Recovery of conformational continuum from single-particle cryo-EM data: Optimization of ManifoldEM informed by ground-truth studies

Evan Seitz,[1] Peter Schwander,[2] Francisco Acosta-Reyes,[3] Suvrajit Maji,[3] and Joachim Frank[1,3]
[1)] *Columbia University, Department of Biological Sciences, New York, NY 10032, USA*
[2)] *University of Wisconsin-Milwaukee, Department of Physics, 3135 N. Maryland Ave, Milwaukee, WI 53211, USA*
[3)] *Columbia University Medical Center, Department of Biochemistry and Molecular Biophysics, New York, NY 10032*

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

This work is based on the manifold-embedding approach to the study of biological molecules exhibiting conformational changes in a continuum. Previous studies established a workflow capable of reconstructing atomic-level structures in the conformational continuum from cryo-EM images so as to reveal the latent space of macromolecules undergoing multiple degrees of freedom. Here, we introduce a new approach that is informed by detailed heuristic analysis of manifolds formed by simulated heterogeneous cryo-EM datasets. These simulated models were generated with increasing complexity to account for multiple motions, state occupancies and CTF 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 the quasi-continuum of conformational states in the form of an energy landscape and respective 3D maps for all states therein. This framework offers substantial enhancements relative to previous work, for which a direct comparison of outputs has been provided.

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 cycles[1]. 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, a path is singled out on this energy landscape along which the machine performs its metabolic function[2].

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 machine undergoing this process. When the number of snapshots is sufficiently large – at least 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, up to a constant, into the corresponding 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 machines and other biological assemblies is of unparalleled importance in modern structural biology.

The way to utilize the data from a cryo-EM experiment is not easy, however. Ideally, we would wish to compare 3D structures, but only 2D projections are accessible experimentally. When each 2D projection is assigned a set of angles to define its viewing direction on the 2-sphere (S^2), a set of projections in close proximity on this space can be assigned to a unique *projection direction* (PD)[†]. For any 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 projection, 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, which has an intrinsic dimension n equal to the number of independent molecular degrees of freedom, is called a *manifold* (for expediency, we will use the same term for the instantiation of a manifold with a given data set). By representing the high-dimensional manifold from experimental data of a molecular machine in a low-dimensional vector subspace, we create the foundation for the analysis of the molecule's free-energy landscape. 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 PDs, with the subset of images within each PD analyzed via manifold embedding, and the resulting representations combined into a consolidated conformational

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

occupancy map. 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 ribosome[1], ryanodine receptor[10] and SARS-CoV-2 spike protein[12] - 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 engineering[13], 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 a manifold in terms of its eigenvalues and eigenvectors gives valuable insights into its intrinsic structure - with these components known to characterize physical properties such as the frequencies and modes of vibration, or energy levels and electron wave functions of atoms[14]. 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 Analysis[23] (PCA) and Diffusion Maps[24,25] (DM), respectively, with each allowing an analysis formed in $\mathbb{R}^N$ (for any PD, typically $N \ll P$). Some techniques are not so easily classified, however, such as the method of Laplacian spectral volumes[4], which relies on both linear and nonlinear dimensionality reduction. These methods can also be divided based on their type of data input – generating manifolds from either 2D projections straight from a cryo-EM experiment (i.e., the ManifoldEM approach), or 3D maps already reconstructed from those projections[4,26–29]. Regardless of the approach, since the intrinsic structure of such manifolds is unknown, these competing dimensionality-reduction methods cannot be immediately validated or informatively compared using experimental data alone. Instead, the manifolds obtained from biologically-relevant systems must first be carefully evaluated 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 controlled movable domains, each having undergone a series of independent *conformational motions* (CMs). The number of CMs constructed for each state space defines its intrinsic dimensionality. We then determine how these quasi-continuous motions are reflected in low-dimensional representations of manifolds obtained by either linear or nonlinear dimensionality reduction techniques. For either construction, we demonstrate that the point cloud representing the PD-manifold is organized with remarkable closeness to the analytical eigenfunctions of the homogeneous Laplace-Beltrami operator[30] (LBO). We next describe how to interpret this complex 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 these techniques under realistic experimental conditions, while exposing any intrinsic uncertainties that may arise. Recently, several issues and limitations have been documented[11] in the ManifoldEM framework, further amplifying the motivation of our current pursuits.

For our second endeavor, we introduce a novel methodology (which we will term Direct Manifold Subspace Analysis, or DMSA) for extraction of conformational information as obtained from specific subspaces of high-dimensional manifolds, to generate an occupancy map and the energy landscape of a molecular machine, and to reconstruct 3D movies depicting its function. Whereas the previous approach[1,10,11] reconstructs images via Nonlinear Laplacian Spectral Analysis[31] (NLSA) in a space spanned by one or more CMs, DMSA 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 manifold transformation that we deem essential for 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, with its advancements further enabling the use of efficient strategies for generating multidimensional energy landscapes not accessible via the founding ManifoldEM framework. Ultimately, we will show that, given certain requirements are met in the quality and structure of a dataset, DMSA offers an alternative strategy with many benefits exceeding 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 Coulomb potential maps arising from a quasi-continuum of atomic structures[32,33]. 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 Bank[34] (PDB). Using 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-shape[35]. *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 crystallography[36].

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 CM_1, CM_2 and CM_3. Each CM was designed to cover a unique range of motions, with the cascade of overlaid states making up CM_1 occupying the largest spatial region, followed in magnitude by CM_2 and then by CM_3. Using combinations of these CMs, three synthetic state spaces were generated with intrinsic dimensionalities of $n = 1, 2, 3$. Specifically, 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.

In the following analysis, these state spaces are termed SS_1, SS_2, and SS_3, and defined by: (1) 20 states exhibiting one degree of freedom (CM_1); (2) 400 states (20×20) with two degrees of freedom (CM_1, CM_2); and (3) 1000 states ($10 \times 10 \times 10$) with three degrees of freedom (CM_1, CM_2, CM_3), respectively. As a specific example of the ranges of motion present in SS_3, the Root-Mean-Square Deviation (RMSD) was calculated[37] between neighboring states in each CM, yielding the values of 1.8 Å, 1.3 Å and 0.3 Å along CM_1, CM_2 and CM_3, 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 section "Simulation of cryo-EM ensembles from atomic models". 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 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 supplementary materials. We first utilize the pristine *data-type I*, which is given no experimental artifacts or occupancy assignments. Within this construction, we examine the manifolds for five example PDs 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 quantify the trajectories of our simulated conformational changes in each manifold in analytical form.

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 the geometric structure of the manifolds, and to quantify the robustness of this structure 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 (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 applying the PD-manifold approach on each of these three data-types, we finally introduce an overview of the DMSA 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 an energy landscape (*final analysis*). All Python scripts for reproducing this workflow, including extensive documentation therein, have been made available online[33,38].

RESULTS

I. Diffusion Maps

A. Data-type I in State Space 1

For its illustrative qualities, we first analyze the manifolds 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[24,25,39], we summarize these procedures here while emphasizing the most important characteristics in supplementary materials (see section "Overview of Diffusion Maps", where we define such parameters as the Gaussian Bandwidth, ε). Using the DM framework, we ultimately generate a different manifold for each of the five PDs, with each of these 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 PD showed

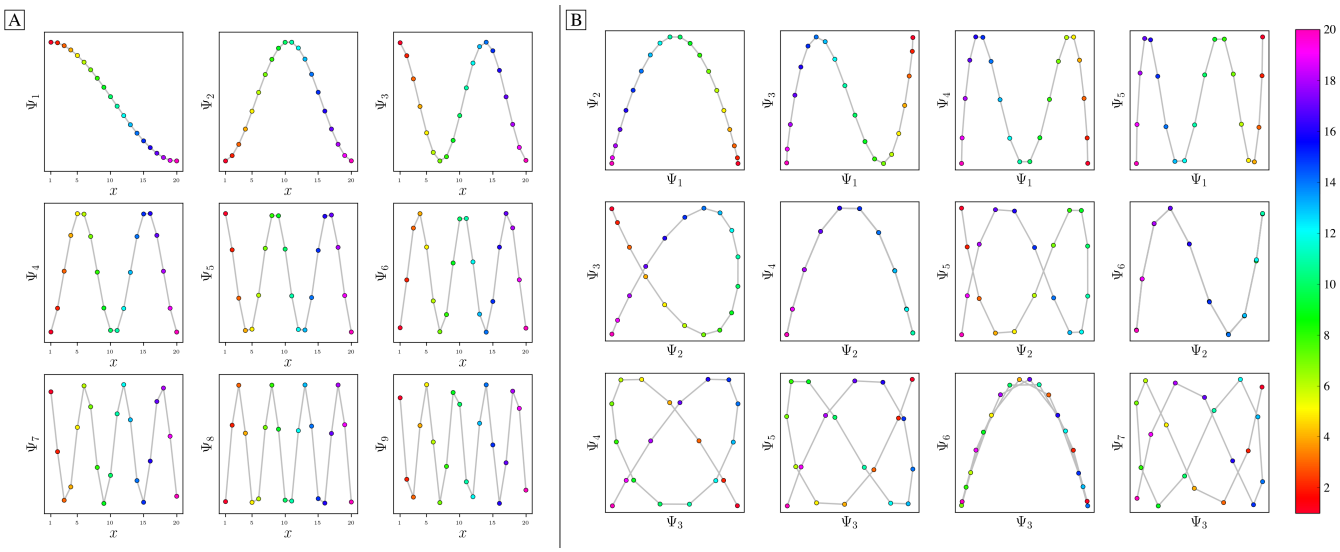

FIG. 1: Analysis of eigenfunctions for PD_1 in SS_1 (i.e., 20 states total making up one degree of freedom). On the left [A] are the sinusoidal forms $\{\cos(k\pi x) \mid k \in \mathbb{Z}^+ \leq N\}$ of each eigenvector Ψ_k that emerge when points (corresponding to images) in each Ψ_k are ordered precisely in the sequence in which their ground-truth images were constructed. Regardless of any knowledge of such a sequence, these eigenfunctions will always form sensical composites (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.

a staggered falloff in the significance of leading eigenvectors which decayed slowly to zero. Projecting the resultant manifold 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 manifold onto its first two eigenvectors $\{\Psi_1, \Psi_2\}$ - forming a 2D subspace - the conformational information existed clearly on a parabolic trajectory (as shown in the first subplot of Fig. 1-B), confirmed by the proper ordering of indices of points along the curve.

Following this analysis, we next proceeded to investigate all other unique combinations of 2D projections of each manifold. Mathematically, each such mapping to a 2D vector subspace is the restriction to the N-dimensional manifold of the projection of $\mathbb{R}^N$ onto $\mathbb{R}^2$; given by $\{\Psi_1, \Psi_2, ..., \Psi_N\} \mapsto \{\Psi_i, \Psi_j\}$, where $i < j$. As seen in Fig. 1-B, a subset of the canonical Lissajous curves[40] emerged across the 2D subspaces of each PD-manifold, 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 product[41].

