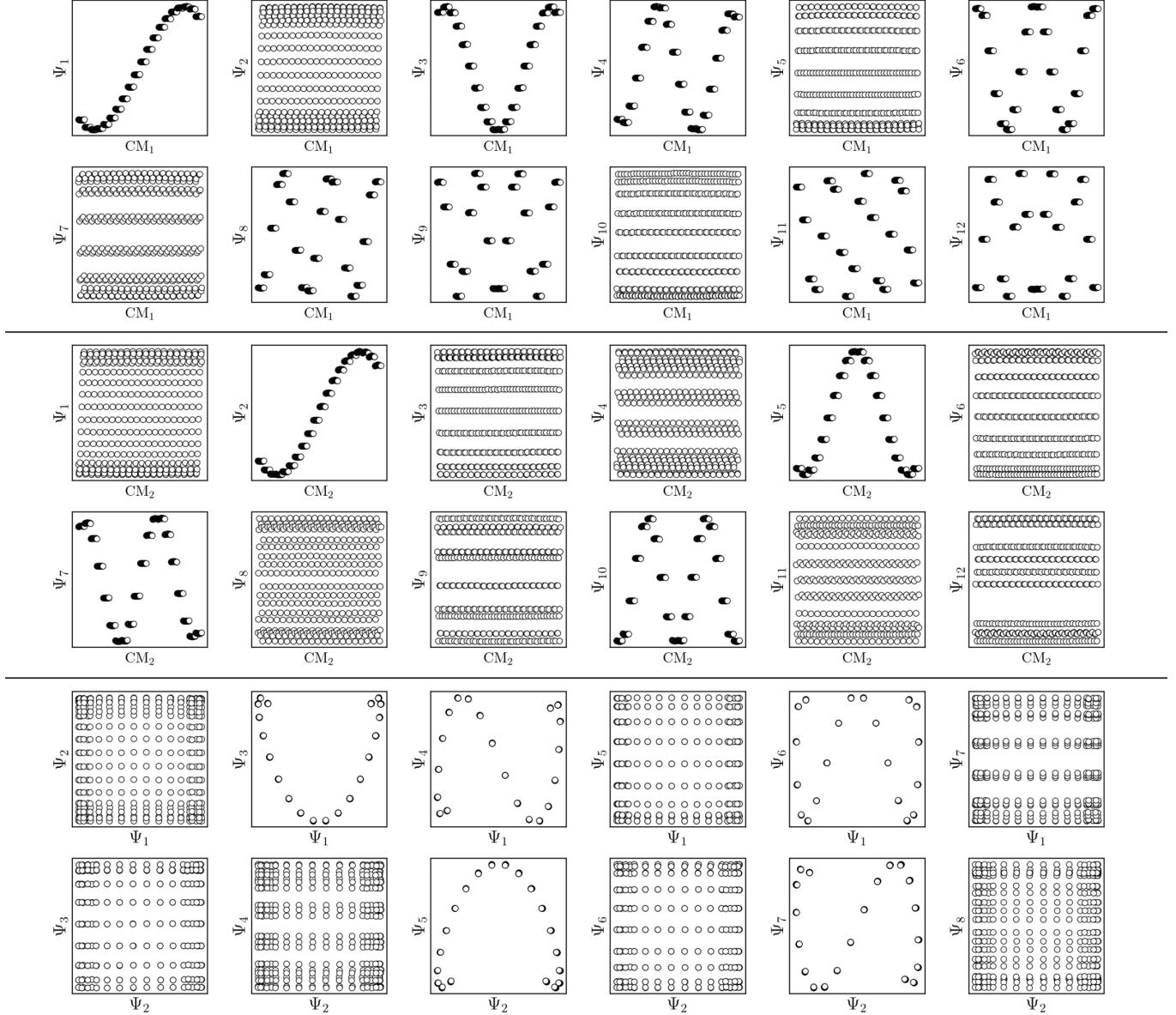


FIG. S23: The results of the DM embedding of EDMs (pure signal) from SS₂ are shown. Leading eigenfunctions as indexed by CM₁ and CM₂ are displayed in the first four rows, followed by their composites. Overall, there is near-perfect alignment of these eigenfunctions with the initially-obtained eigenvector basis, such that no rotations are required. As a final note, the pronounced inward curling at the boundaries of certain subspaces (e.g., {Ψ₁, Ψ₃}) is due to insufficient sampling, as was identically encountered in our analysis of the embeddings of manifolds obtained from subsequent projections of these EDMs.

numerous CM₁ parabolic harmonics occupying the subspaces in between. In addition, eigenvectors with cross terms $\Psi = \{\psi_{vw} \mid v, w \neq 0\}$ were found scattered mostly in mid-range positions (e.g., Ψ₁₂ and higher). Since these properties were not observed in the embeddings of the atomic-coordinate structures, we conclude they are a result of a change in the metric.

In contrast, for larger Gaussian bandwidth regimes (i.e., near and significantly above ε_*), eigenvectors with cross terms were buried in much deeper subspaces (e.g.,

Ψ₃₄ and higher), with the majority of leading eigenvectors housing content exclusively for either CM₁ ($w = 0$) or CM₂ ($v = 0$). Thus, the corresponding CM parabolas for these ε_\uparrow regimes typically occupied the first two subspace rows ($\{\Psi_1, \Psi_i\}$ and $\{\Psi_2, \Psi_j\}$), with respective harmonics positioned in trailing subspaces. These CM parabolas also had a near-perfect distribution of points, whereas for the ε_\downarrow regime, the distribution of points had noticeably less precision to the ideal form. We note that the embeddings obtained above and below these regimes

were 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_\downarrow consistently appears much closer to the ideal form. In Fig. S23, 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 (rows one and two) and the CM_2 reference frame (rows three and four) are displayed, as well as a set of leading composites of these eigenfunctions in the rows that remain.

Importantly, as there was no Gaussian bandwidth value that could “recover” the preferred Legendre-like form, it appears that this feature is “lost in translation” upon transformation from atomic models to EDMs, since the metric is changed. (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 of the 3D atomic-coordinates. Instead of the 3D coordinate points that stand for the atomic positions of each structure, the data for each EDM is represented by a 3D array of values, one at each voxel. A key difference then, is that in the latter case, the displacement of atoms are no longer accounted for individually. Instead, every voxel in the data of one state is now compared to those same voxel locations in the data structure of 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. S24 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). As a consequence, we observed that 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 the corresponding data live in an altogether different manifold. Although the explicit geometric form of Ω_{EDM} is unknown, we have shown that the spectral properties of the Laplacian in 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 since our analysis only requires an un-

derstanding of the eigenfunctions of a manifold, and not necessarily its exact shape.

D. Eigenfunctions of the 2D projections

Since a detailed description of the eigenfunctions of 2D projections is provided in the main text, we continue this current narrative only as it pertains to the relationship of the eigenfunctions of the LBO on $\Omega_{PD} \subset \mathbb{R}^P$ with those from previously-established models (i.e., rectangular Euclidean latent space, atomic-coordinate structures and EDMs). For similarities, as was observed for the EDMs, we found that eigenfunction characteristics could be broadly classified into two classes via either a small or large Gaussian bandwidth regime. In either regime, the eigenfunctions of the PD manifolds were again of the form ψ_{vw} , such that only cosines emerged. 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 was also much more impactful for PDs than for the EDMs. In the small Gaussian bandwidth regime, CM_2 subspaces had a severely suboptimal point distribution, such that, in some PDs, identification of the CM_2 parabola was completely obstructed. These CM_2 subspaces were also buried in trailing eigenvectors, and interspersed among those with cross terms (i.e., $\Psi = \{\psi_{vw} \mid v, w \neq 0\}$). We also note that the value determined by the bandwidth estimation method (ε_*) fell within this regime, making it a suboptimal choice for pristine data. In contrast, the large Gaussian bandwidth regime (i.e., one order of magnitude larger than ε_* and spanning numerous orders of magnitude above it) was superior in every sense, with CM_1 and CM_2 eigenfunctions having ideal point distributions and corresponding subspaces occupying leading eigenvectors. As well, the cross-term eigenfunctions were present only in far trailing eigenvectors (e.g., Ψ_{31} and higher), and would not be obstructive during the analysis. Briefly, we note that upon introduction of noise (SNR = 0.1) and duplicate states ($\tau = 5$), ε_* was instead an optimal choice (along with numerous orders of magnitude above it), while anything below this range was completely inadequate.

