Analysis of Conformational Continuum and Free-energy Landscapes from Manifold Embedding of Single-particle Cryo-EM Ensembles of Biomolecules

Evan Seitz

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Analysis of Conformational Continuum and Free-energy Landscapes from Manifold Embedding of Single-particle Cryo-EM Ensembles of Biomolecules

Evan Seitz

Biological molecules, or molecular machines, visit a continuum of conformational states as they go through work cycles required for their metabolic functions. Single-molecule cryo-EM of suitable *in vitro* systems affords the ability to collect a large ensemble of projections depicting the continuum of structures. This information, however, comes buried among typically hundreds of thousands of unorganized images formed under extremely noisy conditions and microscopy aberrations. Through the use of machine-learning algorithms, it is possible to determine a low-dimensional conformational spectrum from such data, with leading coordinates of the embedding corresponding to each of the system's degrees of freedom. By determining occupancies—or free energies—of the observed states, a free-energy landscape is formed, providing a complete mapping of a system's configurations in state space while articulating its energetics topographically in the form of sprawling hills and valleys. Within this mapping, a minimum-energy path can be derived representing the most probable sequence of transitions taken by the machine between any two states in the landscape. Along this path, an accompanying sequence of

3D structures may be extracted for biophysical analysis, allowing the basis for molecular function to be elucidated. The ability to determine energy landscapes and minimum-energy paths experimentally from ensemble data opens a new horizon in structural biology and, by extension, molecular medicine.

The present work is based on a geometric machine-learning approach using manifold embedding to obtain this desired information, which has been shown possible on two experimental systems—the 80S ribosome and ryanodine receptor—through a previously-established framework termed ManifoldEM. First, this framework is incorporated into an advanced graphic user interface for public release, and augmented with a new method, POLARIS, for determining minimum-energy pathways. ManifoldEM is next applied on two new systems: vacuolar ATPase and the SARS-CoV-2 spike protein, and for both systems, several novel aspects of the machine's function are observed. During this exposition, critical limitations and uncertainties of the framework are also presented, as have been found throughout its extended development and use. However, in the absence of ground-truth data, testing and validation of ManifoldEM is infeasible. As recourse, a protocol is next proposed for generating simulated cryo-EM data from an atomic model subjected to multiple conformational changes and experimental conditions, with several Hsp90 synthetic ensembles generated for analysis by ManifoldEM. Guided by results of these ground-truth studies, new insights are made into the origin of longstanding ManifoldEM problems, further motivating and informing the development of a new, comprehensive method for correcting them, termed ESPER. The ESPER method operates within the ManifoldEM framework and, as will be shown using both synthetic and experimentally-obtained data, ultimately results in substantial improvements to the previous work. Finally, numerous recommendations are laid out for guiding future work on the ManifoldEM suite, particularly aimed at its next public release.

Table of Contents

List of Charts, Graphs, Illustrations

Acknowledgments

There are several individuals who deserve acknowledgement, and my thanks, for their contribution to this endeavor. First of all, I am extremely grateful to my supervisor and mentor Joachim Frank for providing his wisdom, guidance, inspiration and unceasing support throughout this journey; the culmination of which has had an indelible influence on my scientific character and goals. I am further indebted to Peter Schwander and Rommie Amaro for providing their crucial knowledge, ingenuity and enthusiasm throughout the bulk of this project. As well, I am thankful to Liang Tong and John Hunt for their counsel on this thesis and outstanding tutorship even before its conception. My thanks also to so many members who worked on individual pieces of this work, or offered guidance along the way. **The ManifoldEM team:** Joachim Frank, Suvrajit Maji, Peter Schwander, Ali Dashti, Hstau Liao, Abbas Ourmazd, Ghoncheh Mashayekhi, Sonya Hanson, and our alpha testing group. **The POLARIS method:** Joachim Frank, Debashish Chowdhury, and Annwesha Dutta. **ManifoldEM analysis of V-ATPase:** Joachim Frank, John Rubinstein, Rommie Amaro, Suvrajit Maji, Clara Altomare, and Francisco Acosta-Reyes. **ManifoldEM analysis of S Protein:** Joachim Frank, Rommie Amaro, Jason McLellan, Francisco Acosta-Reyes, Suvrajit Maji, Lillian Chong, Terra Sztain and the Amaro lab, Abbas Ourmazd, and Ghoncheh Mashayekhi. **The Synthetic Continuum framework:** Joachim Frank, Francisco Acosta-Reyes, and Peter Schwander. **The ESPER method:** Joachim Frank, Peter Schwander, Francisco-Acosta Reyes, Suvrajit Maji, and Hstau Liao. My gratitude goes out also to **all Frank lab members**, especially Harry Kao, Masgan Saidi, Zuben Brown, Bob Grassucci, Sergey Vorobiev, and Jeevan GC; and to the **GSAS faculty and staff**, notably Ron Prywes and Sarah Kim Fein. Finally, I take this opportunity to express my appreciation to my family and friends for their continued support and weathered ear, and to my wife Kaitlin and our newborn son Felix for lighting the way.

For Debbie and Greg

Preface

This dissertation is submitted for the Doctorate of Philosophy degree at Columbia University, and has been written to fulfill the graduation requirements of the Graduate School of Arts and Science's Department of Biological Sciences. The research described here was conducted under the guidance of Professor Joachim Frank in the Department of Biochemistry and Molecular Biophysics and the Department of Biological Sciences at Columbia University, between January 2018 and January 2022.

Over that time period, research conducted was adventurous, forging ahead into the unknown, winding and turning with each new discovery made, and ultimately clearing a path entirely unpredictable from the start. The work was difficult and uncertain, but interspersed with moments of realization that were nothing short of exhilarating for all those involved. The aim of this thesis is to recapitulate that journey; aiming not just for a regurgitation of previous publications, but to form a condensed narrative of our timeline, while building strong intuition for concepts and results along the way.

To the best of my knowledge, this work is original, with references made to previous efforts. Parts of this work have also been presented in several publications, with each instance marked. These efforts owe their existence to the contributions of many individuals, with specific insights provided in acknowledgements.

Evan Seitz

January 2022

Chapter 1: Function from Structure

1.1 Molecular Machines

Molecular machines, consisting of assemblies of proteins or nucleoproteins, are an essential component of all living organisms and play a central role in every one of their cellular processes (Frank, 2011; Roux, 2011). Due to the scale of these machines within the cellular environment, they are significantly influenced by both random collisions between particles (thermal forces) and any deterministic forces present, such as electrostatic interactions (Phillips et al., 2008, chap. 5). As a result, these machines have evolved to efficiently transform random environmental fluctuations and energy supplied by chemical reactions (e.g., ATP or GTP-hydrolysis) into desired outputs necessary for metabolism (Phillips et al., 2008, chap. 5). Examples of metabolic functions are: tissue construction and repair, secretion and transportation of hormones and other essential chemicals, cellular signaling, and even the production of other molecular machines. In order for a molecular machine to perform its specialized function, it must take on a range of unique configurations or *conformational states* in an appropriate sequence (Dashti et al., 2014). These states are typically characterized by different spatial constellations of the machine's relatively rigid domains, and can be organized based on local affinity in a *state space* according to the continuous motions of each active domain along a unique coordinate. The number of these independent, collective motions (termed *conformational motions*[1], CMs) defines the intrinsic dimensionality n of the state space.

Specific sequences of states in this n-dimensional space form continuous pathways along which the molecular machine may transform, such that hops between neighboring states take on

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

the notion of time. A subset of these paths—known as the *work cycle*—accounts for the sequence of states required to facilitate metabolic function. Progress along these pathways, however, is not deterministic, since molecular machines are constantly buffeted by the random motions of nearby solvent molecules. Each thermally-driven collision can reversibly alter the machine's current configuration, delaying or reversing progress along an intended path. Without any impetus—in thermal equilibrium—molecular machines would follow Brownian pathways on the state space, and rarely ever perform their function. To consistently produce useful work, these machines must impose a bias within their thermally-driven environment.

To understand this process, the role of electrostatic interactions must first be accounted for, which act internally among all atoms in the machine's structure, and externally between those atoms and the atoms in the surrounding solvent. Based on the arrangement of atoms in each state, certain states are more energetically favored than others. This preference can be defined for each state by a *free-energy* value—the amount of energy available to perform thermodynamic work—such that a low free energy represents a highly favorable conformation. When mapped onto the state space, the collection of these values forms a *free-energy landscape* (Figure 1). In this landscape, the sprawling layout of free-energy hills and valleys characterizes the machine's navigational probabilities, with deep wells representing highly favorable conformations and ridges acting to constrain the transitions between them.

As a result, the geometric features of the landscape act to channel thermal influences, such that the molecular machine is probabilistically driven into nearby free-energy basins. While rare sequences of thermal fluctuations may propel the machine from a valley over a ridge into a "forward" basin, other sequences also exist to instead drive the machine into a "backward" basin, with the probability of transition along either of these available routes (e.g., "forward" or

"backward" in 1D) dependent on the highest free-energy peak (Munro et al., 2009) and the total integrated free energy along each path's diverse range of intermediate states. An animated schematic of this process—depicting a machine's current state as a marble randomly perturbed by orthogonal 2D forces in a bowl-like basin—has been provided in Movie A2 in Appendix A. Without intervention, all motions are reversible in the thermal environment, such that on the whole, no useful work is performed (Feynman et al., 1965, chap. 46).

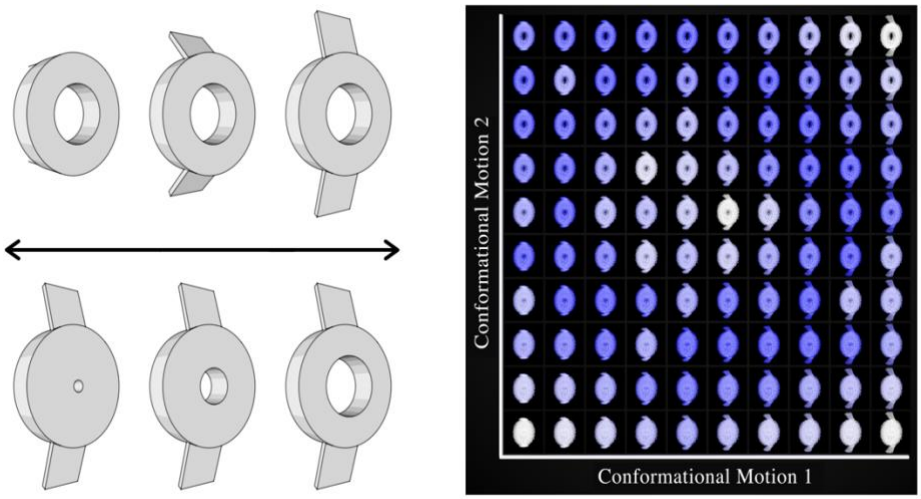


Figure 1: Mouth-wings toy model schematic. Free-energy landscape for a toy model capable of independently exercising two conformational motions: (i) flapping of "wings" and (ii) opening/closing of "mouth". The underlying state space is described by all pairwise combinations of these 10×10 domain motions, such that 100 states are realized in total. Color has been applied to represent the energetics of each state, demonstrating presence of an energetically-favored trough along which consecutive state transitions are most probable. A more elaborate toy model exercising $n = 5$ CMs is visualized in Movie A1 in Appendix A.

Thus, while thermal energy alone is sufficient for observance of all possible conformations at a given temperature, it is not sufficient to drive work cycles, since these pathways must be executed unidirectionally in contrast to the randomness of Brownian motion. When coupled with extraneous energy supply (by means of electrochemical gradients or ATP hydrolysis, for example), a molecular machine can impose this bias in a specific direction on the state space—understood as

a shifting of the free-energy landscape topography—towards an intended state. For example, in order to execute translocation during the ribosomal work cycle, free-energy ridges are erected by interaction with EF-G (which triggers release of energy upon hydrolysis of GTP; Noller et al., 2017) that prohibit reverse motions into a previously-accessible valley. As a result, this reaction constrains thermal fluctuations to impose a preferred directionality to the ribosomal work cycle.

This energetically-driven diffusive process allows molecular machines to exploit thermal and mechanistic forces to predictably transform across specific sequences of states in time (Phillips et al., 2008). As well, such a transformation can occur through a multitude of structural pathways between any two configurations in the free-energy landscape (Whitford et al., 2011; Figure A1 in Appendix A). These pathways are defined by long-lived transit states, called *macrostates*, that correspond to a local free-energy minimum of the system, and are often associated with a biomolecule's function or as an element of a more complex, multistep functional process (Kolimi et al., 2021). An exploration and analysis of pathways linking together macrostates is thus essential for understanding the complex dynamics of molecular systems (Wales, 2003).

Additionally, direct comparisons of landscapes can be performed between identical molecular machines imaged under a different experimental condition. As the condition is changed, the probabilities of landscape pathways can change with respect to one another. For example, previously-unfavorable pathways may become more favorable under specific buffer or temperature conditions, allowing for the modulation of a machine's reaction rates based on fluctuating environmental signals. Similarly, for systems with ligands, entirely new regions of the landscape may become available upon binding. This concept can be extended to the study of disease caused by mutation by comparing, for example, the landscapes of machines differing by a single base pair.

So long as the disease-causing mutation is not so deleterious as to altogether change the conformational coordinates of the state space (i.e., drastically altering the type of CMs and wild-type functionality), its presence may result in energetic barriers that either (1) halt the progress of previously accessible work-cycle conformations entirely, or (2) allow completion of the work cycle, but along previously less favorable routes and thus starkly different rates. If the free-energy landscape were known, further studies could be done by adding specific drugs to see how the free-energy landscape "recovers" (Figure 2). Such knowledge can shed light on the dynamics of disease in both time and conformational space, creating a feedback loop essential for the structural understanding of disease and rational drug design. For all of these reasons and more, an accurate estimation of the free-energy landscape and corresponding atomic-coordinate structures for molecular machines and other biological assemblies is of unparalleled importance in modern structural biology and by extension, molecular medicine.

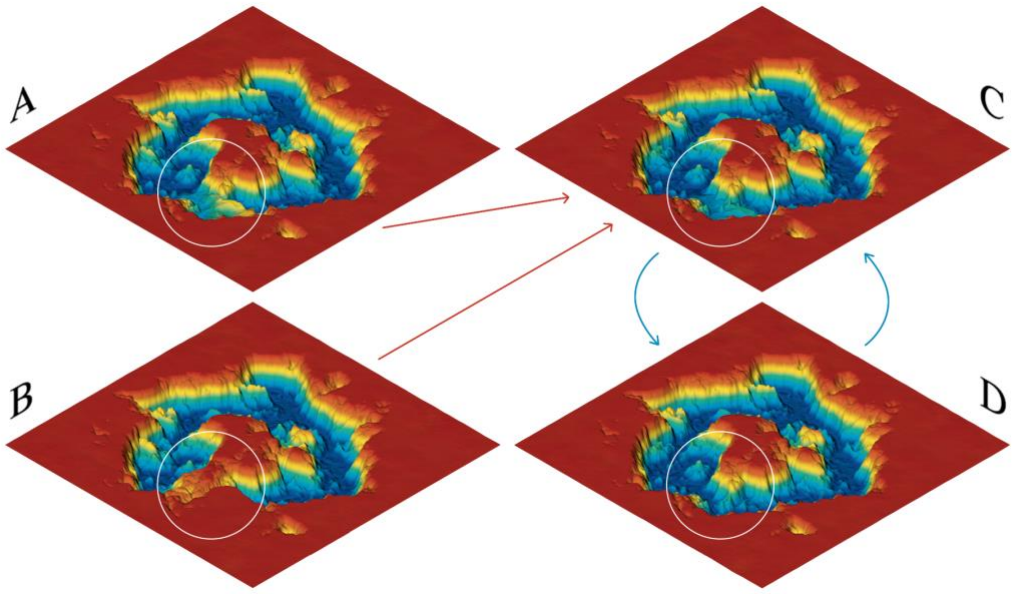


Figure 2: Multiple free-energy landscapes schematic, comparing wild type (WT) and mutated energies of states within the same coordinate system, with regions of free-energy alteration encircled. The molecular machine's work cycle is represented by an energetic trough of minimum free energy (blue), forming an uninterrupted loop on the WT landscape [D]. The free-energy landscape of the nonlethal mutation [A] demonstrates an alteration in the geometry of the

landscape that would increase the time required for the work cycle's completion. The lethal mutation [B] demonstrates presence of an insurmountable barrier inhibiting the completion of the work cycle. Upon addition of a drug [C], changes in the topography can be directly compared to the WT. Further, in each of these encircled areas, the sequence of 3D atomic-coordinate structures occupying these regions can be directly analyzed and compared.

1.2 Capturing the Continuum

For several molecular machines such as the ribosome, it has been shown possible to develop *in vitro* systems with all components necessary to run work cycles delivering the product; in this instance a protein. In order to construct the free-energy landscape and corresponding molecular structures at any point in the landscape from such data, information on each of the conformational states must be acquired. Several methodologies exist to visualize macromolecular conformations, with each approach having spawned an entire branch of structural research with thousands of contributions. Most significantly are the methods of electron crystallography (Zou, 2006), X-ray crystallography (Jaskolki et al., 2014), nuclear magnetic resonance (NMR; Anklin, 2019), and single-particle cryogenic electron microscopy (cryo-EM; Frank, 2006; 2016); with the latter of these recently highlighted by the award of the 2017 Chemistry Nobel Prize (Frank, 2017). Ideally, the goal of applying these techniques is to form a three-dimensional (3D) model of a macromolecule in its entirety, at the highest possible resolution. However, not all are conducive for isolating structures from within a heterogeneous ensemble.

In the case of electron and X-ray crystallography, crystallization of the sample constrains the macromolecular domains by crystal contacts. As a result, the molecular machine only takes on the lowest-energy conformation allowable by the lattice. Since only a small range of physiologically-relevant conformations are observable by these techniques (Frank, 2006), and because of the overwhelming evidence that molecular machines actually exist in a continuum of

structural configurations (Frank, 2018), both of these methods oversimplify the complexity of the molecular machine and thus greatly limit our understanding of its function.

In contrast, single-particle cryo-EM overcomes the necessity of conformational homogeneity by making use of the innate heterogeneous information in a sample, for the purpose of elucidating all structures contained within. This representation is achieved by capturing large numbers of high-resolution two-dimensional snapshots (i.e., projections) from suitable ensembles of macromolecules in an *in vitro* suspension occupying different states (Frank, 2016). From these ensembles, cryo-EM has proven capable of providing an entire inventory of *in vitro* conformers of a molecule (Scheres, 2007; Agirrezabala, 2012; Frank 2017). When the number of snapshots is sufficiently large—typically several hundred thousand—they capture virtually the entire range of conformations, which coexist due to thermal fluctuations in the equilibrated sample prior to vitrification (Whitford et al., 2011). During vitrification, molecules are frozen at a rate assumed to be faster than the overall reconfiguration time of the system. As a result, the ensemble closely approximates the Boltzmann distribution of states immediately prior to being frozen, such that the relative number of sightings in each of these states can be translated into changes of free energy (Fischer et al., 2010; Agirrezabala et al., 2012). Thus, under assumption of thermodynamic equilibrium, cryo-EM allows a molecular machine's free-energy landscape to be obtained from an experiment. Further, once properly organized, the 2D snapshots can be combined to form a collection of 3D Coulomb potential maps, forming a quasi-continuum of structures spanning the machine's state space.

Still, single-particle cryo-EM is not without competition. A promising new route has emerged for obtaining diffraction patterns of biomolecules exhibiting structural variability using an X-ray free-electron laser (XFEL). Single-particle imaging with an XFEL can be used to

determine biomolecular structures at high resolution and ultra-short timescales (Spence et al., 2012). As well, it has the advantage that samples can be examined at room temperature without requiring crystallization (Peplow, 2017). In a recent study, Hosseinizadeh et al. (2017) demonstrated that this technique affords the ability to obtain information beyond static structures, and were able to capture conformational changes of the Coliphage PR772 virus, including reorganization of the internal membrane and genome, the growth of a tubular structure from a portal vertex, and the release of the genome. So far, the proposed technique has several caveats, such as requiring a large number of single-particle diffraction patterns for high-resolution imaging of a single conformation, which is needed to overcome the low XFEL signal levels (Poudyal et al., 2020; Cruz-Chú et al., 2021). While these findings show promise, bottlenecks must still be addressed, such as the sample delivery, which gives a very low hit rate (i.e., number of samples).

Nuclear magnetic resonance spectroscopy is also capable of depicting the conformational variations of molecules in an ensemble, but due to faster magnetization-relaxation rates in large structures—which allow less time to detect signal—it is limited in the number of atoms of a molecule that can be studied. As is, NMR is mainly used for small proteins with molecular weights typically in the range of 50-100 kDa, but given additional techniques—such as transverse relaxation optimized spectroscopy and deuteration of proteins—complexes have been resolved up to ten times this size limit (e.g., 900 kDa; Fiaux et al., 2002). As a measure of scale, the human ribosome (resolvable by cryo-EM) has a molecular weight of 4.5 MDa, or five times the NMR upper bound. While such upper-limit restrictions do not exist for single particle cryo-EM, it is of interest to note that the typical range of an NMR experiment (up to around 50 kDa) is at the lower limit of molecular weight applicable in cryo-EM (Murata and Wolf, 2018).

After acquisition of a foundational atomic model by one of these aforementioned methods, advanced computational simulations can also be performed in isolation using molecular dynamics (MD; Hospital et al., 2015). MD simulations seek to apply simple approximations of quantum mechanics based on Newtonian physics to simulate atomic motions while reducing computational complexity (Durrant and McCammon, 2011). Energy terms in the corresponding equations are parameterized to fit quantum-mechanical calculations and experimental data—such as the ideal stiffness of a chemical bond or proper van der Waals atomic radii—which are collectively called a "force field" (Durrant and McCammon, 2011). Overall, a number of studies have shown good agreement between computational and experimental measurements (such as obtained using NMR) of macromolecular dynamics (Durrant and McCammon, 2011). Later in this work, a recent comparison (Sztain et al., 2021) of the results from MD with those from cryo-EM will be described, which also demonstrated a good agreement.

The utility of MD simulations is limited by two principal challenges (Durrant and McCammon, 2011): (i) force fields require further refinement, and (ii) high computational demands. For the latter, this restriction routinely prohibits simulations greater than a microsecond in length, such that a relatively small system (approximately 25,000 atoms) running on 25 processors takes several months to complete (Durrant and McCammon, 2011). Much like NMR, this last point then further sets restrictions on the size of the system, and thus also the complexity of its interactions. A range of techniques have been developed to counter this limitation, including coarse-grained MD (Arkhipov et al., 2006) and weighted ensembles (Huber et al., 1996; Zhang et al., 2010), with each offering tradeoffs for several unavoidable drawbacks.

Ultimately, it is clear that the choice of experimental technique—from XFEL, to NMR, or single-particle cryo-EM—must be chosen judiciously based on the sample and target of study,

with further considerations given for a subsequent application of MD. Of these, cryo-EM does, however, cast a wide enough net to cover the dynamics of a majority of the larger systems, to a degree of completeness that is unfeasible by other means. Given its unmatched capability, the cryo-EM imaging process will next be examined in some detail, with the analysis of cryo-EM data obtained by those means serving as the primary focus of this work.

1.3 Imaging by Cryo-EM

In application, the cryo-EM imaging process is quite complex—with a history spanning decades of engineering and research—while here, only a brief synopsis is provided in order to touch on some of the most important aspects. A more detailed account of the imaging process has been provided by Frank (2006) and Reimer and Kohl (2008), among many others. Single particle cryo-EM samples are typically prepared by applying a small volume of purified macromolecule in solution to a hydrophilic grid, which is then blotted and plunged into a liquid cryogen (e.g., liquid ethane) for vitrification (Owens et al., 2021, chap. 12). The rapid cooling rate during vitrification prevents the water from turning into crystalline ice, which—by means of a volume change—normally damages the biological specimens (Frank, 2006). The role of liquid cryogen is also essential for sustaining the initially-hydrated machine's atomic configurations, and is able to preserve all interior features as they exist in an aqueous environment, up to atomic resolution (Frank, 2006). The vitrified sample is then loaded into the Transmission Electron Microscope (TEM) to begin data collection.

During data collection in the TEM bright-field mode, the specimen is irradiated with an electron beam of uniform current density, with an acceleration voltage of typically 100 keV or higher (Reimer and Kohl, 2008). In route to the image plane, electrons either pass through the

sample (unscattered) or directly interact with the sample, and scatter either *elastically* or *inelastically*. The electrons which scatter over a narrow angular distribution form a diffraction pattern of the specimen in the focal plane of the objective lens, where each scattering angle corresponds to a spatial frequency k as a result of a periodic spacing in the specimen (Reimer and Kohl, 2008). In the absence of aberrations introduced by the imaging process, the amplitude distribution of the electron wave in the back focal plane is the Fourier transform of the specimen transparency (Reimer and Kohl, 2008). For weakly scattering objects—which is the case for most biological samples—the weak-phase object approximation (WPOA; Vulovic et al., 2014) holds, which implies a decomposition of the wave behind the sample into a primary (unscattered) and a scattered wave (Frank, 2006). Because of the dominance of the linear term in the scattered wave amplitude in bright-field mode, linearity between the projected object potential and the image intensity is ensured (Frank, 2006). As a result—and simplifying this process immensely—the observed image intensity (with magnification) on the final image plane is the Fourier transform of the (complex) electron-wave-amplitude distribution in the back focal plane.

En route from the back focal plane, the complex electron wave is focused by a magnetic lens system (in several stages by the projective lenses) onto the final image plane. This simple depiction of the imaging process is incomplete, however, since there is no image contrast in the absence of aberrations. In actuality, the phases of the electrons are altered along the way to the image plane due to wave aberrations and defocusing by the TEM. The entire rationale of image formation in the bright-field mode is based on the resulting interference of the (twice) phase-shifted elastically-scattered electrons and the primary beam, which creates *phase contrast*. Strictly given the WPOA linear relationship, the contrast transfer function (CTF) can be used to relate the projected object potential to the observed image contrast.

The CTF follows an oscillatory pattern between negative and positive transfer as a function of spatial frequency k, with these oscillations creating zero crossings where no contrast is transferred and information is inevitably lost. The approximate form of the CTF is described by the expression $H(k) = \sin \chi(k) - A \cos \chi(k)$, where A denotes the fraction of amplitude to phase contrast (Frank, 2006). Here, $\chi(k) = (-\pi \Delta z \lambda k^2 + ½\pi C_s \lambda^3 k^4)$ is called the *wave aberration function*; Δz is the defocus value; λ is the wavelength of the electron, as set by the voltage of the TEM; $(½\pi C_s \lambda^3 k^4)$ is the phase shift introduced by the spherical aberration of the lens; and $(-\pi \Delta z \lambda k^2)$ is the phase shift introduced by defocusing (Frank, 2006). Thus, given knowledge of the CTF modifying an image, this relationship can be applied to recover a more accurate representation of the original 3D Coulomb potential distribution. Due to the presence of zero crossings, however, capturing images with only a single CTF curve is insufficient. Instead, experimentalists typically alter Δz over a wide interval, which acts to enhance or suppress different features in the image, and alter the locations of zero crossings. Through image processing, contributions to the reconstruction from images with CTFs for different defocus settings are integrated to effectively cover all spatial frequencies.

When electrons are recorded in the image plane, they are registered by metal-oxide semiconductor (CMOS) cameras ("direct detectors"), resulting in high-contrast images recorded at a high-frame rate (i.e., movies; Owens et al., 2021, chap. 12). These frames are obtained over extremely short exposure times—up to camera saturation—to reduce radiation damage, resulting in very low-contrast images (Owens et al., 2021, chap. 12). As a result of the collision of the electron beam with the sample, particles also undergo a motion which blurs the images, resulting in a loss of information (Owens et al., 2021, chap. 12). Hence, movies have to be corrected for distortions caused by both the camera and the induced motion of the particles during observation

(Owens et al., 2021, chap. 12). This process involves an intricate pipeline, featuring estimation of several parameters for acquiring each image, including among others (Owens et al., 2021, chap. 13): gain of the camera, beam-induced movement, aberrations of the microscope—most importantly the defocus—for each micrograph, the orientation of each projection, and changes in magnification.

The beam-induced motion can be corrected via automation schemes (e.g., MotionCorr; Li et al., 2013), which also incorporate dose-weighting to account for the effect of radiation damage on each frame of the movie, to ultimately output an aligned stack of frames as well as their sum (i.e., a micrograph). The micrograph can then be used to estimate the CTF, which affects the phase and amplitude of its Fourier transform (Wade, 1992). Automated programs (such as CTFFIND; Rohou et al., 2015) fit CTFs to the Thon rings visible in the power spectra of patches of the micrographs (obtained via Fourier transforms of the patches), which are used to then estimate the defocus, size of axial astigmatism, and astigmatism angle (Owens et al., 2021, chap. 12). A theoretical CTF curve can then be produced to CTF correct the Fourier transform of the patch, followed by an inverse Fourier transform to recover the CTF-corrected image. As a note, in practice, the Fourier transforms of images with different CTFs are merged using a Wiener filter (Frank, 2006).

Particle picking is next performed to isolate objects of interest (i.e., macromolecular particles) in the micrograph. For this process, there are an assortment of approaches, including visual assessment, template-matching, and neural networks (Owens et al., 2021, chap. 12). Regardless of technique, these particles are then extracted (boxed) from the micrographs to form a stack of images. 2D classification procedures, such as the RELION maximum-likelihood method (Scheres, 2012), can next be used to group particles with similar shape, and clean the data set by

removing noisy or suboptimal images and those featuring aberrant particles from the stack. A *de novo* 3D initial model is next generated after 2D classification (e.g., as obtained via a stochastic gradient descent algorithm; Punjani et al., 2017), to assign each image—understood as a 2D projection—a set of angles that define its viewing direction on the angular space (S^2). The collection of these microscopy and alignment parameters are conventionally stored for each image in an "alignment file" associated with the corresponding image stack. Along this pipeline, there are numerous other (optional) preprocessing techniques available to further refine and optimize these procedures. As all of the parameter choices during preprocessing arise from noisy environments, these estimations are prone to error. The importance of accurate estimations cannot be understated, with certain errors (such as significantly misassigned angular assignments) having an ability to critically undermine all procedures that follow.