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 set of eigenfunctions $\psi_k = \{\cos(k\pi x) \mid 0 \leq x \leq 1; k \in \mathbb{Z}^+\}$ satisfies the conditions of the LBO, which requires separable solutions in the form of products of complete, orthogonal functions[30]. 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 this demonstration, we plot each of the 1D points in a given eigenvector as a function of a uniform index (i.e., $I \in [1, 20]$ 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 the CM. 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.

Of course, as the points in a real dataset naturally arrive in unordered sequence, one would have to properly rearrange each image index to envision 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 experiment), each sinusoid would be irregularly stretched along the x-axis where those duplicate states occurred, forming an unwieldy non-sinusoidal form. However, as the points in each Ψ_k are always scrambled in the same order as in all other eigenvectors in the manifold, in the presence of duplicate or missing states, the composite of any two will always exist in a readily identifiable $L_{p,q}$ form. Of these composites, we identified 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

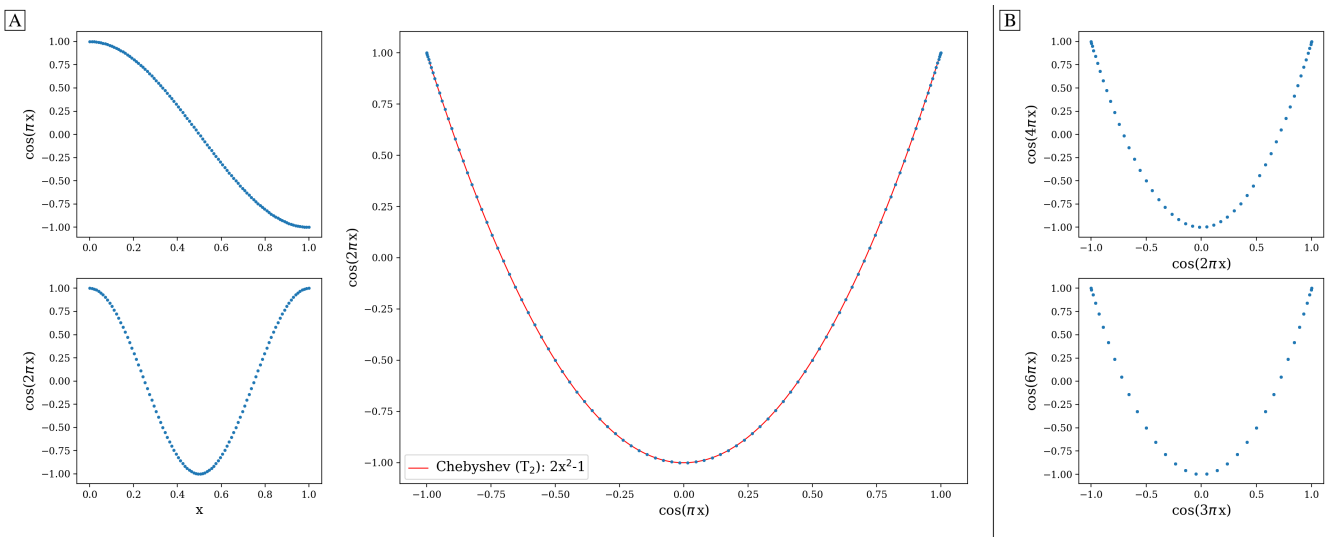

FIG. 2: The analytical generation of the Lissajous curve $L_{1,2} = \{\cos(\pi x) \times \cos(2\pi x) \mid \text{uniform } x \in [0,1]\}$, where $L_{1,2} \in ([-1,1] \times [-1,1])$ is shown in [A]. Note the naturally-induced nonuniform spacing between points near the boundaries and vertex of the parabola.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

eigenvector; i.e., $(\{\Psi_1, \Psi_2\}, \{\Psi_1, \Psi_3\}, ..., \{\Psi_1, \Psi_g\})$, 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 kind[42], of which we observed that the parabolic form is the lowest-order member present in each PD-manifold.

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 $X = \prod_{a \in A} X_a$, and naturally arises when taking the Cartesian product of a set (indexed via A) of 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 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 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, as 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 PD-manifold. 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 that conformational signal non-monotonically. 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 reflect 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)[41], 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 pre-

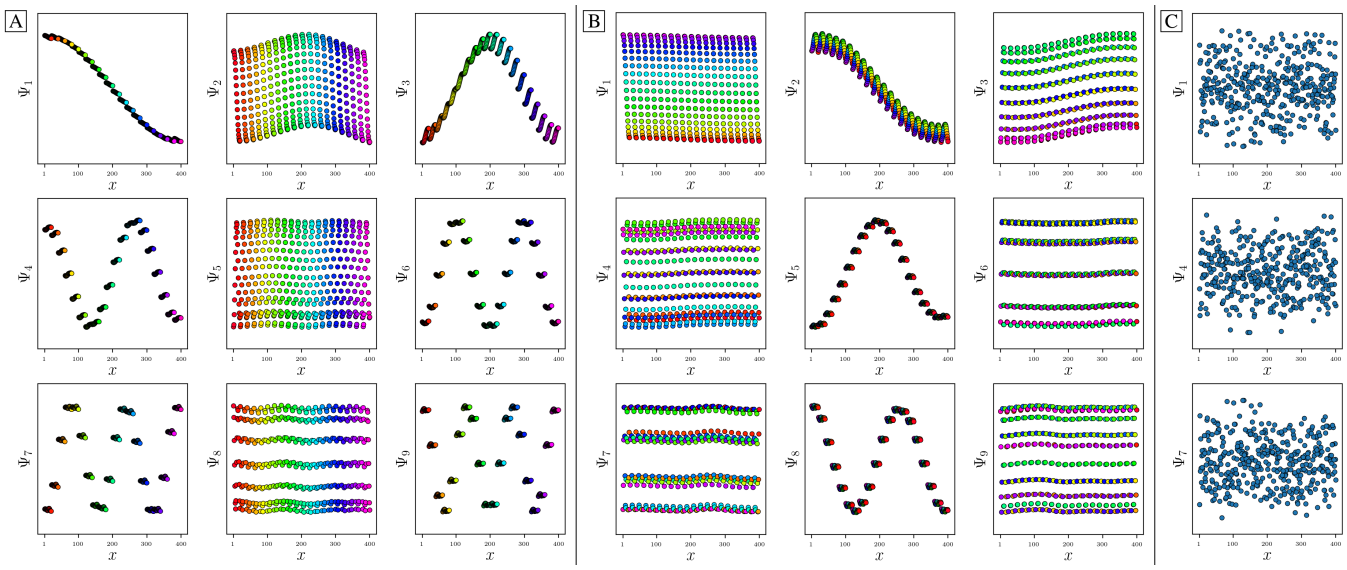

FIG. 3: As described in subsection B below, we provide here a visualization of eigenfunctions for PD_1 in SS_2 (i.e., $20 \times 20 = 400$ states total making up two degrees of freedom). On the left [A] are the sinusoidal forms $\{\cos(k\pi x) \mid k \in \mathbb{Z}^+\}$ that emerge for only a specific subset of eigenvectors $\{k = 1, 3, 4, 6, 7, 9, ...\}$ 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 $\{k = 2, 5, 8, ...\}$ 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 completely unavailable *a priori*. For example, in the above case, there are more than 400! possibilities for such sequences, of which [C] shows one such example obtained when the index of points in each eigenvector was randomized. To note, such indices, which do not follow ground-truth knowledge of CM_1 or CM_2, make up the vast majority of possible sequences.

serves the topological structures involved[41]. 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 difficult, as we will see, as more degrees of freedom are added to the system.

Finally, for our analysis of SS_1, we compared these sets of 2D subspaces among the five example PDs, and found only subtle differences in the distribution of their point clouds. It is important to underscore here the natural discrepancies between PD-manifolds 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 map, depending on the type of motion and its component along the projection direction 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 exists for all PD-manifold characteristics, from eigenfunctions to their spectra, and even the embedded distances between consecutive states. As seen in the next section, these trends will become more apparent as we investigate the manifolds obtained for datasets generated with multiple degrees of freedom.

B. Data-type I in State Space 2

To further understand these conformational-variation signals in the presence of increasing intrinsic dimensionality n, we next investigated the manifolds generated for SS_2. As seen in Fig. 3-A, plotting each eigenvector's 1D points in the specific ground-truth sequence constructed for CM_1 against a uniform index (now for the 400 states in SS_2) presented a similar but now interspersed pattern of sinusoids. 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, and again plotted these points against a uniform index $I \in [1, 400]$. 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, we conclude that for n degrees of freedom in a given dataset, there are n independent sets of sinusoids, with each set interspersed throughout the collection of available eigenvectors.

Following our previous discovery of a single set of or-

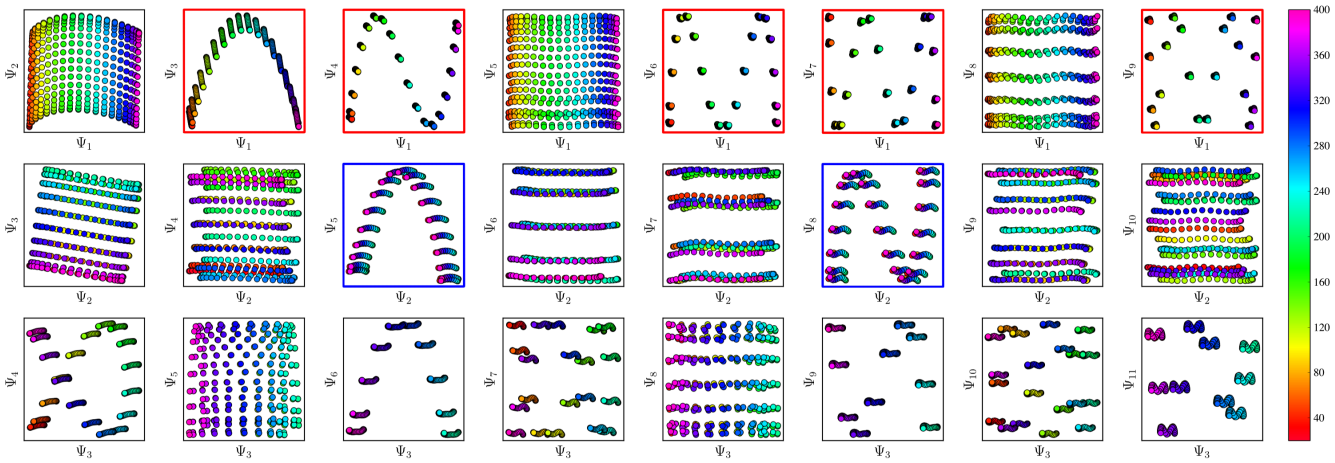

FIG. 4: A subset of the space of 2D manifold projections for PD_1 in SS_2. As demarcated in red and blue boxes, a unique set of conformational modes exists for both CM_1 (red boxes, $\{\Psi_1, \Psi_i\}$) and CM_2 (blue boxes, $\{\Psi_2, \Psi_j\}$; where $i > 1$ and $j > 2$), interspersed throughout each 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.

thogonal Chebyshev polynomials spanning specific 2D projections of SS_1, we next investigated the existence of similar patterns 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 manifold. Specifically, in the case of PD_1, independently projecting the manifold 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 sinusoidal 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*).