For either Gaussian bandwidth regime, we found that the eigenfunction misalignments can emerge—and with varying magnitude—depending on the projection direction. Since previously we have shown that no such property is apparent in the manifold embeddings generated from the 3D EDMs from which these PDs originate, it is clear that the emergence of these eigenfunction misalignments is tied to PD disparity. We hypothesize that, as different 2D projections are taken of the EDMs via $p : \Omega_{EDM} \mapsto \Omega_{PD}$, the geometry of Ω_{EDM} can become contorted due to the change of pairwise interatomic distances resulting from foreshortening in projection (Fig. S25), such that the apparent span of one CM

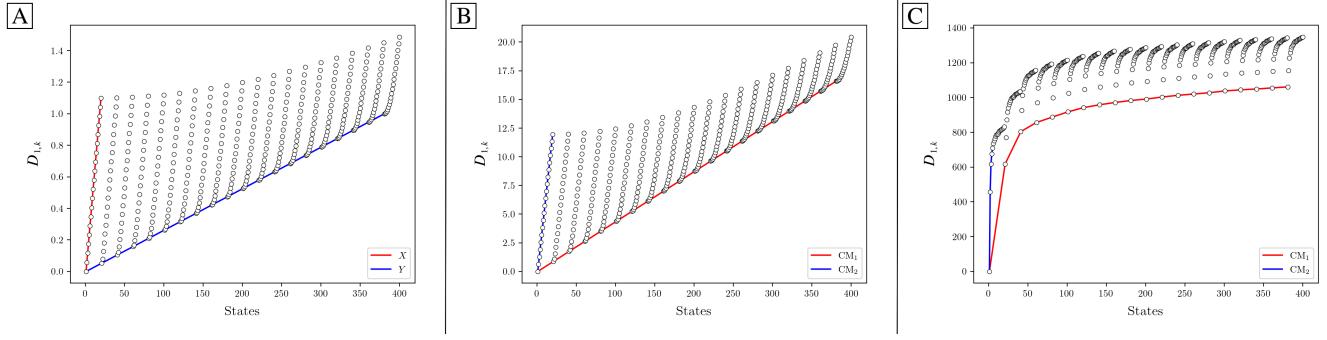


FIG. S24: The first row of the distance matrix \mathbf{D} is plotted for the rectangular Euclidean space [A], the 3D atomic-coordinate structures in SS₂ [B], and the EDMs in SS₂ [C]. Given our ordering of states, the first row $D_{1,k}$ corresponds to the pairwise distance calculated between state 01_01 and all 400 states. For [A], which was calculated for a rectangular domain $\Omega_R \in [0, 1.1] \times [0, 1] \subset \mathbb{R}^2$, 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 base 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 left-hand side boundary (with maximum distance $D\{(x_1, y_1), (x_{381} = 0, y_{381} = 1)\} = 1$). 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.

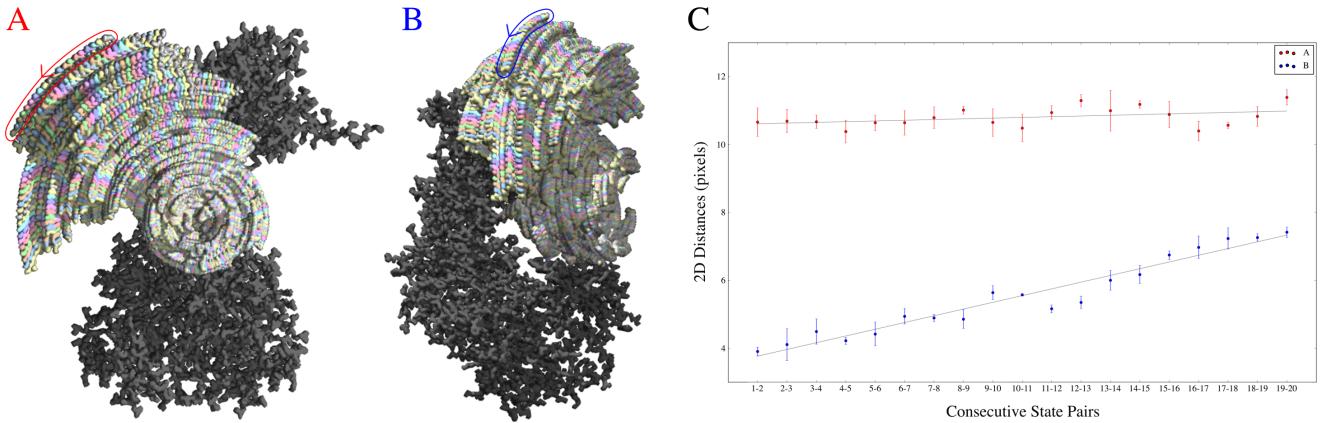


FIG. S25: Here we provide intuition for the emergence of foreshortened distances due to taking 2D projections of 3D EDMs. Two orthographic views of 3D models in the directions of two PDs are shown in [A] and [B], each composed of 20 overlaid 3D volumes from CM₂ (i.e., one degree of freedom). The 2D distances (in pixels) were measured between the peripheral ends of each consecutive states' rotated subunit (as seen in the red and blue encircled regions). After conducting three sets of 2D distance measurements on each region in image [A] and [B] independently, the mean distances were plotted with error bars representing standard deviation, followed by linear regression [C]. As can be seen, although the distances between states in the object's 3D form is constant, when projections are taken, these distances can strongly vary based on the current 2D view. While the Euclidean distance matrix calculated in the DM framework is less intuitive and records these changes on a pixel-by-pixel basis for the entire image, we anticipate analogous relationships to emerge there based on PD and CM.

to another depends on PD.

Thus, throughout these sections, we have demonstrated how the embeddings of the manifolds containing the same conformational information change depending on how the data is represented. Nevertheless, we were able to closely approximate the observed Ω_{PD} eigenfunctions when allowing for eigenbasis rotations of the form

$$\Psi_i = \cos(\theta)\cos(v\pi x) + \sin(\theta)\cos(w\pi y) = A\psi_v + B\psi_w$$

Using this expression, we generate graphs in Figure S26 which are able to analytically reproduce the heuristically-

derived subplots in Figure 3 and Figure 4. Apart from a few minor discrepancies, we observed an outstanding agreement between our analytical functions and the findings from our heuristic analysis. The discrepancies that do emerge can be understood as additional, small-scale perturbations which are currently unaccounted for in our general eigenfunction expression.