1.4 Reconstruction Algorithms

By performing a cryo-EM experiment, all projections of a molecular machine may be obtained for each conformation in its state space. However, this information arrives buried among typically hundreds of thousands of images in a completely unorganized form. Thus, the way to utilize the data from a cryo-EM experiment is not easy, with this problem having inspired the development of numerous competing approaches. As described in the previous section, the first step in any such approach is to preprocess the data, ultimately resulting in an image stack and corresponding alignment file housing particle and microscopy information for each snapshot.

The conformational relationships between the molecules represented by the aligned images, however, are unknown, with numerous approaches developed in recent years that attempt to describe them. Of much prominence is the method of Bayesian estimation (Scheres, 2007;

Scheres, 2012; Punjani et al., 2017), which aims to organize the images based on their innate similarities into a limited set of subsets (classes) by use of maximum-likelihood classification. The images belonging to each class are then used to reconstruct a 3D Coulomb potential map[2] (Scheres et al., 2007; Scheres, 2012; Grigorieff, 2016; Punjani et al., 2017), which can be modeled to yield atomic coordinates. While the set of such volumes—one for each class—is indicative of regions of the multidimensional state space (i.e., macrostates), as states they are fundamentally disjointed, since their relative positions with respect to one another are completely unknown. Recently, the Bayesian approach has been adapted to extract free-energy paths between stable classes (Giraldo-Barreto et al., 2021), but by only singling out most-likely pathways, the technique is unequipped to describe the landscape's multidimensional form.

To reconstruct a molecular machine's quasi-continuum state space, alternative techniques have been proposed that aim to map the cryo-EM images onto a latent coordinate space that represents each image's most fundamental relationship to all others. Of these, *dimensionality reduction* has gained considerable traction. By means of this approach, a suitable embedding can be constructed that maps the data points embedded in a high-dimensional manifold Ω into a low-dimensional Euclidean space, creating a foundation for the approximation of the characteristic properties of the molecular machine's state space. In this state space, the occupancy of each state is directly tied to its free energy via the Boltzmann relationship. The resulting n-dimensional state space, populated with experimentally-determined energies at each state, constitutes the machine's free-energy landscape.

As manifolds are encountered in many domains of mathematics, science and engineering (Maaten et al., 2009), the aim of dimensionality reduction has been widely pursued and given rise

[2] For the remainder of this document—and as is common in the cryo-EM community—the term "volume" will be used interchangeably with "3D Coulomb potential map" resulting from 3D reconstruction in cryo-EM.

to a number of well-established techniques to analyze large and complex data sets. Representing data points on Ω in terms of leading eigenvalues and eigenvectors gives valuable insights into its intrinsic structure, with these relationships having been well studied in the context of spectral geometry (Craioveanu et al., 2001). In the analysis of cryo-EM data, both linear (Liu et al., 1995; Penczek, 2002; Penczek et al., 2006a,b; Liao and Frank, 2010; Penczek et al., 2011; Schwander et al., 2014; Punjani and Fleet, 2020) and nonlinear (Dashti et al., 2014; Schwander et al., 2014; Frank and Ourmazd, 2016; Dashti et al., 2020) dimensionality-reduction methods have been applied, primarily principal component analysis (PCA; Pearson, 1901) and diffusion maps (DM; Coifman et al., 2005; Coifman and Lafon, 2006), respectively. Both approaches allow an analysis of the data points in Ω as embedded in $\mathbb{R}^N$, whose entries are the first N eigenvectors of the respective graph; noting that only a leading subset of these are needed for retrieval of the conformational spectrum.

In the PCA approach, eigenvectors are obtained from the covariance matrix, whereas DM approximates the eigenfunctions of the Laplace-Beltrami operator (LBO) on Ω, sampled at the given data points. Some techniques are not so easily classified, however, such as the method of Laplacian spectral volumes (Moscovich et al., 2020), which relies on both linear and nonlinear dimensionality reduction. The application of these methods can further be classified based on their type of data input: generating embeddings from either 2D projections straight from a cryo-EM experiment (i.e., the *PD-manifold approach*; Dashti et al., 2014; 2020) or 3D volumes which have been reconstructed from those projections (Scheres, 2016; Punjani et al., 2017; Nakane et al., 2018; Zivanov et al., 2018; Moscovich et al., 2020). It is expected that these competing manifold embedding methods should deliver equivalent information when cross-validated, and likewise for

alternative techniques, which extend now into work using deep neural networks (e.g., cryoDRGN; Zhong et al., 2021).

In the following chapter, the aforementioned PD-manifold approach is explored in greater detail, as it is one of the founding frameworks in the field with a growing number of studies performed. This approach was first introduced by Dashti et al. in 2014 and is now broadly termed ManifoldEM (Dashti et al., 2020; Mashayekhi, 2020). Results from previous ManifoldEM studies on biological systems—including the ribosome (Dashti et al., 2014), the ryanodine receptor (Dashti et al., 2020) and most recently, the SARS-CoV-2 spike protein (Sztain et al., 2021)—have proven its viability and its potential to provide new information on the functional dynamics of molecules. Further, there is still room for innovation and growth, with new and future advancements explored.

Chapter 2: The ManifoldEM Framework

2.1 Conceptual Outline

ManifoldEM is one of several dimensionality-reduction techniques that is most distinguished in its analysis of individual manifolds, each corresponding to a set of projections grouped together in close proximity in angular space. Each of these sets can be assigned to a unique *projection direction* (PD $\subset S^2$), having the condition that changes in the appearance of images therein due to angular shifts are negligible compared to conformational change. Similarity between molecules belonging to the same PD, but having 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 under thermal equilibrium are arranged—by virtue of their similarities—according to the extent of continuous motions of the molecule's domains.

The density of points ("occupancy") in this state space can be related to the free-energy landscape sampled by the system through the Boltzmann factor. Hence, by representing the high-dimensional data of a molecular machine in a low-dimensional subspace with intrinsic dimension equal to the number of independent molecular degrees of freedom, a foundation is created for the estimation and analysis of the machine's free-energy landscape. Within this landscape, the loci of minimum energy form the most occupied, statistically-dominant trajectory traversed by the machine, as seen from a given PD. Of most importance, the ManifoldEM framework is founded on the assumption that each of these PDs carry the same conformational spectrum, and thus free-energy profile. This information—contained in each PD—can then be compiled across S^2 on the basis of the similarity metrics between conformational motions of adjacent PDs. Once this information is properly organized, a consolidated occupancy map and corresponding free-energy

landscape can be produced, from which 3D structures may be extracted along a minimum-energy pathway representing metabolic function (Dashti et al., 2014).

2.2 Previous Studies

The ManifoldEM method was first applied in 2014 on a set of experimental cryo-EM snapshots of 80S ribosomes from yeast (Dashti et al., 2014). During the ribosome's elongation work cycle, it translates genetic information residing on messenger RNA (mRNA) into a polypeptide. During this process, the ribosome repeatedly binds an amino acid (carried by a transfer RNA molecule) to the nascent, elongating polypeptide chain, producing a protein with structure dictated by the sequence of mRNA codons. In eukaryotes, this cycle is facilitated by GTPase elongation factors eEF1A and eEF2. However, even in the absence of these factors, idle motions of the ribosome have been observed under thermal fluctuations alone, which resemble conformational changes of the intersubunit previously observed during mRNA-tRNA translocation (Ermolenko et al., 2007; Agirrezabala et al., 2009). Specifically, as the empty ribosomes idled between two states—with intermediates—suitable data-analytic techniques were expected to show these states and how they were connected.

Following this rationale, the ManifoldEM framework was applied on a set of experimental cryo-EM snapshots of a sample of 80S ribosome of yeast, devoid of mRNA, tRNA and elongation factors. From this data set, a 2D free-energy landscape was constructed (Figure 3), from which the conformational changes were analyzed along a minimum-energy pathway (Dashti et al., 2014). Detailed examination of the 3D conformational changes along this trajectory revealed intersubunit rotation, small subunit head closure, head swivel and L1-stalk closing, and their reversal, which are all domain motions known to be associated with the work cycle of the ribosome during protein

synthesis (Dashti et al., 2014). These observations led to the conclusion that the closed minimum-energy trajectory obtained from snapshots of idling ribosomes corresponded to the protein synthesis work cycle of the ribosome.

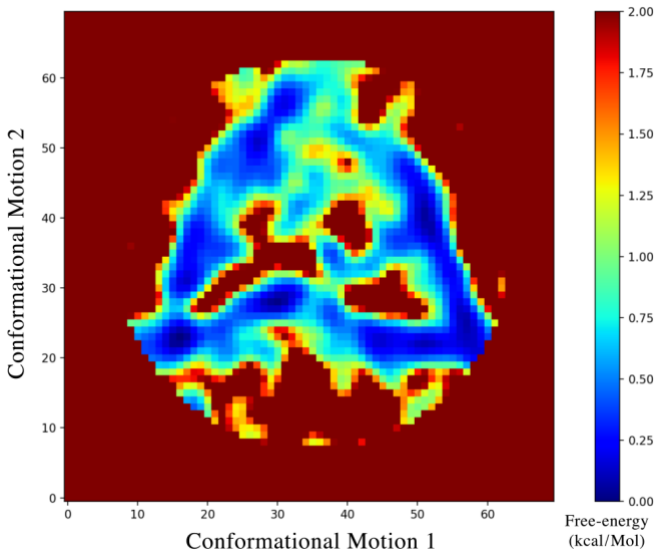


Figure 3: Free-energy landscape obtained from ribosomal cryo-EM data, as determined by NLSA conformational coordinates using ManifoldEM (Dashti et al., 2014).

In 2020, the ManifoldEM technique was expanded upon by Dashti et al. using the ryanodine receptor type 1 (RyR1; Zalk et al., 2015; des Georges et al., 2016) to consider free-energy landscape alterations in the presence of ligands. The ryanodine receptor is a calcium-activated calcium channel critical to excitation and contraction coupling in skeletal muscle (RyR1), heart muscle (RyR2) and brain (RyR3). Its activation core binds activating ligands (Ca^{2+}, ATP and caffeine are well-characterized activators), which synergistically activate the channel by inducing a rotation of its activation core and pore opening (des Georges et al., 2016; Dashti et al., 2020). To understand the allosteric mechanism by which rotation of the activation domain induces pore opening, two sets of cryo-EM snapshots were obtained, with and without these three ligands present in the sample, and analyzed via the ManifoldEM framework.

Two free-energy landscapes (i.e., ±ligands) were constructed, one for each data set, with each one defined by the same two conformational coordinates. Broadly, the first of these coordinates expressed the motion of the entire assembly of cytoplasmic shell, activation core and pore, while the second coordinate involved only the motion of the cytoplasmic shell (Dashti et al., 2020). Between these two landscapes, the probability of transition was estimated between iso-conformational points via a master-equation approach (Dashti et al., 2020), rendering the most probable route on the two landscapes to ligand association. The set of 3D structures modeled from the 3D density maps along the path inferred were then compared with results obtained from the same data set by interpolating between discrete RyR1 classes obtained by maximum-likelihood clustering techniques. Major differences between these two methods were reported, including the structural domains involved in motion, as well as the nature, sequence and range of motions necessary for RyR1 functionality. Overall, the interpolated results, which included a traversal through territory devoid of experimental data, failed to reveal convincing conformational changes. While linear interpolation between volumes is often anticipated to provide a qualitative understanding of function, the comparison with ManifoldEM revealed flaws in this philosophy.

2.3 Overview of Method

The following sections provide a detailed summary of steps employed during the ManifoldEM framework for determination of cryo-EM conformational continua (Dashti et al., 2014; Dashti et al., 2020; Mashayekhi 2020). The first of these steps assumes previous acquisition of an experimental data set, in the form of image stack, alignment file, and related microscopy parameters.

Division into PDs. ManifoldEM first uses the orientation for each snapshot in the alignment file to bin images within local clusters defined by the boundaries of tessellated regions (i.e., a mosaic pattern of many virtually identical shapes fitted together; Lovisolo et al., 2001) in angular space (S^2). The size of the shapes used within this spherical tessellation is determined by the angular aperture width. As a result, snapshots are partitioned into a collection of orientational apertures of fixed diameter (the so-called projection directions, PDs), centered around regular grid points in angular space. This organization allows the group of images to be analyzed as sets independently from all other PDs. For sufficiently small apertures, variations in the images due to changes in orientation are in first approximation decoupled from variations of conformations.

Embedding PDs. The challenge is that the relationship among the N images within a given PD—each represented as a P pixel array—requires an analysis of the point cloud formed in vector space $\mathbb{R}^P$. To make this problem tractable, ManifoldEM uses dimensionality reduction to reduce the data set into a representation existing in a much lower-dimension space capable of accounting for a molecular machine's most essential, collective motions. The geometric structure formed by this collection of images is an n-dimensional manifold Ω embedded in a high-dimensional Euclidean space $\mathbb{R}^P$, with an intrinsic dimension n equal to the number of the system's independent molecular degrees of freedom. To achieve this embedding, a distance matrix is first constructed for defining the pairwise similarity between all N images. First, all images are standardized and low-pass filtered. However, since the collection of images has a range of different CTFs, they are not directly comparable. To address this issue, a "double-filtering kernel" is applied (Dashti et al., 2014), whereby the Fourier transform of each image is multiplied by its CTF to approximately equalize all data for rational pairwise analysis. After this modification, the Euclidean distance (Li et al., 2011) between any two images $X = (x_1, x_2, \dots, x_P)$ and $Y = (y_1, y_2, \dots, y_P)$ is defined by

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

where x_i and y_i respectively denote the intensities at pixel i in images X and Y (each having P pixels). The pairwise distances form a symmetric $N \times N$ distance matrix $\mathbf{D}$, where a single row represents the distance of the corresponding row-indexed image to each column-indexed image.

Using this matrix as input, the nonlinear-dimensionality method diffusion maps (DM; Coifman and Lafon, 2006) is next applied to generate the corresponding conformational manifold for each PD. To begin, an isotropic Gaussian kernel is applied to these distances to create a real, symmetric similarity matrix

$$A_{ij} = \exp(-D_{ij}^2 / 2\varepsilon) \quad (2)$$

The similarity matrix $\mathbf{A}$, calculated using a suitable ε value (i.e., the Gaussian bandwidth), is then divided by a diagonal matrix of its row sums to construct a symmetric, positive semidefinite stochastic Markov transition matrix $\mathbf{M}$. This matrix represents the relative pairwise affinity among all images and is closely related to the normalized graph Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{M}$ (Ferguson et al., 2010), where $\mathbf{I}$ is the identity matrix. Eigendecomposition of the matrix $\mathbf{M}$ is then performed to retrieve an ordered set of N orthogonal eigenvectors ranked by eigenvalue, which define a nonlinear spectral embedding of the data (Coifman et al., 2008).

The Gaussian bandwidth in the above expression has a strong influence on the definition of similarity between images. An optimal Gaussian bandwidth value, henceforth denoted $\varepsilon_\star$, can be determined by a prominent routine—the "bandwidth estimation" method—which uses the correlation dimension as a measure of fractal dimensionality (Grassberger and Procaccia, 1983; Coifman et al., 2008; Ferguson et al., 2010). The ManifoldEM framework provides an automation strategy, which selects $\varepsilon_\star$ using the inflection point of a fitted hyperbolic tangent. At small

Gaussian bandwidths, the system takes on a relatively fine-grained definition of similarity (i.e., data points only see their direct neighbors). Increasing ε transforms this relationship into a more coarse-grained notion of similarity. These notions of similarity govern the behavior of all subsequent steps, and ultimately impact the geometric structure of the resultant manifold embedding.

Particularly, in the limit $\varepsilon \to 0$ and $N \to \infty$, and with an appropriate normalization of the similarity matrix (Coifman and Lafon, 2006), the DM eigenvectors converge to the eigenfunctions of the Laplace-Beltrami operator (LBO; Coifman et al., 2008). The LBO acting on a scalar function f on a compact (closed and bounded; Munkres, 2000) Riemannian manifold is given by

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

where $g = \det(g^{ij})$ and g^{ij} are the components of the metric tensor (Buser, 1992; Jost, 2009). Specifically, the eigenfunctions of the LBO, $\nabla^2 f = \lambda f$, form a complete basis in the function space $L_2(\Omega)$ of measurable and square-integrable functions on the manifold Ω (Grebenkov and Nguyen, 2013). For a bounded manifold, the eigenfunctions must further satisfy boundary conditions; for example, DM requires the Neumann boundary conditions (Coifman and Lafon, 2006), such that the derivatives on the boundaries vanish. Therefore, the eigenfunctions depend also on the boundary of Ω.

It is well understood that the eigenfunctions of the LBO on Ω carry useful information about its intrinsic geometry, and are thus important for understanding many systems. For compact manifolds with a boundary, as an example, these eigenfunctions are the modes of vibration of a 1D string or a 2D membrane. For compact manifolds without a boundary (i.e., closed manifolds), the well-known spherical harmonics are eigenfunctions on the surface of the hypersphere. In the field of structural biology, the eigenfunctions of the LBO on $SO(3)$, which are the Wigner-D

functions, have been used for retrieving the unknown orientations of single-particle X-ray and cryo-EM snapshots (Giannakis et al., 2012b). In general, the eigenfunctions of the LBO on different manifolds are fundamental to mathematics and sciences, and describe a wide diversity of seemingly disparate phenomena—reflecting the so-called "underlying unity of nature"—from quantum mechanics to gravitational fields (Feynman et al., 1965, chap. 12).

NLSA Mapping. Since the geometric structure of these DM embeddings is in general unknown, Dashti et al. (2014) resorted to an additional embedding of the images as arranged within the initial DM embedding. Their approach includes use of nonlinear Laplacian spectral analysis (NLSA; Giannakis and Majda, 2012a) to produce a sequence of images (a 2D NLSA movie), expected to show a conformational change along a given path on the manifold. To note, one of the rationales for the use of NLSA is that it normalizes the initially unknown rates of change in different CM directions (Dashti et al., 2014). As well, NLSA is used for noise reduction during the extraction of the conformational signal contained in each PD. For this procedure, NLSA is first applied independently on a leading set of eigenvectors to construct a new representation for each. This representation is then analyzed in order to assess the "meaning" of each eigenvector in terms of housing a potential CM of interest.

To achieve this NLSA representation, the raw images are concatenated along a chosen eigenvector to produce so-called "supervectors" (Dashti et al., 2014). Nonlinear singular value decomposition (SVD) is then used to extract noise-reduced NLSA images ("topos") and their evolutions ("chronos") from these supervectors. These NLSA images are then embedded to form a new set of eigenvectors in a different space, which results to high accuracy in a 1-manifold with known eigenfunctions $\{\cos(k\pi\tau') \mid k \in \mathbb{Z}^+\}$, parameterized by a conformational parameter τ' (separate from the forthcoming use of τ in Chapter 10). This enables the estimation of the density

of points as a function of τ' together with an ordered sequence of 2D NLSA images. These 2D images can be arranged to form a 2D NLSA movie, designed to represent the conformational signal corresponding to the eigenvector from the initially-embedded PD manifold. Once a set of 2D NLSA movies have been constructed along each of the leading Ω_{PD} eigenvectors independently, supervised identification of "meaningful" CM information is next required.

When knowledge of only one degree of freedom is desired (or available), the 2D movies corresponding to the same CM in different PD manifolds can be further compiled across S^2 (as described in the next section) to reconstruct NLSA volumes and thus a 3D movie representing the CM. The NLSA procedure is more complicated when knowledge of two (or more) degrees of freedom is desired. After supervised identification of two CMs, their respective eigenvectors for the current Ω_{PD} are used to isolate a 2D subspace therein. On this $\{\mathrm{CM}_1, \mathrm{CM}_2\}$ subspace, NLSA is performed independently along the directions of (typically) 180 uniformly-spaced radial lines in the range $\theta \in [0, \pi]$. This yields a collection of point densities (i.e., 1D occupancy maps) $n(\tau', \theta)$ for each θ. The collection of these 1D maps for all θ constitutes the 2D Radon transform of a yet unknown 2D density map (i.e., the desired 2D occupancy map). An inverse Radon transform is then applied to reconstruct the 2D density map. In addition, NLSA also retrieves the noise-reduced images at each point in this map. As in the 1D case, this procedure must next be performed for the eigenvector pairs corresponding to $\{\mathrm{CM}_1, \mathrm{CM}_2\}$ in all other Ω_{PD} embeddings.

Compiling PDs. Recall that one of the main assumptions of the ManifoldEM framework is that the same conformational spectrum is viewed in all of the PDs. In other words, the conformational variability of a molecule and frequency of sightings (occupancies) of the different states is assumed to be independent of its orientation on the EM grid. The difficulty in using this correspondence is that the order (eigenvector ranking) and the directionality (*sense*) of the

conformational coordinates in one PD may be different from those in another PD, even if it is adjacent. The former difference is due to the fact that the conformational change of a given coordinate is easier to observe at certain viewing angles than at others. The latter difference is due to the sign ambiguity (i.e., a movie reverses its sense when the corresponding coordinate flips its sign) in standard eigendecomposition solvers. The task of propagation of CMs is to match the CMs in adjacent (and also, by inference, farther located) PDs, such that they express the same or closely similar motions with matching directionality. A successful outcome of this matching procedure will ensure that compilation is done over corresponding conformational occupancies from all PDs, from which the consolidated occupancy map (and thus the free-energy landscape) of the molecule may be obtained.

Naturally, this is a problem that can be manually solved with heightened user involvement, requiring an assessment of (i) the number and kinds of CMs to consider; (ii) the indices or rankings of these CMs in each PD-manifold; and (iii) the sense of these coordinates. In actuality, as there are typically hundreds to thousands of PDs to consider (and each with several CM coordinates to assess), this is an exceedingly laborious endeavor, fraught with uncertainties and possibilities for error. Instead, an automated approach is highly preferred, such that user involvement is minimized. In the original ManifoldEM framework (Dashti et al., 2014; Mashayekhi, 2020), the correlation coefficient between the conformational spectra of two adjacent PDs is used to propagate the conformational information to all PDs, such that the aforementioned decisions need only be assessed for a small subset of PDs, termed *anchor PDs*. A more reliable approach has been recently developed using optical flow and belief propagation algorithms (Maji et al., 2020), and will be described in more detail in Section 2.4.

Free-energy Landscape. When the user-determined CMs have been labeled across all PDs, a consolidated free-energy landscape is generated, with the dimension of this landscape defined by the number of CMs chosen. The number of NLSA images falling within each bin in the landscape is tallied towards its state's occupancy, with these occupancy values then transformed via the Boltzmann factor to derive the corresponding free energies. Specifically, in thermal equilibrium, differences in occupancy can be attributed to differences in the molecule's free energy ΔG via the mapping $\Delta G/k_B T = -\ln(n_c/n_0)$, where n_c is the occupancy of the current state and n_0 is the occupancy of the maximum-occupancy state (Fischer et al., 2010; Agirrezabala et al., 2012). Here, the lowest observable occupancy in the ensemble represented by the data set—of one particle in a state—defines the highest measured free energy, while the highest observable occupancy defines the lowest measured free energy. The Boltzmann constant k_B is a physical constant that relates the average relative free energy of particles to their bulk temperature T.

In the case of a 2D free-energy landscape, a set of pathways of significant biological interest are anticipated, defined by a minimum-energy trajectory between states, which approximate the path of least action. In the following section, an automation strategy (POLARIS) will be introduced for quantitatively determining this correct sequence of states. Along this pathway—or any pathway; including the single conformational coordinate in the case of a 1D free-energy landscape—a series of 3D Coulomb potential maps (NLSA volumes) can be reconstructed, representing the 3D molecular structure as it transforms.

2.4 Recent Advancements

A number of recent advancements have been made by the Frank lab (2016−2021) as it relates to the ManifoldEM framework outlined above. As previously mentioned, this includes an

automation strategy to connect the CMs across angular space using optical flow and belief propagation (Maji et al., 2020). Together, these two methods are used to propagate assignment of CMs—obtained through ManifoldEM analysis, or any dimensionality reduction method—from a given PD to all other PDs across the angular space (Maji et al., 2020). Without such an automation scheme, this subproblem could previously only be done by tedious visual comparisons between adjacent PDs, with high probability of user error. Instead, only a small subset of PDs, called "anchor PDs", require supervision. These decisions include assignment of the eigenvector corresponding to each sought-after CM and its sense, and are made by assessing each eigenvector's 2D NLSA movie for a given PD. Once these anchor PDs have been accurately assigned, the unsupervised procedure aims to select the correct CM-sense combinations across all other sufficiently-populated PDs.

To achieve this, optical flow is first used to define the most prominent motions in the NLSA movies in terms of feature vectors based on a histogram of oriented gradients (e.g., Figure B1 in Appendix B). Next, the set of all feature vectors in each PD is compared to the set in each of its immediate neighbors on S^2. Following this, an algorithm based on belief propagation assesses the affinity between all pairwise combinations of feature vectors between all neighboring PDs, ultimately giving each comparison a likelihood score for how well it conserves the intended conformational information (defined by the anchor PDs). These probabilities are propagated across the network of PDs until uncertainty is minimized on the global scale. The accuracy of assignments afforded by this strategy is essential, since this information is relied upon for constructing both the free-energy landscape and 3D Coulomb potential maps of the molecule.

A second advancement was made towards navigation of free-energy landscapes. Seitz and Frank (2020) devised an automated strategy for deriving the most probable sequence of transitions

taken between any two states on a 2D landscape. Here, the assumption is made that a free-energy landscape has already been successfully determined (or estimated) using ManifoldEM or any other method proposed. The method (POLARIS: *Path of least action recursive survey*) constructs an optimal minimum-energy pathway between any collection of user-defined states, and provides tools for analyzing and comparing alternative branches of bifurcations. A detailed overview of the POLARIS method is provided in the following chapter.

Finally, considerable work has gone into translating the ManifoldEM framework (initially coded in Matlab) into a comprehensive Python repository and graphic user interface (GUI) for public release, which includes many modifications and improvements. This package is detailed fully in the following section.

2.5 Public Release

The original ManifoldEM methodology (Dashti et al., 2014) was initially implemented using Matlab code, which was later made available for Matlab users (Mashayekhi, 2020). More recently, the Python implementation of the ManifoldEM framework has been released in its "beta" form (Seitz et al., 2021b), mirroring an earlier 1D version of the Matlab code (2019), except for a number of additional enhancements, discussed below. In this beta release, only tools for the generation of 1D energy paths have been made available. Future releases will extend this software package to incorporate 2D energy landscapes as done in the current Matlab package, and in addition, incorporate a new, alternative method—ESPER: *Embedded subspace partitioning and eigenfunction realignment* (Seitz et al., 2021a; properly introduced in the forthcoming Chapter 11)—for recovery of the free-energy landscape in conjunction with conformational movies. As will be discussed, the ESPER method is able to combine with the current NLSA-based

ManifoldEM framework and, given certain prerequisites, enhance the quality of its final outputs. As well, there are several limitations to the current ManifoldEM method that have been outlined by Mashayekhi (2020) and Seitz et al. (2021a,b)—and detailed later in Chapter 7—which the ESPER method is able to circumvent.

The Python code comprises two principal features not contained in the original Matlab code: (i) an implementation of the automated angular propagation algorithm by Maji et al. (2020); and (ii) an advanced GUI that integrates all Python code, and greatly facilitates the execution of essential operations, while providing informative visualization of important intermediate and final outputs. This second feature is imperative for the level of complexity in decisions demanded throughout the framework. In contrast, many advanced software systems now employed in cryo-EM are of the "black box" type, allowing only the specification of a command stream without feedback to the user, and without providing intuition about the way parameter choices affect the outcome (Maji and Frank, 2021). However, the diversity in the nature of the input data and in the kinds of questions asked in the ManifoldEM approach make the black-box type of user control difficult to sustain even by experts, and virtually impossible in the hands of non-expert users. By providing instant feedback on the impact of every parameter choice, the ManifoldEM Python GUI puts the user firmly in the driver's seat, allowing them to steer their research in the desired direction.

The ManifoldEM Python GUI was developed by this author with help from others[3] and is coded using Python 3 with PyQt5 and Mayavi libraries, among many others. All in all, the GUI

[3] Contributions: E. Seitz conceptualized and implemented the features present in the user interface, and made numerous modifications to the backend code necessary for their use. Additionally, he worked with other members of the ManifoldEM team (H. Liao and S. Maji) to add several enhancements to the original Matlab framework. A tremendous amount of troubleshooting and testing was also performed, which included over 300 open issues solved in the main repository. Finally, he was the lead author on a comprehensive ManifoldEM Python user manual packaged along with the beta release (Seitz et al., 2021b).

spans over 10,000 lines of code, providing a complex, multi-tab experience enabling intuitive and dynamic navigation across each module in the ManifoldEM framework. The vast majority of the user experience with ManifoldEM revolves around the use of this GUI, presenting an interface for importing data, analyzing results, making pertinent strategic decisions, and exporting final outputs. In Figure 4, the "Eigenvectors" tab of the GUI is shown, within which users can view the embeddings and corresponding NLSA outputs for each PD to make a number of important decisions. On this tab, user decisions include determination of the set of anchor PDs required for subsequent steps, and for each anchor, the eigenvector corresponding to each sought-after CM and its sense. This decision is aided by means of an internal video player for viewing the 2D NLSA movies for each PD, or optionally, side-by-side with the 2D NLSA movies from an adjacent PD to compare.