With these relationships established, we returned back to the seemingly-arcane grid-like patterns obtained in Fig. 3, which occupied the complement of each sinusoidal eigenfunction set in Fig. 3-A and Fig. 3-B, and found them to be the composite of many orthogonally aligned sinusoids. As a proof of concept, we were able to independently reproduce the $\{x, \Psi_1\}$ grid-like plot in Fig. 3-B by (1) forming a 2D array X_0 containing $\eta = 25$ copies (one copy per row) of η points uniformly sampled from $[0, 1]$; (2) forming a 2D array Y_k containing η copies (one copy per row) of η points uniformly sampled from $\psi_k = \{\cos(k\pi x) \mid x \in [0, 1]; k \in \mathbb{Z}^+\}$; and then using these to (3) plot the composite space $\{X_0, Y_1^T\}$ shown in Fig. 5-A. Likewise, we could reproduce the singular cosine plot $\{x, \Psi_1\}$ in Fig. 3-A via $\{X_0, Y_1\}$. As evident by these constructions, these two spaces exist as projections from a higher-dimensional space, namely $\{X_0, Y_1^T, Y_1\}$, as is shown in Fig. 5-D. Even though these sinusoids remain unseen outside of ground-truth studies, our analysis confirms that their 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. It remains to be seen, however, what role these grid-like sinusoids (which arise exclusively for $n > 1$) have on the structure of the manifold.

To understand this relationship, we graphed the first three eigenvectors of PD_1 in SS_2, and found that the resulting manifold took on the form of a 400-point (20×20) parabolic surface made up of an array of parabolic curves; having a proper arrangement of points as confirmed by the correct order of indices $I \in [1, 400]$ across its structure. As seen in Fig. 5-E, this result could be reproduced using composites of our constructed spaces, via $\{Y_1^T, Y_1, Y_2^T\} = \{Y_1, Y_1^T, Y_2\}$. Meanwhile, selectively plotting the Cartesian product of only those subspaces exhibiting non-grid-like patterns in a given row (e.g., $\{\Psi_1, \Psi_3, \Psi_4, \Psi_6, \Psi_7, \Psi_9, ...\}$ of PD_1, as demarcated by red outlines in Fig. 4) resulted in a Lissajous curve of respective dimensionality. Specifically, in the case of PD_1, the Cartesian product of $\{\Psi_1, \Psi_3, \Psi_4\}$ formed the trajectory $L_{1,2,3}$ (an open 1-manifold[41]), while the Cartesian

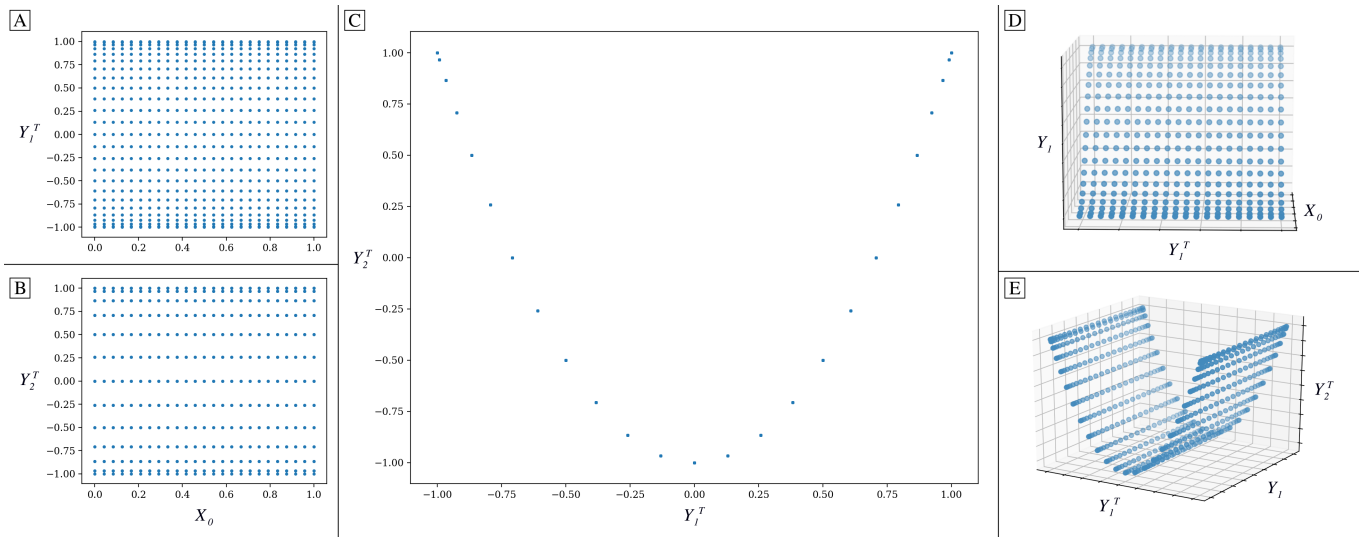

FIG. 5: Subplots [A] and [B] showcase elements of the previously-described construction: specifically $\{X_0, Y_1^T\}$ and $\{X_0, Y_2^T\}$, respectively. Using elements from [A] and [B], $\{Y_1^T, Y_2^T\} = \{Y_1, Y_2\}$ is constructed in [C] to form a parabola. In [D], the grid-like arrangement of sinusoids is displayed at a tilted angle in its 3D space (before projection to form [A]) to better emulate the perspective seen in $\{x, \Psi_1\}$ of Fig. 3-B. Finally, [E] shows the composite of three constructed elements $\{Y_1^T, Y_1, Y_2^T\} = \{Y_1, Y_1^T, Y_2\}$, which forms a parabolic surface, as is similarly seen when taking the Cartesian product of eigenfunctions in each PD-manifold. The Python code used for generating these figures is available in our online repository[38].

product of $\{\Psi_1, \Psi_2, \Psi_3\}$ formed the previously-described parabolic surface (an open 2-manifold, locally homeomorphic to $\mathbb{R}^2$). Hence, it can be seen that hypersurfaces arise for composites of eigenfunctions corresponding to different degrees of freedom.

When either these trajectories or surfaces are aligned with the plane of their 2D subspace (as seen, for example, in comparison of the parabola in Fig. 5-C with the selected view shown for the parabolic surface in Fig. 5-E), the projected structure is again that of the 2D Chebyshev polynomial, which carries 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 subspace representations. However, the existence of this preferred alignment is not always guaranteed. As seen most strikingly in $\{\Psi_3, \Psi_4\}$ of PD_3 (Fig. S5), the parabolic surface described by the Chebyshev polynomial is tilted out of alignment with the plane of the 2D subspace containing it. Similar instances, albeit more subtle, also arise for surfaces in the remaining three PDs of Fig. S5. From this observation, it appears that depending on the PD, CM information – pertaining to each of the system's degrees of freedom – can lie on some linear combination of the manifold's orthogonal eigenvectors.

When these aberrant surfaces arise, they have the appearance of having been rotated in their 2D subspace by some fixed amount, such that - in the higher-dimensional space - the set of orthogonal eigenvectors Ψ are unaligned from the set of orthogonal eigenfunctions ψ. Naturally, there exist innumerable eigenvector bases from which even perfectly-sampled eigenfunctions may appear obfuscated from their ideal form. Likewise, one can also induce a preferred eigenvector basis on this space, such that its set of coordinates are perfectly aligned with the eigenfunctions of the LBO. Hence, this aberration can be resolved by replacing the eigenvector basis we obtained *in situ* (via diagonalization) with the most desired one: in other words, by rotating the original coordinate system to align with the LBO eigenfunction basis. As described formally in the supplementary section "Eigenfunctions of the Laplace-Beltrami operator", the justification for an eigenspace rotation is granted by the presence of non-zero off-diagonal components in the metric tensor g^{ij} of the Riemannian manifold.

In the supplementary section "Manifolds from 3D density maps", we show that this disconnect between eigenvectors of one space and eigenfunctions of another is a result of PD disparity. Specifically, we generate a DM manifold from the SS_2 collection of 3D Coulomb potential maps (from which the PD datasets originate), and demonstrate near-perfect alignment of its *in situ* eigenvector basis with the LBO eigenfunctions. Thus, it is clear that the need for eigenspace rotations is explicitly tied to PD disparity, and specifically, the introduction of altered, foreshortened distances, compared to the original 3D distances, in each 2D projection. Thus, *in situ*, what we obtain is a set of PD-manifolds originating from