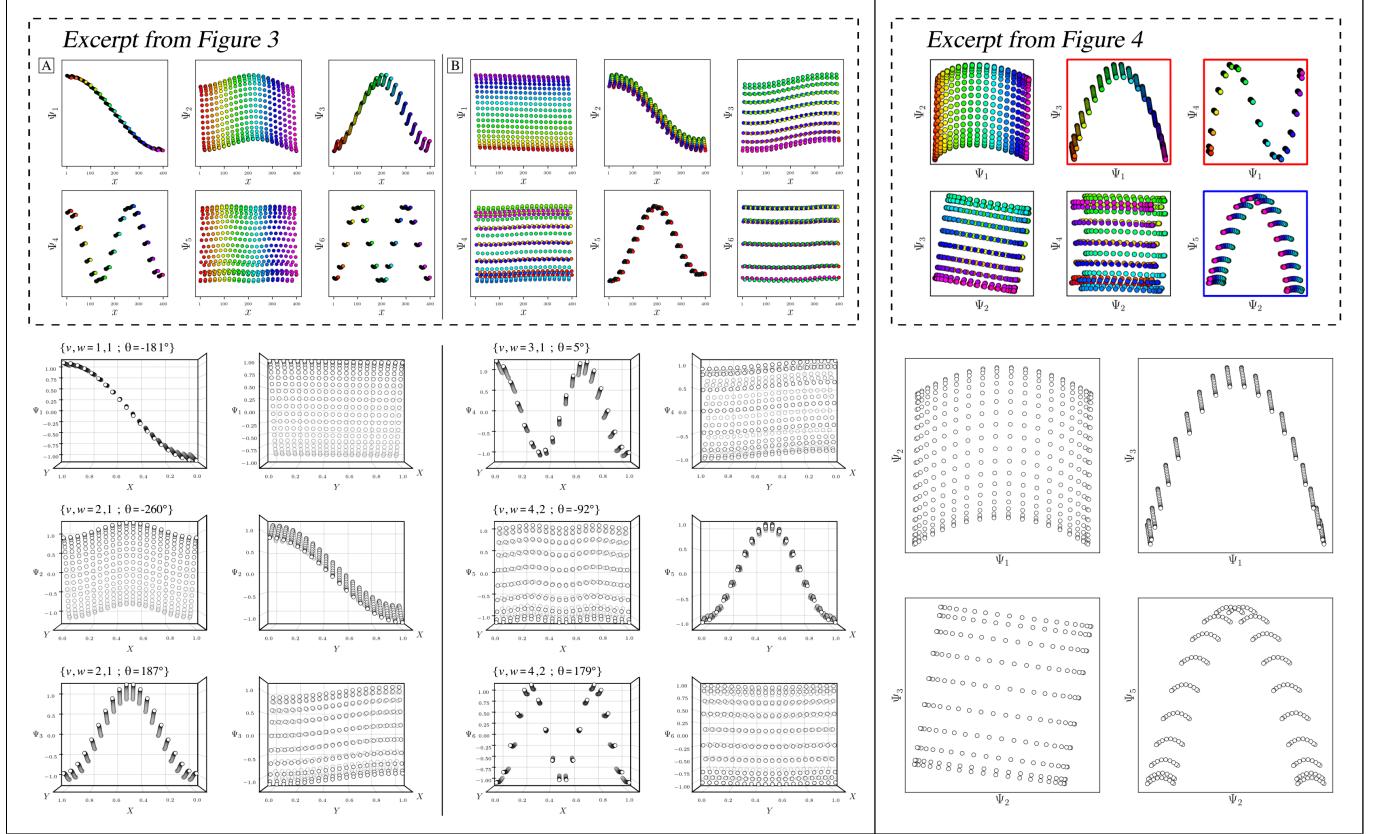


FIG. S26: Comparison of analytically-generated functions with the eigenfunctions previously obtained for PD₁ (Fig. 3 and Fig. 4). For each pair of subplots, values for θ were approximated by eye. Our approximations share a remarkable similarity with earlier heuristic results, and are able to account for geometric minutiae previously unaccounted for, as well as larger-scale rotations seen in the composite of eigenfunctions. Discrepancies can be seen in the slightly tilted appearance of parabolas in Ψ_3 and Ψ_5 of Fig. 3, as well as the clumping of points as observed in the CM₂ reference frame of Ψ_6 .

SM-XV. BOUNDARY CONDITIONS OF THE LBO

Here we lay out a strategy for dealing with molecular machines that exhibit each of their domain motions along an independent and mutually unrestricted sequence of quasi-continuous states. The set of all n -wise combinations of these bounded intervals (one for each conformational motion) produces an n -dimensional shape with a rectangular boundary. In section SM-XIV, we have shown that the corresponding Laplacian eigenfunctions are well defined. However, in general, analytically solving the Laplacian for any arbitrary boundary is impossible. Eigenfunctions can change drastically depending on the boundary, and are analytically only known for certain elementary shapes, such as rectangles, discs, ellipses and special triangles^{28,31}. On the other hand, geometric machine learning approaches can obtain solutions numerically, in principle for any boundary. However, such geometric machine learning methods still require the boundary to be known *a priori*. For systems with unknown boundaries, the problem is intractable.

As the set of all possible molecular machines is unfathomably complex, it is unlikely that one single algorithm could ever be so versatile as to anticipate every possible instance. Instead, we are interested in casting a wide enough net to capture the dynamics of a large portion of these systems, which we surmise operate within rectangular boundaries of an n -dimensional latent space of multi-body motions. 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 (Fig. S27-B). Importantly, our requirement of adequate coverage (as detailed in Discussion) excludes the case of obtaining poorly sampled data from a rectangular domain, which would allow any number of arbitrary shapes to emerge. This exclusion also holds for state spaces with “holes” (i.e., interior boundaries)²⁸, where the occurrence of certain states is forbidden due to energetic restraints. To better understand the effects of these boundary challenges, we have created a 2D state space with an octagonal domain (noting the Laplacian eigenfunctions of the octagon is an analytically unsolved problem), which was achieved by eliminating states at the four corners of our standard rectangular domain (Fig. S27-A). To circumvent the occurrence of eigenfunction misalignments due to PD disparity, which may complicate the interpretation of the boundary influence, we embed the 3D electron density maps instead of 2D projections.

The corresponding manifold embedding obtained from this octagonal state space is shown in Fig. S27-C, which features a number of deviations from the canonical rectangular eigenbasis (Fig. S23). Manually, we attempt to find a transformation from the octagonally-derived eigenbasis (Fig. S27-C) to the rectangular form (Fig. S23) by intuiting 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. S27-D).

Thus, it is not that the eigenfunctions are dramatically changed by the imposed boundaries, but, instead, that the eigenvectors can now contain multiple cosine terms. We note that both the indices and number of rotation operators required for this transformation deviated from our findings on eigenfunction realignment performed on rectangular state spaces (see section SM-XVI), with the collection of decisions required now more complex. Thus we believe this instance only further motivates the need for a future comprehensive method for estimating the preferred eigenbasis rotations. Given our own observations, a maximum-likelihood approach may be better suited for these demands, with such a study deserving of the scale delivered for other ManifoldEM subproblems, such as Belief Propagation³².