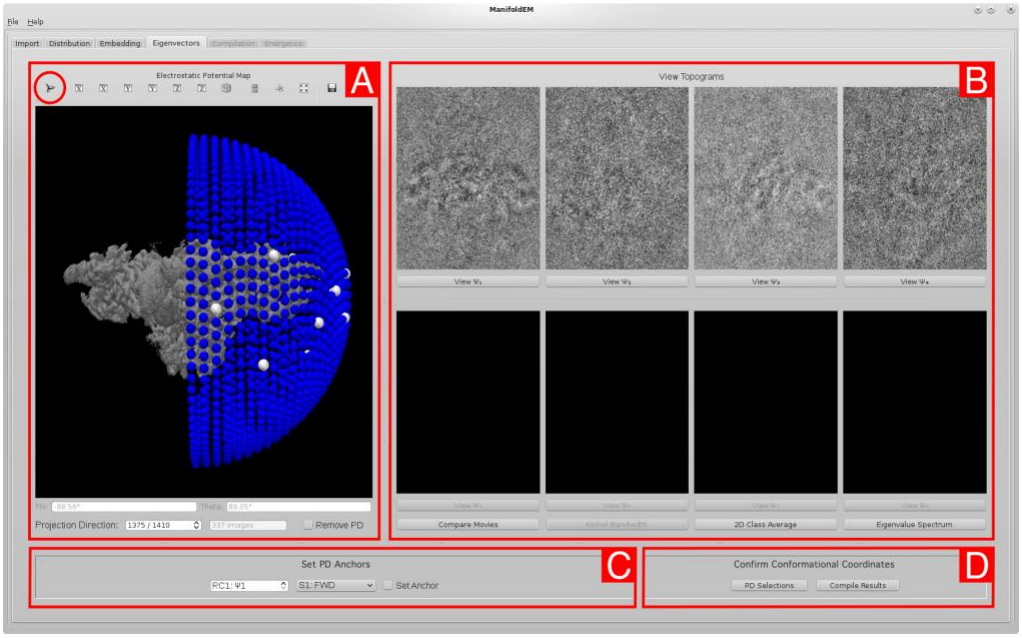


Figure 4: ManifoldEM Python GUI "Eigenvectors" tab, showing outputs of embedding for RyR1. This tab is segmented into four main compartments. [A] allows users to rotate the angular PD distribution (shown with blue nodes) in conjunction with the average volume (shown in gray) to select and highlight an individual PD for detailed analysis. For each PD chosen in this way (via mouse clicks), the information in compartment [B] will automatically update, displaying that PD's set of eigenvectors and corresponding image averages. The panels in [B] can further be opened to view the corresponding 2D NLSA movies and related information. In compartment

[C], anchor PD assignments are recorded for a selection of PDs (allowing also for PDs to be removed for subsequent calculations), with this information viewable in [D].

These user interactions are further supplemented by a detailed manual and demo tutorial data set (RyR1; Zalk et al., 2015) included along with the release. Since the dynamic control of the GUI cannot be illustrated by static figures alone, a comprehensive video demonstration of the ManifoldEM GUI has been provided in Movie B1. As well, while not present in this video of the beta version, the final release will include functionality for up to two degrees of freedom. As is planned, depending on the user's choice of dimensionality, the "Energetics" tab will display either a 1D or 2D free-energy landscape. In the case of 2D, the user will be presented with a subsequent choice to define a pathway through the landscape (Figure B2). Using this interface, points can be added onto this 2D plot to form a custom path, or alternatively, this energy landscape can be exported for analysis with external pathfinding programs—such as POLARIS (Seitz and Frank, 2020)—to define a minimum-energy path of presumed biological significance. These trajectories can then be imported into the ManifoldEM GUI for use in its final computations, as previously described.

Throughout its development, internal testing and debugging of code has been conducted by members of the Frank lab (E. Seitz, H. Liao and S. Maji). In order to expand our awareness of potential problems in software, performance and interface, we also organized and conducted an "alpha" release of the ManifoldEM Python suite, with scope limited to acquisition of a 1D free-energy landscape. The length of our alpha trial spanned approximately eight months—beginning in February and ending in late August, 2019—during which this author acted as a liaison with the alpha-user group. This group encompassed 10 individuals, and comprised scientists and engineers from various cryo-EM related fields. Extensive communications with these users were conducted by our team via email and private GitHub forum to provide guidance on operations and errors

encountered. Alpha users were specifically instructed to test the software using their own or previously published cryo-EM data sets. Both options were exercised, and as these users worked with a variety of operating systems and hardware, we encountered a unique range of problems to explore. This experience made us aware of shortcomings in documentation and errors in the code, helped guide our team towards optimizing software performance, and provided opportunities to enhance useability through the GUI. We also discovered several limitations and uncertainties during this process, which will be discussed with more context in subsequent chapters.

Chapter 3: The POLARIS Method

3.1 Conceptual Outline

Within a free-energy landscape, the *path of least action* is the most probable sequence of transitions taken between any two states. Deriving this minimum-energy path presents much difficulty, however, since there are an infinite number of competing pathways to consider. The founding ManifoldEM study (Dashti et al., 2014) coarsely approximated this pathway by hand, which is a suboptimal strategy in general, and insufficient for analyzing landscapes with competing pathways. Seitz and Frank[4] (2020) presented a dynamic, global approach for efficiently automating this discovery on previously determined 2D free-energy landscapes. The published technique—termed POLARIS: *Path of least action recursive survey* (Seitz and Frank, 2020)—outperformed another algorithmic approach MEPSA (*Minimum energy pathway analysis*; Marcos-Alcalde et al., 2015) on the ribosomal free-energy landscape (Dashti et al., 2014), and additionally offered several built-in features for public release (Seitz, 2020), including bifurcation analysis tools that provide downstream versatility for comparing most probable paths and associated reaction rates.

3.2 Overview of Method

POLARIS provides an alternative approach to preexisting minimum-energy pathfinding algorithms by avoiding the arbitrary assignment of edge weights, physics-based optimizers, or globally restrictive strings (Seitz and Frank, 2020). Instead, POLARIS aims to prioritize and isolate the energetically most-favorable coordinates in a given landscape, as these represent highly occupied transit regions on the macroscopic scale. This initial method reflects the paradigm used

[4] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; software; validation; manuscript draft, review and editing. J. Frank: Conceptualization; direction and project administration; manuscript review and editing.

by hierarchical pathfinders (Botea et al., 2004), where the high-level overview route is assigned first by analyzing the most favorable transit locations in between. Between these anchors in the landscape, all permutations of a set of higher-order lines are drawn, with the net energies compared from each. The algorithm then takes the resulting minimum-energy discoveries as new inputs to itself to repeat this procedure recursively: breaking down each best global line approximation into continuously finer, mesoscopic subsections until the finest-granular, microscopic path is resolved. POLARIS is thus designed for biological context, since in the case of a macromolecular assembly driven by thermal motion, such a metric should reflect the total energy of each possible path, while also analyzing these paths within the landscape's hierarchy of distinguished scales (macroscopic, mesoscopic and microscopic; Munro et al., 2009). In the following, the POLARIS method is described in four main steps: (i) Image segmentation; (ii) permutational analysis; (iii) branching recursion; and (iv) pathway pruning.

Image Segmentation. In the first step, the landscape is divided into a number of equally sized squares, with the local minimum-energy values recorded within each. For this procedure, the segmentation depth η defines the number of subdivisions created and, thus, the number of local minimum-energy coordinates recorded (to note, the set of η_i are user-defined). For each η_i, $H_i = 4^{n_i}$ image subdivisions are created, with an equivalent number of minimum-energy points stored. Within this set, the highest allowable value of η_i (η_{max}) is defined via that subdivision where further images divisions are impossible, so that every segmented grid spans an area of exactly one pixel. As the segmentation depth is increased from η_1 to η_{max}, previously overlooked higher-energy minima (relative to the system's global minima) are geometrically separated from the global minimum values and placed into their own neighboring grids. Since these locations become

newly defined local energy-minima at higher depths, this method ensures that all points are eventually considered as potential nodes in the permutational analysis, outlined next.

Permutational Analysis. Following this procedure, the set of all possible permutations are created between the previously-isolated minima over a range of increasing line orders. To initiate this task, each set of H_i minimum-energy coordinates obtained during image segmentation are used as transit options for comparing potential paths across the free-energy landscape. For each η_i, permutations of straight lines are connected between the set of H_i coordinates and the user-defined start points S and end point E. A partnering value for the permutational order (r_j) is also defined for each value of η_i. Here, r_j can range from r_0 to r_{max}, with r_0 representing the line directly connecting S to E (with no midpoints in between), and r_{max} typically no more than five. Within this formulation, the standard permutation notation $P(H_i, r_j)$ represents a unique combination of some η_i and r_j. As an example, for η_2 and r_2, one such path $S \rightarrow M_1 \rightarrow M_{16} \rightarrow E$ is created among 239 other permutations, via $H_2 = 16$ and $P(16, 2) = 240$.

Within each member of a given $\{\eta_i, r_j\}$ permutational pool, each set of ordered transit points is connected by straight lines, drawn using a variant of Bressenham's line algorithm (Kuzmin, 1995) modified for energy awareness. These path approximations act to gather overhead awareness across different regions in the landscape. The total energy across the paths formed by each of these permutations is then independently integrated, with only the transit points from the minimum-energy permutation stored for that $\{n_i, r_j\}$. This process is then repeated for all combinations of $\{n_i, r_j\}$, ultimately only storing the set of transit points belonging to the lowest-energy path approximation discovered between the initial S and E. These minimum-energy coordinates are then introduced as inputs for the subsequent computations.

Branching Recursion. POLARIS next uses the minimum-energy nodes discovered during permutational analysis recursively as inputs to itself, replacing the initial inputs S and E with the set of the newly discovered intermediate, minimum-energy nodes. First, a for-loop is created between each pair of intermediate nodes, such that if the pathway containing points $S \rightarrow M_1 \rightarrow M_2 \rightarrow E$ were found as a minimum among all other permutations in the previous step, a loop containing $S \rightarrow M_1$, $M_1 \rightarrow M_2$, and $M_2 \rightarrow E$ would emerge. Within this loop, POLARIS performs branching recursion, repeating all of these aforementioned procedures on its newly obtained outputs. From here, the permutational steps are repeated for each new set of start and end points discovered—proceeding recursively down each inner branch—until two output points having lowest-line approximation r_0 are found within each one of its individual *leaves* (defining the maximum extent of each *branch*). Upon encountering each leaf-break, the coordinate of that leaf is globally recorded. This process continues with the algorithm navigating throughout its recursive hierarchy until the path of least action is filled in completely with a set of coordinates spanning from the initial, user-defined start point to the initial, user-defined end point. Ultimately, these recursive steps break down the best global line approximation into continuously finer subsections within each recursion until the finest-granular path is resolved, giving rise to both global and local awareness of the landscape.

Pathway Pruning. A final step is applied on the output path to account for artifacts due to aliasing encountered when drawing diagonal lines between coordinates. This procedure includes finite, local perturbations along each coordinate in the completed path. So long as these perturbations preserve the continuity of the overall path (i.e., no breaks), every point is iteratively sent to occupy its set of unrestricted neighboring pixels and reshuffle into those coordinates that ultimately minimize the pathway's global energy.

3.3 Analysis of POLARIS Results

The ManifoldEM-generated ribosomal free-energy landscape (Figure 3; Dashti et al., 2014) was used for testing the performance of the POLARIS method. The ribosome has long been described as a "thermal ratchet machine" (Spirin, 2002) where random energetic perturbations lead to large-scale shifts across its available configurations. An analysis of the ManifoldEM landscape has been shown to account for these coordinated conformational motions (Dashti et al., 2014), and is an ideal setting for deeper analysis via POLARIS. Based on the appearance of several branching routes throughout the ribosomal landscape, we were curious to understand their potential significance in alternating metabolic function (Seitz and Frank, 2020).

In order for the POLARIS method to compare the two branching troughs located in the southwest region of the landscape (Figure 5), midpoints were placed alongside the initial user-defined start and end points, which forced POLARIS to explore both routes independently on separate runs. These two diverging paths go through the center (black points) and southeast (white points) regions of the landscape, and contain path energies 9.0 kcal/Mol (with 16 points) and 8.5 kcal/Mol (with 37 points), respectively. With only a difference of 0.5 kcal/Mol between them, it is possible that such a degenerate least-energy bifurcation could represent novel reaction mechanics that allow for flexibility in macromolecular processes both spatially and temporarily. For example, the bifurcation seen in Figure 5 may represent a shortcut in the ribosomal work cycle (elongation) that only becomes available under specific buffer or temperature conditions, allowing the ribosome to modulate its reaction rates based on fluctuating environmental signals.

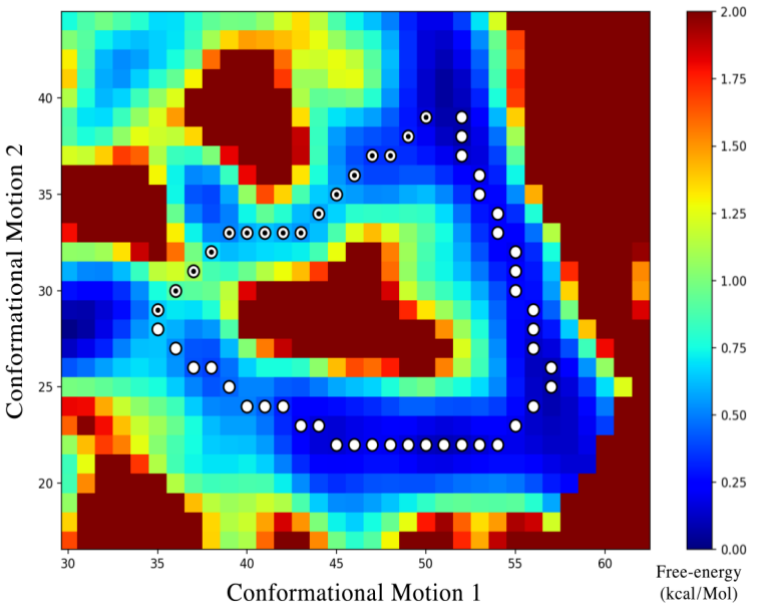


Figure 5: POLARIS bifurcation analysis on the ribosomal free-energy landscape. Comparison of two approximately degenerate, spatially-different paths found by the POLARIS method, differing in free energy only by 0.5 kcal/Mol. The southeast pathway (white points) describes a section of the trajectory analyzed in the original study (Dashti et al., 2014).

Additionally, we were interested to compare the results from POLARIS with another published free-energy landscape analysis algorithm, MEPSA, which uses an approach similar to Dijkstra's algorithm (Dijkstra, 1959) with small differences in the sampling and trace-back. The metric we used to compare the two algorithms is the final integrated free energy from a single continuous list of coordinates spanning between any two user-defined points. The MEPSA tool allows two options for algorithmic comparison. The first of these is the self-defined less-accurate "GLOBAL" option, which is capable of accepting arbitrary user start and end inputs as well as predefined anchor points. The second option, "NODE BY NODE", is more accurate and accepts only predefined global minima as user inputs. In both instances, POLARIS surpassed the performance of MEPSA, with detailed results provided via Seitz and Frank (2020). Here we show the "NODE BY NODE" comparison, where MEPSA's node 2 at coordinate $\{16, 23\}$ and node 41

at coordinate $\{47, 55\}$ were selected from the set of the MEPSA method's predefined nodes (to note, POLARIS requires no such restrictions).

The pathway found by the MEPSA approach returned an integrated free energy of 28.50 kcal/Mol with 98 points (Figure 6, left). POLARIS used the same nodes 2 and 41 as user inputs and identified a 72-point pathway having an integrated free energy of 20.02 kcal/Mol (Figure 6, right). The total difference between the two minimum-energy approximations found by the two algorithms was 8.5 kcal/Mol, favoring POLARIS, with a path-length difference of 26 conformational states. In all comparisons with MEPSA, the POLARIS outputs gave a substantially lower energy path while providing higher flexibility in the choice of user-defined start and end points, as well as in the number of optional user-defined transit locations. As for computation time, it should be noted that MEPSA generated its paths within seconds. However, the MEPSA method's self-defined "global minimum" solutions were undershot considerably by the POLARIS techniques.

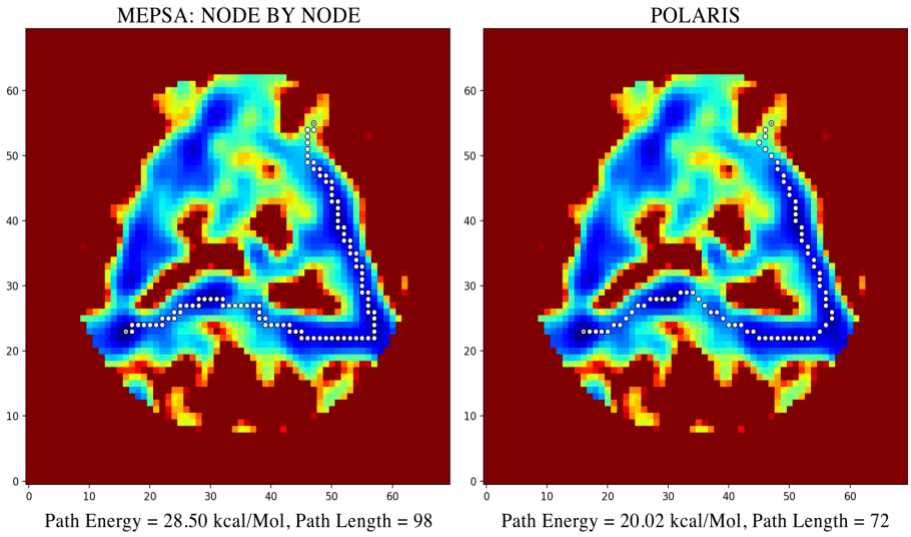


Figure 6: Comparison of paths output by POLARIS and MEPSA, shown on the left and right (using the "NODE BY NODE" option), respectively. While both algorithms found the same approximate global path, POLARIS returned a path of integrated free energy 8.5 kcal/Mol less with a 26-point difference in path lengths. The POLARIS calculation took approximately 5

minutes using multiprocessing with 4 1.2 GHz processors and 8 GB of memory. In other scenarios explored, globally different paths emerged between methods (Seitz and Frank, 2020).

While the POLARIS method is capable of similar speeds at drastically lowered segmentation depth and permutational order, the lowest-energy trajectories found (shown here) required more computation time (on the order of minutes). Thus, when weighing accuracy over timing, POLARIS is considerably more fit. Although these differences may seem trivial from a macroscopic view, such exactitude is essential on the microscopic level for accurately calculating biologically relevant reaction rates via downstream algorithms.

As a final note for this section, the POLARIS package additionally includes a graphic user interface (Figure 7), allowing users to load in free-energy landscapes, select coordinates, and adjust parameters and constraints as desired. After each run, landscape trajectories and transition state theory diagrams are rendered, as well as generation of text files with least action coordinates and respective energies. This interface provides flexibility for user experimentation with transit points and parameters to expedite the discovery of their most significant pathways. As this software relates to the ManifoldEM framework, it is a work in progress to combine the POLARIS algorithm into the final steps of the Python ManifoldEM public release, such that 3D movies may be rendered along a given path of least action.

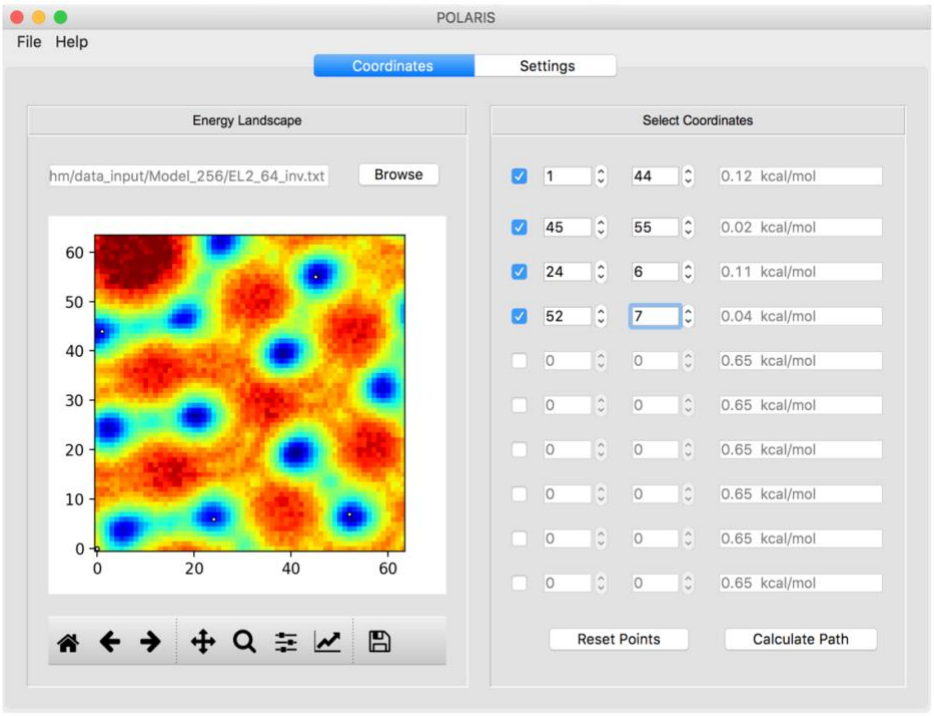


Figure 7: POLARIS Python GUI "Coordinates" tab. Image from the main page of the POLARIS user interface, allowing users to load a valid data file and add up to 10 intermediate transit locations on the free-energy landscape as desired.

Chapter 4: ManifoldEM Analysis of V-ATPase

4.1 Introduction

In the interest of expanding knowledge on the dynamics of molecular machines, and expanding the current pool of machines analyzed by ManifoldEM, an analysis was performed on the rotary nano-motor Vacuolar-type ATPase (V-ATPase; Figure 8). From a mechanistic perspective, V-ATPase presents a novel level of complexity in motions not existing in those studied in the 80S ribosome and RyR1 complexes. This evolutionarily-ancient enzyme is an ATP-powered proton pump composed of two rotary motors that cooperate sequentially—with one motor powering the rotation of another—which, in concert, ultimately serves to acidify a wide array of intracellular organelles. Upon binding with ATP, the ATP-driven motor turns an axle that rotates a second motor embedded in the membrane, rotating it such that protons are pumped across the bilayer. Meanwhile, the peripheral stators (subunits G and E) stabilize the complex during these coordinated rotations.

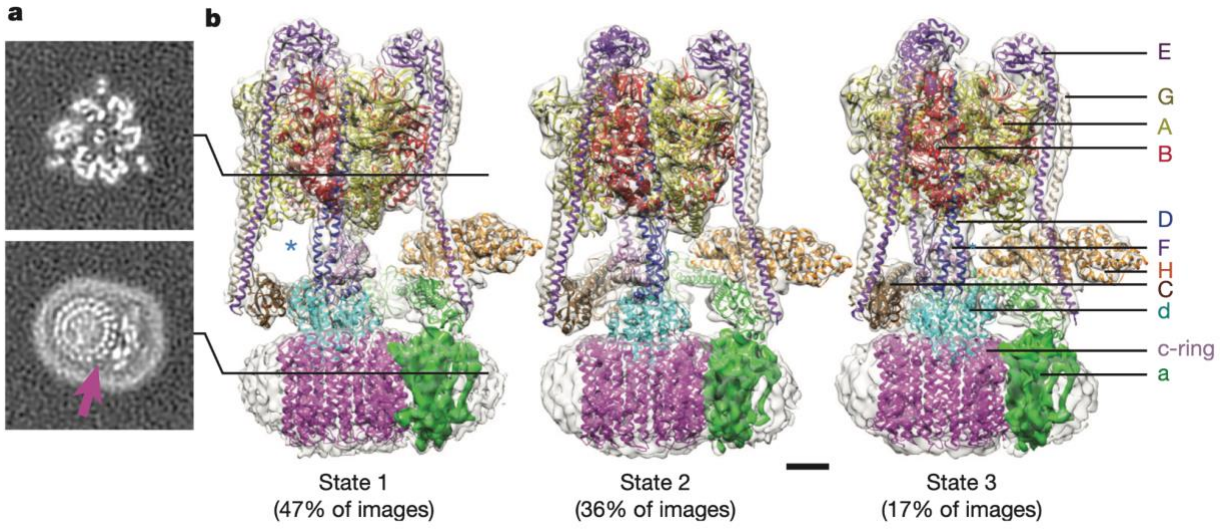


Figure 8: V-ATPase rotational macrostates. Three rotational macrostates of the V-ATPase, as reconstructed by Zhao et al. (2015; from which this figure originates). Cross-sections through a 3D volume of rotational state 1 are shown in [a]. The three different volumes—representing

macrostate 1, 2 and 3—are shown in [b]. Each volume has been presented in semi-transparent grey, overlaid with atomic models of subunits (denoted alphabetically). As a note, these V-ATPase structures are modelled with 116,324 atoms each, making up 7,451 residues.

The eukaryotic V-ATPase is involved in many critical physiological processes (Baker et al., 2015), including energy metabolism, vesicle trafficking, coat formation, endosomal pH sensing, membrane fusion, intracellular signaling, endocytic sensing, cellular aging and rejuvenation (Merkulova et al., 2015). V-ATPase is also known to have a systemic role in renal acid excretion, maintaining blood pH balance, male fertility, bone remodeling, synaptic transmission, olfaction and hearing. Further, loss-of-function mutations have been linked to osteopetrosis, osteoporosis, viral infections, distal renal tubular acidosis, deafness, Parkinsonism, impairment of insulin secretion and cancer. In the latter, V-ATPase activity exacerbates metastasis (Merkulova et al., 2015). Due to these crucial roles across numerous pathways essential for life, a firm understanding of the underlying dynamics of V-ATPase is highly relevant to human medicine.

As V-ATPase is a complex molecular machine that relies on the coordination of many moving parts by functional necessity, it has proven difficult to study (Goodsell, 2018). Currently, the most complete structures have been obtained via cryo-EM, which can naturally account for the machines' innate conformational heterogeneity. For our ManifoldEM analysis, cryo-EM data were obtained from previous work on yeast V-ATPase monomers by Zhao et al. (2015). In their study, V-ATPase molecules were detergent-solubilized without ligands and imaged via the Tecnai F20 TEM. These images were then processed via maximum likelihood procedures in RELION, resulting in three class averages: "State 1" (47% of images at 6.9 Å); "State 2" (36% of images at 7.6 Å); and "State 3" (17% of images at 8.3 Å), each representing one of three 120° power-stroke rotational macrostates about the central D-subunit.

The following analysis is an unpublished work and without formal peer review. Research was conducted within the Frank lab (E. Seitz and J. Frank, 2019)[5], and informed by helpful conversations with John Rubinstein.

4.2 Materials and Methods

As the ManifoldEM framework relies on images of a molecule in a quasi-continuum of states to elicit conformational information, the original image stack was used with minimal pruning as input. Since the ManifoldEM distance calculations can be easily misled by inclusion of aberrant images, special care was taken to remove as many erroneous images as possible. Thus, the original image stack—consisting of 106,406 images—was re-analyzed via 2D and 3D classifications and handpicked assignments to remove 0.83% of particles, resulting in 105,524 images total. This stack was then processed using RELION 3D Auto-Refine with the original State 1 volume (EMD-6284) as reference to produce an alignment file and average volume (8.4 Å resolution).

The ManifoldEM Python repository and GUI (Seitz et al., 2021b) were used for processing of this data. First, a number of initial inputs are required to tessellate the angular space (S^2) into a collection of PDs. These are: (i) Pixel Size: 1.45 Å; (ii) Resolution: 8.4 Å; (iii) Object Diameter: 230 Å (taken as the maximum width of the molecule within the reconstructed volume); and (iv) Aperture Index: 4. The aperture index is a flexible parameter that controls the angular width of each PD, such that a larger aperture index corresponds to more images assigned to each PD from a larger region of angular space. After experimenting with several aperture indices and evaluating the corresponding PD statistics and 2D movie qualities, we chose an aperture index resulting in a

[5] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; manuscript draft, review and editing. J. Frank: Conceptualization; direction and project administration; manuscript review and editing.

total of 212 PDs. These PDs formed a nonuniform coverage across S^2, with a region of heightened PD-occupancy along the equator (Figure 9, left). Such a distribution is a likely byproduct of preferred orientations, where each detergent-solubilized monomer forms favorable interactions with the water's surface. This proclivity leads to an ensemble with unevenly occupied viewing angles; in our case, PDs orthogonal to the axis of the central rotor were most highly populated.

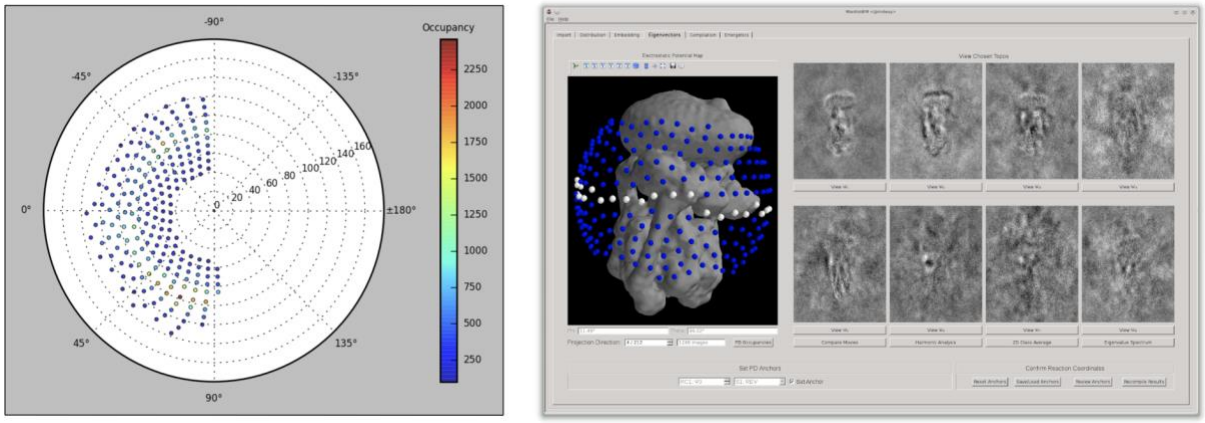


Figure 9: V-ATPase angular distribution in ManifoldEM Python GUI. On the left is a polar coordinates plot of the 212 PDs, with a heatmap displaying the occupancy of each. Note the equatorial ring of high-occupancy projections. On the right is an image from the "Eigenvectors" tab of the Python ManifoldEM GUI, where this high-occupancy equator has been selected in 3D space using white nodes. A collection of eight 2D image averages for PD_4 are also shown on this tab, obtained via manifold embedding of the V-ATPase cryo-EM data set.

Following the ManifoldEM framework, 212 manifolds—each one corresponding to images within one of the available PDs—were analyzed via diffusion maps. NLSA was then performed on the eigenvectors of these high-dimensional manifolds to extract a set of possibly relevant CMs from each. As shown in Movie C1-A and Movie C1-B, a prominent CM was found in the majority of PDs, which took on a unique appearance depending on the direction of 2D projection. From some areas of S^2, this motion resembled a kinking of the V-ATPase structure, sending the central rotor outwards while the V-ATPase base (i.e., c-ring and a-subunit) and top (i.e., AB subunits) collapsed toward each other in the reverse direction. From a different viewing direction, the central

d and F-subunits could be seen twisting, with a pinching motion observed between one of the A-B subunit pairs. By tracing out an equatorial path along S^2, it was observed that these two seemingly-dissimilar motions were in fact the same, forming a twist/kink motion labelled as CM_1. In a handful of PDs, a second CM was observed: a spring-like motion along the direction of the central rotor axis (Movie C1-C). As these trailing CMs could not be identified in a substantial number of available PDs, the ManifoldEM analysis was restricted to a 1D analysis of the dominantly-present CM_1. For a general demonstration of the outputs obtained for each ManifoldEM module in this V-ATPase data set, refer to Movie B1.