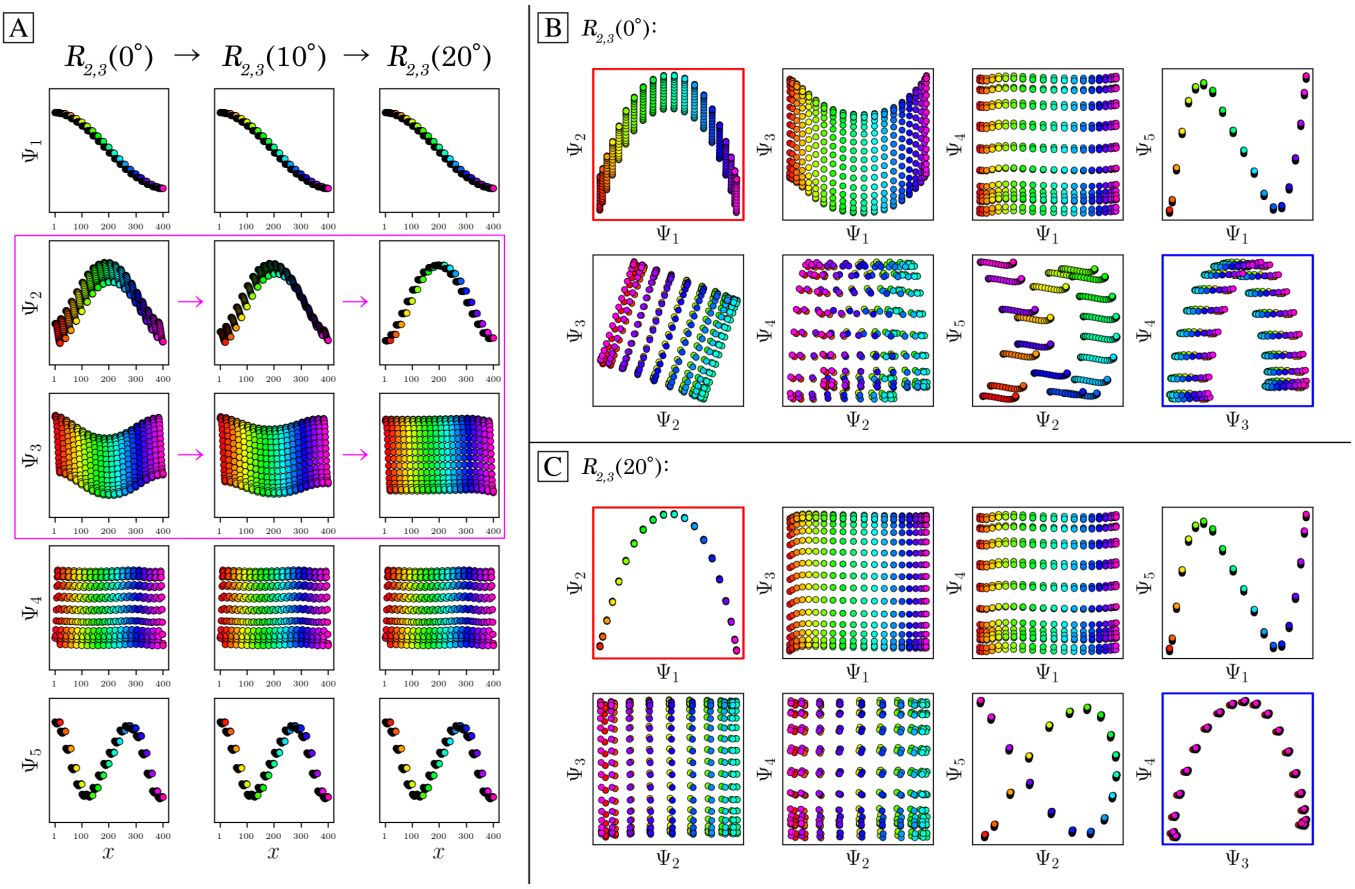

FIG. 6: Manifold for PD_3 from SS_2 under application of a 5D rotation matrix $R_{2,3}(\theta)$. The three columns in [A] display the individual eigenfunctions (as plotted by indices corresponding to the CM_1 frame of reference) before the rotation is applied, at $R_{2,3}(10^\circ)$, and finally at $R_{2,3}(20^\circ)$, 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.

eigenfunctions that map identical 3D conformational information, yet each display that content in a different, uniquely-determined orthogonal coordinate system. This disparity among PDs is inevitable, and it poses a fundamental problem that must be accounted for.

As a remedy to this problem, our aim then is to stitch the PD-manifolds into one consolidated orthogonal coordinate system, by aligning the eigenfunctions of the Laplace-Beltrami operator - housing identical CM information shared ideally by all PDs - with the set of each PD-manifold's *in situ* eigenvectors. 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, mapping each along the direction of an eigenvector defined by the corresponding PD-manifold *in situ* eigenbasis. From our formal description, this is equivalent to diminishing the off-diagonal components of the metric tensor, where the magnitude of these components depend on PD. To provide rationalization for this technique, Fig. 6-A shows the eigenvectors as indexed in the reference frame of CM_1 for the highly-misaligned PD_3 manifold. As seen in the first column of Fig. 6-A, while the sinusoids for $\Psi_1 = \{\cos(\pi x) \mid CM_1\}$, $\Psi_4 = \{\cos(2\pi x) \mid CM_2\}$ and $\Psi_5 = \{\cos(3\pi x) \mid CM_1\}$ are in agreement with expectations, the graphs of $\Psi_2 = \{\cos(2\pi x) \mid CM_1\}$ and $\Psi_3 = \{\cos(\pi x) \mid CM_2\}$ appear deformed with respect to their 2D plane. 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 from the coordinates of its eigenbasis (Fig. 6-B).

Given this insight, we next introduce a method for correcting these misalignments using orthogonal transformations; taking advantage of the fact that the Laplace-Beltrami operator is invariant to isometries[30]. Specifically, we apply a d-dimensional rotation operator (having enough dimensions d to encompass each leading Chebyshev parabola for all CMs present) to single-handedly reorient all aberrant surfaces into the plane of their respective 2D subspaces. The results of this operation on

the manifold associated with PD_3 can be seen in Fig. 6-B and Fig. 6-C; before and after applying a 5D rotation matrix, respectively. Mathematically, this d-dimensional rotation is a special 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. 6, there exists 10 rotation sub-matrices in total, with each corresponding to a specific planar rotation on the manifold. Of these 10 matrices, we found that only one had to be altered to achieve the results shown, having general form

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

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. 6-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 (via vector additions and scalar multiplications), so too is Ψ_2 weighted with Ψ_3, and vice versa. As seen in Fig. 6-A, along the way in reaching $R_{2,3}(20°)$, 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. After this exchange, this 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. 6-C), as desired. Later in our analysis, we will return to this topic under the moniker *eigenspace rotations*, and describe a strategy for automating these decisions 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 the dataset (this time for CM_1, CM_2 and CM_3), a set of unique Lissajous curves were again found spanning specific 2D projections of the manifold, with the Chebyshev subset explicitly describing the corresponding CM along a 2D trajectory. 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 manifold, 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, both the matrices ZZ^T and $Z^T Z$ are symmetric and positive semi-definite (i.e., all eigenvalues are non-negative)[43], which is similarly the case for the Markov transition matrix used in the DM framework.

A set of different projections of this manifold via selected principal components (i.e., eigenvectors) {PC_1, PC_2}, {PC_1, PC_3}, and {PC_1, PC_4} 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 were 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 geometries for positive semidefinite matrices, as described in the supplementary section "Comparison of PCA and DM"). 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.

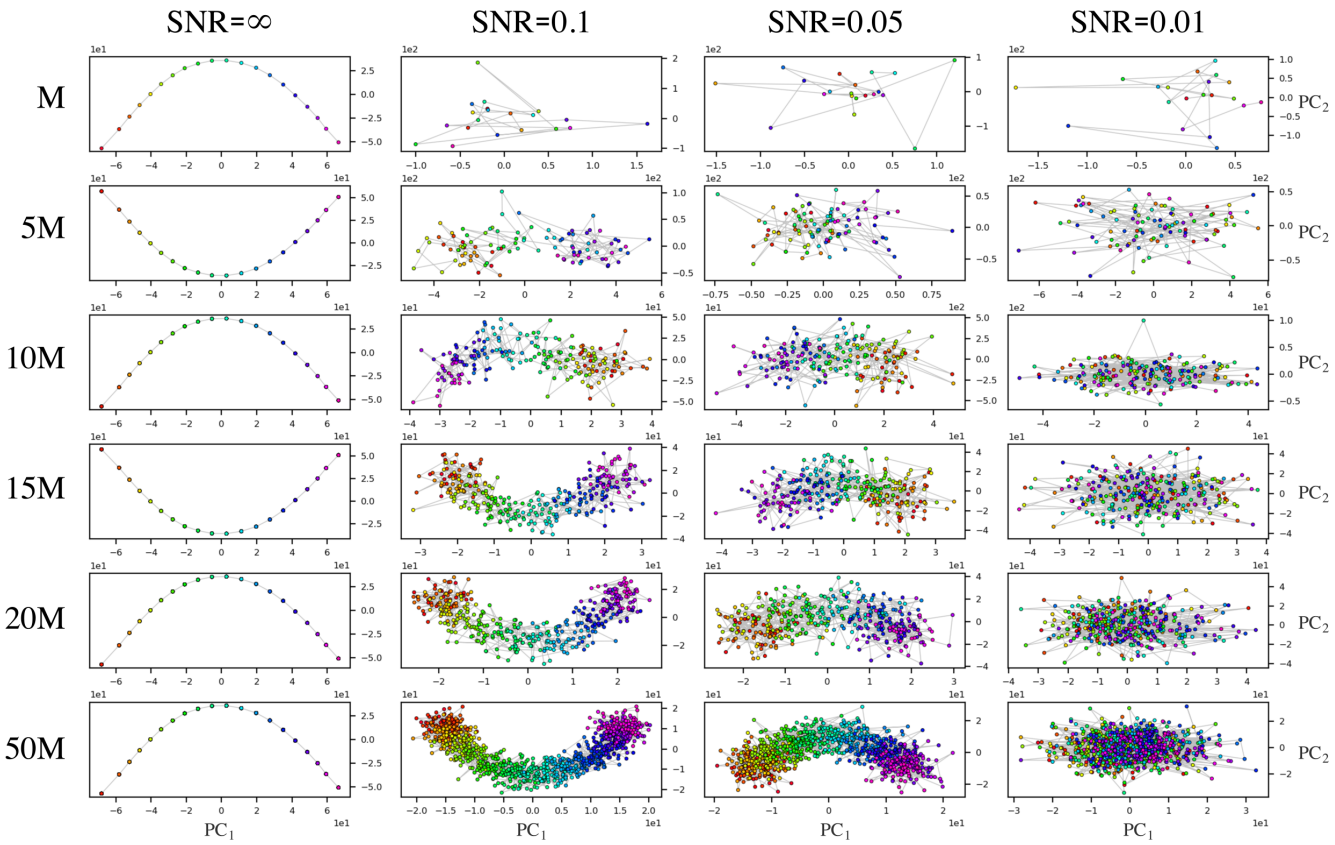

FIG. 7: Set of $\{PC_1, PC_2\}$ manifolds produced by PCA from PD_1 images in SS_1 over a range of SNR values and state space coverage. As can be seen in the first column, the images in the manifold were well-organized in the absence of additive Gaussian noise (SNR_∞). Meanwhile, as seen in the remaining columns, decreasing SNR regimes increasingly deteriorated the manifold's intrinsic structure. However, as the $M = 20$ state space was multiplied by increasing values of τ in each of these SNR regimes, the intrinsic structure of the manifold 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 manifolds, and, as further illustrated by the color mapping of their points, no sensible ordering of snapshots can be ascertained within these subspaces.