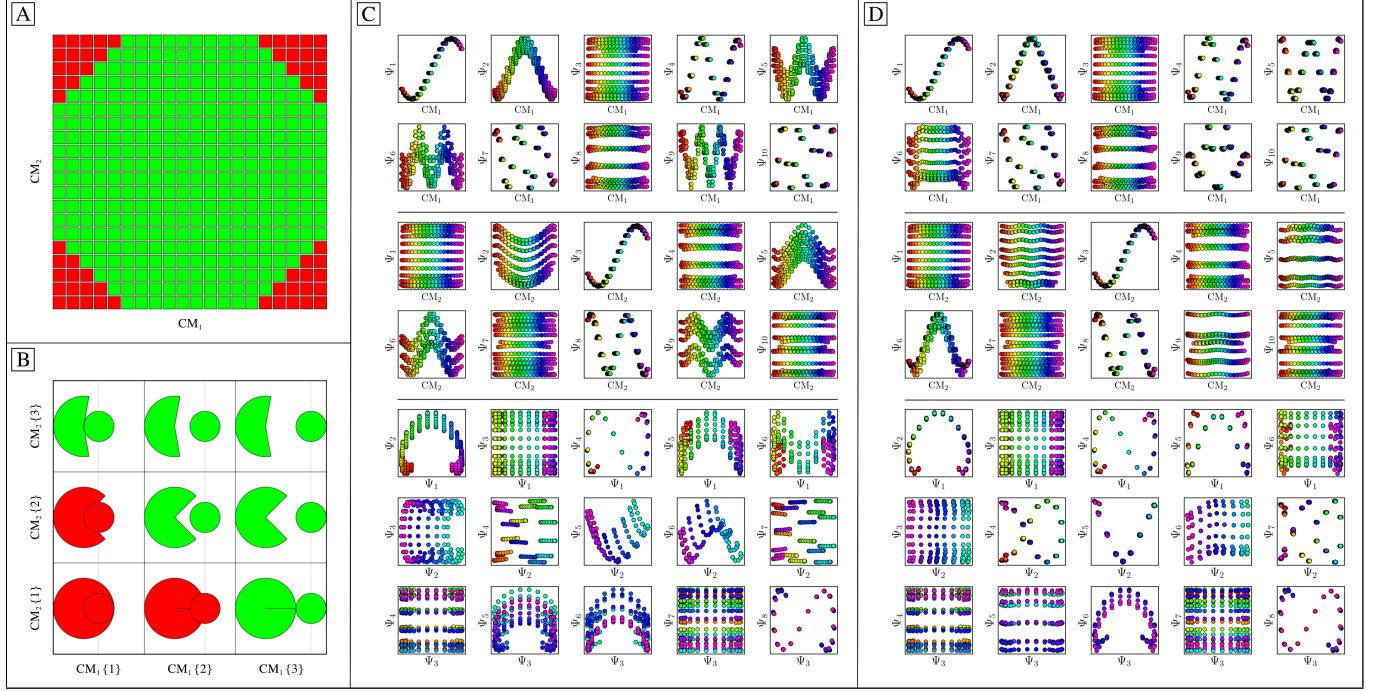


FIG. S27: Analysis of the eigenfunctions [C] associated with the octagonal state space [A] of EDMs. The initial 20×20 rectangular state space is displayed in [A], where red boxes illustrate states that were removed to form an octagonal grid. The schematic in [B] provides some context for the possibility of a non-rectangular state space, which can be envisioned as a top-down view of (1) a large domain that opens and closes, and (2) a small domain that translates left and right. Naturally, while the larger domain is in a closed or half-closed state, the smaller domain is impeded from accessing a subset of its possible states, and vice versa. The eigenbasis obtained after application of a set of high-dimensional rotations (of dimension $d = 15$) is shown in [D]. The required operators were estimated by hand, and included several large and small transformations: $\{R_{5,6}(40^\circ), R_{2,6}(-15^\circ), R_{2,5}(3^\circ), R_{2,9}(4^\circ), R_{6,9}(20^\circ), R_{6,12}(-25^\circ), R_{2,11}(-6^\circ), R_{9,11}(25^\circ), R_{9,15}(5^\circ), R_{6,11}(3^\circ)\}$. We note these calculations are for completeness, and not a minimal set for a desired subspace. Still, we were unable to perfectly decouple the $\{\Psi_3, \Psi_6\}$ subspace. As Ψ_3 appeared well behaved in the CM₂ reference frame, it is possible that either the Ψ_6 eigenfunction was fundamentally altered due to the new boundaries, or our choice of an earlier operator trapped us in a suboptimal solution.

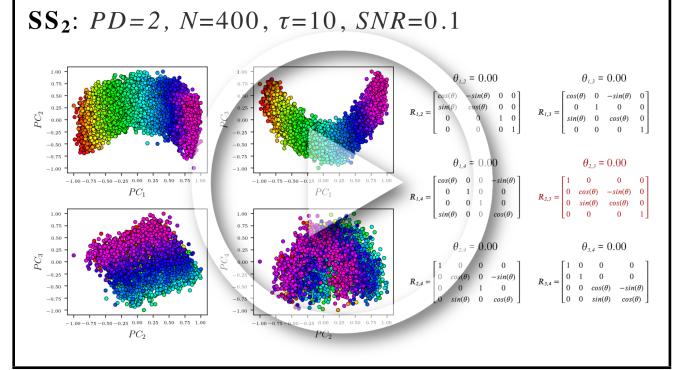
SM-XVI. EIGENFUNCTION REALIGNMENT

In this section, we demonstrate application of our eigenfunction-rotation algorithm and present the operators R_{ij} needed to counter-rotate each CM subspace for our final analysis. In addition, we provide here intuition for the aforementioned elimination procedure, which we apply to remove parabolic harmonics. These strategies are formulated by inductive reasoning based on patterns observed across hundreds of Ω_{PD} embeddings.

We start by identifying the eigenvector indices corresponding to the first CM subspace. Since a macromolecule's most prominent conformational signal as seen from the current viewing angle must always be the leading factor in the embedding of a PD manifold, we expect the leading CM subspace to appear in the set $\{\Psi_1, \Psi_i\}$, and use a least-square fitting strategy to discover the second eigenvector Ψ_A (see movie M7). Once this first parabola-housing subspace is identified, we can eliminate all subspaces $\{\Psi_A, \Psi_i\}$ as housing harmonic information corresponding to this leading CM at $\{\Psi_1, \Psi_A\}$. We can additionally eliminate $\{\Psi_B, \Psi_A\}$ in the next non-harmonic eigenvector row as a candidate for the location of the orthogonal CM parabola, which is easily understood via examination of movie M6. Subsequently, once the second CM subspace is identified by similar least-square measures (e.g., $\{\Psi_B, \Psi_C\}$ if $C \neq A$), the previous elimination procedure can be repeated to identify the location of this CM's corresponding harmonic spanning $\{\Psi_C, \Psi_i\}$, and so on, as is afforded by the geometric fidelity to the LBO (and thus ability for parabolic fitting) of these higher-order subspaces. As it is currently designed, this elimination procedure is applied only to ensure removal of the lowest-order harmonics¹⁹, such that the leading set of potential CMs discovered through this approach should always be unique.

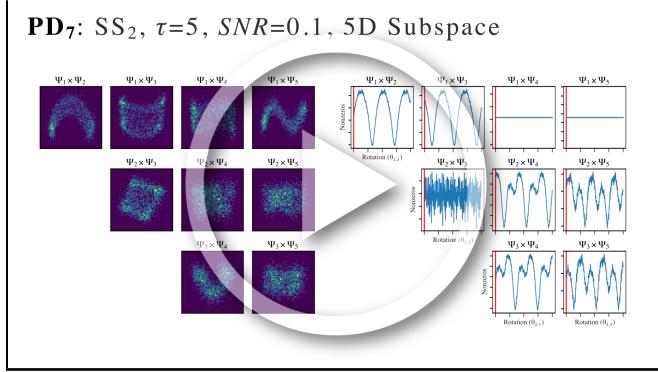
Eigenvector indices can be additionally used to narrow down the rotational operators R_{ij} required to adequately rotate each 2D subspace. Specifically, for two degrees of freedom, all pairwise combinations of the eigenvector indices corresponding to the first and second parabola-housing subspace (excluding parabolic harmonics) determine the set of required R_{ij} operators for $n = 2$. As a concrete example, we consider a 6-dimensional case where the CM_1 parabola is found in the $\{\Psi_1, \Psi_3\}$ subspace, the CM_2 parabola in the $\{\Psi_2, \Psi_5\}$ subspace, and the first CM_1 parabolic harmonic in the $\{\Psi_3, \Psi_6\}$ subspace. Then the 15 available rotation operators can be narrowed down to just the use of some combination of $\{R_{1,2}|R_{1,5}|R_{2,3}|R_{3,5}\}$, and applied in order of highest to lowest eigenvector significance while measuring minima via the aforementioned histogram routine. To note, while it is true that finite rotations about different axes do not commute³³, we found only minor deviations in the final orientation of each eigenbasis, given different permutations in the sequence of these four operators. Once the location of these parabola-housing subspaces are established, harmonics can alternatively be discarded if

duplicate R_{ij} indices emerge under this scheme. The emergence of duplicate indices can be understood in the current example when taking pairwise combinations of the leading parabola-housing subspace with the $\{\Psi_3, \Psi_6\}$ harmonic, which will incur the trivial $R_{3,3}$ operator.

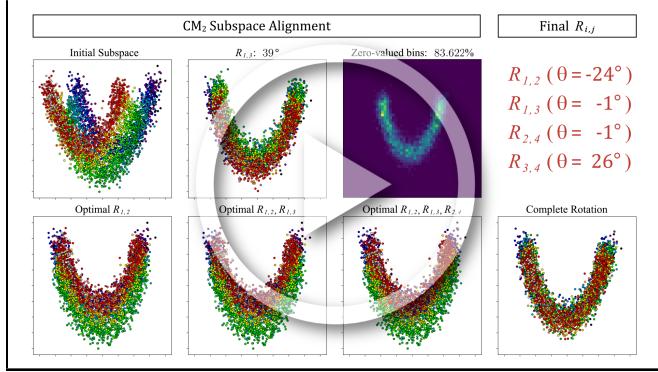


MOV. M5: Effect of applying a 4D orthogonal rotation to the 4D subspace (shown here using four projections of that subspace) obtained from PD_2 (in SS_2 with $\tau = 10$ and $\text{SNR} = 0.1$). The six rotation matrices required for the rotation of a 4D subspace are shown on the right, where the number of matrices for a d -dimensional subspace scales via $d(d - 1)/2$. (Note that the minimum dimensionality of the rotation matrix used must match the dimensionality of the subspace required to encompass all parabolic modes present in the embedding). As can be seen, by only applying rotation operator $R_{2,3}$ with 0.5 radians (28.65°), both CM_1 and CM_2 parabolic modes are corrected (preserving all distances between points) such that they reside completely in the plane of $\{\text{PC}_1, \text{PC}_3\}$ and $\{\text{PC}_2, \text{PC}_4\}$, respectively.
https://www.dropbox.com/s/c9b6vcjeffwziqp/M5_EigRot_4D.mp4?dl=0

As a final note for this section, there exists a rare occurrence that must be accounted for when the initial eigenbasis of the Ω_{PD} embedding is severely misaligned from the preferred coordinate system. In this event, we have observed that CM subspaces cannot be initially located by the presence of favorable R^2 values generated from least-squares fits. We encountered this problem for one manifold of the 504 analyzed (0.2%) across both data-type II (using three great circles) and data-type III (using one great circle). This error is easily recognized by the presence of a significantly low R^2 value (less than 0.1) for each 2D subspace of a given PD manifold. In such an event, we suggest rotating the eigenbasis by 45° using the $R_{1,2}$ operator and recalculating least-squares fits. As additional cases emerge, it is likely that a more comprehensive strategy may be required.