Of the 212 PDs, a small subpopulation of approximately 30 PDs held no valuable conformational information. These PDs were typically located farthest away from the high-occupancy S^2 equator (i.e., nearest either pole) and were removed from use in subsequent calculations. Of the remaining PDs, both the correct eigenfunction and its sense—representing the directionality of the CM_1 2D NLSA movie in each PD—were assigned by eye (with help from S. Maji). As previously addressed, automation techniques do exist (Maji et al., 2020), but were avoided to maximize accuracy, granted the high complexity of the motions as seen from different PDs and the relatively small number of PDs to consider. The assignments for each of these PDs were then used to effectively stitch the PD-manifolds into one coherent whole, resulting in a 1D free-energy path and corresponding sequence of 50 NLSA volumes.

This 1D free-energy path resembled the cross section of a jagged bowl centered about the 25$^{\text{th}}$ NLSA state, with energies near the boundaries notably skewed (strikingly low), such that the first and last five NLSA states were affected. To note, this abnormality has since been well described (Mashayekhi, 2020; Seitz et al., 2021b) under the term “boundary problems”, and will be addressed separately in Chapter 7 on ManifoldEM limitations and uncertainties. When viewed

in sequence, the 50 NLSA volumes displayed the familiar, monotonic motions of CM_1 (different 2D views of this 3D motion are available in Movie C2 in Appendix C). Again, boundary problems were apparent, now manifested as elevated noise for NLSA volumes corresponding to the first and last five states.

Overall, although visually present, an innate difficulty emerged in understanding the conformational changes—in terms of precise displacements—as they appear in the NLSA volumes. Although unavailable, one would ideally prefer to instead interpret the CM in terms of deformations of each state's corresponding atomic models, rather than changes in map density. Following this rational, and given the availability of predetermined classes via the initial study (i.e., three volumes each fit with a corresponding atomic-coordinate structure; Zhao et al., 2015), Seitz and Frank devised an approach to efficiently fit each of the 50 NLSA volumes with atomic models for enhanced conformational study. For this operation, molecular dynamics flexible fitting (MDFF; Trabuco et al., 2008) was a natural choice for automating these (otherwise exhaustive) modeling decisions, while involving minimal parameter tuning.

Using the MDFF method, one can flexibly fit atomic models into Coulomb potential maps by adding a potential term to the standard MD potentials[6]—which is proportional to the gradient of the Coulomb potential map—into a molecular dynamics (MD) simulation of the atomic structure (Trabuco et al., 2008). In a typical MDFF application, an atomic model of a macromolecule is flexibly fitted into a low-resolution Coulomb potential map of that same molecule in a different conformation (Trabuco et al., 2008). For our purposes, this concept can be extended to a simulation of an atomic structure across an entire sequence of continuously

[6]To eliminate any confusion over the overuse of the term "potential" here, the Coulomb potential map constructed from cryo-EM data has nothing to do with the standard potentials used in MD. The former is the structural mathematical function solved by cryo-EM, while the latter provides a description of the terms by which the particles in the MD simulation will interact.

transforming Coulomb potential maps. Hence, by using a guiding potential in the MDFF force field calculations defined by each NLSA volume (one at a time, in sequence along CM_1), the predetermined atomic-coordinate structure (Zhao et al., 2015) is encouraged to repeatedly "fall" into each consecutive NLSA volume. This simulation thus results in a series of atomic models that closely approximate the conformational states of the underlying atomic structures reflected by the original NLSA volumes on which they are based.

Given its prominence in the image stack (corresponding to 47% of images), PDB-3J9T (State 1) was chosen as an initial seed for initiating the series of MDFF simulations. As a preliminary step, all NLSA volumes were first aligned individually with the State 1 volume (EMD-6284) via the Chimera "fitmap" and "resample" commands (Pettersen et al., 2004). Next, by assessment of the position of the two central alpha helices (seen as two high-intensity dots in overhead cross-section), it was found that the 19th NLSA volume was of highest similarity with the configuration of the State 1 volume. Hence, the 19th NLSA volume was chosen as the first MDFF potential for which PDB-3J9T would be flexibly fitted via MDFF. After these simulations converged, the newly fitted atomic structure was then flexibly fitted using the 20th NLSA volume, and so on.

All MD simulations were performed in a vacuum (i.e., no explicit solvent) to prioritize fitting of the NLSA volumes, noting that naturally, explicit solvent was effectively "baked in" for each image captured from cryo-EM experiment. The MDFF procedure included use of visual molecular dynamics (VMD; Humphrey et al., 1996) and nanoscale molecular dynamics (NAMD; Phillips et al., 2020) protocols. In the former, Protein Structure Files (PSF; containing all the connectivity information and partial atomic charges required by NAMD) were created using AutoPSF, NLSA volumes were converted to MDFF potentials, and restraints were applied for

secondary structures, chiral center handedness, and generation of cis peptide bonds. MDFF simulations were then run using NAMD, taking care to supply enough steps such that the root-mean-square deviation (RMSD) of the backbone with respect to the initial structure converged over time. For the initial seed (PDB-3J9T converging onto the 19th NLSA volume), 10 million steps were required. Once converged, the last atomic structure in the resulting trajectory file was used as a seed for the following NLSA volume (in the forward direction; i.e., the 20th NLSA volume). The required number of steps for these subsequent (post-seed) simulations was significantly less (approximately 600,000). To increase consonance between atomic structures and NLSA volumes, the MDFF ξ parameter was set to 10 for the final steps, representing an arbitrary scaling factor for the magnitude of forces. Upon reaching the 50th NLSA volume, this process was then repeated in the reverse direction starting at state 50, so as to better stabilize the fitting of the initial trajectory and remove any bias incurred from the abrupt initiation via PDB-3J9T. Upon completion, an atomic structure was generated for each of the initial 50 NLSA volumes.

4.3 Results and Discussion

In the following, a description of the MDFF results is provided as a proof of concept, with notably less attention given to biochemical legitimacy. Different views of the atomic-coordinate structures and overlaid 3D maps along the MDFF trajectory are shown in Movie C3 and Movie C4, showing a sequence of atomic-coordinate structures transforming in time with corresponding NLSA volumes overlaid. In the former, the results of the initial, forward MDFF simulation are shown. An initial analysis was first performed using the Phenix "Comprehensive Validation" module (Liebschner et al., 2019) to measure the structural quality of the MDFF-produced atomic structure in comparison to PDB-3J9T. This software provides numerous validation tools, including

identification of outliers from restrained geometric parameters (bonds, angles, etc.), Ramachandran plots, and all-atom contact analysis. Overall, the structural integrity was highly consistent between both structures, with favorable metrics defined for each one independently.

Outside of this validation suite, we additionally observed a good visual agreement between 3D NLSA density and atomic-coordinate positioning, with the twisting/kinking attributed to CM_1 much more visually pronounced. Most notably, the presence of an unoccupied region of 3D NLSA density (encircled in Movie C3) allowed for an alpha helix to unwind during the MDFF simulation: starting from its previously-published coordinates (i.e., PDB-3J9T; Zhou et al., 2015) and binding onto an external E/G domain. During this unwinding, the aspartic acid (residue 611)—located on the tip of the alpha helix—forms a new bond with the arginine (residue 25) on the E/G domain, in effect creating an arginine-aspartic acid salt bridge.

To understand whether the observed "jumping" motion was relevant, we next ran the simulation in reverse starting from the 50^{th} state (Movie C4). Since the MDFF simulations are unbiased by directionality, if the jumping is more than an artifact of the initiation, it should occur in both the forward and backward directions. In contrast to Movie C3, the reverse trajectory in Movie C4 shows that the unwound helix remains affixed to the E/G domain across all 50 states, with the majority of its residues resting in a region occupied by a significant amount of 3D NLSA density in the corresponding intersubunit space. Thus, if the NLSA densities are to be believed, there is substantial evidence for the presence of this salt bridge, and further, that the helix "jump" is only due to the abrupt initiation of the NLSA-MDFF simulation from an external PDB.

Notably, a description of this salt bridge is undocumented in the original analysis (Zhao et al., 2015), which was informed by maximum-likelihood classification. A bridge-like attachment has been found in a more recent study on ATP synthase (Guo et al., 2020), which is located in a

similar region—with respect to the overall structure—as the V-ATPase salt bridge. In the case of ATP synthase, this attachment is between a rotor and stator subunit, which suggests a preventative measure used to halt the enzyme from running in reverse, thus inhibiting ATP hydrolysis. In contrast, the V-ATPase interaction predicted by MDFF is between two stator subunits, and would thus have much less control over metabolic function. Instead, it is more likely that the V-ATPase salt bridge may help stabilize those domains that are dispersed during CM_1. In any case, an ability for ManifoldEM to garner such intricate information then, if accurate, proves its potential for guiding structural research.

The remaining content in Movie C4 illustrates the large-scale motions of the different subunits. In Movie C4-B, two side cross-sections are shown, demonstrating prominent motions occurring in synchrony across several domains over the course of the 50-state trajectory. Most notably, as seen in both Movie C4-B and C4-C, a distinguished A-B subunit pair can be seen clinching/relaxing, with this interaction potentially triggered by hydrolysis of ATP. Also note the relative stability of several other domains, such as the large H-subunit arm and c-ring. As seen in the top-right of Movie C4-C, the latter domain shows only a translational drift, without any rotational activity.

4.4 Closing Remarks

Our analysis of V-ATPase using ManifoldEM and MDFF provides a solid proof of concept, which has the potential to expediate structure determination for ManifoldEM-output volumes, and advance the understanding of molecular machines on the atomic scale. Due to the Covid-19 outbreak (2020), this work was abruptly halted, with efforts turned to the study of the SARS-CoV-2 virus; detailed in the next chapter. As such, finer, biologically-focused details—

including rigorous validation of these V-ATPase results—are left for the possibility of a future analysis and publication.

Chapter 5: ManifoldEM Analysis of SARS-CoV-2 S Protein

5.1 Motivation

Shortly after the Covid-19 worldwide outbreak (2020), all Frank lab efforts shifted to prioritize a structural and functional understanding of the SARS-CoV-2 virus. Our group (E. Seitz, F. Acosta-Reyes, S. Maji and J. Frank) formed a collaboration with the Ourmazd lab (A. Ourmazd and G. Mashayekhi) to apply the ManifoldEM framework on a leading antagonist: the CoV spike (S) glycoprotein. The S protein is a club-shaped macromolecule that is decorated with sugar molecules called glycans. An array of these glycoproteins protrudes from the SARS-CoV-2 virus shell. Cryo-EM studies had been recently conducted on the S protein by Wrapp et al. (March 2020), showing a propensity for its receptor binding domain (RBD) to transition from a glycan-shielded "down" to an exposed "up" state that facilitates its binding to the human ACE2 receptor and infect the cell. Near the onset of the pandemic, a cryo-EM data set representing a large ensemble of S proteins was collected by Jason McLellan and colleagues at the University of Texas at Austin and analyzed (Wrapp et al., 2020). Using the cryoSPARC 3D variability approach (Punjani and Fleet, 2020), Wrapp et al. (2020) recovered an observed "breathing" motion of the S1 subunits as the RBD underwent a hinge-like movement. Working with Jason McLellan, our team acquired the spike data set for further analysis using ManifoldEM.

5.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped RNA virus responsible for the COVID-19 disease, now a worldwide pandemic causing significant morbidity and mortality (Chan et al., 2020; Lu et al., 2020). The virus' main mechanism for infection is facilitated through an array of spike proteins presented on its exterior, which are a

major viral antigen and the first point of contact with a host cell. A number of studies have collectively reported on the proclivity of the spike protein to assume one of two conformations: the "down" or "up" states, which are defined by the position of the RBD. The transition between these two states—from RBD-down to RBD-up—enables binding with the ACE2 receptor located on the surface of the host cell (Benton et al., 2020; Walls et al., 2020; Wrapp et al., 2020). In order for binding with ACE2 to occur, the RBDs must transition from the "down" to "up" state, exposing the receptor binding motif required for cell docking. At the onset of our study, little was known about the continuum of conformational changes that occur between these two well-described states. For these reasons, it was an ideal system to study using ManifoldEM.

5.3 Materials and Methods

Preprocessing: The initial image-stack we received from McLellan and colleagues corresponding to PDB-6VSB (Wrapp et al., 2020) contained 631,920 snapshots. This initial image stack was pruned by approximately 10% (from 631,920 to 578,588 particles) to remove images with artifacts. Additional 3D Auto-Refinement via RELION was performed to realign all images. Next RELION 2D Classification was used to remove an additional 1% of particles, leaving the final count of 574,324. The consensus refinement in RELION displayed a Fourier Shell Correlation ($FSC_{0.143}$) of 4.3 Å. In parallel, this stack was separately refined using cryoSPARC non-uniform refinement with a gold-standard FSC resolution of 3.5 Å.

The volumes resulting from these two refinements were next compared within the preliminary steps of ManifoldEM. Although these volumes appeared highly similar, upon closer examination, we found that the RELION refinement produced a problem of preferred orientations, where thousands of particles had been clumped within nearly the same local area (i.e., nearly

identical angular coordinates) on S^2. In contrast, the cryoSPARC non-uniform refinement produced much more uniformly-distributed angular assignments, albeit with typically less occupancy per PD. 2D conformational coordinate movies obtained in ManifoldEM from the cryoSPARC alignment proved superior to those using RELION. While the cryoSPARC alignment was chosen for all subsequent steps in ManifoldEM, the RELION protocol was not altogether without its own merit. Additionally, F. Acosta-Reyes ran RELION focused 3D Classification using the angular alignment from cryoSPARC with a mask around the RBDs, and obtained classes with different configurations of the RBD, including one class in the RBD-down conformation (Figure 10). The original study (Wrapp et al., 2020), in contrast, found no such particles, nor did any other labs to which the data were sent for further analysis. Importantly, the discovery of these missing particles explains the presence of the RBD-down volumes constructed along the 3DVA "reaction coordinate" discovered in McLellan's study (Wrapp et al., 2020).

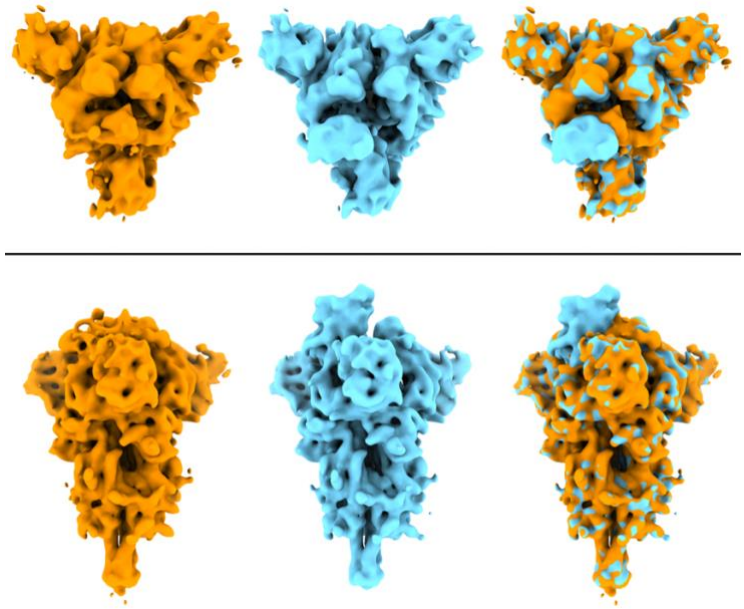


Figure 10: S protein classes from RELION focused classification. Comparison of two S protein classes from the focused classification in RELION with top and side views of the reconstructed classes. Volumes are low pass filtered to 8 Å for display purposes. The class with the RBD "down" conformation is displayed with orange on the left, the class with the RBD "up"

is displayed with cyan in the center, and the superposition of both volumes is shown on the right to highlight their differences. (Figure by F. Acosta-Reyes).

Manifold Embedding. We next set up a more thorough ManifoldEM analysis using the cryoSPARC alignment. First, recall that a number of initial inputs are required for the ManifoldEM pipeline to tessellate the angular space into a finite number of PDs. These are (i) Pixel Size: 1.047 Å; (ii) Resolution: 3.5 Å; (iii) Object Diameter: 335 Å; and (iv) Aperture Index. After experimenting with several aperture indices and evaluating the corresponding PD statistics and 2D movie qualities, we chose aperture index 5 for all future computations. This measure provided us with 1678 PDs thoroughly spread out in angular space, with a handful of regions with heightened PD-occupancy. When displayed as a histogram, the occupancy of PDs exhibited a chi-squared distribution, with the majority of PDs housing around 230 images and a rightward tail reaching approximately 800 images in the most highly-occupied PD.

Following the ManifoldEM framework, 1678 manifolds were constructed from the images in each corresponding PD via the diffusion maps method. NLSA was then performed on the eigenvectors of these embedded manifolds to extract a set of possibly relevant conformational coordinates from each. In sum, these steps were programmed to produce eight 2D movies per PD, with each 2D movie corresponding to one of the PD-manifold's eigenvectors.

Conformational Analysis. Upon completion, our task was to next classify the type of motions seen in each 2D movie per PD, noting that not all 2D NLSA movies extracted must necessarily correspond to valid conformational information; this is especially true of those obtained with smaller eigenvalues. Our approach was to initiate a search to detect all PDs housing 2D NLSA movies with above-average visual appearance. In this search, many PD-manifolds were found to have extremely noisy or otherwise insensible content (e.g., as shown in Movie D4-D). This was a predictable scenario given the known deficiencies in the Wrapp et al. (2020) data set

(i.e., orientational bias leading to low occupancies in many PDs), and beyond remediation by ManifoldEM. As a result, only a subset containing 1678 PDs—where the images therein met the prerequisites for the manifold-embedding approach—could be sensibly analyzed for conformational information. Of these, we found that 216 PDs (13%) were above-average quality, while 73 PDs (4%) were high-quality; as judged by visual inspection relative to the whole (e.g., Movie D4-A, D4-B and D4-C). Thus, overall, a relatively small percentage of the data as partitioned into these PDs met the prerequisite conditions for displaying the informative conformational variation signals.

We next organized all above-average PDs into 22 well-spaced groups on S^2, and selected several of the best PDs from each angular region. Detailed analysis was performed on the 64 PDs chosen, including classification of conformational motion type in each of the eight 2D NLSA movies per PD (Movie D4). As shown in Movie D1 and Movie D2, we predominantly observed two conformational motions: (i) RBD-down to RBD-up; and (ii) trimer-claw close to open, which we call CM_1 and CM_2, respectively. However, PDs where a clear distinction existed between CM_1 and CM_2 were rare. Specifically, CM_1 alone could only be clearly established in 31 of 64 PDs (48%); while both CM_1 and CM_2 were found occupying separate 2D movies in only 6 of 64 PDs (9%). In the remaining PDs, these conformational motions were not cleanly separated but were present in hybrid form (e.g., Movie D3-A and Movie D4-C). As previously addressed, such ManifoldEM anomalies will be discussed separately in Chapter 7 on ManifoldEM limitations and uncertainties.

Hybrid motions notwithstanding, PD-manifold discrepancies can arise from the nature of the ManifoldEM analysis, where Euclidean distances are defined between images that are projections of the molecule. As a result, from a given viewing direction, a 3D motion projected

onto 2D will appear more or less pronounced than it does in some other, depending on the type of motion and viewing angle. For example, we found that the CM_2 trimer claw motion was most pronounced only when observed from the "top-down view"; i.e., the PD aligned with the axis of the protein's central alpha helices (Movie D2). In all other PDs, this motion was exceptionally rare, and it is possible that its observance—which was not described by other groups using the same data set—is only afforded by careful dissection of the image ensemble into discrete PD-manifolds. Throughout this work, this notion will be referred back to under the label *PD disparity*.

The feasibility of 3D reconstruction of CM_1 NLSA states alone was investigated by S. Maji using the belief propagation algorithm, which was informed by detailed analysis of all 2D NLSA movies as assigned by E. Seitz, F. Acosta-Reyes and S. Maji. A collection of 2D NLSA movies corresponding to the CM_1 anchor PDs chosen are displayed in Movie D3-B. In the end, however, this data set proved to have far too few PDs of sufficient quality to recover informative 3D Coulomb potential maps for either CM_1 or CM_2. This realization provided an insurmountable challenge, since a small collection of 2D projections uniquely displaying conformational motions is insufficient for accurately elucidating 3D relations between structural domains.

5.4 Cross-Validation

In parallel to our work, the lab of Rommie Amaro had been using an entirely different method, performing all-atom MD simulations of the spike glycoprotein with experimentally accurate glycosylation. These simulations indicated an extensive shielding by spike glycans and a mechanistic role for glycans in supporting the RBD-open conformation (Casalino et al., 2020), and used full-length models of the glycosylated SARS-CoV-2 S protein as built using the structural information derived in the Wrapp et al. (2020) spike data set (the same that was used in our

ManifoldEM analysis). More recently, their work had shifted to characterizing the spike RBD opening pathway for the fully glycosylated S protein, so as to gain a detailed understanding of the activation mechanism (Sztain et al., 2021). Their group had used the weighted ensemble (WE) path sampling strategy (Huber et al., 1996; Zhang et al., 2010) to great effect on the S protein, which generates continuous pathways with unbiased dynamics as achieved by running multiple trajectories simultaneously, and pruning suboptimal routes. In total, WE simulations were computed over the course of approximately 50 days using supercomputers at the San Diego Supercomputer Center and Texas Advanced Computing Center. Outputs were extensively analyzed to characterize an as-of-yet unpublished series of transition pathways of the spike opening, which included identification of key residues that participate in the activation mechanism (Sztain et al., 2021).

The motions analyzed by the Amaro lab shared a striking resemblance to the subset of high-quality 2D NLSA movies generated by ManifoldEM. A collaboration was hence formed to make a formal comparison of each group's results, with the aim to potentially cross-validate findings from two disparate fields: simulation and experiment. Typically, in science, the fact that two entirely independent studies using different methods come to the same conclusion makes the results ironclad. However, these two techniques could not be immediately compared, since the output of a MD simulation is a trajectory of (several thousand) 3D atomic structures while, via ManifoldEM, only 2D NLSA movies were acquired from specific views of a 3D Coulomb potential profile. To bridge this gap, a direct comparison was conducted, such that the PDB files from the WE spike-opening trajectory were converted into a collection of 2D projections.

Generation of WE Projections. First, 20 atomic structures from the WE trajectory were selected and imported into Chimera (Pettersen et al., 2004) along with a coarse volume obtained

from ManifoldEM to be used for alignment reference. In order to place both frameworks in the same coordinate system for subsequent analysis, these PDB files were translated and rotated to coincide with the ManifoldEM volume, using the Chimera "fitmap" command. Each PDB was then saved in Chimera. Next, these fitted PDBs were re-centered using Phenix (Liebschner et al., 2019) "pdbtools" and converted into MRC-formatted 3D electron density maps[7] via EMAN2 (Tang et al., 2007) "e2pdb2mrc". For this last step, a resolution of 5 Å was chosen based on visual assessment of the EMAN2 outputs relative to those from ManifoldEM. Projections of these 20 density maps were then taken using the standard projection operator in "e2project3d" with C1 symmetry in EMAN2. Importantly, the angular coordinates for these projections were supplied by those representing the 64 ManifoldEM anchor PDs (after correcting for a coordinate transformation from ManifoldEM to ZXZ' convention). Finally, these projections were combined into sequences for each PD to form 64 20-frame 2D movies of the WE trajectory.

Comparison of Method Outputs. As shown in Movie D1 and Movie D2, a striking visual resemblance emerged between conformational motions obtained by WE simulation and experiment. To aid the visual comparison, 2D movies from the WE simulation and from ManifoldEM corresponding to the same PD and CM were next overlaid to directly highlight similarities and differences. For this procedure, the ManifoldEM movie was first layered over a homogenous red backdrop with a "linear dodge" blend mode applied, and the same was done for the WE movie over a blue backdrop (Figure 11). The ManifoldEM composite image and the WE composite image were next multiplied together. As an outcome of this multiplication, pixels that are white (signal) in both movies retain their whiteness in the composite. In this way, whiteness in

[7] An inconsequential mistake is present in the original work (Sztain et al., 2021), where these outputs were erroneously defined as Coulomb potential maps.

the composite movie becomes a qualitative measure of similarity between conforming domains, while non-white regions emphasize differences.

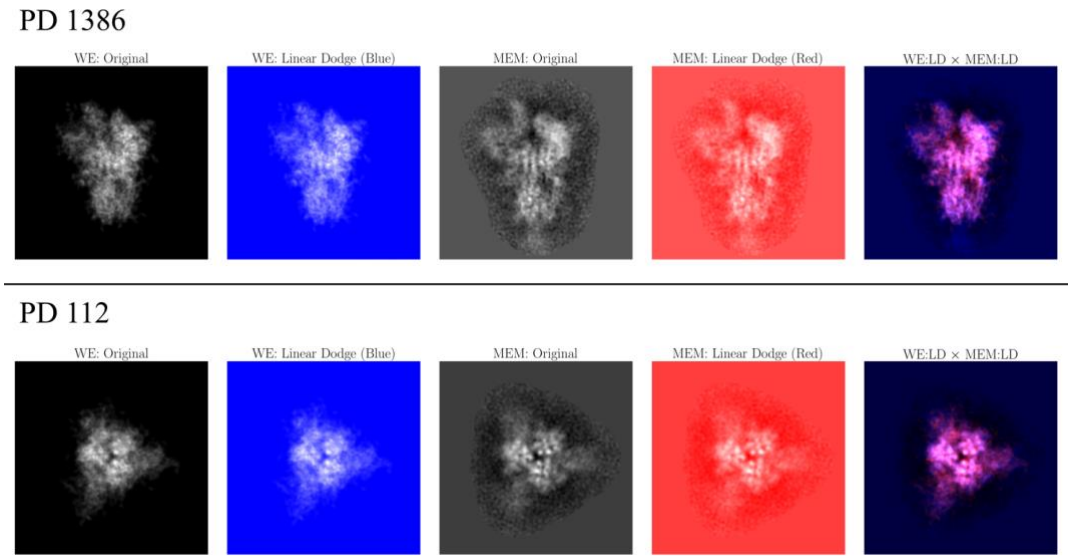


Figure 11: S protein comparison of frames from WE and MEM trajectory, as seen from a side view (PD_{1386}) and top-down view (PD_{112}). For this comparison, image compositing techniques are applied on the outputs of each method as shown in the columns, including Linear Dodge and Multiply. As an example of its utility, after performing this operation on CM_2 from a top-down view (PD_{112}), it can be seen that a collection of white pixels emerged in the composite movie (bottom-right entry), which strongly emphasize the similarities in positions of RBD and spike core helices between frameworks.

Finally, this overlaying approach was used to estimate the total extent of the RBD motion as expressed in the ManifoldEM and WE frameworks. For this comparison, CM_1 from a side view (PD 1386) was chosen based on its highly prominent view of RBD-up and RBD-down motion. Next the ManifoldEM movie was time-remapped to align it optimally with the motions observed in the corresponding WE movie (Movie D1). Using the multiplication-composite as a guide, it was determined that the ManifoldEM RBD domain reaches its full extent in the "up" position at the 14th frame out of the 20 frames from the WE trajectory, before the WE trajectory moves onward to a more fully open state. With this knowledge, the total difference in conformational extents was estimated at 11 Å as calculated via RBD—core distance (Sztain et al., 2021).

The results of this cross-validation have been formally published in Sztain, et al. (2021), which includes an in-depth analysis of the cross-validated WE results. Overall, there was very good agreement between the ManifoldEM conformational motions and WE trajectory, aside from two discrepancies. First, as described, the WE trajectory ultimately opens to an RBD—core distance 11 Å greater than the most open conformation in ManifoldEM. This is likely because the simulations sample the changes of the S1 subunit en route to the post-fusion conformation, whereas the experimental data set does not (Sztain et al., 2021).

Second, CM_2, as observed in top-down PDs produced by ManifoldEM, included concerted opening of all three RBDs (i.e., trimer opening), while the WE focused sampling on the opening of a single RBD. However, from side-view PDs produced by ManifoldEM, CM_1 and CM_2 corresponded only to the motion of one RBD, in agreement with the WE trajectory. Several aspects challenge the functional validity of the trimer-opening event: (i) the fully-open three-RBD end state is traditionally considered an energetically-unstable configuration; (ii) the trimer motion was localized to PDs positioned within a very small angular region on S^2; and (iii) in these top-down PDs, all three RBDs are symmetrically aligned orthogonal to the air-water interface. Thus, there is enough evidence to suggest that the trimer opening is a rare event resulting from symmetric air-water interface interactions with the peripheral ends of each RBD. In any case, this uncertainty underscores the importance of properly accounting for all experimental interactions when interpreting ManifoldEM outputs.

In summary, by examination of over 300 continuous, kinetically unbiased RBD opening pathways—supported by ManifoldEM analysis and biolayer interferometry experiments—a novel, atomic-level characterization of the glycosylated spike activation mechanism was achieved. Structural analysis of these ~130 μs WE simulations determined (among many features) a gating

role for the N-glycan at position N343 which facilitates RBD opening, alongside participation from residues D405, R408 and D427 (Sztain et al., 2021). These results are a significant achievement for both ensemble pathway simulations and ManifoldEM. For the latter, our approach opens up new avenues for ManifoldEM application as a cross validation tool, even in circumstances where there is a paucity in quality of conformational information.

5.5 Closing Remarks

In the midst of a devastating pandemic sweeping the globe, the readiness of these methods and the preparation of those involved proved vital for our understanding of the main antagonist, the SARS-CoV-2 spike protein. This understanding was enabled by the speed and breadth of each approach, and of the cryo-EM methodology in general, which has revolutionized our ability to quickly visualize molecular machines occupying numerous states, alongside interactions with receptors and antibodies (Rapp, 2021). All in all, this work was a small but important contribution alongside a growing body of scientific achievements which, as a whole, have greatly increased our chances of success in fighting this pandemic.