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 supplementary materials: "Simulation of cryo-EM ensembles for data-type II"), such that the images in each dataset had unique SNR $\in \{1, 0.1\}$ consistently applied for 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 manifolds to the state space ordering decayed with increasing noise level. Within low-SNR regimes, the behavior of the manifolds from PCA and DM were highly consistent, with DM generated within its optimal range of Gaussian bandwidth. Overall, the corresponding geometric structures of manifolds 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 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

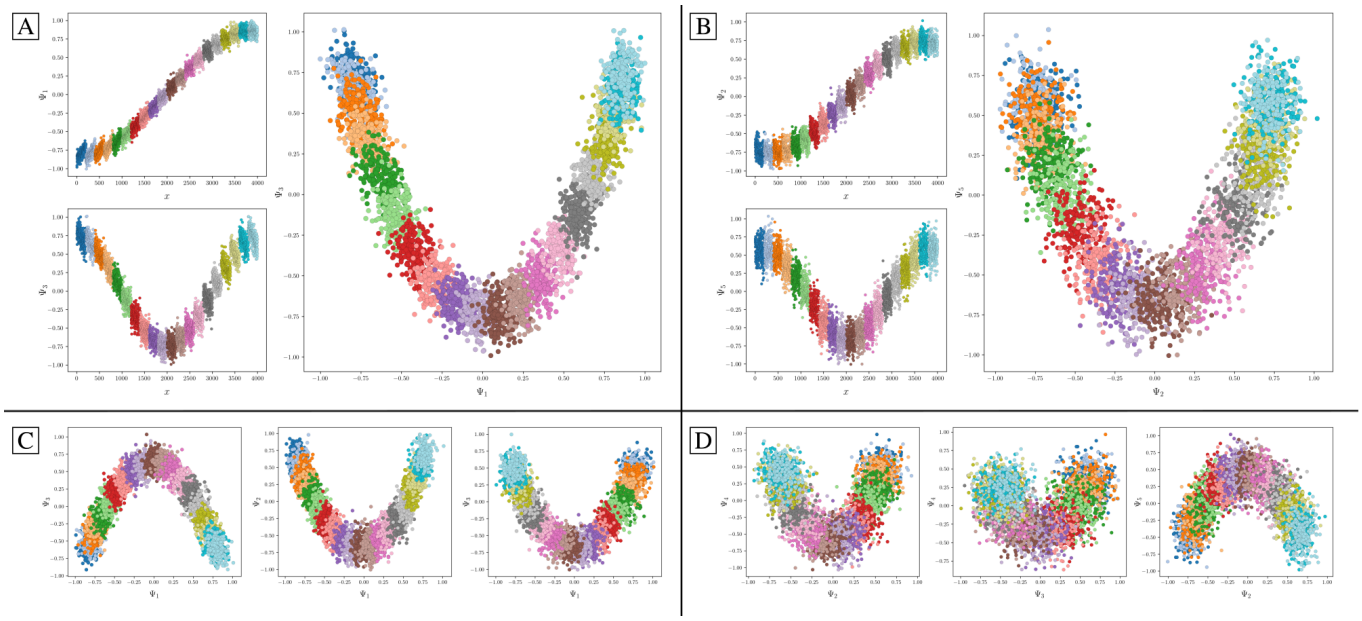

FIG. 8: CM subspaces for a set of five PDs generated with SNR = 0.1 and $\tau = 10$ and embedded via PCA. The coordinates within each point cloud are colored to signify their ground-truth CM state assignment, such that each point belongs to one of 20 CM bins, and each bin contains 200 points (with the same coloring scheme used regardless of CM). In [A], the parabolic CM_1 subspace of PD_1 is shown along with its two leading cosine eigenfunctions (with each cosine ordered according to its ground-truth sequence). Similarly, in [B], the parabolic CM_2 subspace of PD_1 is shown with its own set of leading cosine eigenfunctions. The remaining subplots show a variety of CM_1 [C] and CM_2 [D] subspaces for three randomly-oriented PDs, so as to emphasize the variability in features prevalent in manifolds obtained from noisy images.

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. 7, 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 the manifold's intrinsic geometric structure; i.e., the set of Lissajous curves and corresponding Chebyshev polynomials.

To quantify these relations, each member of the set of PCA manifolds in a τ-series was fitted with a set of leading Chebyshev polynomials, as seen in Fig. S9 for SNR = 0.1. The coefficient of determination (R^2), which can be interpreted as the proportion of variance in one variable accounted for by another[44], 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 each manifold 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 fidelity of the manifold.

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), 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 not only include unknowns such as the number of ground-truth states M and SNR regime, but 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 SNR = 0.1 and $\tau = 10$, which will guide several choices made for our framework in the following section. Fig. 8 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 eigenfunction point clouds maintains well-defined order after introduction of SNR, disorder in CM state partitioning is introduced in their composite parabolic point cloud.

Additionally, due to PD disparity, the characteristics of each CM-parabola can be seen to vary significantly depending 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 conformational maps, certain clusters of 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 manifolds obtained using DM (Fig. S18).

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 supplementary materials: "Simulation of cryo-EM ensembles for data-type III". Specifically, using images from PD_2 of SS_2 with $\tau = 10$, we assign to each image a random defocus value from the interval [5000, 15000] Å. This wide range is 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. As such, 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 defocus[1]. 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 change. The corresponding manifold is shown in Fig. 9, 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. 9, there is a noticeable inward-curling at the ends of the CM-subspace parabolas generated using the double-filtering kernel. In addition, the priority of the leading conformational motion has been flipped in eigenvector significance; noting that CM_2 is the most visually pronounced of the two motions as viewed in this PD. 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 obtaining manifolds using a standard kernel from sets of CTF-corrected images. To note, perfect defocus assignments were used for generating the manifold, when in reality these values would be estimated first using established algorithms[45–47].

V. Overview of the DMSA Framework

Having explored all three data-types including CTF and noise, we now lay out the DMSA strategy for recovery of conformational motions in the form of 3D movies and corresponding energy landscape. This methodology requires several steps that will be introduced in turn, with the entire process schematized in Fig. 10.

The general intuition for our approach is as follows. Ideally, for each PD-manifold, 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 PD-manifold 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 belonging to each compiled bin can be combined to reconstruct a 3D 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 energy landscape) can be formed, along with a set of corresponding 3D 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

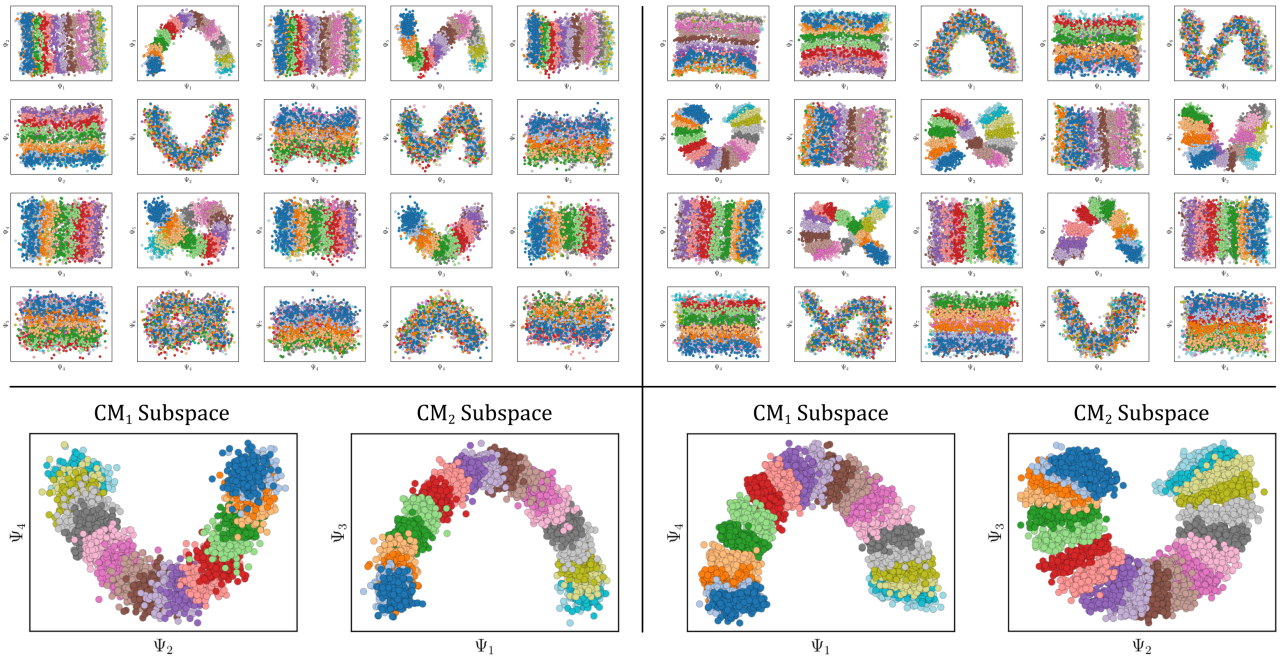

FIG. 9: SS_2 manifolds (PD_2, $\tau = 10$, SNR $= 0.1$) obtained with and without CTF applied, as shown on the right and left-hand side, respectively. For the case of the manifold obtained from images without defocus, protocols for synthetic generation follow those established in Fig. S12 (A, B). Likewise, on the right, protocols follow synthetic generation of images with microscopy parameters as shown in Fig. S12 (D, E). The non-CTF manifold was generated via DM with the standard Gaussian kernel, while the CTF-manifold 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).

and direct mapping. Since there are many such potential subspaces housing parabolic surfaces (even for $n = 2$) within a given PD-manifold, 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 stratify 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 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 DMSA 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 eigenspace 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 stratified into contiguous equal-area bins [step 5]. Each bin is then appended 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 Coulomb potential 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 eigenspace rotations and manifold binning (subsection A and B, respectively), followed by use of our final analysis data-type (i.e., SS_2 via data-type III with energy landscape) in subsection C to furnish final outputs and assess their validity.