MOV. M6: A movie displaying the 2D histogram approach for finding optimal angles and corresponding parabolic modes. Specifically, the effect of an incremental 4D rotation operator $R_{2,3}$ on each 2D subspace is shown. During these rotations, each 2D subspace exhibits a unique profile which can be characterized by the number of nonzero bins in the corresponding 2D histogram as a function of angle. Note the appearance of specific patterns that emerge between these 2D subspaces as rotations are performed, which we leverage in our algorithm to procedurally eliminate subspaces from potential misuse. https://www.dropbox.com/s/ekh44n66x5j7sz5/M6_EigRot_Hist.mp4?dl=0



MOV. M7: An example PD from datatype II is chosen to demonstrate the inner workings of our eigenfunction realignment algorithm. Here, a $d = 5$ dimensional subspace is first isolated, with each 2D subspace therein assigned an R^2 -value based on least-square fits. Given presence of adequate fits, the parabola-housing subspace in each eigenvector row is determined via the best R^2 value, with the corresponding eigenvector indices used to procure the four rotational operators (of 10, for $d = 5$) required to align each point cloud with the plane of its subspace. We next demonstrate the generation of 2D histograms as these operators are exercised to determine the optimal angles, as previously detailed in Movie M6. As can be visually assessed, slight inaccuracies may emerge during the histogram optimization (typically no more than 5°), but prove insignificant for downstream procedures.

https://www.dropbox.com/s/u6qqrc6bmm5881t/M7_EigRot_Rij.mp4?dl=0

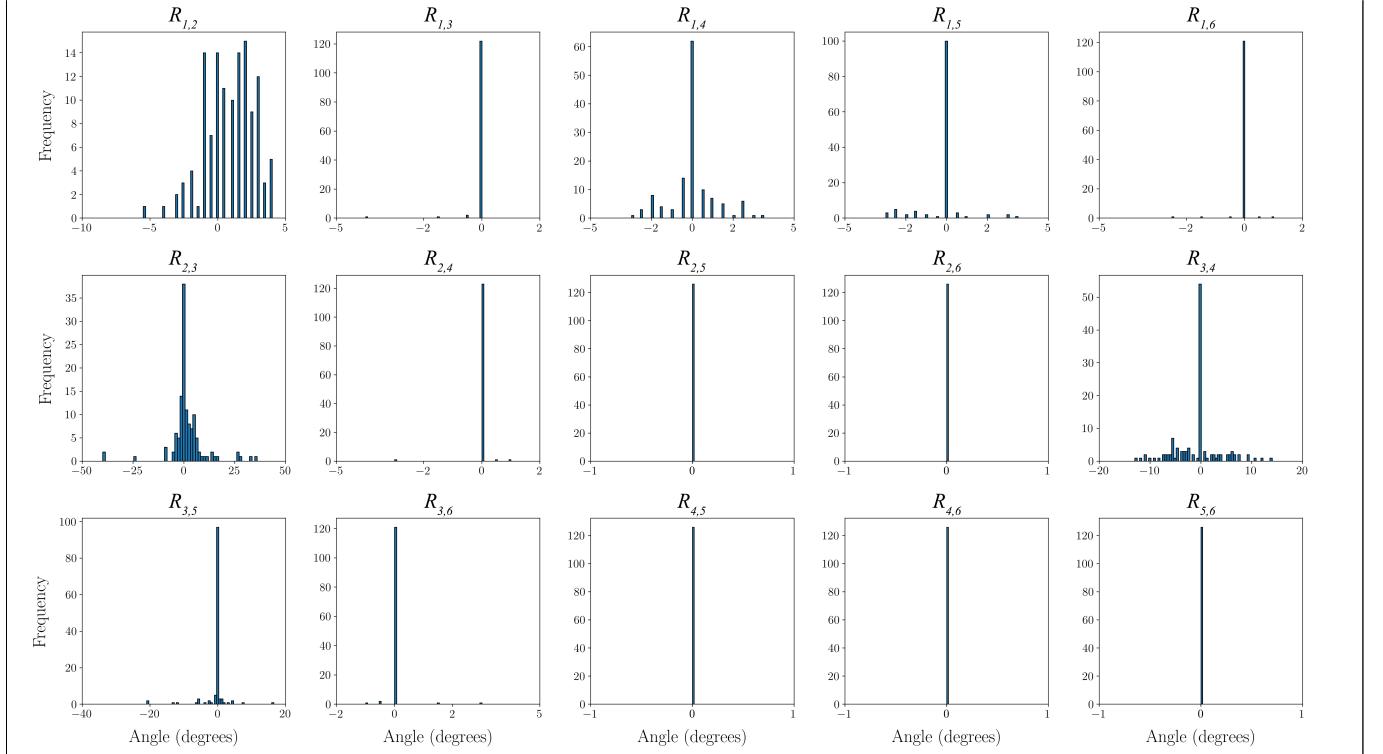


FIG. S28: Frequency of $R_{i,j}$ angles ($d = 6$) required to counter-rotate each CM_1 subspace within the 126 PDs along the S^2 trajectory spanning half of one great circle. A total of 15 subplots are displayed corresponding to the $d(d - 1)/2$ rotation operators necessary for this lower-dimension analysis. As can be seen, the magnitude of rotations required for a significant portion of PDs was substantial. To note, angular trends for each operator are highly subject to the choices of S^2 trajectory and CM.

SM-XVII. RECOMBINATION OF CONFORMATIONAL STATES

The following figure provides intuition for our method of generating multidimensional free-energy landscapes and corresponding 3D movies using the ESPER intersection of image-indices within each PD manifold. For this schematic, we have simplified the problem from that of accounting for a parabolic surface (as observed throughout this study) to a plane; however, the following intuition remains the same in either case.

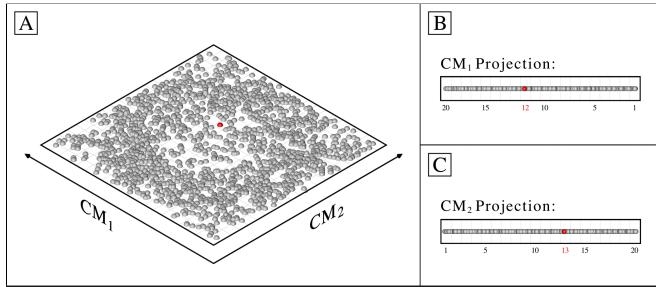


FIG. S29: A schematic to provide intuition for our intersection of image-indices approach, which simplifies the complex parabolic features in our observed PD embeddings into linear ones, on a plane. Let the red point near the center of the plane represent an image with image index p_i , such that p_i belongs to both $CM_1\{12\}$ and $CM_2\{13\}$, and thus $p_i \in CM_1\{12\} \cap CM_2\{13\}$. This point, along with all others in the intersection $CM_1\{12\} \cap CM_2\{13\}$, is used to define an occupancy and reconstruct a 3D density map for the respective state 12_13 .

Just like the plane in Fig. S29-A, PD images (represented by points) are organized along n orthogonal degrees of freedom (CMs) on a higher-dimensional hypersurface. For our needs, this hypersurface can be approximated as a parabolic surface. Given the aforementioned uncertainty and difficulty in identifying and mapping this surface directly, we can instead refer to its set of n orthogonal projections (e.g., Fig. S29-B and Fig. S29-C), which can be found and mapped with less difficulty. In the case of the plane—as in the case of our simplified illustration—these subspaces are 1D, while for Ω_{PD} embedding, a 2D subspace is required to adequately capture each parabolic component. Recall that we identify these subspaces after performing eigenvector rotations to align the parabolic surface, such that only the CM parabola is visible in each of the respective 2D subspaces. Once located, we straighten each CM trajectory in each of these lower-dimensional projections into a 1D trajectory, such that the parabola is transformed into rectilinear form. Next, we partition the points separately into β contiguous bins (here, $\beta = 20$), and collect the set of image indices falling into each bin. Note that the size of the bin effectively defines the precision to which we can locate each point on the plane, and determines the range of images falling within each state for our final outputs.