Chapter 6: The Synthetic Continuum Framework

6.1 Motivation

Previous chapters have addressed several limitations and uncertainties identified for the ManifoldEM framework during analysis of V-ATPase and SARS-CoV-2 S protein. In the latter study, some of these problems were critical enough to disallow the attainment of 3D conformational information, while in the former, cohesive information could still be obtained for one CM. Several of these issues were foreshadowed by previous uncertainties uncovered during the initial testing phase of the Python software (2018-2019). During that period, our team consistently encountered a common set of problems when analyzing a number of experimental data sets—including the ribosome (Dashti et al., 2014) and RyR1 (Zalk et al., 2015)—which were broadly categorized. Indeed, all of the issues found in the analysis of V-ATPase and SARS-CoV-2 S protein are also well described by these categories. While some of these problems can be attributed to potential shortcomings in certain data sets, others are specific to the methodology employed by the ManifoldEM framework, regardless of data quality. The ability to distinguish between these two scenarios is crucial for researchers, and essential for increasing the validity and relevance of ManifoldEM results.

However, in the absence of ground-truth data, adequate testing and validation of these methods is impossible, since the free-energy landscapes of molecular machines are not experimentally known. The ManifoldEM framework must instead be tested first on simulated cryo-EM images of a synthetically-designed ensemble having known structural and statistical properties. To this end, members of the ManifoldEM group at Columbia University (E. Seitz, F. Acosta-Reyes and J. Frank) and University of Wisconsin-Milwaukee (P. Schwander) presented a workflow in late 2019 for creating a synthetic quasi-continuum of structures from an

experimentally-relevant ground-truth model (Seitz et al., 2019; Seitz 2019)[8]. This synthetic generation protocol broadly emulates the statistical ensemble of macromolecular projections obtained from single-particle cryo-EM experiments, including: (i) the possibility of numerous independent, easily identifiable conformational motions; (ii) an explicit distribution of occupancies across all states; (iii) additive Gaussian noise with experimentally-plausible signal-to-noise ratio (SNR); and (iv) experimentally-accurate microscopy parameters for each image, represented by the CTF. To note, in the time since its conception, this synthetic framework (Seitz et al., 2019) has already been used as a benchmark in two external cryo-EM studies (Gupta et al., 2020; Giraldo-Barreto et al., 2021).

Given such absolute control and knowledge of the baseline object's mechanics, the accuracy and appropriateness of the ManifoldEM approach can be quantified, providing an assessment of its inherent strengths and limitations. Additionally, those uncertainties or limitations that instead arise due to experimental shortcomings—or by decisions made external to the ManifoldEM methodology—can be isolated. In the following sections, this synthetic generation protocol is presented, with a more detailed account available in Seitz et al. (2019) and Seitz et al. (2021a).

6.2 Simulation of Cryo-EM Images of an Ensemble of Molecules

Here the protocol for simulating cryo-EM ensembles from atomic models of molecules exhibiting continuous conformations is described. Our procedure for the creation of a synthetic

[8] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; software; validation; manuscript draft, review and editing. F. Acosta-Reyes: Methodology; software; validation. P. Schwander: Methodology; software; validation; manuscript review and editing. J. Frank: Conceptualization; direction and project administration; manuscript review and editing.

continuum is outlined in Figure 12. 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 (PDB; Berman et al., 2003). 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 CMs 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.

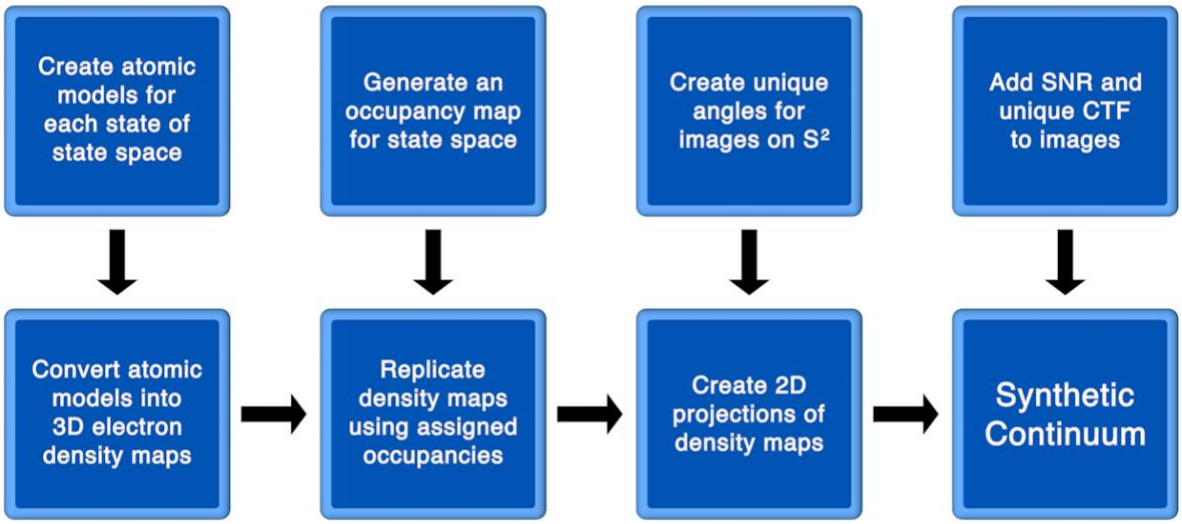


Figure 12: Flowchart for synthetic continuum generation protocol. Note again the erroneous use of "Coulomb potential map" in the original manuscript (Seitz et al., 2019), which has been replaced here—and in all similar uses that follow—with "electron density map" (EDM). For the purposes of manifold embedding, this difference is negligible, but may prove undesirable for studies requiring a heightened realism of interatomic structure.

For use in describing this protocol, 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 (Schopf et al., 2017). 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 (Ali et al., 2006).

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 are referred 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 1, 2 and 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. For use as a benchmark, ideally only one of these such state spaces would be analyzed at a time.

To establish a notation for use here and in future chapters, these state spaces are termed SS_1, SS_2 and SS_3, defined by: (i) 20 states exhibiting one degree of freedom $\{\mathrm{CM}_1\}$; (ii) 400 states (20×20) with $n = 2$ $\{\mathrm{CM}_1, \mathrm{CM}_2\}$; and (iii) 1000 states $(10 \times 10 \times 10)$ with $n = 3$ $\{\mathrm{CM}_1, \mathrm{CM}_2, \mathrm{CM}_3\}$, respectively. As a specific example of the ranges of motion present in SS_3, between neighboring states in each CM, the Root-Mean-Square Deviation (RMSD) was calculated (Schrödinger LLC, 2015), 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.

It is also important to describe the order in which the states for each state space are indexed, with this ordering repeatedly visualized by color maps throughout this work and used to locate each state's coordinates within a given embedding. The ordering of SS_1 is trivial, with states following a sequence showing CM_1 transition from closed to open. Movie E1 should be referenced for the ordering of SS_2, which provides an explanation of the "clockwork" indexing used, while providing an animation of the molecule transforming through images in this sequence. A similar pattern holds for SS_3, which now additionally includes an hour hand in the analogy (CM_1), followed by a minute hand for CM_2, and a second hand for CM_3.

Next, an exact description of the atomic-coordinate displacements between states for the specific case of SS_2 is provided. CM_1 and CM_2 were designed so as to be fully decoupled from each other, such that no overlap occurs between the two sets of distinct atomic displacements (Figure 13). Atomic manipulations of the original PDB coordinate file were done using PyMOL (Schrödinger LLC, 2015). For the first motion (CM_1), the chain A arm was rotated outwards (directly away from chain B) from its central hinge at residue 677, in 1° increments until a series of 20 rotational states were defined (Figure 14-A). For each of these 20 CM_1 states, all residues above chain B's elbow (residues $12-442$) were rotated perpendicularly to CM_1 in 2° increments until a series of 20 states were defined for the second motion (CM_2) as seen in Figure 14-B.

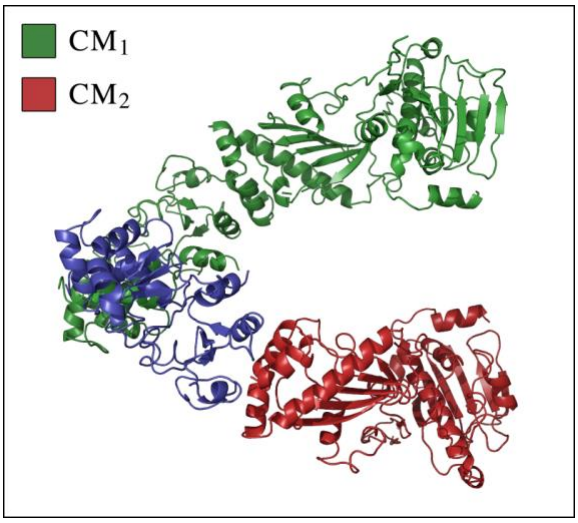


Figure 13: Hsp90 atomic-coordinate structure displaying CM regions. In this cartoon representation of state 20_01, residues are demarcated for CM_1 in green and CM_2 in red. The CM_1 central hinge can be found near the intersection of the blue and green regions at residue 674, while the CM_2 hinge is found near the intersection of the blue and red regions at residue 443. For SS_2, the residues making up the blue region are immobile throughout the entire state space.

These operations resulted in the creation of a total of 400 unique conformational states, as visualized in Movie E1. The ensemble of these states can be organized in a 20×20 state space, where each entry represents one of the possible combinations of CM_1 and CM_2. This state space covers our synthetic model's complete ensemble of physically allowable conformations (i.e., a quasi-continuum). The specific size of the state space (400 states) was chosen based on the relative scale of the protein and range of its motions, providing approximately 3 Å and 2 Å gaps between states over a total arc length of 60 Å and 40 Å for CM_1 and CM_2, respectively (as visualized in Figure 14). To note, geometry correction and energy minimization of generated states were skipped to avoid introduction of unintentional coupling of CMs.

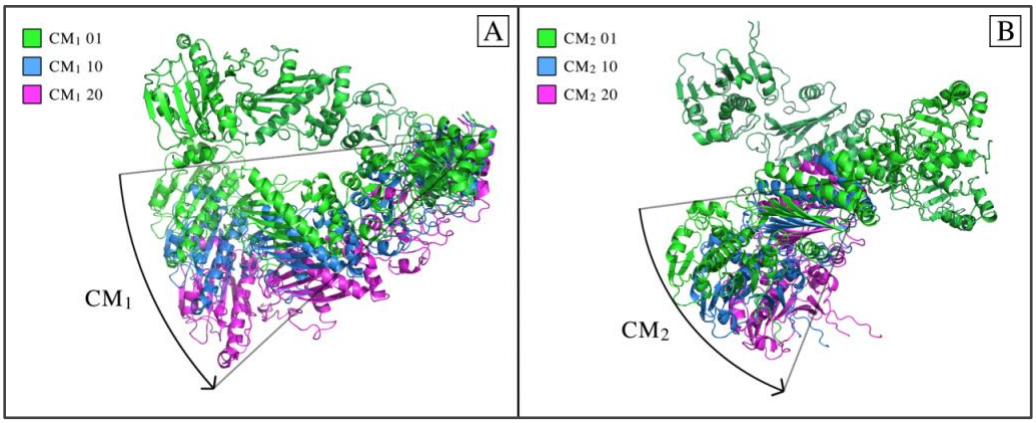


Figure 14: Hsp90 atomic-coordinate structures displaying two CMs. In figure [A], a comparison of the CM_1 first, middle, and final state is shown. Similarly, in figure [B], a comparison of the CM_2 first, middle, and final state is shown. The CM_2 rotation was performed perpendicularly to the CM_1 rotational hinge, such that only those atoms above chain B's elbow-region were displaced. As a note, to remove the potential for overlapping residues within certain states, the loose tail formed by residues 1−11 was removed from both chain A and B.

These 400 structures represented by PDB files were then each transformed into 3D electron density maps (formatted as MRC volume files) with a sampling rate of 1 Å per voxel and simulated resolution of 3 Å (chosen based on the resolution determined for PDB-2CG9) using the EMAN2 (Tang et al., 2007) module "e2pdb2mrc" (Figure 15). As a note, this module excludes calculations such as atomic form factors and molecular orbitals, and while more accurate maps can be constructed using other methods, this does not affect the results of our forthcoming heuristic analysis. Projections of each 3D EDM can then be obtained via standard parallel line integrals along the direction of the electron beam so as to simulate images generated in a TEM operated in the bright-field mode (Frank, 2006). See Movie E1 for a conformational animation of the molecule as viewed from five example projection directions. This core framework can next be used as a basis for the creation of images accounting for numerous statistical properties, including modification of each image to incorporate SNR and CTF, and construction of unique free-energy landscape and angular distributions on S^2. In the following sections, the execution of several of

these options will be detailed, with corresponding code for each of these steps—and all previous ones—available in our online repository (Seitz, 2019).

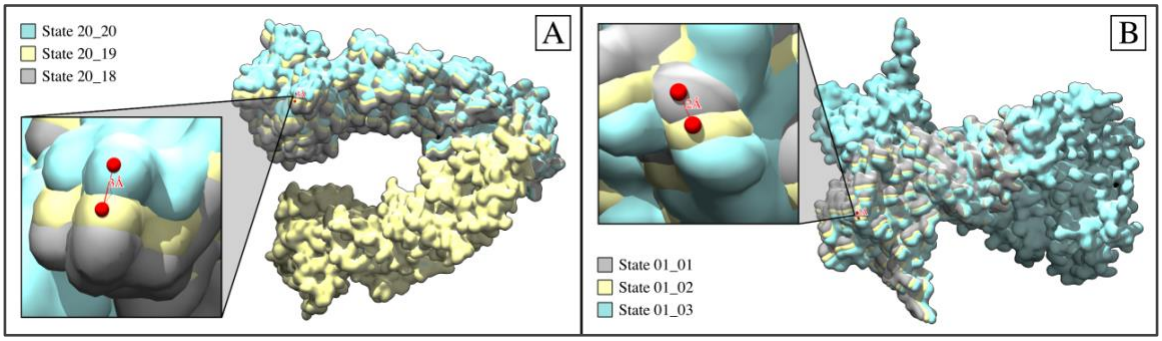


Figure 15: Hsp90 electron density maps displaying two CMs. In [A], a volumetric overlay of the first three rotational states of CM_1 (state 01_01, state 02_01, state_03_01) is presented, visualized as EDMs (MRC format) via Chimera. As can be seen, only the upper arm (chain A) has been rotationally displaced along CM_1, with 3 Å gaps measured between each consecutive state at the peripheral ends of this rotated region. In [B], a volumetric overlay of the last three rotational states of CM_2 (state 20_18, state 20_19, state 20_20) is shown. Only the upper region of chain B (above the elbow) has been rotationally displaced along CM_2, with 2 Å gaps measured between each consecutive state at the peripheral ends of this rotated region.

6.3 Simulation of Noise

Additive Gaussian noise can be applied to each image individually to grant it a specific SNR that is the same for all images in the set. We define the SNR by the ratio of each image's signal variance (σ^2_{signal}) to its noise variance (σ^2_{noise}), as described by J. Frank (2006). Here, signal represents the 2D region of pixels corresponding to the average area occupied by the macromolecule. This region is obtained by masking out all pixels within one standard deviation of each image's mean intensity value; in effect, excluding the approximately uniform-intensity background. This process is thus performed by first finding the mean pixel intensity (μ_{signal}) and variance (σ^2_{signal}) of the signal, and then calculating

$$\sigma^2_{noise} = \sigma^2_{signal} \,/\, SNR \tag{3}$$

Using this parameter, additive Gaussian noise is applied to each image in order to obtain an output image having the desired SNR. In this process, a sample from the Gaussian distribution is added to each pixel's intensity. Each resulting image is then normalized such that the average pixel intensity and standard deviation of pixel intensities is approximately 0 and 1, respectively. As a note on experimental relevance, SNR of 0.1 has been previously established as a suitable choice for experimental SNR in images obtained by cryo-EM, and can be attributed to the low contrast between macromolecules and their surrounding ice, as well as the limited electron dose required to avoid radiation damage (Baxter et al., 2009).

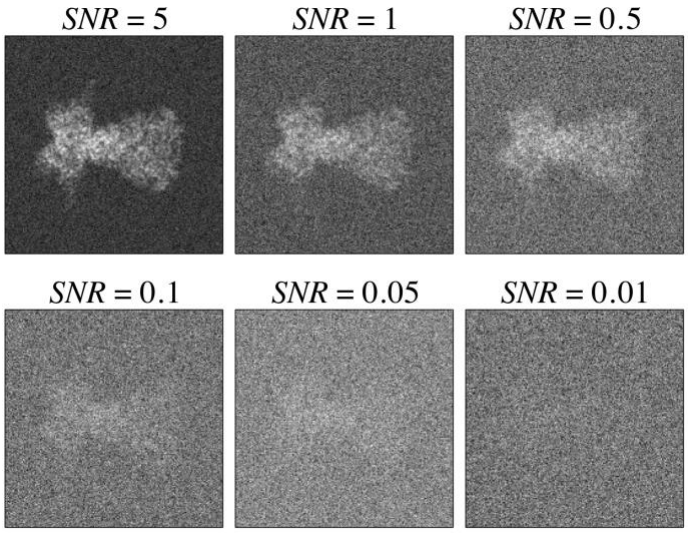


Figure 16: SNR simulation on projections of Hsp90. First image in SS_1 from a PD with different values of SNR, as obtained via additive Gaussian noise with appropriate statistics.

6.4 Simulation of CTF

The presence of experimental aberrations can be further increased through application of a contrast transfer function (CTF) with realistic microscopy parameters and random defocus (within the typical range expected in a TEM). For this scenario, each image is first filtered by the TEM's CTF (via the equation introduced in Chapter 1) in an experimentally-relevant range. As an example

of use, we assigned each image a random defocus value from the interval $[5000, 15000]$ Å (positive is underfocus). This wide range was chosen to compensate for the zero-crossings of CTFs where no information is transferred, with similar intervals typically used in modern cryo-EM experiments. Likewise for each image, constant values were 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. This CTF is next applied through scalar multiplication with the Fourier transform of the image, followed by an inverse Fourier transform of the product. After this procedure, additive Gaussian noise is uniquely applied to each image (SNR of 0.1) following protocol in the previous section. Figure 17 provides a demonstration of this workflow as applied on an image stack pertaining to a given PD, which further includes results of CTF correction.

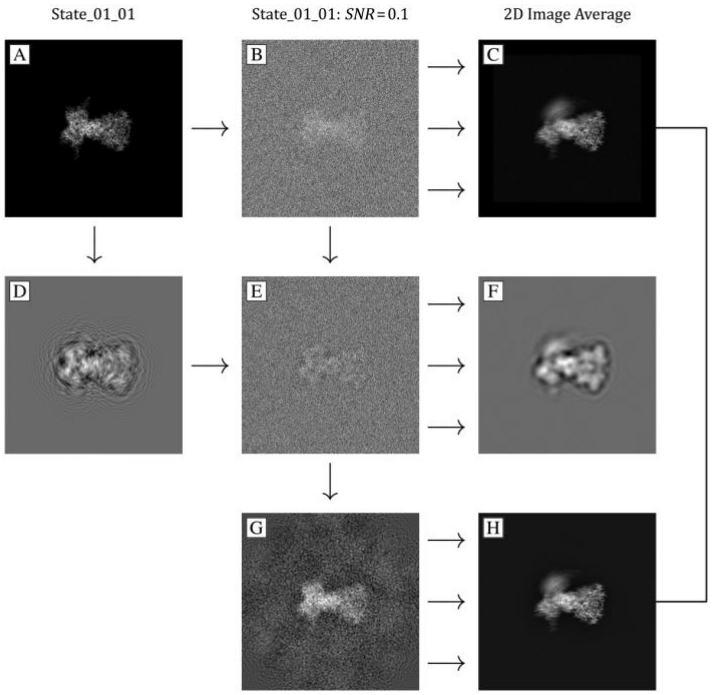


Figure 17: CTF simulation and correction on projections of Hsp90. In the first column, image [A] is a 2D projection of state 01_01 of a given PD taken without CTF parameters using the

EMAN2 module "e2project3d". Image [D] is that same 2D projection but generated with CTF signal using parameters using the following parameter values: defocus 13,500 Å; amplitude contrast ratio 0.1; spherical aberration 2.7 mm; and voltage 300 kV. For both the first and second row, there are 399 other images each, created with respective attributes corresponding to the remaining SS_2 states. All of the 399 images generated similar to [D] were likewise given a random defocus value in the interval [5000, 15000] Å. Image [B] and [E] are the results of adding Gaussian noise to each image [A] and [D], respectively, such that the SNR is 0.1. As depicted with three cascaded arrows in the diagram, image [C] is the 2D average of all states generated similar to image [B], and likewise for the relationship of image [F] to image [E]. The image size is 320×320 pixels, with this decision guided by [D] such that the broadening of the point-response function stays within the image bounds. Image [G] is the CTF-corrected version of image [E], as depicted by J. Frank (2006) using a Wiener filter with SNR of 0.1 and exact assignment of known CTF parameters. Finally, image [H] is the 2D average of all CTF-corrected images similar to image [G]. Note the similarity of [H] with [C], which is a result of filling the missing CTF zero-crossings during the averaging of all CTF-corrected images.

6.5 Simulation of Occupancy Map

For the synthetic ensemble, an occupancy map must be created to represent the preference of a molecule to occupy each state, so as to simulate the number of sightings of each conformation that would be obtained in thermal equilibrium via cryo-EM. Each entry in this occupancy distribution is an integer used to define the number of clones produced for a given state. Recall that since the number of sightings of each conformation observed in experiment directly relates to its underlying energetics, this distribution can be transformed via the Boltzmann factor into a corresponding free-energy landscape. Figure 18 demonstrates one such arbitrary, nonuniform distribution chosen for assignment of SS_2 occupancies.

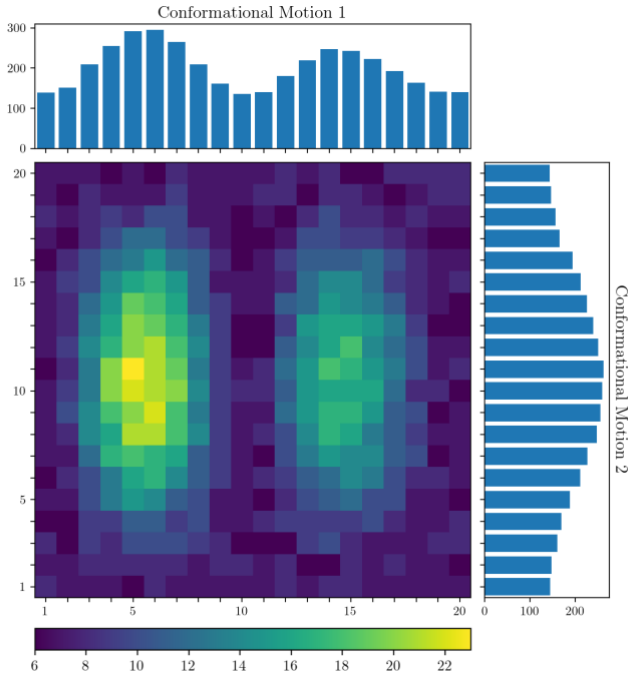


Figure 18: 2D occupancy map generated for Hsp90 for each of the 400 states in SS_2, assigned identically for each PD. The net occupancy of this state space is 4000, such that the total number of images in the complete data set is the product of 4000 and the number of PDs. The characteristics of this occupancy map were chosen to provide easily distinguishable features along both 1D and 2D CMs: specifically, bimodal and unimodal distributions for CM_1 and CM_2, respectively.

6.6 Simulation of Randomized Uniform Angular Distribution

As each experimentally-obtained image naturally captures a macromolecule from a given viewing direction, angular assignments must be given to map the position of each image within the angular space S^2. The location of each snapshot in this space is represented by a combination of orientational angles—each combination identifying a single point on S^2—which need to be defined in the columns of each row within the alignment file (e.g., Euler angles if using the RELION STAR format). The distribution of the points on S^2 depend on the needs of the analysis. As such, we developed a method to produce evenly clustered images on S^2, which additionally

emulates the naturally-occurring variations in orientational coverage of a cryo-EM experiment (Figure 19). Since each local cluster of images is given the same occupancy statistics, the PD-manifold technique can be applied to investigate the corresponding manifold of each cluster, with results compared to ground truth (and similarly, as it applies, to global-manifold approaches).

To generate this image distribution, we first form a polyhedron with a set number of segments (e.g., 48 are shown in Figure E1), which in turn creates a corresponding number of evenly spaced vertices (e.g., 812) on S^2. Each of these vertices defines the location of a set of images (i.e., a projection direction). For each image, in order to vary its position with respect to the assigned vertex, two sequential quaternion transformations are performed. These transformations were designed to preserve uniform rates of angular changes for all images across S^2 (Figure E2). First, the initial combination of Euler angles for a given image's corresponding PD (Figure E1) is transformed into an equivalent unit vector $\mathbf{v_0} = (v_{0_x}, v_{0_y}, v_{0_z})$. Next, a unit vector, $\mathbf{w} = (w_x, w_y, w_z)$, orthogonal to $\mathbf{v_0}$, is generated. A unit quaternion is then generated to represent a rotation about the axis $\mathbf{w}$ by an angle θ_1 as follows:

$$\mathbf{q_1} = \cos(\theta_1/2) + w_x \sin(\theta_1/2)i + w_y \sin(\theta_1/2)j + w_z \sin(\theta_1/2)k \qquad (4)$$

The rotation angle θ_1 for each image is chosen from a random Gaussian distribution with zero mean (the center of the PD) and a standard deviation of 1.25°. The vector $\mathbf{v_0}$ is then rotated according to $\mathbf{v_1} = \mathbf{q_1} \cdot \mathbf{v_0} \cdot \mathbf{q_1}^{-1}$, which effectively rotates $\mathbf{v_0}$ radially outward from its original location (the PD center) on a great circle, spanning a range of approximately 6° on S^2. Next, $\mathbf{v_1}$ is randomly rotated by a second unit quaternion with axis of the PD's original location, $\mathbf{v_0}$, by θ_2:

$$\mathbf{q_2} = \cos(\theta_2/2) + v_{0_x} \sin(\theta_2/2)i + v_{0_y} \sin(\theta_2/2)j + v_{0_z} \sin(\theta_2/2)k \qquad (5)$$

The angle θ_2 is chosen randomly from a uniform distribution $[-180°, 180°]$ for each image, such that the initial Gaussian distribution across each PD (formed from numerous possible random

samples of $\mathbf{v_1}$ locations for each image) is transformed into an isotropic Gaussian distribution about the center of each PD (Figure 19). This second transformation, $\mathbf{v_2} = \mathbf{q_2} \cdot \mathbf{v_1} \cdot \mathbf{q_2}^{-1}$, defines the final vector for each image, which is stored in the alignment file for calculating subsequent projections of the 3D EDMs.

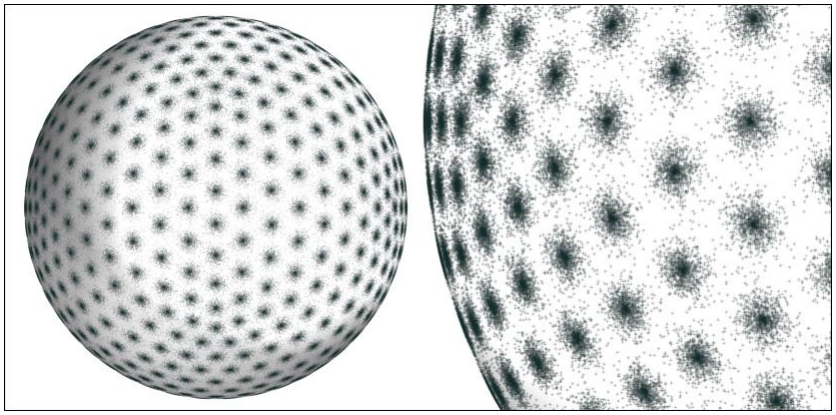


Figure 19: Angular distribution generated via quaternion transformations, starting with a base polyhedron with 48 segments. The final distribution of 812,000 images on S^2 after the back-to-back quaternion transformations with randomized sampling of θ_1 and θ_2 for each image. As can be seen, the Gaussian distribution defining the radial distance of the first transformation has been chosen here such that little overlap occurs between the locations of images in neighboring PDs. Importantly, these perturbations are designed to reproduce the naturally-occurring orientational variations of a real experiment.

6.7 Simulation of Great Circle Angular Distribution

Alternatively, it may prove useful—in terms of minimum information required for an adequate tomographic reconstruction—to only sample a subset of the angular space. For synthetic data created for the purpose of computing final 3D reconstructions, a sufficient number and arrangement of PDs is required. Specifically, the minimum number of equispaced PDs (P_{min}) on a great circle required for 3D tomographic reconstruction at a given resolution is obtained from the Crowther criterion (Gilbert, 1972)

$$P_{min} = \pi D / r \tag{6}$$

In this expression, 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. As used later in this work, $r = 3$ Å is chosen to match the resolution of the ground-truth EDMs. According to this criterion, we generate 126 equidistant PDs spaced 1.5° apart along one half of a great circle, chosen so as to avoid redundant information due to diametric mirroring.

6.8 Closing Remarks

After completion of all (or some subset) of the above subprocesses, images are combined into a single image stack. Finally, an alignment file is produced containing all relevant parameters (i.e., orientations and microscopy information) used to synthesize each image within the stack. Importantly, this workflow can be completely customized to meet the needs of any user, including choice of atomic model, number and types of CMs, density and distribution of states, occupancy of state space, distribution of projection directions, microscopy parameters and noise. This flexibility allows for comprehensive investigations of published data-analytic techniques, highlighting the performance and limitations of these approaches under a number of complex conditions. All scripts for reproducing and customizing this workflow, including additional steps not described here, are available online (Seitz, 2019).

Chapter 7: ManifoldEM Limitations and Uncertainties

7.1 Background

In the introduction of the previous chapter, a collection of ManifoldEM limitations and uncertainties—as discovered for a number of experimental data sets—were alluded to, ultimately motivating an investigation of the ManifoldEM framework using a synthetic quasi-continuum of structures from an experimentally-relevant ground-truth model (Seitz et al., 2019; Seitz, 2019). In this chapter, an examination of the results of ManifoldEM on this synthetic data is conducted to untangle the previously convoluted trends seen in experimental data sets, allowing for the prominence and significance of these deviations to be quantitatively assessed.