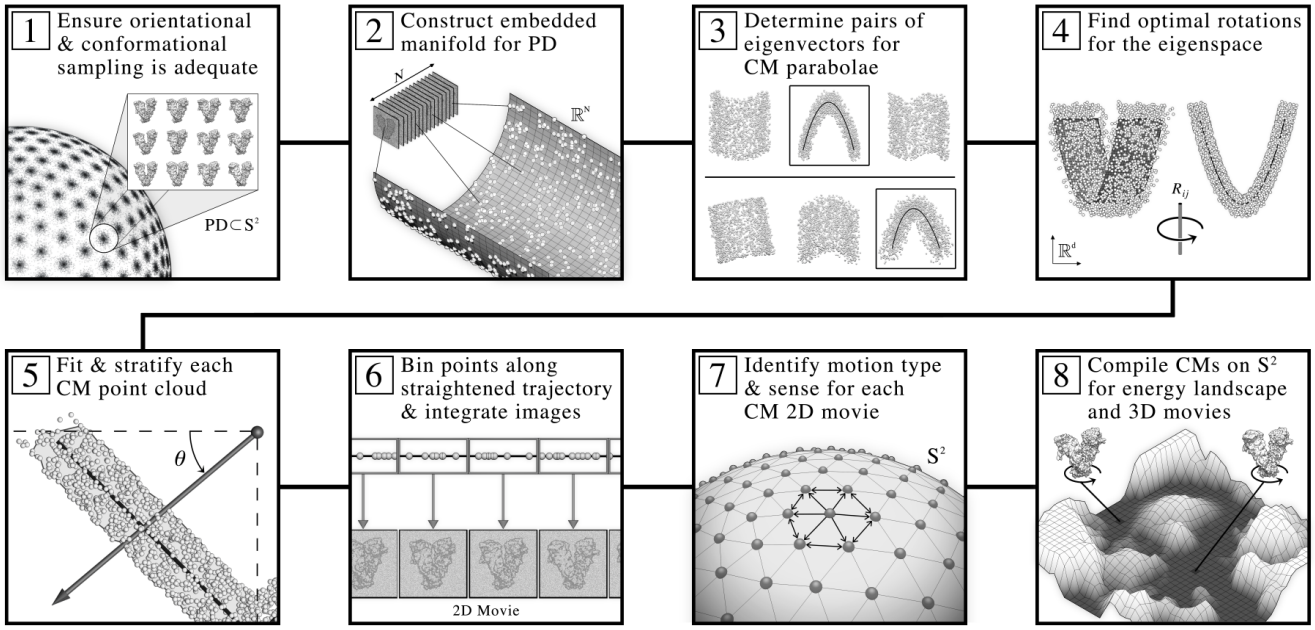

FIG. 10: Schematic detailing the DMSA workflow for recovery of conformational continuum as informed by our heuristic analysis. Through this framework, 3D movies and corresponding energy landscapes are obtained for the set of conformational motions in a given dataset. Note that the previous ManifoldEM workflow branches off after completion of step 2 above, and after performing a series of alternative steps required by NLSA, enters again with our pipeline at step 7, before splitting off again to form final outputs as we achieve independently 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.

A. Eigenspace Rotations for Data-type II

We describe here our methodology for calculating the required eigenspace rotations for each PD-manifold in the presence of noise with experimentally-relevant SNR, which we deem the initial step in the DMSA framework (i.e., [step 3] in Fig. 10). To note, after generation of each manifold 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 eigenspace rotations on each PD-manifold. First, recall that depending on the PD, the observed conformational eigenfunctions may be misaligned with respect to the eigenfunctions of the ideal LBO, requiring application of a d-dimensional rotation matrix to align the subspace (Fig. 6). As an example, the effects of applying a 4D rotation to the 4D manifold 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 required rotations so as to counter-rotate each PD-manifold. This algorithm is informed, first and foremost, by our heuristic findings of the existence of parabolic surfaces in each PD-manifold, 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, which 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 supplementary section "Eigenspace rotations", 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. 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 parabolae, ellipses or hyperbolae (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 manifolds with nonzero discriminant, which are encountered for manifolds with boundary aberrations and especially those obtained from images modified by CTF. In this form, the xy-term rotates the graph, providing for the possibility of encountering manifolds with an axis of symmetry unaligned from the 2D eigenbasis. This equation can thus be rewritten with a new set of coefficients[48] to effectively rotate the coordinate axes such that they come to alignment with the axis of symmetry.

B. Manifold Binning for Data-type II

Once all PD-manifolds 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 DMSA 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 the supplementary section "Overview of NLSA".

The motivation for the DMSA manifold binning approach stems from our analysis of PD disparity in the presence of noise (as shown in Fig. 8), 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 manifold-fitting approach able to correctly chart spatial discrepancies while remaining unencumbered by changing densities (i.e., occupancies) along each trajectory. Fig. 11 provides an overview of our novel strategy for splitting up each CM-manifold 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 PD-manifold, 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 manifold 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 half-manifold with a polygon enclosing a majority of our remaining points (Fig. 11-B and Fig. 11-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 half-manifold 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 equal[52]. For our purposes, 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 $\{\Phi_i, \Phi_j\}$ manifold. 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. 11-B and Fig. 11-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 $\theta = 0°$ (Fig. 11-D). For each angle $\theta \in [0°, 90°]$, the area of the lower sub-polygon formed by the intersection of the alpha-shape with the ray is recorded[54] to define the overall ratio of that sub-polygon to the whole. We form 10 bins in this fashion along the Φ_{fit} trajectory (per half-manifold), 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. 11-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. 11-E and Fig. 11-G, showing overall agreement with expectations. To note, we have programmed

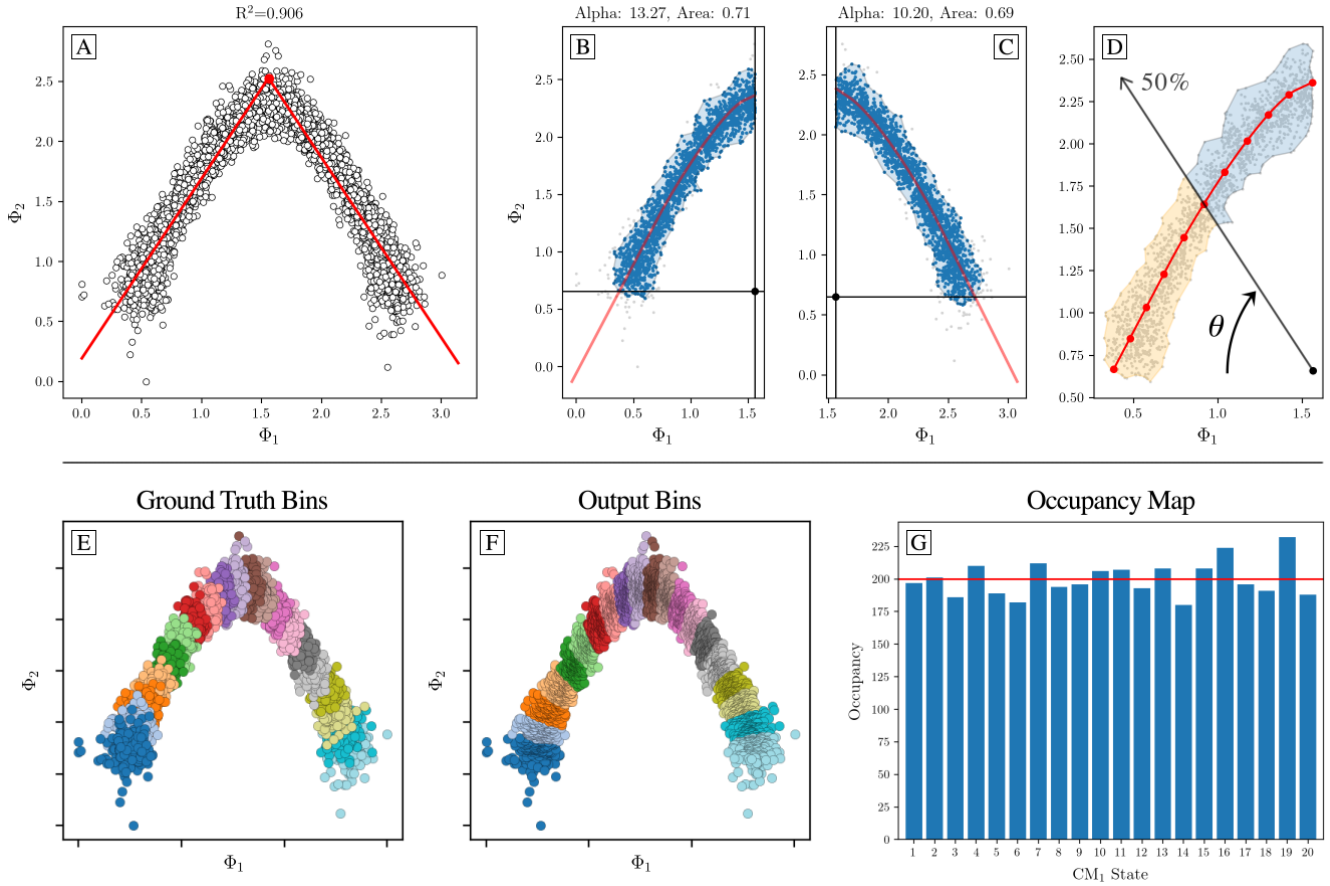

FIG. 11: Overview of area-based method for extracting sequential conformational information from a given CM subspace. Subplots [A] through [G] display our algorithm's outputs on the CM_1 subspace of an arbitrary PD from data-type II. First, [A] shows the inverse-cosine transformation and corresponding preliminary fit using an absolute value function. Subplots [B] and [C] demonstrate the alpha-shape polygon and Φ_{fit} trajectory defined on each half-manifold, 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) CM_1 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 CM_1 state.

these steps to require no intermediate supervision via a robust automation strategy, with details for specific subtasks available in the comments of our corresponding code[38]. As a result of this procedure, the points in the manifold corresponding to independent conformational motions are all lined up via the 2D subspace projection, such that averaging them together only reveals the conformational-variation signal corresponding to the current 2D subspace. 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 movie M2, where we have provided 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 DMSA methodology [steps 3-6] which are required for handling our final analysis dataset - which, recall, now includes introduction of CTF as well as noise. Finally, once 2D movies have been obtained from each PD-manifold, we describe here the conclusion of the DMSA framework via an efficient method for compiling CM information from all PDs [step 7] to create an 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 tomo-

graphic reconstruction at a given resolution is defined by the Crowther criterion[55] $\mathrm{PD}_{\min} = \pi D/r$. 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 from a fictitious occupancy map (see the supplementary section "Occupancy assignments for final analysis") 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 between 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 PD-manifold analysis.

As was discovered for data-type I and II, we found that a substantial number of these 126 PD-manifolds had a misaligned *in situ* eigenbasis from our preferred, common coordinate system, with the magnitude of counter-rotations required varying significantly for each PD. For all 126 PDs, our previously-described rotation-automation strategy (see DMSA 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 CTF (Fig. S14). A histogram of the magnitude of rotations used across all rotation operators - which varied nontrivially for each PD - is provided in Fig. S20.

With the 126 PD-manifolds correctly rotated into a common eigenbasis and each 2D CM subspace identified, we next proceed to partition each 2D subspace into bins representing conformational states. This procedure follows our previous description (see DMSA section B) with only slight modifications required for the case of images modified with CTF. First, to combat the effects of inwards curling (Fig. 9) 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 resides 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 binning 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 more globular 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. 11-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 for the creation of these bins (i.e., nearing that of vertically-aligned projections). This added flexibility made our binning algorithm robust in the presence of both parabolic and 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-filter[5] each image before integration within each CM bin. A subset of the final 2D movies produced through this framework are available to view in movie M3. We have also provided a comparison of these final DMSA outputs to NSLA for one degree of freedom in the supplementary section "Comparison of results from DMSA and NLSA". Through analysis shown in several movies, we found that DMSA 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 NLSA pipeline, such as an inability to capture certain conformational motions captured properly by DMSA and 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 DMSA.