As a result, we are left with $n\beta$ sets of image indices

combined across each set of CM coordinates for each PD. For ease of explanation in the following notation, assume $n = 2$. Next, we construct an empty $\beta \times \beta$ (i.e., β^n) array \mathbf{P} and fill each element $P_{x,y}$ with the set of all image indices in the intersection $CM_1\{x\} \cap CM_2\{y\}$, where x and y are bins from CM_1 and CM_2 , respectively. Since manifolds from each PD were obtained independently, we must also correct for *sense* (the directionality of the CM coordinates) as we accumulate indices in \mathbf{P} . At the end of this procedure, we sum the total number of entries in each $P_{x,y}$ to form a $\beta \times \beta$ occupancy map (which can then be converted into a free-energy landscape via the Boltzmann relation). We additionally use the indices of images within each $P_{x,y}$ to reconstruct a 3D density map for the set of corresponding images; in this example, producing 400 3D density maps in total. Naturally, this construction can be easily extended to three or more degrees of freedom.

Thus, given only a set of CM subspaces—each a parabolic trajectory defining an orthogonal degree of freedom—and with knowledge of the higher-dimensional relationship between them (i.e., the parabolic surface, as determined throughout our heuristic analysis), we can reconstruct that joint geometrical relationship using only the intersection of image indices obtained in all pairwise combinations of bins from straightened CM coordinates. In effect, this procedure only requires that we collect the indices of images, without the need to integrate them into 1D occupancy maps. This is in contrast to the previous ManifoldEM methodology employing NLSA, which discards the original image indices. This action carries a price, and must be reversed by performing a lengthy tomographic reconstruction using the 1D occupancy maps to obtain the 2D distribution (for more information on the NLSA strategy, see the following section SM-XVIII).

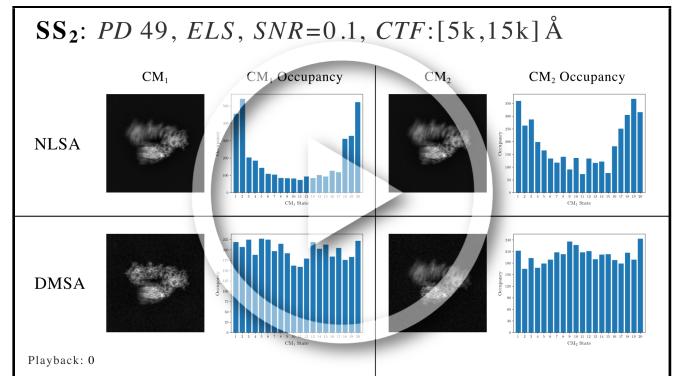
SM-XVIII. OVERVIEW OF NLSA

Nonlinear Laplacian spectral analysis³⁴ (NLSA) is primarily used for noise reduction during the extraction of the conformational signal contained in each PD. NLSA is applied independently on each member of a leading set of Ω_{PD} eigenvectors in order to assess the “meaning” of each in terms of housing potential CMs of interest. For each of these eigenvectors, NLSA is performed as follows. First, the raw images are concatenated along a chosen eigenvector to produce so-called “supervectors”³⁵. These supervectors are then embedded to form a new set of eigenvectors in a different space. This results to high accuracy in a 1-dimensional manifold, with known eigenfunctions $\cos(k\pi\tau)$ parameterized³⁵ by a conformational parameter τ (separate from the use of τ in our analysis). This enables the estimation of a density of points as a function of τ together with an ordered sequence of noise-reduced (i.e., interpolated, via the supervectors) 2D images. These 2D images can be arranged to form a 2D NLSA movie, designed to represent the conformational signal corresponding to the eigenvector chosen from the initially-embedded PD manifold. Once a set of 2D NLSA movies have been constructed along each of the leading Ω_{PD} eigenvectors independently, supervised identification of “meaningful” CM information is next required.

When only one degree of freedom is desired (or available), the 2D movies corresponding to the same CM content in different PD manifolds can be further compiled across S^2 to reconstruct 3D density maps and thus a 3D movie representing the CM. The NLSA procedure is more complicated when two (or more) degrees of freedom are desired. After supervised identification of two CMs, their respective eigenvectors for the current Ω_{PD} are used to isolate a 2D subspace therein. On this $\{\text{CM}_1, \text{CM}_2\}$ subspace, NLSA is performed independently along the directions of (typically) 180 uniformly-spaced radial lines in the range ($0 \leq \theta \leq \pi$). This yields a collection of point densities (i.e., 1D occupancy maps) $n(\tau, \theta)$ for each θ . The collection of these 1D maps for all θ constitutes the 2D Radon transform of a yet unknown 2D density map (i.e., the desired 2D occupancy map). An inverse Radon transform is then applied to reconstruct the 2D density map. In addition, NLSA also retrieves the noise-reduced images at each point in this map. To note, one of the rationales in the way NLSA-based retrieval of images is organized is that it normalizes the initially unknown rates of change in different CM directions³⁵. As in the 1D case, this procedure must next be performed for the eigenvector pairs corresponding to $\{\text{CM}_1, \text{CM}_2\}$ in all other embeddings of Ω_{PD} , from which noise-reduced 3D density maps can be reconstructed to form 3D movies of concerted conformational motions.

SM-XIX. COMPARISON OF RESULTS FROM ESPER AND NLSA

Here we compare the outputs of ESPER and NLSA for a few example PDs from our final analysis dataset (126 PDs, SNR, CTF and nonuniform occupancy map, as previously detailed), with these PDs selected based on the properties of both their visual appearance and respective embedded geometries. Importantly, the same preliminary steps were performed for both frameworks (i.e., generation of identical manifolds and corresponding embeddings) before a branch in workflows. Also recall that ESPER includes the use of our eigenfunction re-alignment methodology on these manifold embeddings, whereas ManifoldEM with NLSA currently does not.



MOV. M8: For our comparison of ESPER and NLSA, three example PDs were chosen from our final analysis dataset for direct comparison of their 2D movies and corresponding occupancy maps. In the first segment of the movie M8, PD₂ represents the class of extremely well-behaved manifolds: near-perfect eigenfunction pre-alignment and irrelevant inward curling of either CM₁ or CM₂ subspaces. Next, PD₃₃ is a representative from the class of manifolds with appreciably unaligned eigenfunctions from the ideal eigenbasis, with the subspace of CM₂ here requiring a larger counter-rotation than CM₁. Finally, PD₄₉ required a minimal d -dimensional rotation (much like PD₂), but exhibited significant inward curling at the boundaries of its subspaces. The last segment of movie M8 demonstrates the final 2D movies (one per eigenfunction) output by NLSA for PD₃₃ and PD₄₉. https://www.dropbox.com/s/qvpwsyqq6zc95qg/M8_Comparison_NLSA.mp4?dl=0

As can be seen in movie M8, for the most well-behaved PD manifold, PD₂, there is general visual agreement in the 2D movies obtained from NLSA and ESPER for CM₁ and CM₂. However, there are noticeable discrepancies in the outputs between the results from these two techniques. Immediately apparent is the difference in quality of the domains under motion corresponding to the given CM. For ESPER, these domains are highly resolved across all frames produced, while for NLSA these regions are much less resolved and noticeably smeared out. Second, while the visual differences between frames of ESPER appear to evolve at an even pace, differences

in frames appear less emphasized near the beginning and end of the NLSA movies, as if the CM movies were decelerating near these regions. In addition, the NLSA occupancies share little resemblance to our ground truth, with accentuated errors near the boundaries. Similar boundary problems exist but are significantly less present for ESPER occupancy maps, as each map shows reasonable agreement with ground truth (i.e., bimodal for CM_1 and unimodal for CM_2 , as expected via Fig. S13).

The NLSA outputs for CM_1 in PD_{33} follow similar trends to those described for PD_2 , with the exception that the overall range of motion for this CM is noticeably reduced compared to outputs from ESPER. For CM_2 , matters are much worse. While our procedure using ESPER correctly charted a rotated, properly-aligned set of eigenfunctions, ManifoldEM employing NLSA used the existing manifold embedding without applying the essential eigenfunction realignment. As a result, the 2D movie produced by NLSA having closest resemblance to CM_2 (i.e., Ψ_4) demonstrated a physically-impossible sequence of motions: the splitting of the CM_2 domain into two separate domains. At the end of movie M8, the NLSA 2D movies obtained for the leading four eigenfunctions are shown for comparison. Here, both (1) a physically-impossible splitting of the CM_1 domain, and (2) a subdued CM_2 motion can be seen in the 2D movies obtained for both Ψ_2 and Ψ_3 .