In the time since that analysis (conducted by E. Seitz, S. Maji and H. Liao), several of these issues have been briefly documented by G. Mashayekhi (2020) in the ManifoldEM Matlab repository, involving a mixture of limitations due to experiment and methodology. A more thorough discussion of those issues has been recently provided in the user manual of the ManifoldEM Python suite (Seitz et al., 2021b), with a similar narrative to follow. By examining both experimental and synthetic data sets, our group determined a set of common issues stemming directly from the ManifoldEM approach. These are: (i) Boundary problems; (ii) CM ambiguity; (iii) CM propagation; (iv) dampening of 3D motion amplitudes; and (v) inaccurate determination of occupancies. For completeness, the details of these individual issues, including examples of their occurrences, are described in the sections that follow.

7.2 Boundary Problems

For a typical PD embedding, the 1D occupancy map and corresponding free-energy path obtained for a given CM displays a prominent, irregular pattern near the boundaries, regardless of

data set. Specifically, there is a substantial, unaccounted increase in occupancies of states along each CM nearest the peripheral regions. The inconsistencies near each boundary of the 1D occupancy map first emerge in each PD individually, and thus also accumulate in the overall, combined occupancy map for all PDs and corresponding free-energy landscape. In experimental data sets—such as RyR1 (−ligand), V-ATPase, and the S protein—we have found that roughly 5% to 10% (on average) of the states near the boundaries are affected. As an explicit example, the 1D occupancy map obtained for V-ATPase is presented at the end of Movie B1. For synthetic data (Hsp90; Seitz et al., 2019), these occupancy misassignments were found to be nearly four times the expected value at the first and last state, and decreasing in severity over a region spanning around 15% of states on each side (30% total).

Importantly, these boundary characteristics result in biophysically-unrealistic deep free-energy wells at the border regions of the landscape. When generating 2D occupancy maps, with two such CMs in one map, these trends occur mainly in the four corners of the plane. In the event that an energetic route encroaches into these regions, care must be given to these discrepancies. As shown in the NLSA volumes of V-ATPase (Movie C2), these boundary aberrations further manifest as confused conformational variation signals in the corresponding NLSA volumes, which is a consequence of images in these high-occupancy states carrying conflicting conformational information (i.e., the image content is not monolithic). In the analysis of V-ATPase, these peripheral NLSA volumes additionally proved to be inadequate for modeling a set of cohesive atomic-coordinate structures using MDFF, ultimately serving to only obstruct the analysis of structure in these regions (Movie C3).

7.3 CM Ambiguity

The following prongs pertain to limitations and uncertainties encountered when comparing the NLSA outputs for each PD. Recall that NLSA produces an independent set of outputs for each embedded PD-manifold eigenvector (for a delimited set of leading eigenvectors; e.g., eight). These outputs include—but are not limited to—a 2D NLSA movie and corresponding 1D occupancy map. By inspection of these 2D NLSA movies, decisions must be made to decide which eigenvectors represent a biologically-relevant CM, and assign to each of them a sense.

Duplicate CMs. In the set of NLSA outputs for each PD, the 2D NLSA movies along two different eigenvectors can sometimes be indistinguishable from one another—both visually and computationally—while providing significantly different occupancy distributions (e.g., as seen in Movie C1-A). Computational attempts to parse these CMs have included analysis of optical flow vectors and of the distance matrices formed from the furnished NLSA images. The appearance of duplicate CMs with different energies is neither desirable nor biophysically feasible. As the ranking of each eigenvector (via the eigenvalue spectrum) represents the significance of each conformational variation signal in a given data set, the duplicate CM corresponding to the highest eigenvalue index has been typically chosen. The case of identical types of CM content varying only in 2D motion amplitude has also been occasionally observed. For example, in the RyR1 (−ligand) data set, both a large and small "wing"-like motion of the peripheral subunits was present in two different 2D NLSA movies belonging to the same PD; occurring for several PDs.

Aberrant CMs. Certain 2D movies may showcase hybrid and even physically-impossible motions, and must be avoided. For PDs where these occur, the biophysically correct CM choice is often (but not always) available. NLSA movies that show a "hybrid" motion appear to display a combination of two different motions, with each corresponding to two different CMs, as are

observed from that same PD or neighboring PDs. As previously described, these motions were substantially present in our study of the SARS-CoV-2 S protein (e.g., Movie D4-C and Movie D3-A), and critically detracted from our ability to reliably assign CMs. As an example of a physically-impossible motion, we have observed 2D movies—each corresponding to one eigenvector—obtained from synthetic data (Seitz et al., 2019) showing structural domains moving in different directions at the same time, which is physically impossible. User discretion—and subsequent validation—is therefore essential when considering the motions observed in each 2D NLSA movie.

Erroneous Movies. For certain eigenvectors corresponding to potential CMs within a given PD, the NLSA movies may appear to stall while broadcasting a series of identical frames over the course of its duration (e.g., Movie D4-D). The occurrence of NLSA movies with numerous pauses usually occurs due to confused NLSA manifolds; possibly due to insufficient image counts or the presence of aberrant or misaligned snapshots. For these eigenvectors, while the NLSA approach produces smoothly-varying and well-connected 1-manifolds, the resultant 2D movie is not smooth (i.e., featuring abrupt pauses). As an automated remedy, during analysis of the SARS-CoV-2 S protein, S. Maji and E. Seitz introduced a strategy to specifically target and remove these eigenvectors as possibilities using τ' statistics during belief propagation.

7.4 CM Propagation Across the Angular Sphere

Errors in assignment of CMs and their senses may emerge during automation strategies present in the ManifoldEM Matlab (Mashayekhi, 2020) and Python (Seitz, 2021b) frameworks, with the latter (i.e., optical flow and belief propagation) observed to be significantly more accurate than the former (i.e., correlation coefficients). These errors can affect the quality of the final free-

energy landscape and reconstructed 3D motions. As the CM propagation step (Maji et al., 2020) implements feature extraction methods which are applied to each of the NLSA movies to determine their similarity, its performance is thus naturally complicated by the presence of duplicate and aberrant CMs, as described above. In the absence of upstream errors, the optical flow routine has been found to capture motions with good accuracy, even with significant noise present.

However, for data sets where the NLSA movies show relatively small, complex or obscure motions (i.e., obscure as seen from a specific PD), the performance of optical flow computations can suffer, and as a consequence, the choice of CMs may be inaccurate (Maji et al., 2020). To remedy these problems, the user-defined anchor PDs must be chosen as carefully as possible in order to obtain an accurate occupancy map and reconstructed 3D movies. As an added layer of supervision, the outputs of the Python automation strategy can be spot-checked after completion. If erroneous 3D motions are observed, the Python GUI allows previous decisions to be revisited, such that more anchor PDs (or more accurate assignments) can be added to better guide these decisions.

7.5 Dampening of 3D Motion Amplitudes

In several data sets, the first and last reconstructed NLSA volumes produced for a given CM have been observed to show a dampened range of motion compared to the corresponding 2D NLSA movies from various PDs. The extent of this dampening has been found to vary with the data set. For example, in comparison to 2D NLSA movies, we observed a noticeably smaller "wing"-like motion of the peripheral domains in the RyR1 ($-$ligands) NLSA volumes and, in the case of the spike protein, this dampening severely limited CM_1 motion: providing a movie showing the appearance of a single structure vibrating under thermal influence. The extent of the 3D motion

can appear dampened due to various factors in the ManifoldEM framework. As a rough estimate, we have found the dampening to result in a third or less of the motion range as observed in the corresponding NLSA movies or via external studies. To note, for larger motions, we have also observed a smearing effect in the 3D structure of the domain under motion, which was most significant for our synthetic data (Seitz et al., 2019).

7.6 Inaccurate Determination of Occupancies

Since the free-energy landscape of a molecular machine is in general unknown, an analysis of experimental data is unable to validate the landscapes generated by ManifoldEM. Instead, only synthetic data sets will suffice for this purpose. A comparison of the ground-truth 2D free-energy landscape of our synthetic continuum data set—with SNR and CTF in the range of experimental values—was thus performed with the corresponding results of ManifoldEM. Overall, the NLSA-derived occupancies shared little resemblance to our ground-truth distributions (even for individual PDs), and also displayed accentuated errors near the boundaries (i.e., "boundary problems"). A more detailed review of all aforementioned errors will be provided in Chapter 11 after introduction of the ESPER methodology, from which outputs will be compared directly with the NLSA approach for both synthetic and experimental data.

7.7 Closing Remarks

The occurrence of these limitations and uncertainties significantly complicates the analysis of molecular machines using ManifoldEM. The results of synthetic data sets are perhaps the most telling. Even given the availability of such perfectly-obtained data, these results show that there are fundamental errors in the ManifoldEM methodology that act to confuse (in the case of "CM

ambiguity"), deteriorate ("dampening of 3D motion amplitudes"), and even prevent ("inaccurate determination of occupancies") the ability to accurately acquire its contents. Notwithstanding the latter however, cohesive 3D motions are still obtainable by the ManifoldEM framework which, given the right circumstances, may be good enough to elucidate a significant amount of biological function. Such sequences of motions have been shown on the 3D level for V-ATPase (Movie C2), for example, and for previous studies on the ribosome and RyR1. For the study of such structures on energetically-favorable paths, these ManifoldEM limitations and uncertainties must be better understood and corrected.

Additionally, it is important to consider the potential for several upstream problems that—for all practical purposes—are beyond remediation by ManifoldEM, while still affecting the quality of its outputs. These externally-induced uncertainties may arise from the possibility of (i) suboptimal cryo-EM data quality, and (ii) misestimations of image parameters calculated in upstream algorithms. For the former, experimental data collected in cryo-EM can suffer from a wide range of nuisances, such as occurrence of aberrant particles (such as ice shards or foreign bodies), low image counts, preferred orientations, uncertainty in CTF estimation, and uncertainty in angular assignments (which are more pronounced in heterogeneous data).

For instance, the ManifoldEM issue of "erroneous movies" can occur due to the presence of aberrant cryo-EM images or low image counts. Likewise, "dampening of 3D motion amplitudes" will become more severe for data sets with low image counts or preferred orientations, which effectively limit the number and distribution of PDs required for adequate 3D reconstruction of conformational states. Further, a sparsity in PD coverage would also affect the efficacy of "CM propagation", since belief propagation would have fewer immediate neighbors in close proximity to each PD, resulting in less reliable propagation of conformational information. These issues

become increasingly significant as the quality of the cryo-EM data decreases, and can combine to amplify preexisting ManifoldEM errors observed for synthetic data. Naturally, these problems only become worse in the presence of the second form of uncertainty (ii), which is analyzed in detail using the Hsp90 synthetic continuum in the following chapter.

Chapter 8: Heuristic Analysis of Upstream Methods

8.1 Motivation

In the previous chapter, several problems encountered in the ManifoldEM methodology were described, potentially arising in part due to (i) suboptimal cryo-EM data quality; (ii) misestimations of image parameters calculated in upstream algorithms; and (iii) suboptimal performance of procedures within the ManifoldEM framework itself. The first and last of these uncertainties require more groundwork, and will be discussed in chapters to follow. Now, the effectiveness of upstream algorithms—specifically maximum-likelihood methods—will be explored, which are essential for initiation of ManifoldEM.

To conduct an analysis of these methods, the previously-described Hsp90 synthetic data set is used as input into prominent Bayesian structural-determination techniques—including RELION and cryoSPARC—with outputs contextualized within the scope of ground truth. The Bayesian (maximum-likelihood) approach has been widely used in the field of cryo-EM and consistently demonstrated the ability to recover structures at high resolution. However, while both RELION and cryo-SPARC contain techniques to handle some degree of heterogeneity, it is unclear what limitations may exist when dealing with large, highly heterogeneous data sets. Importantly, since the alignments estimated by these Bayesian approximations are prerequisite for application of ManifoldEM—and are beyond remediation by ManifoldEM—any upstream errors produced will ultimately affect the accuracy of all downstream ManifoldEM outputs. Therefore, our investigation served to assess the magnitude of such upstream uncertainties using our highly-heterogeneous synthetic data.

Briefly, the parameters for creating the Hsp90 continuum for this analysis exactly match those provided by Seitz et al. (2019). Images are generated from SS_2 with SNR of 0.1, CTF with

$\Delta z \in [2000, 10000]$ Å, and occupancies defined by a nonuniform occupancy map (modal and bimodal for CM_1 and CM_2, respectively). In total, 812 PDs were created—each housing 1000 images—with the ground-truth angular assignment for generating each image defined as is shown in Figure 19. The following results are unpublished, with analysis performed by E. Seitz and F. Acosta-Reyes (2019)[9].

8.2 Analysis of Bayesian Approach

To start, we ran our 812,000 synthetic images and alignment file (without ground-truth Euler angles) in the RELION "3D Auto-Refine" module, which aimed to define the angular distribution of our data set using maximum-likelihood classifications. For the required reference map, we used the state 01_01 volume generated in our synthetic workflow, with C1 symmetry, 30 Å low-pass filter, 200 Å mask diameter, and 7.5° initial angular sampling. Figure 20 shows the average volume and angular distribution obtained from this process after 23 iterations. From a qualitative perspective, the outputs appear well behaved, with the reconstructed volume sharing an overall agreement with the ground-truth volumes initially constructed.

Noticeably, however, low resolution information and map anisotropy due to domain motility is present strictly on one side of the molecule, instead of on both the *arm* and *elbow* domains (corresponding to CM_1 and CM_2, respectively) as one might anticipate given the earlier presentation (e.g., Figure 15). This trend indicates that CM_1 and CM_2, originally constructed on chain A and B independently, have been realigned to a new frame of reference (chain A). Thus, an important distinction must be made in how RELION aligns these motions. As the decision in

[9] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; manuscript draft, review and editing. F. Acosta-Reyes: Methodology; validation; manuscript review and editing. J. Frank: Manuscript review and editing.

RELION is based on maximizing the amount of signal in the image stack, immobilizing one of the largest chains (arm or elbow subunits) greatly outweighs our preferred choice of the smaller region in between. This apparent change due to alignment is only superficial, and as shown in Movie F1, a one-to-one structural mapping still exists between all states, regardless of the frame of reference used to view them.

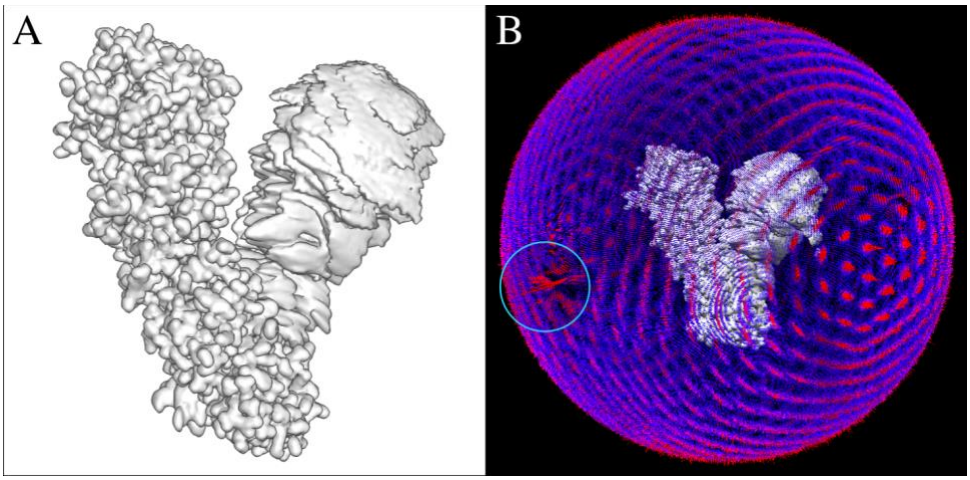


Figure 20: Hsp90 average volume and angular distribution from RELION 3D auto-refinement is shown in [A]. In this depiction, the lowest resolution features reside on the right-hand side domain, while the resolution across the rest of the structure is consistently uniform with higher-resolution information. As a result, atomic features on the mobile region are less defined. In [B], the RELION estimated angular distribution of particles on S^2 is shown, with 0.12° orientational sampling, 1320 orientations, 0.15 translational sampling, 36 translations, 0.2 pixel offset, and an estimated angular accuracy of 0.27°.

Assessment of Angular Assignments. We next sought to quantitatively assess the accuracy of each of the 812,000 angular assignments estimated by RELION. Generally, the distribution in Figure 20-B shares many similarities with the ground-truth model, which included 812 PDs with a Gaussian distribution for scattering the coordinates of each of the PD's 1000 images about the corresponding PD center (Figure 19). However, large-scale discrepancies did emerge, such as the region of heightened relative occupancy encircled in cyan (Figure 20-B), indicating many mismatched coordinates lumped together erroneously on S^2. To gain an

understanding of the finer discrepancies present, for each image, we calculated the geodesic distance (i.e., great-circle distance) between its ground-truth S^2 coordinates and those chosen by RELION 3D auto-refinement, as obtained in the corresponding output alignment file. The resulting distributions are presented in Figure F1-A. Additionally, we repeated this analysis for the outputs obtained via cryo-SPARC non-uniform alignment, which is presented in Figure F1-B.

As seen in Figure F1, compared to ground truth, the RELION angular assignments had an average error of approximately 7.9° with a standard deviation of 11.8°, while the cryo-SPARC outputs had an average error of 4.7° with a standard deviation of 9.8°. In the RELION case, 4.15% of the image stack's angular assignments (33,733 of 812,000 images) were significantly misplaced, while for cryo-SPARC, we observed only 2.70% (21,901 images). These outliers may be attributed to fundamental uncertainties introduced at certain viewing angles, with errors further amplified by our addition of noise and CTF on each image. As shown in Figure 21, a more intricate understanding of these trends can be gained by comparing the ground-truth and output angular distributions on S^2 for each PD individually. For PD assignments estimated by either RELION (Figure 21-A, B, C) or cryo-SPARC (Figure 21-D, E, F), the distribution of images in each PD was noticeably smeared (i.e., oblong) compared to the ground-truth coordinates (i.e., circular), and in many cases with clustered outliers occupying distant regions on S^2.

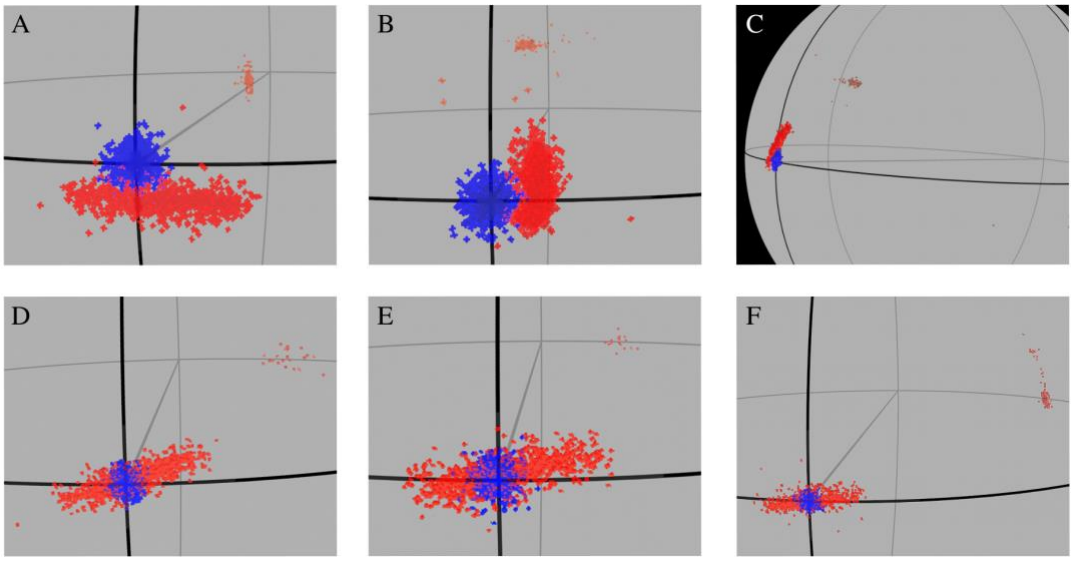


Figure 21: Hsp90 comparison of estimated angular assignments to ground truth. For all subplots [A-F], the blue points represent the ground-truth angular coordinates of images in a single PD (one PD per subplot), as presented globally in Figure 19. In contrast, the coordinates estimated by the maximum-likelihood approaches are displayed in red, and correspond to RELION outputs in the top row [A, B, C], and cryo-SPARC outputs in the bottom row [D, E, F]. In each subplot, a diametric line has been drawn to orient the reader, and extends out from each PD center to the point diametrically opposed on S^2.

Assessment of Reconstructed Volume. We next investigated the characteristics of the 3D reconstructed volume obtained via RELION 3D auto-refinement. Ideally, one might expect the correct 3D reconstructed volume to correspond to a weighted average of volumes as defined by the original occupancies—or equally energies—of each state (Figure 22-A). To compare the RELION outputs to this intuition, we calculated the cross correlation (using the Chimera "fitmap" command with translational shifts and rotations) between the reconstructed volume and each of the ground-truth volumes used to initially generate the synthetic continuum's 2D projections. As expected, the resulting correlation matrix (Figure 22-B) closely resembled the centralized, dumbbell-shaped occupancy distribution of the continuum (Figure 22-A). This result provides a clear interpretation for the 3D reconstructed volumes, idealizing them not as discrete states—since there is no on-to-one mapping to any one volume in the correlation matrix—but as an

amalgamation of states weighted based on the energetics of the corresponding biophysical system. This amalgam is increasingly delocalized as the total number of conformational states, degrees of freedom, and overall complexity of the free-energy distribution are increased.

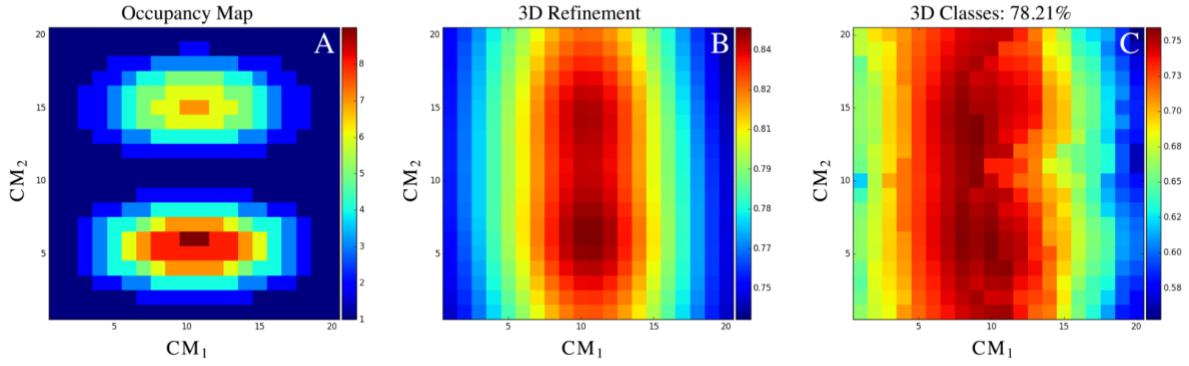


Figure 22: Hsp90 comparison of average volume to each of the ground truth volumes (e.g., as seen in Figure 15). In [A], the original occupancy map for our published synthetic continuum is shown. [B] shows the correlation matrix obtained from calculating the correlations between RELION's final reconstructed volume and the set of original state space volumes. Overall, there is a general agreement between these two graphs, with discrepancies attributed to the addition of noise and CTF in the images used for reconstruction, and errors in angular assignments. Subplot [C] shows the average of the class correlation matrices presented in Figure 23. As expected, the average correlation matrix for these 12 classes shares a high resemblance to the distribution seen in [B]. Color bars correspond to occupancy of states [A], correlation [B], and average correlation [C].

Assessment of Class Averages. We next analyzed the outputs of the RELION "3D Classification" procedure, which aims to distinguish projections originating from different 3D structures (Scheres, 2016). The output from RELION 3D auto-refinement was provided as a reference volume, with the regularization parameter set to T4. After 30 iterations, 40 3D class averages (Figure F2) were obtained, covering a wide range of resolutions. To note, 3D classification has limitations—due to memory and exacerbated by increasing box size—in the number of classes achievable, and in practice, independent rounds of classification are often explored with a unique number of classes requested in each. As was done in the previous step, we next created a correlation matrix between all 400 ground-truth volumes and each of these

reconstructed classes individually. As seen in Figure 23, the resulting matrices—one for each of the highest-occupancy reconstructed classes—corroborated our initial hypothesis, with the volume from each class partially corresponding to one of the two deepest free-energy wells in the ground-truth energy landscape (Figure 22-A), with the similarity validated further by eye.

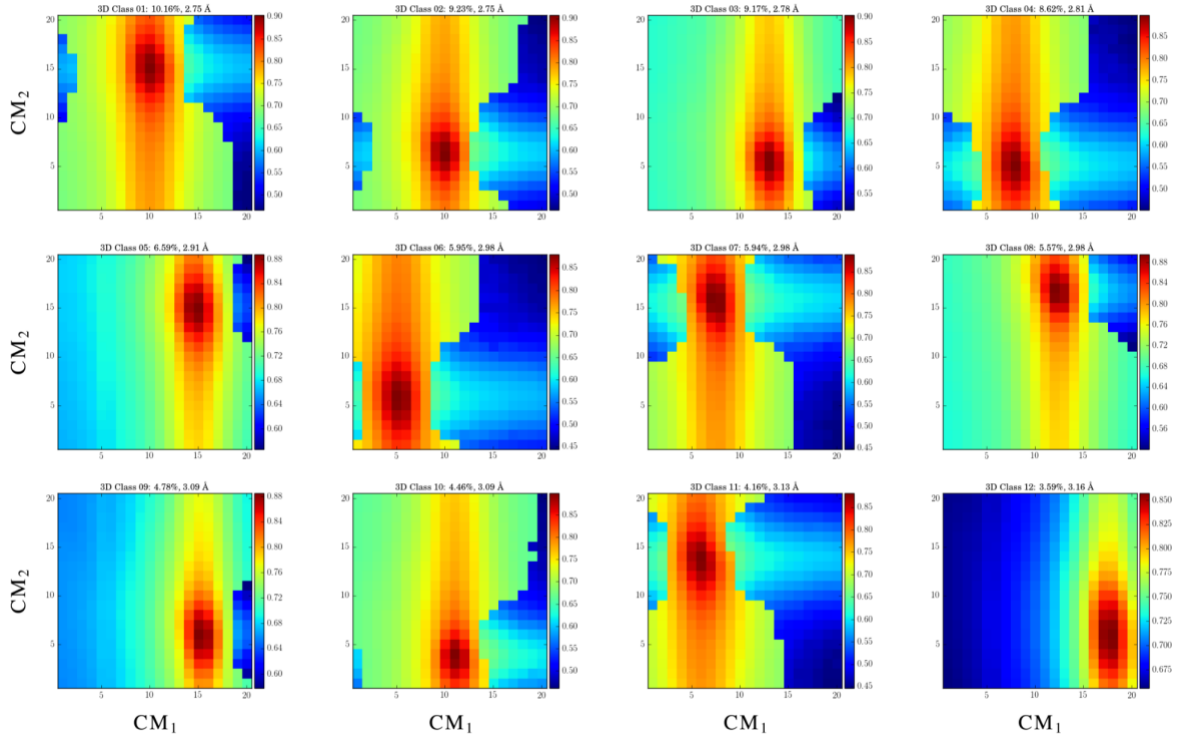


Figure 23: Hsp90 correlation between RELION 3D class averages and ground truth. Here, the correlation matrices are obtained from calculating the correlations between each of the top 12 RELION 3D class average volumes (Figure F2; as ranked by "Class Distribution") and the set of the 400 original state space volumes (e.g., as shown in Figure 15). Based on this ranking, it can be seen that the depth and width of the free-energy wells defines the ordering of the classes (and their resolution), with the first class centered in the upper free-energy well (a secondary minimum in the original landscape; Figure 22-A) while classes 2, 3 and 4 fill in the bottom free-energy well (the primary minimum in the original landscape). All other classes in the set can be found near the boundaries of these two wells.

8.3 Closing Remarks

Overall, cryo-SPARC non-uniform alignment consistently produced more accurate angular assignments than RELION 3D auto-refinement for the Hsp90 synthetic data; as was observed also in our analysis of the spike protein. However, recall that, for either method, the distribution of images in each PD was noticeably smeared (oblong) compared to the (circular) ground-truth assignments, and further obfuscated by clustered outliers occupying distant regions on S^2 (Figure 21). While these behaviors may prove to be innocuous for the purposes of broader classifications in cryo-EM, they carry significantly more weight on the downstream PD-analysis performed via ManifoldEM.

Given presence of these characteristic misestimations, PDs with circular outer boundaries—as formed in angular space by ManifoldEM—would only contain some subpopulation of relevant ground-truth conformational information. If the angular widths of the ManifoldEM PDs were to be increased—i.e., in an attempt to compensate for angular estimation errors—a larger number of misassignments would be incorporated into the boundaries of each PD from neighboring regions of S^2, and further exacerbated by the presence of outlier clusters from distant PDs. Indeed, since either situation critically undermines the efficacy of the ManifoldEM framework, these misassignments present an unavoidable conflict that can only be addressed at the source.

Following this assessment, our analysis of 3D class averages illustrates the outputs of maximum-likelihood classification schemes in the context of ground truth. Specifically, we observed that the fundamental state space relationships, while inaccessible without ground-truth awareness, are still conserved by the RELION "3D Classification" algorithm. In practice, however, one would have no knowledge of how to properly arrange these classes with respect to one another

in the underlying state space, or of the transitional structures in between. As a result, the Bayesian classification strategy is understood as a powerful method for reconstructing disjointed pieces of the continuum, with limitations in its ability to provide deeper knowledge of the relationship between states therein. To move beyond this hurdle, other techniques—such as manifold embedding—must be enlisted.

Chapter 9: Groundwork for Spectral Geometry

Notwithstanding the possibility of unavoidable (at least at present) upstream uncertainties, this narrative now shifts towards forming an understanding of problems arising strictly due to the ManifoldEM methodology. In Chapter 7, several systematic errors were observed when treating ManifoldEM as a "black box"; that is, by either inputting data sets with known (i.e., synthetic) or unknown (i.e., experimental) properties, and simply categorizing the results. While this understanding was helpful in addressing several longstanding issues arising in the ManifoldEM outputs, it is still unsatisfactory for the purposes of correcting them.