After generating all 2D movies (one per CM for each PD), both the type of conformational motion present in the 2D movie (e.g., CM_1 or CM_2) as well as its *sense* must be determined individually for each PD. As to the meaning 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 path[56]. For example, for CM_1, 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 signs which naturally arise via eigendecomposition[57]. 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 algorithms[56], 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 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 CM_1 and CM_2 independently, with corresponding occupancy statistics accumulated for each state therein, and compared with ground-truth knowledge (as shown on the left in Fig. 12). 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 manifold trajectory and having a relatively small spread of uncertainty for each bin. However, 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 CM_1 and CM_2 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 for all occupancy maps observed, a problem seemingly unavoidable based on our observations of corresponding manifold 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-manifolds 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 DMSA 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 DMSA 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 CM_1 and CM_2 trajectories: effectively reconstructing the hypersurface on which they jointly reside. For more information on this procedure, see the supplementary section "CM composition". The efficiency and simplicity 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 occupancy map. The result of our DMSA operation is shown on the right in Fig. 12, 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 $\{CM_1, CM_2\}$ plane, 3D 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` module[46] to create a 3D reconstruction for each of the 400 states. These 3D maps were loaded in sequence to create 3D movies from different views using Chimera[58]. As shown in Fig. 12 and qualitatively expressed in movie M4, these 3D 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 correlations[5] (FSC) curves between several ground-truth maps (in MRC format) and their corresponding DMSA output 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 maps. These curves confirmed a good recovery of our different states up to a resolution near 3 Å, the ground-truth value. Additionally, Q-scores[59] 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 DMSA output maps, we found highly favorable agreement across all residues for each state. Results for an example state are available in the supplementary section "Structural validation of DMSA outputs".

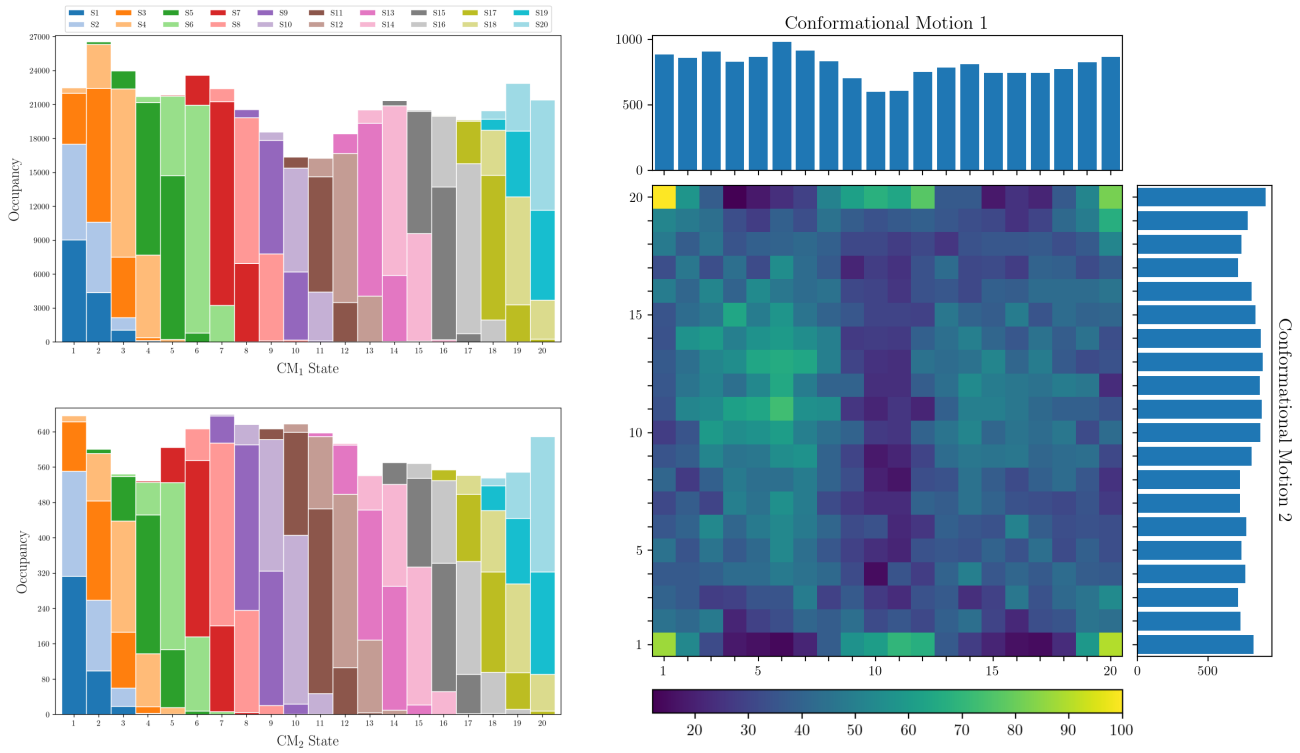

FIG. 12: On the left, the final occupancy maps for the 20 states in CM_1 (top) are shown alongside an equivalent representation for the 20 states in CM_2 (bottom). Each plot was obtained by integration of the corresponding 20 bins (corrected for sense) in each of the 126 PDs. The total number of images as assigned to each state via our manifold 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 for an assessment of the True Positive rate. On the right, the final 2D occupancy map for the 400 states formed by CM_1 and CM_2 is shown; obtained via the intersection of image indices in all pairwise combinations of CM_1 and CM_2 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-manifolds 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.

DISCUSSION

Through our analysis, we have identified the way sets of images originating from a varying atomic structure are represented in the manifolds 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 manifold-fitting procedure using pairwise combinations of eigenfunctions of the homogeneous Laplace-Beltrami operator, 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 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 the supplementary section "Comparison of results from DMSA and NLSA").

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 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 com-

plexities existent in an experimental system. These complexities can be considered as introducing higher-order terms in our parameter space, where we have attempted to emulate all lower-order terms up to a threshold we deem 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 also limited to a set of simplified data models originating from a molecule undergoing collections of rigid-body motions, which we believe should suffice for a subset of molecular machines, but falls 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. By this motivation, a study of ligands has been performed with the founding ManifoldEM approach[10] using two state spaces, and could similarly be explored in an extension of the DMSA framework.

It is also worth noting that while the results of our heuristic analysis are most relevant to methods requiring the PD-manifold approach, several portions of our analysis can be extended to other techniques dealing with alternative manifold inputs (such as the use of 3D density maps in cryo-ET). In our supplementary materials, we detail how the structure of manifolds obtained from a conformational state space transforms as the data type is translated from 3D maps to 2D projections, showing that the eigenspace rotations are a consequence of PD disparity.

With that said, we return to the flowchart presented in Fig. 10, 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 views across the 2-sphere; 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 manifolds of the PDs 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 predictor for the fidelity of each PD's manifold in rendering the 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 the ground-truth structure of the manifold (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 manifold 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. 7).

However, even 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. 7). 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 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. τ_c 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 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 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 manifold of the ribosome[1]. As one potential scheme, for each PD-manifold 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 DMSA 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 DMSA is by far the better choice for manifolds with geometrically-structured subspaces, only NLSA can be applied in the regime of manifolds completely lacking discernible form, while still potentially incurring its known limitations and uncertainties[11].

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 bodies); uncertainty in CTF estimation; and uncertainty in angular alignments, which are more pronounced in heterogeneous data. It is obvious that such artifacts and errors will 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 method of 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). However, 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[1,4,24,39,62]. In contrast, a comparative review of twelve prominent dimensionality frameworks[13] 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 errors[13]. 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 approaches[13] using our simulated datasets.

Finally for this section, we discuss the use of the defocus-tolerant kernel[1] for calculating Euclidean distances between images on which a CTF was imposed. In summary, its success dwarfed that of two other techniques explored. For comparison, we first observed that manifolds obtained using the standard kernel without any CTF correction were completely incoherent in structure, as expected. Better behaved but still flawed were manifolds obtained from sets of CTF-corrected images (Fig. S12-G) using the standard kernel, which ultimately provided suboptimal structure to those obtained using the double-filtering kernel. We consider this study the first demonstration that the previous double-filter kernel introduced[1] is more effective than CTF-correction of individual images. However, as previously noted, the defocus-tolerant kernel is not free of artifacts. Most significantly, CM subspaces typically had a strong proclivity of curling inwards near the outer edge, with states clumped more densely in these regions (Fig. 9). 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 manifold subspace is greatly exacerbated. There are additional higher-order aberrations of the CTF relevant for high-resolution cryo-EM[63] (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 manifold geometry if not accounted for properly.

3. Determine pairs of eigenvectors for CM parabolae

Specific techniques must next be applied to discover the set of conformational modes corresponding to each CM within the corresponding PD-manifold, with the number of these sets defined by the dataset's intrinsic dimensionality. However, as described in the supplementary

section "Estimation of Gaussian bandwidth and intrinsic dimensionality", there exists an initial uncertainty on 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 potential study.

Our analysis demonstrated that the minimum information 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 manifold 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 projections of the manifold. However, care must still be taken to prevent overfitting these manifolds, since blindly performing the subsequent steps in this framework on subspaces that do not actually correspond to one of the dataset's CMs (of which there are fundamentally many) 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, with further descriptions provided in the supplementary section "Comparison of results from DMSA and NLSA". Likewise, we demonstrated how sets of parabolic harmonics for each CM naturally exist in all manifolds, and, due to the overlapping nature of their point clouds, are unviable for mapping conformational information.

In contrast, the DMSA 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 the supplementary section "Eigenspace rotations". 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 uncertainties[11] in the founding ManifoldEM framework (see the supplementary section "Comparison of results from DMSA and NLSA"). 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 manifold. In a similar vein, the possibility of underfitting the manifolds (analyzing too few CMs) must also be taken into account, though it carries significantly less risks than overfitting.

4. *Find optimal rotations for the eigenspace*

As seen to some extent in every manifold 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 the supplementary section "Manifolds from 3D density maps", this effect arises from the introduction of altered, foreshortened distances in each 2D projection (i.e., PD disparity). As detailed in the supplementary section "Eigenfunctions of the Laplace-Beltrami operator", and justified further through our heuristic accounts, this behavior could be countered via rotation of the eigenspace to ease retrieval of the CM. Given this knowledge, we presented an automated procedure for the discovery of these optimal eigenspace rotation angles for each PD. Despite its remarkable performance on our four 126-PD datasets (Fig. 12 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 as well as datasets with $n > 2$ degrees of freedom (see the supplementary section "Eigenspace rotations"). Furthermore, for noisier, less-structured manifolds, 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 PD-manifolds, we claim it is effectively impossible to find the proper counter-rotation such that all manifolds are displayed in a common coordinate system defined by CMs. We have further observed that the inability to align PD-manifolds onto this common eigenbasis can subvert the quality of 2D movies and occupancy maps (see PD_{33} in movie M8) - and thus ultimately the final 3D reconstructions, to varying extents. This fact explains several of the limitations[11] recently documented in the founding PD-manifold approach[1,10], which does not account for the tendency of PD-manifolds to be unpredictably unaligned from a common coordinate system (e.g., Fig. S20). For a more detailed analysis of limitations and uncertainties observed, see supplementary section "Comparison of results from DMSA and NLSA". Thus, in order to maximize fidelity of final outputs with ground truth, a dataset must first fulfill a minimum set of required properties that ultimately provide structured manifolds when embedded. The performance of DMSA

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 given its availability.