In the third scene of movie M8, the NLSA outputs of PD_{49} can be described most similarly to those obtained in PD_2 , and aside from flaws previously listed, are in general visual agreement with ESPER outputs. Again, the last segment of movie M8 showcases the alternative NLSA outputs obtained from Ψ_3 and Ψ_4 , for which physically-impossible conformational information is apparent corresponding to CM_1 and CM_2 domains, respectively. Note that while NLSA and ESPER have been provided the exact same data—even up to generation of identical manifold embeddings—only ESPER is able to fully leverage the geometric structure present to consistently recapitulate ground-truth conformational motions and occupancies from a variety of PD manifolds. Further, while ESPER offers strategies to procedurally avoid introduction of nonsensical contextual output, NLSA can generate 2D movies with a wide range of defects¹⁸, with each having the potential of appearing as a likely CM candidate to the naive eye.

Finally, we note the total computation time for performing these two techniques on the same CM-eigenvector (Ψ_1) from the same PD manifold (PD_2), with final output a single 2D movie (as seen in movie M8). The ESPER approach required approximately one minute for finding the optimal d -dimensional rotations and CM subspaces for this Ω_{PD} , followed by approximately two minutes to perform subspace partitioning on the leading CM parabola. Meanwhile, the NLSA approach, which does not use our eigenfunction-rotation technique, took 4 hours and 37 minutes to furnish a 2D movie. The total computation time for NLSA was thus

over 90 times longer than ESPER, with both methods having been run using a single-processor on the same workstation (3.8 GHz 8-Core Intel Core i7; 8 GB 2667 MHz DDR4). Although not investigated further for this study, we additionally note that in the current ManifoldEM framework¹⁸, it is required that this time-expensive NLSA computation is repeated in its entirety for every Ω_{PD} eigenvector chosen during final compilation of the free-energy landscape. Meanwhile, using our intersection of image-indices approach, the ESPER algorithm compiles all PD-eigenfunction content and generates free-energy landscapes within minutes. Recall that these high computational demands were rationalized for NLSA as a way to handle the unknown manifold structures, noise-reduce images and normalize unknown rates of change. Meanwhile, as our heuristic analysis directly informed us of anticipated manifold characteristics and spectral structure, we were able to design ESPER to circumvent these previous unknowns, and perform the necessary operations required to accurately retrieve high-resolution images and corresponding occupancy map of all CM states. The results of our analysis show that ESPER produced appreciably more accurate outputs than the previous technique in a fraction of the time.

SM-XX. STRUCTURAL VALIDATION OF ESPER OUTPUTS

In the following we provide a validation for the fidelity of the ESPER outputs as they relate to the ground-truth electron density maps obtained from PDB-formatted atomic-coordinate structures. (Recall that, following our synthetic-generation protocol, these EDMs are then projected to form images—and these images then duplicated via multiple occupancy assignments and modified by CTF and noise—before being processed via the ESPER framework). Specifically for our validation, we generated a Fourier Shell Correlation (FSC) curve⁷ for the ground-truth simulated map against the corresponding ESPER recovered map. As seen in Fig. S30 for state 05_10, we found favorable global agreement between maps up to a resolution near 3 Å (i.e., the value used to generate each ground-truth EDM).

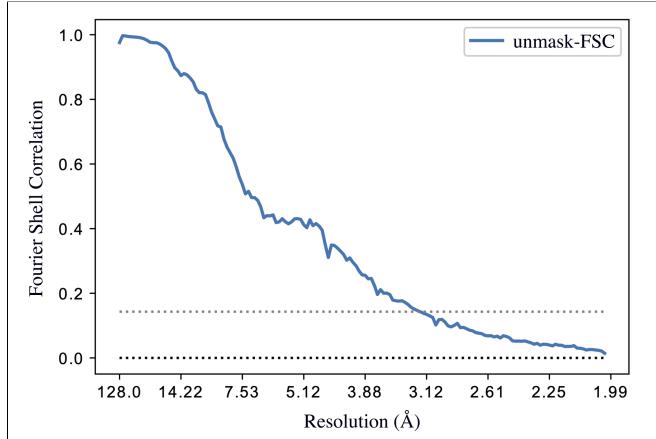


FIG. S30: FSC curve comparing the state 05_10 input (ground-truth) and output (ESPER) maps. Specifically, the FSC measures the normalized cross-correlation coefficient between the two maps as measured over a series of shells in Fourier space. As one proceeds along the *x*-axis from the left (representing the center of the FT) to the right, increasingly larger shells are compared in Fourier space, such that the largest shells (far right) correspond to the highest resolution features. The FSC curve thus provides a global measure of how well one 3D density map matches the other.

For molecules that exhibit domain motions, the global resolution is no longer a good indicator of how well-resolved these regions are in the reconstructed EDM. Thus, a complementary local validation was done by calculating the Q-scores³⁶ for the same example EDM (state 05_10) output from ESPER; with the use of Q-scores serving as a local indicator of how well our final 3D density maps recover the ground-truth atomic information. The results of this analysis are shown in Fig. S31.

For these statistics, the mean Q-score for the (1) backbone, (2) side chains, and (3) residues was 0.784, 0.768 and 0.778, with standard deviation 0.042, 0.042 and 0.039, respectively. Almost all Q-scores obtained were well above the expected Q-score value (0.5862), which is

calculated³⁶ based on correlations to the reported resolutions of maps in the EMDB. On average, these Q-scores were approximately 1.3 times that of the expected value. We note that exceeding these expectations is anticipated, since our study was initiated with pristine structural information (i.e., atomic models). Outliers were found predominantly in the periphery of the domain corresponding to CM₂ and included side chains (minimum Q-score: 0.401) and three residues (minimum Q-score: 0.377). While the FSC provides an indicator for the global agreement between two maps, we believe the Q-score is more appropriate for interpretation of our outputs, especially given the design of our study; i.e., a synthetic-generation protocol from ground-truth atomic coordinates. Similarly favorable structural fidelity was found for the other outputs of ESPER in relation to their corresponding ground-truth atomic representations.