In order to embark on such a task, an extensive heuristic analysis was conducted of ManifoldEM procedures using the Hsp90 synthetic data, which entailed several modifications including experimentally-relevant noise and CTF (Seitz et al., 2021a). This heuristic analysis is presented in its entirety in Chapter 10. First, however, for these heuristic findings to make sense in terms of established theory, a much simpler data set must be introduced, which we term the *latent space*. It is formed by a collection of coordinates in a Euclidean space; forming, for example, a line of points for one degree of freedom or an array of points forming a rectangle for $n = 2$. At first glance, the latent space has seemingly little relation to our relatively complex Hsp90 synthetic-continuum data set. As will be shown, however, many striking similarities do emerge between the two; so many so that when differences emerge, the information is revelatory. In the following sections, an exposition of these findings will be detailed, which boils down to an analysis of the spectral geometry (i.e., eigenfunctions of the LBO) on a set of distinct manifolds[10], as is described in the Seitz et al. (2021a) supplementary material.

[10] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; validation; manuscript draft, review and editing. P. Schwander: Methodology; validation; manuscript review and editing. J. Frank: Manuscript review and editing.

First, in Section 9.1, we use the DM method to investigate the known eigenfunctions of the LBO on the interval and rectangular domains, and compare these results to the manifolds formed by a quasi-continuum of atomic-coordinate structures. Following this analysis, we ultimately detail how the structure of manifolds Ω obtained from a conformational state space transforms as the data type is translated stepwise from atomic-coordinate structures (Ω_{ACS}) in Section 9.2, to 3D density maps (Ω_{EDM}) in Section 9.3, and finally to 2D projections (Ω_{PD}) in Section 9.4. (In the last case, recall that 2D projections are the only form of data readily accessible in a cryo-EM experiment). Ultimately, we will use this analysis as a vehicle to build intuition for many of the properties observed in our heuristic analysis of Ω_{PD} embeddings, presented in Chapter 10.

9.1 Eigenfunctions of the Latent Space

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

In the 1D space, a set of pairwise distances between a collection of equispaced coordinates on a line carries all essential information necessary to model the pairwise distances between a sequence of atomic models with molecular domain rotated by a constant angular increment. To represent our SS_1 data set, we uniformly sample $N = 50$ equispaced points from a 1D interval $X \in [0, \ell = 1] \subset \mathbb{R}$, with each of these points representing a unique state of the molecule. Following the DM method, we then calculate the distance matrix for this collection of points and embed the data in a low-dimensional space. As a note, for all embeddings that follow, we will show that two characteristic regimes emerge depending on choice of Gaussian bandwidth, which we will denote with $\varepsilon_{\downarrow}$ and $\varepsilon_{\uparrow}$ for the small and large regime, respectively.

For the small Gaussian bandwidth, a cosine series emerged for all eigenfunctions (Figure 24-A), in very good agreement with the Laplacian on a 1D Euclidean interval with Neumann boundary conditions. Specifically, we anticipate—and retrieve—canonical (Grebenkov and Nguyen, 2013) eigenfunctions of the form $\psi_v(x) = \{\cos(v\pi x/\ell) \mid v \geq 1\}$. As the Gaussian bandwidth was incrementally increased from $\varepsilon_{\downarrow}$ to $\varepsilon_{\uparrow}$, this cosine series smoothly transformed into a different complete, orthogonal set: the Legendre polynomials (Figure 24-B). However, we note that these polynomials only occur for hyperrectangles, which are n-dimensional Cartesian products of orthogonal intervals.

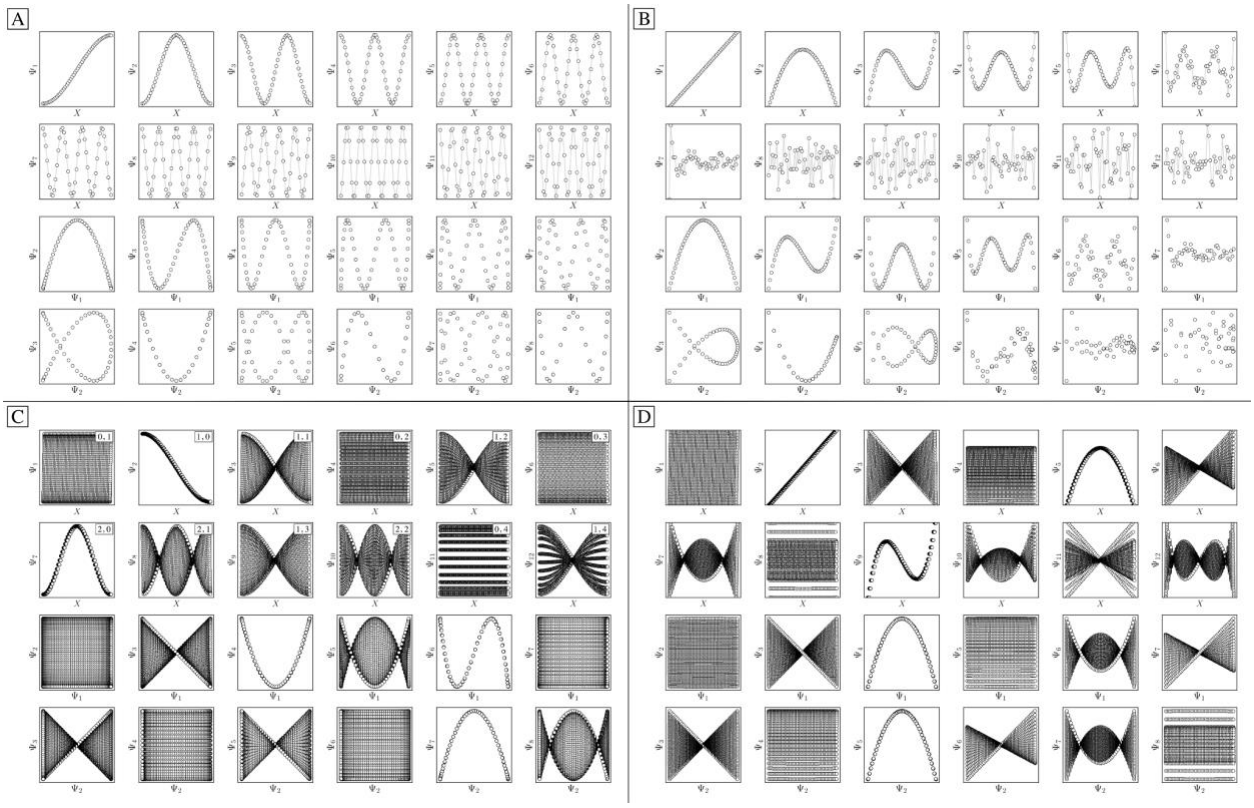


Figure 24: DM eigenfunctions in the 1D and 2D nondegenerate latent space. The DM eigenfunctions of the 1D interval for small ($\varepsilon_{\downarrow} = 5 \times 10^{-5}$) and large ($\varepsilon_{\uparrow} = 10$) Gaussian bandwidths are shown in [A] and [B], respectively. Likewise, eigenfunctions of the $N = 50 \times 50$ rectangular (nondegenerate) domain for small and large Gaussian bandwidth are shown in [C] and [D], respectively. As will be done throughout this text, eigenfunctions have been independently displayed by indexing each by its ground-truth ordering (here via sequential x-coordinates). For [C] and [D], a similar appearance of eigenfunction plots, albeit interchanged, would be seen when indexing instead via sequential y-coordinates. In [C], each eigenfunction's corresponding modes $\{v, w\}$ have also been provided in the top right-hand corner. For all four subplots, pairwise combinations of eigenfunctions are additionally shown, which can be visualized after an embedding without any ground-truth knowledge.

Next, to represent our SS_2 data set, we uniformly sample $N = 50 \times 50$ points from a 2D interval $X \times Y \in [0, \ell_x = 1] \times [0, \ell_y = 1.1] \subset \mathbb{R}^2$, where the operator $\times$ denotes the Cartesian product. For its illustrative properties, we avoid degeneracy by ensuring $(\ell_x/\ell_y)^2$ is not a simple ratio (Grebenkov and Nguyen, 2013). Again, we follow the DM method by calculating the pairwise distances between these points and embedding the data in a low-dimensional space. As demonstrated in Figure 24-C, the set of eigenfunctions obtained in the smaller Gaussian bandwidth

regime matched our *a priori* expectations for the Laplacian on a rectangular domain with Neumann boundary conditions. These canonical eigenfunctions are

$$\psi_{vw}(x,y) = \left\{\cos(v\pi x/\ell_x)\cos(w\pi y/\ell_y) \middle| \, v,w \geq 0\right\}, \tag{7}$$

following the same pattern for higher-dimensional domains $\Omega_{\mathrm{R}} = [0,\ell_1] \times \cdots \times [0,\ell_n] \subset \mathbb{R}^n$ (with $\ell_i > 0$; Grebenkov and Nguyen, 2013). Again, as we incrementally increased the Gaussian bandwidth from $\varepsilon_{\downarrow}$ to $\varepsilon_{\uparrow}$, this set of complete and orthogonal cosines smoothly transformed into the orthogonal Legendre polynomials set, which are now a function of both x and y, as expected (Figure 24-D). Importantly, the leading Legendre polynomials provide a direct linear map of the input data points, which is a consequence of the linear terms $P_1(x)$ and $P_1(y)$. While these linear relationships are easier to analyze than the cosine form, in Section 9.3 and 9.4, we will show that they are unavailable in the embeddings of 3D EDMs and 2D projections.

It is important to take notice that, alongside leading eigenfunctions in the first two rows of each subplot in Figure 24, the leading composites of these eigenfunctions $\{\Psi_i \times \Psi_j \mid i < j\}$ have also been plotted in the rows that remain, with each composite forming a unique 2D subspace. Mathematically, each such mapping to a 2D vector subspace is the restriction to the N-dimensional embedding of the projection of $\mathbb{R}^N$ onto $\mathbb{R}^2$; given by $\{\Psi_1 \times \Psi_2 \times \ldots \times \Psi_N\} \to \{\Psi_i \times \Psi_j\}$. (For expediency, we will use the term *subspace* to specifically refer to a subspace of an embedded manifold). Of interest, among the available subspaces, a leading parabolic trajectory exists for each degree of freedom present; for example, $\{\Psi_1 \times \Psi_4\}$ and $\{\Psi_2 \times \Psi_7\}$ in Figure 24-C, which correspond to the sequence of states along Y and X, respectively. While less significant in the current scope, the study and use of these 2D subspaces will be crucial in the chapters to come.

Returning to the smaller of the two Gaussian bandwidth regimes, we next compare the previous nondegenerate rectangular results to those from a degenerate square domain, with $N =$

50×50 points equispaced identically along X and Y (Figure 25-A). Due to the presence of degenerate eigenvalues, which can arise for domains with a rational ratio $(\ell_x/\ell_y)^2$, we encounter pairs of eigenfunctions that appear different from the non-degenerate case of the rectangle (Grebenkov and Nguyen, 2013): as seen, for example, by eigenvector pairs $\{\Psi_1, \Psi_2\}$ and $\{\Psi_4, \Psi_5\}$ in Figure 25-A. In Figure 25-C, we illustrate that these eigenfunctions are just rotated within their degenerate space, exactly as expected. We note that an eigenfunction associated with a degenerate eigenvalue is a linear combination of the degenerate eigenfunctions (Grebenkov and Nguyen, 2013), where the normalization of the eigenfunctions restricts this linear transformation to a rotation and reflection (i.e., the group of orthogonal transformations). For example, the $\{\Psi_1, \Psi_2\}$ eigenvector pair is of form $\Psi' = R^T\Psi$ such that

$$\begin{bmatrix}\Psi_1'(\theta)\\ \Psi_2'(\theta)\end{bmatrix} = \begin{bmatrix}\cos(\theta)\cos(\pi x) + \sin(\theta)\cos(\pi y)\\ -\sin(\theta)\cos(\pi y) + \cos(\theta)\cos(\pi x)\end{bmatrix} \quad (8)$$

As seen for eigenvector Ψ_6 in Figure 25-A, these summands can also have the form of two products: $\Psi_6 = A\cos(\pi x)\cos(2\pi y) + B\cos(2\pi x)\cos(\pi y)$, with any A and B such that $A^2 + B^2 \neq 0$. Hence, it can be seen that these aberrant eigenfunction pairs are defined by an admixture of cosines in a higher-dimensional space, with form

$$\Psi_i = A\cos(v\pi x)\cos(w\pi y) + B\cos(w\pi x)\cos(v\pi y) = A\psi_{vw} + B\psi_{wv}$$

By using an appropriate rotation operator $R_{i,j}$, the summands within each eigenfunction pair can be maximally separated among both members $\Psi_i = \psi_{vw}$ and $\Psi_j = \psi_{wv}$, such that the canonical eigenbasis is recovered (Figure 25-B). As demonstrated using analytical expressions $\psi_{1,0} = \cos(\pi x)$ and $\psi_{0,1} = \cos(\pi y)$ in Figure 25-C, this separation occurs multiples of $\theta = 90°$ apart. In the Figure 25-C example, at $R_{1,2}(45°)$, these eigenfunctions have form

$$\begin{bmatrix}\Psi_1'(\theta = 45°)\\ \Psi_2'(\theta = 45°)\end{bmatrix} = \begin{bmatrix}\sqrt{2}/2\cos(\pi x) + \sqrt{2}/2\cos(\pi y)\\ -\sqrt{2}/2\cos(\pi y) + \sqrt{2}/2\cos(\pi x)\end{bmatrix} \quad (9)$$

which decouples back into two distinct modes (i.e., $\cos(\pi y)$ and $\cos(\pi x)$ for Ψ_1 and Ψ_2, respectively) at $R_{1,2}(90°)$. A similar result is obtained by applying this operation on the appropriate eigenfunctions obtained via DM, with each initially assuming a random rotation angle (Figure 25-A) requiring a specific correction $R_{i,j}(\theta)$, as seen in Figure 25-B.

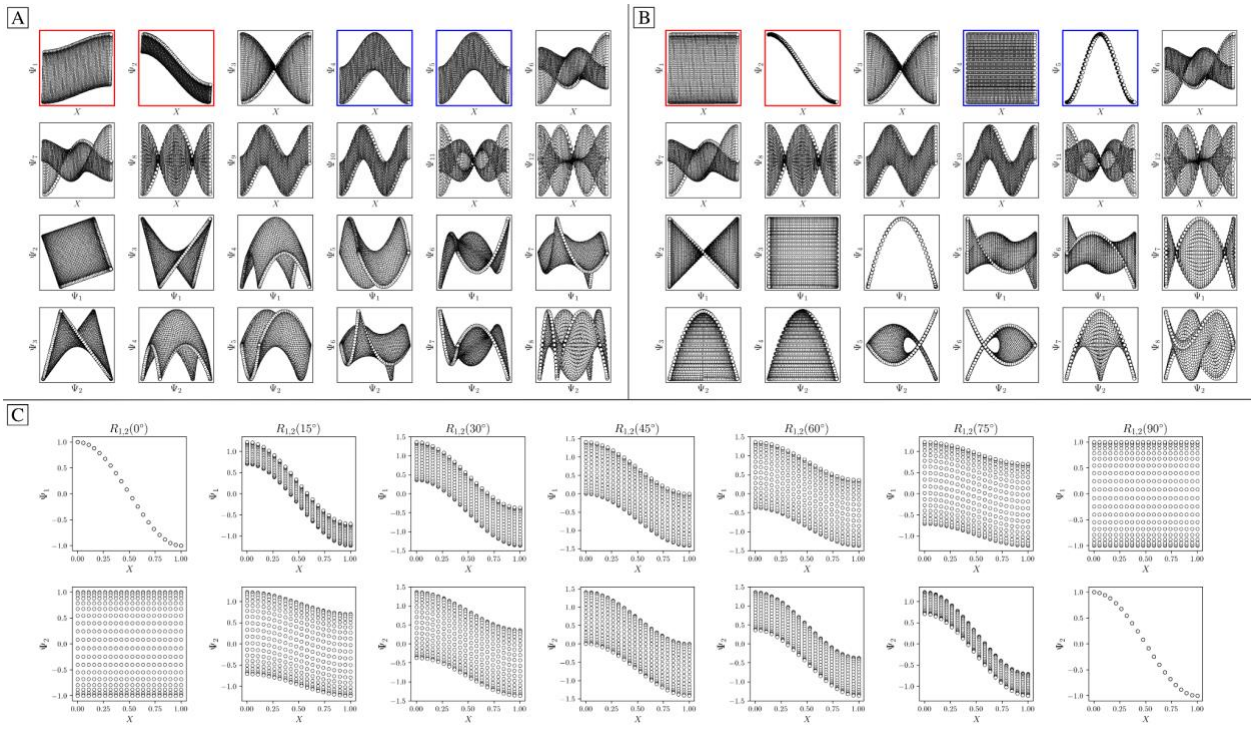


Figure 25: DM eigenfunctions in the 2D degenerate latent space. DM eigenfunctions of the $N = 50 \times 50$ square domain for small Gaussian bandwidth ($\varepsilon_{\downarrow} = 5 \times 10^{-5}$) are shown in [A] and [B] before and after high-dimensional rotations, respectively. It can be seen here that pairs of eigenfunctions exist that contain relationships aberrant to the canonical eigenfunction form seen in Figure 24-C. Two such pairs have been highlighted in red and blue, respectively, with the members of each pair always rotated 90° apart. To note, as any rotation can happen in the presence of degeneracy, this initial rotation is an arbitrary one. We demonstrate this property via the schematic in [C], which shows the angular relationship between two analytically-generated functions ($\cos(\pi x)$ and $\cos(\pi y)$, each displayed in the reference frame of states in X) as they are jointly rotated 90°. By applying rotation operators $R_{1,2}(\theta) = -19°$ and $R_{4,5}(\theta) = 45°$ independently to two such aberrant pairs in [A], the canonical eigenfunction form begins to recover in [B], and more so as additional operators are intelligently applied.

While degeneracy is a rather rare event that can be identified from the eigenvalue spectrum, a similarly-rotated appearance (i.e., *eigenfunction misalignment*) will later turn up during our investigation of PD manifolds. Pairs of misaligned eigenfunctions, at least approximately, can also be mimicked when domains have undergone certain elementary geometric transformations. For example, by performing an affine transformation on a rectangle Ω_R to form a parallelogram Ω_P, we observed a rotation of the first two eigenvectors, as similarly seen in Figure 25-A. Recall that an affine mapping preserves collinearity and ratios of distances, but in general not distances and angles. In Section 9.3, we will explore the possibility of other classes of transformations.

As a final point in this section, we illustrate our method for retrieving the canonical eigenfunctions buried within an embedding, which has been used in Figure 24 and Figure 25, and extensively throughout the remainder of this work. Figure 26 provides a schematic using the known analytical eigenfunctions (Figure 26-B and Figure 26-C) chosen so as to match the results from DM on the square (degenerate) Ω_R. In Figure 26-A, we display the eigenfunctions from Figure 25-A in two different reference frames corresponding to our ground-truth knowledge. Specifically, we plot the points in each eigenvector in a sequence corresponding to their initial ground-truth arrangement along each degree of freedom (for the rectangular Euclidean space, along either X or Y), which is shown in the first and second row of Figure 26-A, respectively. As shown in Figure 26-B and Figure 26-C, a given reference frame captures the eigenfunction on a projected plane in the n-dimensional space where it resides.

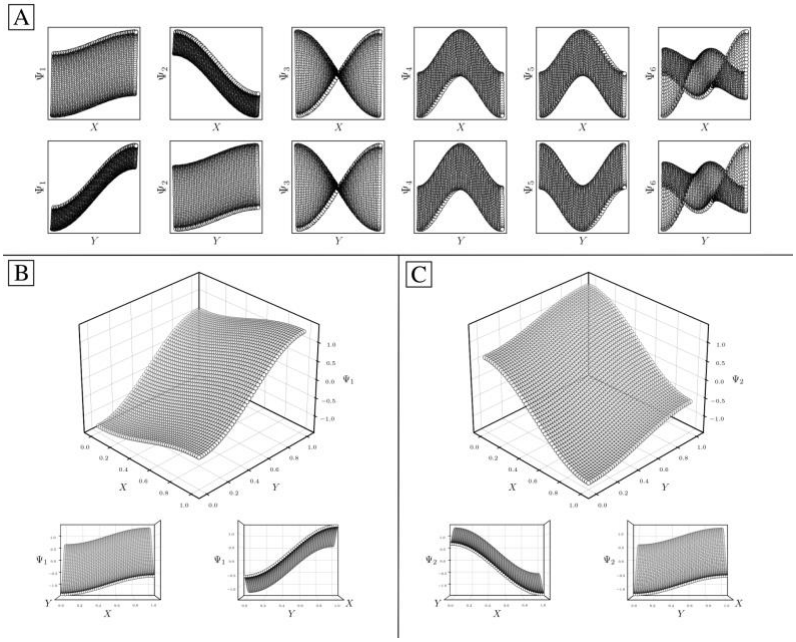


Figure 26: Intuition for sequential ordering of eigenfunctions based on ground truth. DM eigenfunctions from Figure 25-A are shown again in the first row of [A], which were displayed by ordering the points of the embedding in a sequence based on the ground-truth x-coordinates. The second row of [A] displays these eigenfunctions instead via the sequential ordering of ground-truth y-coordinates. Subplots [B] and [C] are analytically generated so as to match the appearance of Ψ_1 and Ψ_2 in the first and second row of [A], respectively. For this presentation, the equations $\Psi_1 = \cos(\theta)\cos(\pi x) + \sin(\theta)\cos(\pi y)$ and $\Psi_2 = -\sin(\theta)\cos(\pi x) + \cos(\theta)\cos(\pi y)$ were used, with $\theta = 250°$. As can be seen in [B] and [C], each eigenfunction exists in an n-dimensional space defined by the n degrees of freedom of the system. By displaying points in sequence corresponding to a known degree of freedom, we are effectively viewing each eigenfunction on a projected plane in its n-dimensional space.

9.2 Eigenfunctions of the Atomic Models

We next investigate the manifolds obtained from the state spaces formed from a quasi-continuum of atomic-coordinate structures, each represented by a set of 3D atomic-coordinates $3m$ (e.g., as visualized in Figure 14). We generate these structures as described in the first step of our synthetic-generation protocol in Chapter 6, which are subsequently used to produce a corresponding set of 3D electron density maps and 2D projections. Importantly, the set of these 3D atomic-coordinate structures in $\Omega_{\mathrm{ACS}} \subset \mathbb{R}^{3m}$ represent the fundamental biophysical identity of

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

$$\mathrm{RMSD}(X, Y) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \|(x_i - y_i)\|^2} \tag{10}$$

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

The resulting DM embeddings for the small and large Gaussian bandwidth regimes are shown in Figure 27-A and Figure 27-B, respectively, and share a strong resemblance to those found for the latent rectangular domain (Figure 24). Again, we note the presence of cosine eigenfunctions for the small Gaussian bandwidth regime, and a near-perfect linear form (via leading Legendre polynomials) in the large Gaussian bandwidth regime ($\{\Psi_1, \Psi_2\}$ in Figure 27-B). For the latter, we will show this feature is a luxury not obtained in the other data types to be explored. In the small Gaussian bandwidth regime, we can identify both CM_1 and CM_2 parabolas residing in the subspaces $\{\Psi_1 \times \Psi_3\}$ and $\{\Psi_2 \times \Psi_8\}$, respectively. Similar results—albeit for different dimensions—were found for the Ω_{ACS} embeddings from SS_1 and SS_3.

The striking similarity between the eigenfunctions of the latent space and the eigenfunctions of the atomic models can be rationalized as follows. If the range of a single body rotation is moderate ($\lesssim 30°$), the distance D_{ij} between any two states i and j within this range is to high accuracy $D_{ij} = \Theta_{ij}(\sum_{k=1}^{m} r_k^2)^{1/2}$, where r_k is the distance of atom k away from the rotation axis, m the number of atoms of the body, and Θ_{ij} the angular difference between the states.

Therefore, D_{ij} is directly proportional to Θ_{ij}. (We note that this approximation is justified since the rotation matrix has only linear terms in rotation angle, provided the rotation angle is small). If there are multiple independent body rotations (i.e., CMs) present, the individual distances add in quadrature as in a Euclidean space. While not investigated in this paper, the linearity also holds for body translations as well, where the distance is directly proportional to the magnitude of the translation. Thus, the agreement between the eigenfunctions of the latent space and the ones of the atomic models is a consequence of the linear relationship between distance and the multi-body motions, rotations, and translations.

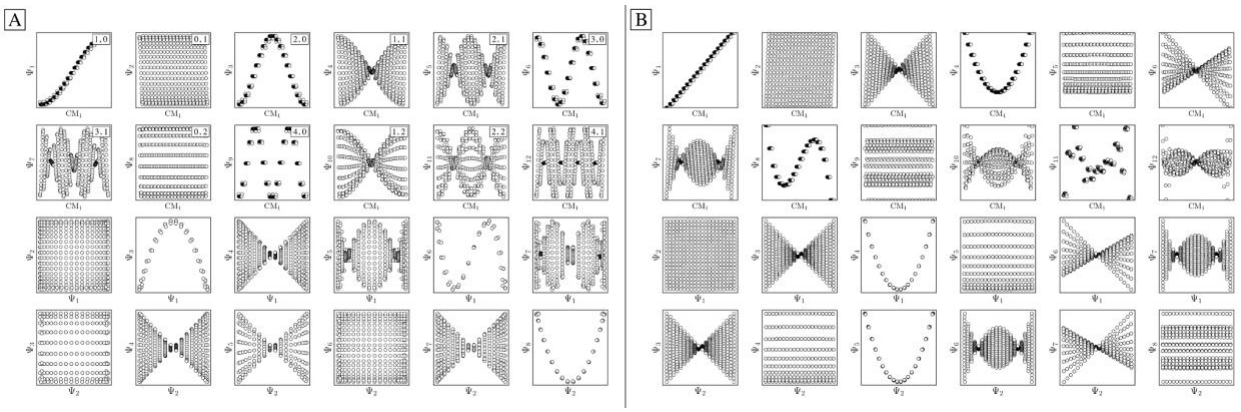


Figure 27: DM eigenfunctions for the quasi-continuum of atomic-coordinate structures. Eigenfunctions obtained for $20 \times 20 = 400$ atomic models occupying SS_2 for small ($\varepsilon_{\downarrow} = 0.1$) and large ($\varepsilon_{\uparrow} = 1000$) Gaussian bandwidths are shown in [A] and [B], respectively. Leading eigenfunctions are displayed in the first two rows via sequential indexing along the ground-truth CM_1 coordinates (i.e., equispaced rotations of chain A). In [A], each eigenfunction's corresponding modes $\{v, w\}$ are provided in the top right-hand corner, showing exceptional agreement with the LBO eigenfunctions on a rectangular domain with Neumann boundary conditions. We additionally note the absence of any significant eigenfunction misalignments.

9.3 Eigenfunctions of the 3D Density Maps

We next demonstrate that the properties of manifolds (as seen in the previous section for the 3D atomic-coordinate structures) significantly transform when the data representation of their underlying states is altered. Specifically, we investigate how the conformational relationships

between states are changed when representation by atomic coordinates is transformed into one by 3D electron density maps (EDMs; as visualized in Figure 15). To this end, we generated the EDMs for each of the 3D atomic-coordinate structures for all previously defined state spaces, as is described in our synthetic-generation protocol (Chapter 6). We next calculated the pairwise Euclidean distances between these EDMs in $\Omega_{\mathrm{EDM}} \subset \mathbb{R}^V$, with V the number of voxels in an EDM, and performed an embedding via the DM method. Overall, over a wide range of Gaussian bandwidths, the structure of the resulting eigenfunctions was very similar to the structure of eigenfunctions retrieved for the atomic models in the small Gaussian bandwidth regime. Importantly, as in Section 9.2, these eigenfunctions were still of the form ψ_{vw}, with subspaces having no significant appearance of eigenfunction misalignments.

However, there are a few attributes to consider that distinguish the manifolds obtained for EDMs from those retrieved for the previous data types. First, the difference between small and large Gaussian bandwidth regimes was much less drastic, such that the cosine eigenfunctions appeared in both regimes. For small Gaussian bandwidth regimes (i.e., a few orders of magnitude below the optimal value $\varepsilon_\star$ determined by the bandwidth estimation method), we found that the leading CM_2 eigenfunctions were buried deeply in low-ranking eigenvectors (e.g., Ψ_8 and higher), with numerous CM_1 eigenfunctions occupying the eigenvectors in between. In addition, eigenvectors with cross terms $\Psi_i = \{\psi_{vw} | \, v, w \neq 0\}$ were found scattered mostly in mid-range positions (e.g., Ψ_{12} and higher). Since these properties were not observed in the embeddings of the atomic-coordinate structures, we conclude they are a result of a change in the metric.

In contrast, for larger Gaussian bandwidth regimes (i.e., near and significantly above $\varepsilon_\star$), eigenvectors with cross terms were buried in much deeper subspaces (e.g., Ψ_{34} and higher), with the majority of leading eigenvectors housing content exclusively for either CM_1 ($w = 0$) or

CM_2 ($v = 0$). These CM eigenfunctions also had a near-perfect distribution of points, whereas for the $\varepsilon_{\downarrow}$ regime, the distribution of points had noticeably less precision to the ideal form. Notably, the embeddings obtained above and below these regimes were incoherent in form.

We conclude that the eigenfunctions obtained from the larger Gaussian bandwidth regime would be preferred for several reasons. First, the desired CM_1 and CM_2 parabolas occupy leading subspaces and are thus easily identifiable. The paucity of leading cross-term eigenfunctions is also convenient, since they provide no useful information for our analysis, while also obfuscating our search for desired subspaces. Additionally, the geometric structure of all subspaces obtained via $\varepsilon_{\uparrow}$ consistently appears much closer to the ideal form. In Figure 28, we display the DM eigenfunctions obtained from this regime for the $20 \times 20 = 400$ EDMs occupying SS_2. Subspaces indexed in the CM_1 reference frame (rows one and two in Figure 28) and the CM_2 reference frame (rows three and four) are displayed, as well as a set of leading composites of these eigenfunctions—forming 2D subspaces—in the rows that remain.

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

those same voxel locations in the data structure of another state, with only changes in the value at each voxel entering the distance measure.