5. Fit and stratify each CM point cloud

Once the set of conformational modes are identified and counter-rotated for each PD-manifold, a further obstacle is encountered in the heightened variability of these point clouds (Fig. 8) – which vary in average thickness, density, length and type of trajectory, and spread of data points – depending on CM and PD. These complications were addressed via use of a robust automation strategy, including least-squares fitting and area-based stratification of the manifold. 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 SS_2 CM_1 and CM_2 subspaces from data-type II and III. Detailed notes on these procedures have been provided in our published repository[38], 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.

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 the supplementary section "CM composition", note that the bin size effectively controls the precision to which we can locate each point in the higher-dimensional surface, and influences the range of images falling within each state that we group together as 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. 11-G, Fig. 12, 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 manifold crudely into only two or a handful of states, as the number of states requested is increased, the amount of sophistication in mapping and segmenting each manifold 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 solving two manually-performed subproblems: (i) identifying each set of CMs across S^2 (e.g., such that CM_1 in PD_i is matched with CM_1 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 algorithms[56], and strongly advise against making such assignments arduously by hand.

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

After CMs have been properly identified and matched amongst 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 map. Once these previous steps were performed for our dataset, we compared the state assignments of each image from S^2 to its ground-truth index. We found that the majority of images in each bin were correctly assigned, with each bin also encom-

passing a small subpopulation of images that were actually ground-truth members of neighboring bins (Fig. 12 and Fig. S11). Typically, these misassigned images were members from a neighborhood spanning a limited range of states in close proximity to the bin they were erroneously assigned to, and increasingly so for bins in higher proximity.

Our final 2D occupancy map (also seen in Fig. 12), 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 maps output by DMSA (movie M4) and their respective ground-truth 3D maps. As well, 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 quality of final 3D 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 via an assessment of the structural fidelity of DMSA outputs with ground-truth atomic-coordinate structures and 3D Coulomb potential maps using Q-score calculations on residues and FSC curves, respectively (see the supplementary section "Structural Validation of DMSA Outputs").

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 manifold-fitting routines. Our ability to leverage the final image content to further improve 3D maps and corresponding occupancy distributions stands in contrast to the founding PD-manifold approach[1], which relies on histogram equalization (technically *histogram matching*) to match the occupancy maps across PDs. Although not pursued here, one possible choice 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 heterogeneity[66], as does a method based on neural networks[67].

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 heuristic analysis, we have introduced a novel workflow with several enhancements that substantially improve the recovery of conformational-variation information from single-particle cryo-EM, while also describing the challenges, fundamental limitations and uncertainties that emerge. We hope that the insights we have gained from the manifolds of ground-truth data for a molecule with several rigid-body motions will help in the development of techniques aimed at untangling more complex experimental datasets from molecular machines performing their work cycles – as we discussed, these often present complications such as more involved domain motions and the binding and release of ligands. Finally, we would like to point out that our ground-truth data, which are available for download, or similarly constructed ground-truth data for any other molecule in the PDB, may be useful to examine and validate other approaches to the recovery of energy landscapes.

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 Institutions 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 DMSA algorithm suite have been supplied in the header of all scripts in our online repository[38].

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The Python software repository[38] is available at the following link: `https://github.com/evanseitz/cryoEM_DMSA`. 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 eigenspace rotations and manifold binning algorithms presented here.

Appendix: Description of symbols and abbreviations

Name	Description
S^2	2-sphere: set of Euler angles for 3D viewing orientations
PD	Projection direction, PD $\subset S^2$
N	Number of images in PD-manifold of dimension $\mathbb{R}^N$
P	Number of pixels per image
CM	Conformational motion
n	Intrinsic dimensionality of the state space: one degree of freedom per CM
SS_n	State space of dimension n
ψ_k	Eigenfunction of Laplace-Beltrami operator; $\psi_k = \{\cos(k\pi x) \mid x \in [0,1]; k \in \mathbb{Z}^+\}$
Ψ_i	Eigenvector of N-dimensional manifold; $i \in \{1, 2, ..., N\}$
λ_i	Eigenvalue of N-dimensional manifold; $i \in \{1, 2, ..., N\}$
Φ_i	Inverse cosine of Ψ_i
L	Lissajous curves; $L_{p,q} \subset L$
T_k	Chebyshev polynomial of the first kind
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

REFERENCES

[1] A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseinizadeh, H. Y. Liao, J. Pallesen, G. Sharma, V. A. Stupina, A. E. Simon, J. D. Dinman, J. Frank, and A. Ourmazd, "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.

[2] E. 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.

[3] P. 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).

[4] A. Moscovich, A. Halevi, J. Andén, and A. Singer, "Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes," Inverse Probl **36** (2020).

[5] J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State* (Oxford University Press, Oxford, New York, 2006).

[6] J. Frank, "Generalized single-particle cryo-EM–a historical perspective," Microscopy (Oxf) **65**, 3–8 (2016).

[7] J. Frank, "Single-particle reconstruction - story in a sample." Nobel Lecture (2017).

[8] N. 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).

[9] P. C. Whitford, R. B. Altman, P. Geggier, D. S. Terry, J. B. Munro, J. N. Onuchic, C. M. T. Spahn, K. Y. Sanbonmatsu, and S. C. Blanchard, *Dynamic views of ribosome function: Energy landscapes and ensembles* (Springer, Vienna, 2011).

[10] A. Dashti, G. Mashayekhi, M. Shekhar, D. Ben Hail, S. Salah, P. Schwander, A. des Georges, A. Singharoy, J. Frank, and A. Ourmazd, "Retrieving functional pathways of biomolecules from single-particle snapshots," Nat Commun **11**, 4734 (2020).

[11] G. Mashayekhi, "Manifoldem matlab repository," `https://github.com/GMashayekhi/ManifoldEM_Matlab` (2020).

[12] T. Sztain, S.-H. Ahn, A. T. Bogetti, L. Casalino, J. A. Goldsmith, E. Seitz, R. S. McCool, F. L. Kearns, F. Acosta-Reyes, S. Maji, G. Mashayekhi, J. A. McCammon, A. Ourmazd, J. Frank, J. S. McLellan, L. T. Chong, and R. E. Amaro, "A glycan gate controls opening of the sars-cov-2 spike protein," bioRxiv (2021), 10.1101/2021.02.15.431212.

[13] L. V. D. Maaten, E. Postma, and J. Herik, "Dimensionality reduction: A comparative review," (2009).

[14] C. B. Moler, "Numerical Computing with Matlab," in *Numerical Computing with MATLAB, Revised Reprint* (SIAM, Philadelphia, 2008) Chap. 0.

[15] W. 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).

[16] P. Penczek, "Variance in three-dimensional reconstructions from projections," in *Proceedings IEEE International Symposium on Biomedical Imaging* (2002) pp. 749–752.

[17] P. 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).

[18] P. 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).

[19] H. Y. Liao and J. Frank, "Classification by bootstrapping in single particle methods," Proc IEEE Int Symp Biomed Imaging **2010**, 169–172 (2010).

[20] P. 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).

[21] A. 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.

[22] J. Frank and A. Ourmazd, "Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM," Methods **100**, 61–67 (2016).

[23] K. Pearson, "On lines and planes of closest fit to systems of points in space," Philosophical Magazine **2**, 559–572 (1901).

[24] R. 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).

[25] R. R. Coifman and S. Lafon, "Diffusion maps," Applied and Computational Harmonic Analysis **21**, 5–30 (2006), special Issue: Diffusion Maps and Wavelets.

[26] S. H. Scheres, "Processing of Structurally Heterogeneous Cryo-EM Data in RELION," Methods Enzymol **579**, 125–157 (2016).

[27] A. 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).

[28] T. 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).

[29] J. 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).

[30] P. Buser, *Geometry and Spectra of Compact Riemann Surfaces* (Springer, 1992).

[31] D. 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).

[32] E. 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.

[33] E. Seitz, "Repository: Cryo-em synthetic continua (initial release)," Zenodo , 10.5281/zenodo.3561105 (2019).

[34] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide protein data bank," Nature Structural Biology (2003).

[35] D. Goodsell, "Pdb-101 molecule of the month: Hsp90," (2008).

[36] M. 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).

[37] L. Schrödinger, "The pymol molecular graphics system," (2015).

[38] E. Seitz, "ManifoldEM: Direct manifold subspace analysis (DMSA)," Zenodo , 10.5281/zenodo.4987719 (2021).

[39] A. 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).

[40] H. Cundy and A. Rollet, *Mathematical Models*, 3rd ed. (Tarquin Pub., 1989) pp. 242–244.

[41] J. R. Munkres, *Topology*, 2nd ed. (Prentice Hall, Incorporated, 2000).

[42] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. (Dover, 1972) pp. 771–802.

[43] D. Lay, S. Lay, and J. McDonald, *Linear algebra and its applications*, 5th ed. (Pearson, 2016).

[44] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," Anesth Analg **126**, 1763–1768 (2018).

[45] G. 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).

[46]S. H. Scheres, "RELION: implementation of a Bayesian approach to cryo-EM structure determination," J Struct Biol **180**, 519–530 (2012).

[47]A. Rohou and N. Grigorieff, "CTFFIND4: Fast and accurate defocus estimation from electron micrographs," J Struct Biol **192**, 216–221 (2015).

[48]J. Stewart, *Calculus*, 6th ed. (Brooks/Cole, 2008).

[49]F. 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).

[50]S. Omohundro, "Five balltree construction algorithms," (2009).

[51]H. 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).

[52]J. D. Gardiner, J. Behnsen, and C. A. Brassey, "Alpha shapes: determining 3D shape complexity across morphologically diverse structures," BMC Evol Biol **18**, 184 (2018).

[53]K. Bellock, N. Godber, and P. Khan, "Python alpha shape repository," (2021).

[54]S. Gillies *et al.*, "Shapely: manipulation and analysis of geometric objects," (2007–).

[55]P. 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).

[56]S. 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).

[57]R. 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).

[58]E. 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).

[59]G. 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).

[60]W. 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).

[61]S. 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).

[62]C. 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).

[63]X. Zhang and Z. H. Zhou, "Limiting factors in atomic resolution cryo electron microscopy: no simple tricks," J Struct Biol **175**, 253–263 (2011).

[64]M. 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.

[65]F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," Information Sciences **328**, 26–41 (2016).

[66]J. 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 using a path collective variable," (2021), arXiv:2102.02077 [q-bio.BM].

[67]E. 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.

[68]J. 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).

[69]E. 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).