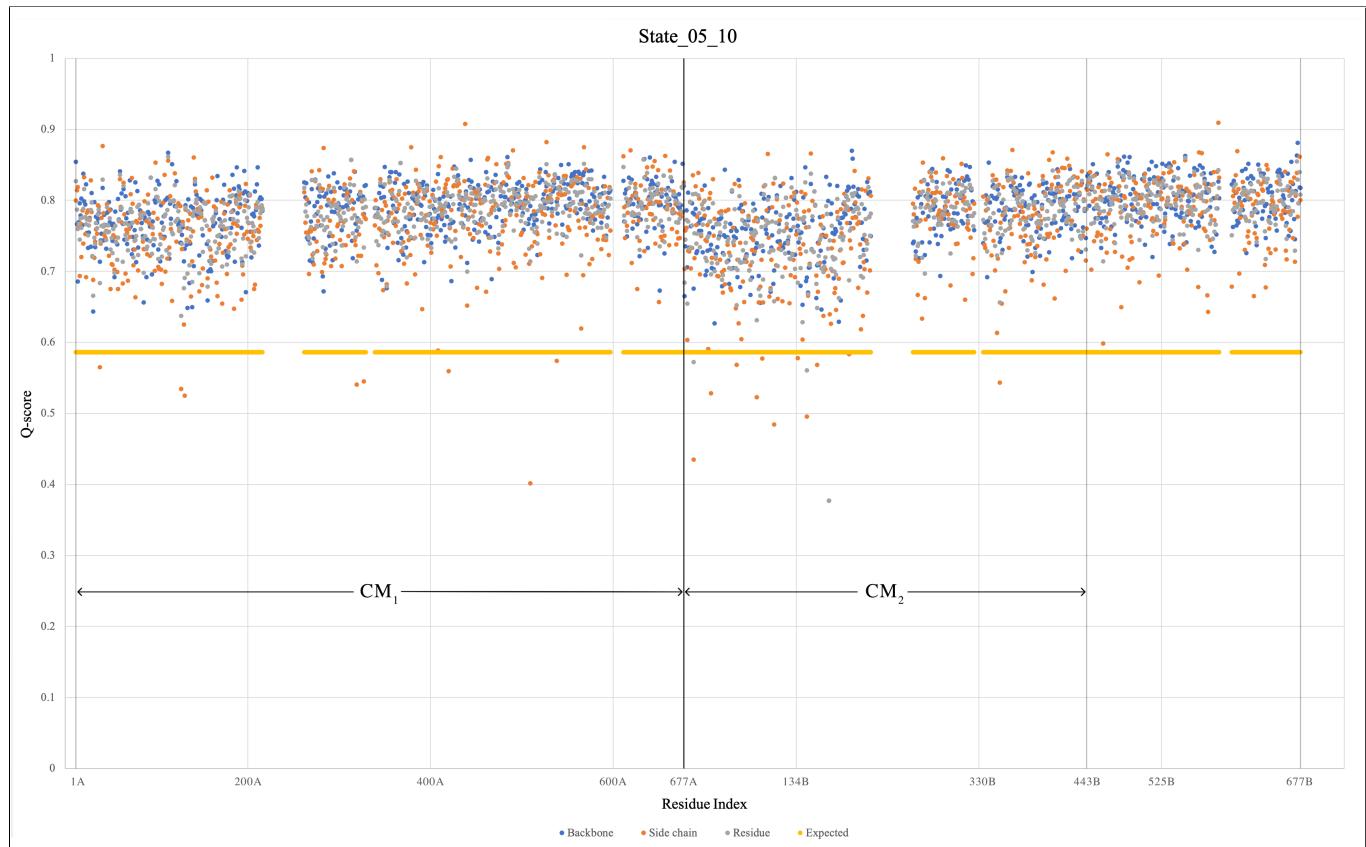


FIG. S31: Q-scores for protein backbone, side chains and residues calculated using ESPER output map 05_10 with corresponding ground-truth atomic-coordinate structure. Q-scores were ascertained using the MapQ³⁷ plugin for Chimera⁶. As can be seen, the range of residues corresponding to each conformational motion (CM₁ and CM₂) are demarcated on chain A and B, respectively. Empty Q-scores correspond to those residue indices missing in the initial crystal structure (PDB 2CG9), whether due to insufficient resolution or electron density in the preceding study. The expected Q-score represents the average Q-score at a resolution of 3 Å, and is calculated via MapQ based on the reported resolutions of 3D density maps in the Electron Microscopy Data Bank (EMDB).

SM REFERENCES

- ¹E. Seitz, F. Acosta-Reyes, P. Schwander, and J. Frank, "Simulation of cryo-em ensembles from atomic models of molecules exhibiting continuous conformations," bioRxiv (2019), 10.1101/864116.
- ²E. Seitz, "Repository: Cryo-em synthetic continua (initial release)," Zenodo , 10.5281/zenodo.3561105 (2019).
- ³J. Giraldo-Barreto, S. Ortiz, E. H. Thiede, K. Palacio-Rodriguez, B. Carpenter, A. H. Barnett, and P. Cossio, "A bayesian approach for extracting free energy profiles from cryo-electron microscopy experiments," Scientific Reports **11** (2021).
- ⁴H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide protein data bank," Nature Structural Biology (2003).
- ⁵L. Schrödinger, "The pymol molecular graphics system," (2015).
- ⁶E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—a visualization system for exploratory research and analysis," J Comput Chem **25**, 1605–1612 (2004).
- ⁷R. Warshamanager and G. Murshudov, "Mapq repository," <https://www2.mrc-lmb.cam.ac.uk/groups/murshudov/content/emda/emda.html> (2019).
- ⁸J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State* (Oxford University Press, Oxford, New York, 2006).
- ⁹G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, "EMAN2: an extensible image processing suite for electron microscopy," J Struct Biol **157**, 38–46 (2007).
- ¹⁰R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph Laplacian tomography from unknown random projections," IEEE Trans Image Process **17**, 1891–1899 (2008).
- ¹¹W. T. Baxter, R. A. Grassucci, H. Gao, and J. Frank, "Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules," J Struct Biol **166**, 126–132 (2009).
- ¹²N. Fischer, A. L. Konevga, W. Wintermeyer, M. V. Rodnina, and H. Stark, "Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy," Nature **466**, 329–333 (2010).
- ¹³S. H. Scheres, "RELION: implementation of a Bayesian approach to cryo-EM structure determination," J Struct Biol **180**, 519–530 (2012).
- ¹⁴Z. Li, Q. Ding, and W. Zhang, "A comparative study of different distances for similarity estimation," in *Intelligent Computing and Information Science*, edited by R. Chen (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011) pp. 483–488.
- ¹⁵A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, "Systematic determination of order parameters for chain dynamics using diffusion maps," Proc Natl Acad Sci U S A **107**, 13597–13602 (2010).
- ¹⁶R. R. Coifman and S. Lafon, "Diffusion maps," Applied and Computational Harmonic Analysis **21**, 5–30 (2006), special Issue: Diffusion Maps and Wavelets.
- ¹⁷P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," Physica D: Nonlinear Phenomena **9**, 189–208 (1983).
- ¹⁸G. Mashayekhi, "Manifoldem matlab repository," https://github.com/GMashayekhi/ManifoldEM_Matlab (2020).
- ¹⁹E. Seitz, "ManifoldEM: ESPER," Zenodo , 10.5281/zenodo.3884160 (2021).
- ²⁰M. Hein and J.-Y. Audibert, "Intrinsic dimensionality estimation of submanifolds in rd," in *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05 (Association for Computing Machinery, New York, NY, USA, 2005) p. 289–296.
- ²¹F. Camastrà and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," Information Sciences **328**, 26–41 (2016).
- ²²D. Lay, S. Lay, and J. McDonald, *Linear algebra and its applications*, 5th ed. (Pearson, 2016).
- ²³V. Iskovskikh, "Quadric," (1994).
- ²⁴J. R. Munkres, *Topology*, 2nd ed. (Prentice Hall, Incorporated, 2000).
- ²⁵J. Jost, *Geometry and Physics* (Springer-Verlag Berlin Heidelberg, 2009).
- ²⁶H. Donnelly, "Eigenfunctions of the laplacian on compact riemannian manifolds," Asian Journal of Mathematics **10**, 115 – 126 (2006).
- ²⁷P. Buser, *Geometry and Spectra of Compact Riemann Surfaces* (Springer, 1992).
- ²⁸D. S. Grebenkov and B.-T. Nguyen, "Geometrical structure of laplacian eigenfunctions," SIAM Review **55**, 601–667 (2013), <https://doi.org/10.1137/120880173>.
- ²⁹D. Giannakis, P. Schwander, and A. Ourmazd, "The symmetries of image formation by scattering. i. theoretical framework," Opt. Express **20**, 12799–12826 (2012).
- ³⁰R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics* (Basic Books, New York, NY, 2010) Chap. 12, originally published 1963–1965.
- ³¹B. McCartin, "On Polygonal Domains with Trigonometric Eigenfunctions of the Laplacian under Dirichlet or Neumann Boundary Conditions," Applied Mathematical Sciences **2**, 2891–2901 (2008).
- ³²S. Maji, H. Liao, A. Dashti, G. Mashayekhi, A. Ourmazd, and J. Frank, "Propagation of Conformational Coordinates Across Angular Space in Mapping the Continuum of States from Cryo-EM Data by Manifold Embedding," J Chem Inf Model **60**, 2484–2491 (2020).
- ³³J. Townsend, *A modern approach to quantum mechanics*, 6th ed. (University Science Books, 2000).
- ³⁴D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability," Proc Natl Acad Sci U S A **109**, 2222–2227 (2012).
- ³⁵A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseiniadeh, H. Y. Liao, J. Pallesen, G. Sharma, V. A. Stupina, A. E. Simon, J. D. Dimman, J. Frank, and A. Ourmazd, "Trajectories of the ribosome as a brownian nanomachine," Proceedings of the National Academy of Sciences **111**, 17492–17497 (2014).
- ³⁶G. Pintilie, K. Zhang, Z. Su, S. Li, M. F. Schmid, and W. Chiu, "Measurement of atom resolvability in cryo-EM maps with Q-scores," Nat Methods **17**, 328–334 (2020).
- ³⁷G. Pintilie, "Mapq repository," <https://github.com/gregdp/mapq> (2021).
- ³⁸D. J. Griffiths, *Introduction to electrodynamics*; 4th ed. (Pearson, Boston, MA, 2013) Chap. 4.