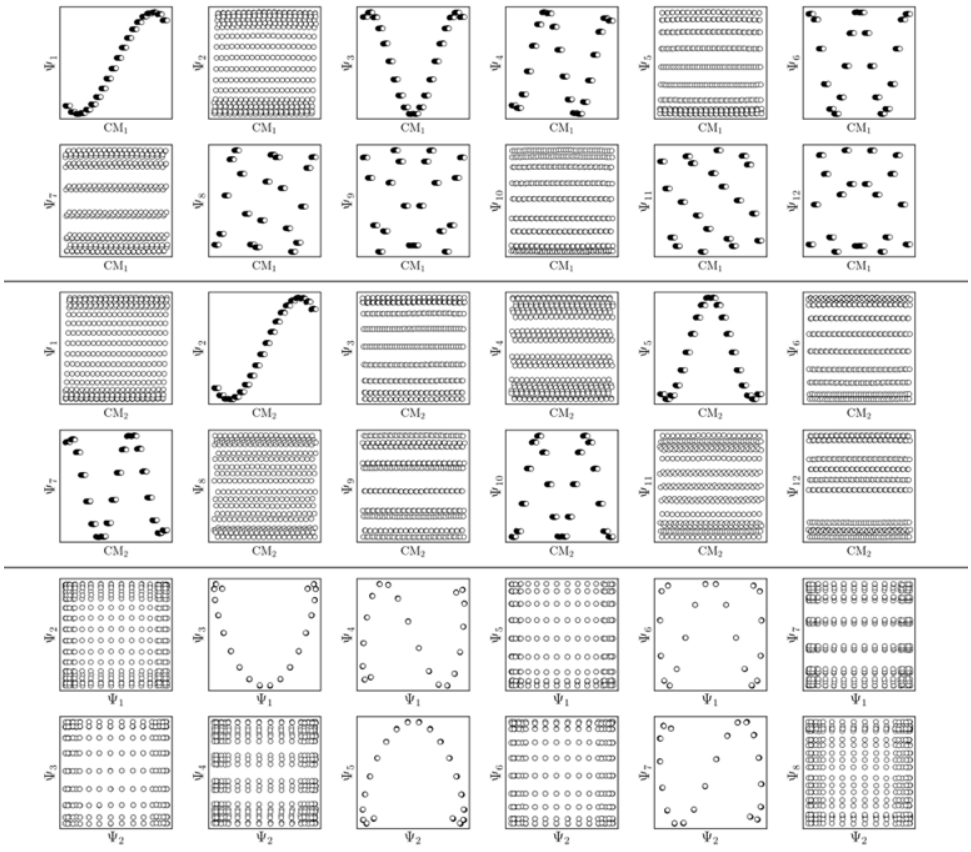


Figure 28: DM eigenfunctions for the quasi-continuum of electron density maps. The results of the DM embedding of EDMs (pure signal) from SS_2 are shown. Leading eigenfunctions as indexed by CM_1 and CM_2 are displayed in the first four rows, followed by their composites. Overall, there is near-perfect alignment of these eigenfunctions with the initially-obtained eigenvector basis, such that no rotations are required. As a final note, the pronounced inward curling at the boundaries of certain subspaces (e.g., $\{\Psi_1, \Psi_3\}$) is due to insufficient sampling.

Hence, while the eigenfunctions are similar, the relationship between states in these two data types is fundamentally different. To demonstrate this change, Figure 29 shows a comparison of the pairwise distances between states as calculated for the rectangular latent space, atomic-coordinate structures, and EDMs. As noted in the caption, by assessment of the close similarity between the distances from the latent space and atomic models, we can infer that these two data

types are both confined to the rectangular manifold Ω_R (albeit of different sizes). As a consequence, we observed that their eigenfunctions are nearly identical.

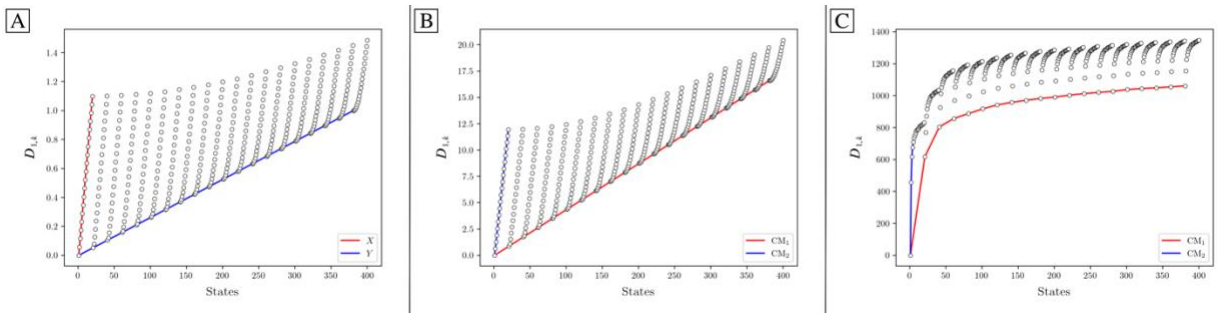


Figure 29: Comparison of induced metric for the three data types. The first row of the distance matrix **D** is plotted for the rectangular Euclidean space [A], the 3D atomic-coordinate structures in SS_2 [B], and the EDMs in SS_2 [C]. Given our ordering of states, the first row $D_{1,k}$ corresponds to the pairwise distance calculated between state 01_01 and all 400 states. For [A], which was calculated for a rectangular domain $\Omega_R \in [0,1] \times [0,1.1] \subset \mathbb{R}^2$, one can identify the distance of the first state $(x_1 = 0, y_1 = 0)$ to all other coordinates, such that the red line depicts the left-hand side of the rectangle (with maximum distance $D\{(x_1, y_1), (x_{20} = 1.1, y_{20} = 0)\} = 1.1$), and the blue line depicts the rectangle's base (with maximum distance 1). In [B], a similar rectangular pattern arises for the RMSD values calculated between atomic models. The pattern in [C], however, is starkly different from [A] and [B], such that no rectangular (or rectangle-like) domain could be drawn to reproduce this trend.

In contrast, we see that the distances from the EDMs are starkly different from the rectangular pattern, where neighboring states are spatially arranged via an asymptotic-like trend. From these findings, we must infer the corresponding data live in an altogether different manifold. Although the explicit geometric form of Ω_{EDM} is unknown, we have shown that the spectral properties of the Laplacian in Ω_{EDM} are essentially preserved via the mapping from the latent space. While detailed knowledge of Ω_{EDM} is certainly of interest, it is inconsequential here since our analysis only requires an understanding of the eigenfunctions of a manifold, and not necessarily its exact shape.

9.4 Eigenfunctions of the 2D Projections

Since a detailed description of the eigenfunctions of embedded 2D projections is provided in the following chapter, we continue this current narrative only as it pertains to the relationship of the eigenfunctions of the LBO on $\Omega_{\mathrm{PD}} \in \mathbb{R}^P$ with those from previously-established models (i.e., rectangular Euclidean latent space, atomic-coordinate structures, and EDMs). For similarities, as was observed for the EDMs, we found that eigenfunction characteristics could be broadly classified into two classes via either a small or large Gaussian bandwidth regime. In either regime, the eigenfunctions of the PD manifolds were again of the form ψ_{vw}, such that only cosines emerged. The lack of the Legendre-like form and a similar asymptote-like appearance of distances between images suggests that the PD states in Ω_{PD} reside on a manifold similar to Ω_{EDM}.

The overall difference between eigenfunctions obtained via $\varepsilon_{\downarrow}$ and $\varepsilon_{\uparrow}$ was also much more impactful for PDs than for the EDMs. In the small Gaussian bandwidth regime, CM_2 subspaces exhibited a severely suboptimal point distribution, such that in some PDs, identification of the CM_2 parabola-housing 2D subspace was completely obstructed. These CM_2 subspaces were also buried in trailing eigenvectors, and interspersed among those with cross terms. We also note that the value determined by the bandwidth estimation method ($\varepsilon_{\star}$) fell within this regime, making it a suboptimal choice for pristine data. In contrast, the large Gaussian bandwidth regime (i.e., one order of magnitude larger than $\varepsilon_{\star}$ and spanning numerous orders of magnitude above it) was superior in every sense, with CM_1 and CM_2 eigenfunctions having ideal point distributions and corresponding subspaces occupying leading eigenvectors. As well, the cross-term eigenfunctions were present only in far trailing eigenvectors (e.g., Ψ_{31} and higher), and would thus not be obstructive during an analysis. Briefly, we note that upon introduction of noise (SNR of 0.1) and

five noisy-duplicates of each state, $\varepsilon_\star$ was instead the most suited choice (along with numerous orders of magnitude above it), with anything below this range completely inadequate.

For either Gaussian bandwidth regime, we found that significant eigenfunction misalignments can emerge—and with varying magnitude—depending on the projection direction. Since we have previously shown that no such property is apparent in the manifold embeddings generated from the 3D EDMs from which these PDs originate, it is clear that the emergence of these eigenfunction misalignments is tied to PD disparity. We hypothesize that, as different 2D projections are taken of the EDMs via p: $\Omega_{\mathrm{EDM}} \to \Omega_{\mathrm{PD}}$, the geometry of Ω_{EDM} can become contorted due to the change of pairwise interatomic distances resulting from foreshortening in projection (Figure 30), such that the apparent span of one CM to another depends on PD.

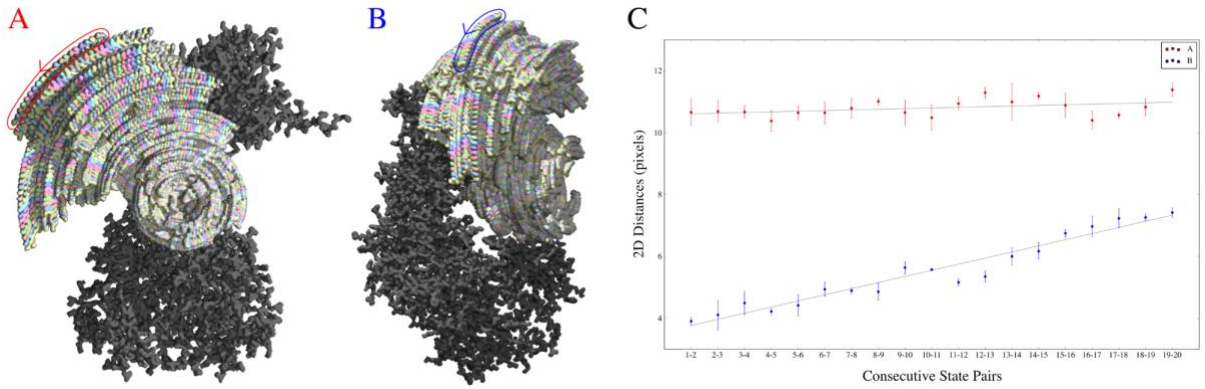


Figure 30: Intuition for emergence of PD disparity due to foreshortened distances when taking 2D projections of 3D EDMs. Two orthographic views of 3D models in the directions of two PDs are shown in [A] and [B], each composed of 20 overlaid 3D volumes from CM_2. The 2D distances (in pixels) were measured between the peripheral ends of each consecutive states' rotated subunit (as seen in red and blue encircled regions). In [C], the mean 2D distance measurements on each consecutive region in image [A] and [B] are plotted with error bars representing standard deviation, along with linear regression. Although the distances between states in the object's 3D form is constant, when projections are taken, these distances can strongly vary based on the current 2D view. While the Euclidean distance matrix calculated in the DM method is less intuitive, and instead records these changes on a pixel-by-pixel basis for the entire image, we anticipate analogous relationships to emerge there based on PD and CM.

9.5 Closing Remarks

Throughout this chapter, it has been shown how the embeddings of manifolds containing the same conformational information change depending on how the data is represented. In the next chapter, we will use this knowledge to gain a more complete understanding of PD manifolds generated with numerous degrees of freedom and experimentally-relevant conditions. This heuristic analysis is introduced as a clean slate free from assumptions, aiming to further investigate—using ideal data—the feasibility of several ManifoldEM-related techniques under realistic experimental conditions, while exposing any intrinsic uncertainties that may arise.

Chapter 10: Heuristic Analysis of PD Manifolds

10.1 Motivation

Following the previous analysis, a collection of Ω_{PD} embeddings are next analyzed in greater detail, with each one obtained by modulating key properties of the synthetic Hsp90 data set. Specifically, several parameter modifications are introduced—including varying type and number of CMs, PDs, SNR, CTF, and occupancy distribution—to determine how these changes are reflected in the low-dimensional representations of the corresponding manifold's spectral geometry. Notably, the explicit expressions derived in Chapter 9 will be used to account for the geometric structures observed in each embedding, which generally describe perturbations of a hypersurface spanned by multiple degrees of freedom.

This work was conducted from 2019 to 2021 in collaboration with F. Acosta-Reyes, S. Maji, P. Schwander and J. Frank. Here, an overview of these findings is provided highlighting our most important discoveries, and is based on the comprehensive description available via Seitz et al. (2021a; currently under review for publication)[11]. Our presentation showcases results from detailed evaluation of four data types—termed data-type I, II, III, and IV—with each step incorporating image artifacts and ensemble statistics in our state-space models as is anticipated in a cryo-EM experiment.

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

[11] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; software; validation; manuscript draft, review and editing. F. Acosta-Reyes: Methodology; validation; manuscript review and editing. S. Maji: Methodology; manuscript review and editing. P. Schwander: Methodology; validation; manuscript review and editing. J. Frank: Conceptualization; direction and project administration; manuscript review and editing.

- **Data-type IV:** nonuniform occupancy map (Fig. 18) with relevant SNR and CTF

To note, a new methodology—ESPER—will be proposed in the next chapter, with the decisions employed there guided by insights from this heuristic analysis of data-type I, II and III. Later, Data-type IV will be used strictly as a benchmark for the performance of that approach.

As an aside, a reference back to Section 6.2 may be required for its detailed description of state space indexing, since this ordering is repeatedly visualized by color maps throughout this current chapter to locate each state's coordinates. Of course, any embedding is independent of the ordering of the data points (Coifman et al., 2008); we have merely chosen one such ordering that heightens our awareness of trends in the subsequent outputs.

10.2 Embeddings for Data-type I

We first investigate the pristine data-type I, which is given no simulated experimental artifacts or occupancy assignments. Within this construction, within each state space (SS_1, SS_2 and SS_3) sets of images in five chosen PDs are first obtained. The first PD was chosen to be normal to the plane of the CM_1 rotation, such that all CM_1 motions from that perspective only underwent changes in the plane of the projection. A similar choice was made for PD_2, which was projected into the plane of CM_2 motions, with the remaining three PDs chosen at arbitrary positions in angular space. For each state space, the set of these images as generated from one of these five PDs forms a high-dimensional manifold Ω_{PD}. We next embed these manifolds via the DM method and, using the eigenfunctions of the LBO, analytically quantify the trajectories of our simulated conformational changes as embodied by the spectral geometry of each Ω_{PD}. In the following section, an overview of these results will be supplied, introduced in sequence of increasing intrinsic dimensionality.

State Space 1. Using the DM method, we first generated a different embedding—using a suitable ε value within the range discovered—for each of the five PD manifolds in SS_1. Each of the resultant point clouds contain 20 points, with each point corresponding to an image of a conformational state from CM_1. We next ordered the eigenvector coordinates in each Ω_{PD} embeddings to correspond to the ground-truth sequence of CM_1 states; i.e., as understood via Figure 26. In doing so— as anticipated given our analysis in Chapter 9—we observed the canonical eigenfunctions of the LBO on the interval $[0, 1]$ subject to Neumann boundary conditions

$$\psi_k = \{\cos(k\pi x) \mid 0 \leq x \leq 1; k \in \mathbb{Z}^+ \leq N\}, \tag{11}$$

where x is the conformational coordinate represented by a number in the interval $[0, 1]$. For demonstration, we plot each of the 1D points in a given eigenvector as a function of a uniform index $I \in [1, 20]$ (for the 20 total states in SS_1), making sure that the ordering of points in 1D follows the sequence assigned by the ground-truth index of its corresponding image along CM_1. As seen in Figure 31-A, when the collection of points in each eigenvector are ordered appropriately, the eigenfunction's sinusoidal form emerges along the full extent of the degree of freedom present (i.e., mapping $I \mapsto x \in [0, 1]$).

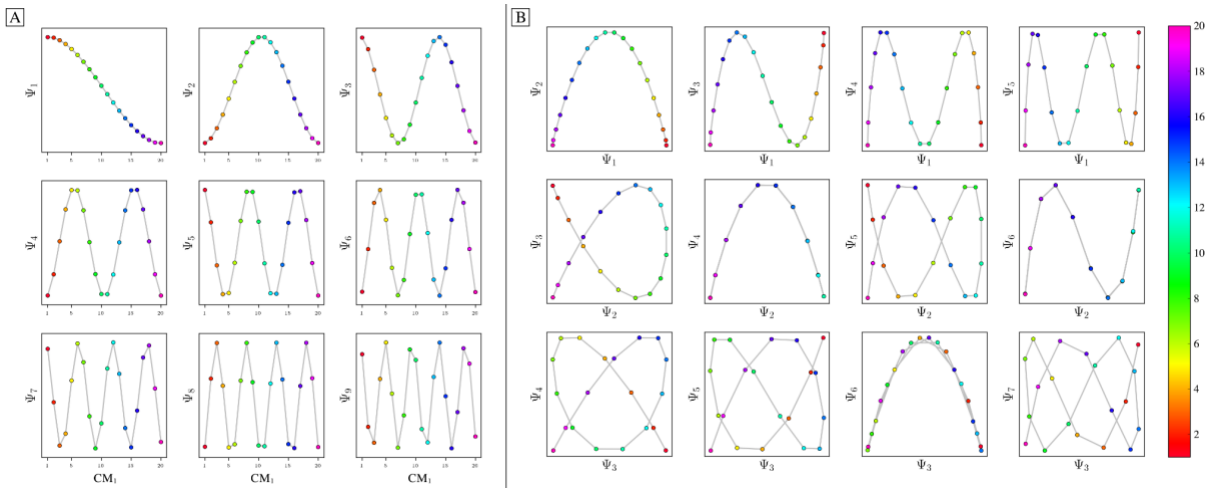


Figure 31: Analysis of eigenfunctions for PD_1 in SS_1 from data-type I. On the left [A] are the sinusoidal forms ψ_k that emerge when points—corresponding to images—in each eigenvector

are ordered precisely in the sequence in which their ground-truth images were constructed. Regardless of any knowledge of such a sequence, the composites of these eigenvectors will always form well-defined geometries (via the Lissajous curves), as shown in [B]. In the first row are the Chebyshev polynomials of the first kind, of which the parabola $\{\Psi_1 \times \Psi_2\}$ is the simplest mapping of the conformational information present.

Of course, as the points in an experimental data set naturally arrive in unordered sequence, one would have to properly sort the image indices to recognize these sinusoids; here, for example, there would be $20!$ sequences to consider. In the application, even if an approximation of this sequence were obtained, then in the presence of duplicate CM states (which we anticipate in an experiment), each sinusoid would be irregularly stretched along the x-axis where those duplicate states occurred, forming an unwieldy distorted sinusoidal form. However, as the points in each eigenvector are always scrambled in the same order for all eigenvectors, the composite of any two will always exist in a readily identifiable form. Specifically, as seen in Figure 31-B, a subset of the canonical Lissajous curves (Cundy and Rollet, 1989) emerges across the 2D subspaces of each Ω_{PD}, with the curves in this set having the form

$$L_{p,q} = \{\cos(p\pi x) \times \cos(q\pi x) \mid 0 \leq x \leq 1; p \neq q \in \mathbb{Z}^+\} \tag{12}$$

For these composites, we found that CM information is portrayed most simply (without overlap) along a specific subset of L, here as seen across the set of 2D subspaces defined in pairwise combination with the leading eigenvector; i.e., $\{(\Psi_1 \times \Psi_2), (\Psi_1 \times \Psi_3), \dots, (\Psi_1 \times \Psi_z)\}$, where z is the index of the smallest non-zero eigenvalue. Specifically, this subset $T_k \in L$ corresponds to the known Chebyshev polynomials of the first kind (Abramowitz and Stegun, 1972), of which we observed that the parabolic form is the lowest-order member present in each Ω_{PD} embedding.

Given their significance, these 2D subspaces have several important properties worth highlighting for their eventual use (or avoidance). First, note that for each sinusoidal subplot in Figure 31-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 a set (indexed via A) of Cartesian products (Π) operating on cosine functions X_a (of form ψ_k) that are each uniformly occupied with a finite number of data points. As shown in Figure 32-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. For reference later, this aspect is denoted with the term *nonuniform rates of change*.

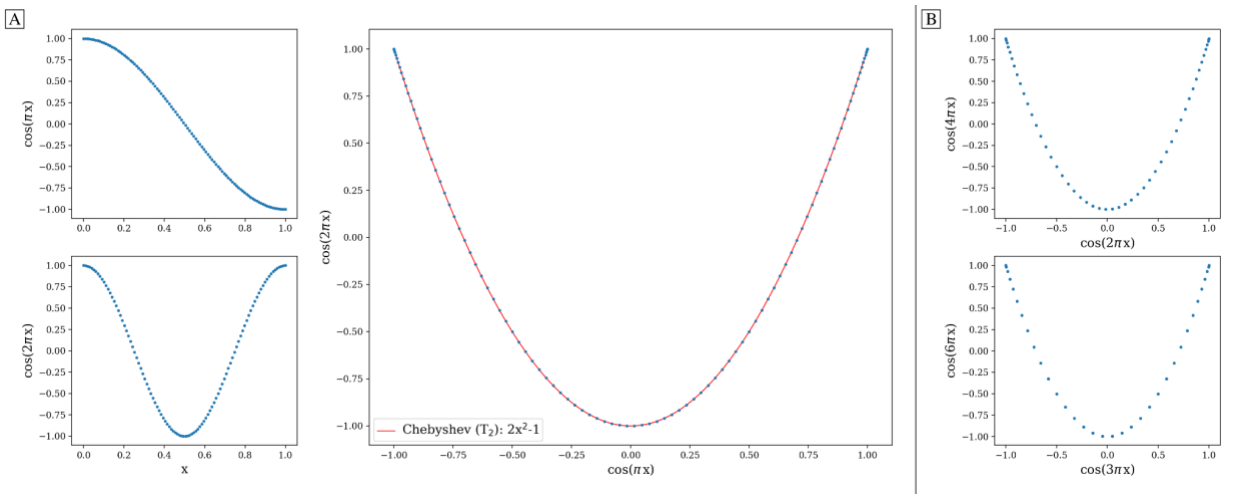


Figure 32: Analytical generation and analysis of Lissajous curves. The analytical generation of the Lissajous curve $L_{1,2} = \{\cos(\pi x) \times \cos(2\pi x) \mid \text{uniform } x \in [0,1]\}$ is shown in [A]. Note the naturally-induced nonuniform spacing between points near the boundaries and vertex of the parabola. As a simple demonstration, we also fit this curve with the Chebyshev T_2 polynomial, which is a subset of the Lissajous curves; however, T_2 does not share the same nonuniformity in spacing as $L_{1,2}$. In [B], parabolic harmonics are likewise generated for $L_{2,4}$ and $L_{3,6}$. While the same x-coordinates were used to generate all underlying cosines for parabolas in both [A] and [B], more than one point in the domain ends up mapping to each coordinate of these parabolic harmonics. As such, these harmonics obfuscate the true conformational information, which is intact on $L_{1,2}$.

Next for consideration, as seen in Figure 31-B, there exist several parabolic trajectories scattered throughout the 2D subspaces of a given Ω_{PD}. As confirmed by the indices of points and the corresponding color map along each curve, only the first of these parabolas (as shown in the first subplot of Figure 31-B) describes the full extent of the conformational motion present monotonically, while all other trailing parabolas display a non-monotonic signal. As a specific example, Figure 32 shows that the first three such parabolas can be generated via $L_{1,2}$, $L_{2,4}$ and $L_{3,6}$. These repeat the conformational information once, twice and three times, respectively, within one span of the parabolic trajectory.

As a consequence, only the mapping from the sinusoids to the first parabola in this set is bijective (injective and surjective; Munkres, 2000), with all other mappings to higher-order parabolas non-injective surjections. Importantly, since the Cartesian product of continuous functions in continuous and projections from product spaces are also continuous, this bijection further meets the requirements of a *homeomorphism*: a bijective correspondence that preserves the topological structures involved (Munkres, 2000). We denote the higher-order parabolas (formed via the non-injective surjections) as parabolic harmonics, which do not preserve topological structure and must be avoided when mapping a given CM; a problem that becomes more challenging as more degrees of freedom are added to the system.

We next compared these sets of 2D subspaces among the five PDs, and found only subtle differences in the distribution of their point clouds. It is important to underscore here the natural discrepancies between different PD manifolds that should be expected—due to PD disparity—which will continue to manifest in several significant forms throughout this analysis. Naturally, as each 2D projection provides an incomplete representation of the underlying 3D density map, depending on the type of motion and its component along the PD under investigation, ground truth

is preserved to different degrees. This disparity affects all Ω_{PD} characteristics, and will become more relevant as we investigate the embeddings of data sets generated from structures with multiple degrees of freedom.

State Space 2. To further understand these trends for embeddings formed with increasing intrinsic dimensionality, we next investigated the embeddings generated for SS_2. As seen in Figure 33-A, by plotting the points in each eigenvector in the specific ground-truth sequence constructed for CM_1 against a uniform index (now for the 400 states in SS_2; i.e., $\{1, 2, 3, \ldots, 400\}$), a similar but now interspersed pattern of sinusoids appeared. Specifically, the appearance of the sinusoids (with increasing $k \in \mathbb{Z}^+$) only manifested in a subset of all eigenvectors present, while for all other eigenvectors outside of this set, grid-like patterns emerged. These findings align with our analysis in the preceding chapter, where it was shown that these patterns arise as a consequence of projection of each eigenfunction on a plane in its n-dimensional space (Figure 26).

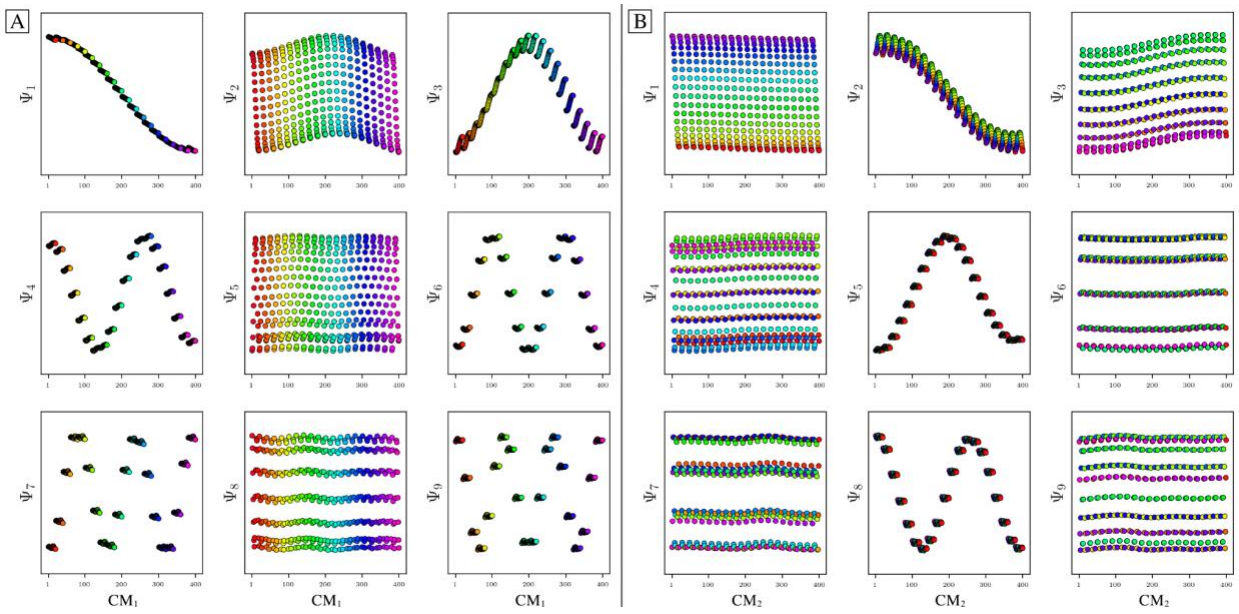


Figure 33: Analysis of eigenfunctions for PD_1 in SS_2 from data-type I. On the left [A] are the sinusoidal forms $\{\cos(k\pi x) \mid k \in \mathbb{Z}^+\}$ that emerge for only a specific subset of eigenvectors $\{k = 1, 3, 4, 6, 7, 9, \ldots\}$ when points in each Ψ_k are ordered precisely in the sequence 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 in the sequence of CM_2 , a new set of sinusoids emerge

$\{k' = 2, 5, 8, \ldots\}$ specifically 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, recall that such frames of reference are unavailable *a priori*.

As a demonstration, we next reordered the indices of points within all eigenvectors to instead correspond with the specific ground-truth sequence constructed for CM_2 (i.e., $\{1, 21, 41, \ldots, 381\}, \ldots, \{20, 40, 60, \ldots, 400\}$). The output of this operation can be seen in Figure 33-B, which manifested a new subset of interspersed sinusoids, with increasing $k' \in \mathbb{Z}^+$ independent from the previous subset; and inhabiting only those eigenvectors in the complement of the CM_1 subset. By induction—based on these observations and those in Chapter 9—we conclude that for n degrees of freedom in a given Ω_{PD}, there are n independent sets of sinusoids $\{\cos(k\pi x_q) \mid q \in n\}$, with each set interspersed throughout the collection of available eigenvectors $\{\Psi_i \mid i \in N\}$.

Following our previous discovery of a single set of orthogonal Chebyshev polynomials spanning specific 2D subspaces of SS_1, we next investigated whether similar patterns existed in SS_2. In doing so, we found that for every CM present in a state space, there exists a corresponding set of Lissajous curves interspersed across specific $\{\Psi_i \times \Psi_j\}$ projections of the embedding. Specifically, in the case of PD_1, independently projecting the data for SS_2 onto the planes spanned by its $\{\Psi_1 \times \Psi_i\}$ and $\{\Psi_2 \times \Psi_j\}$ 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 (Figure 34). With this knowledge in hand, we next compare the subset of eigenfunctions as obtained in either the reference frame of CM_1 (Figure 33-A) or CM_2 (Figure 33-B) with the Chebyshev polynomials in Figure 34.

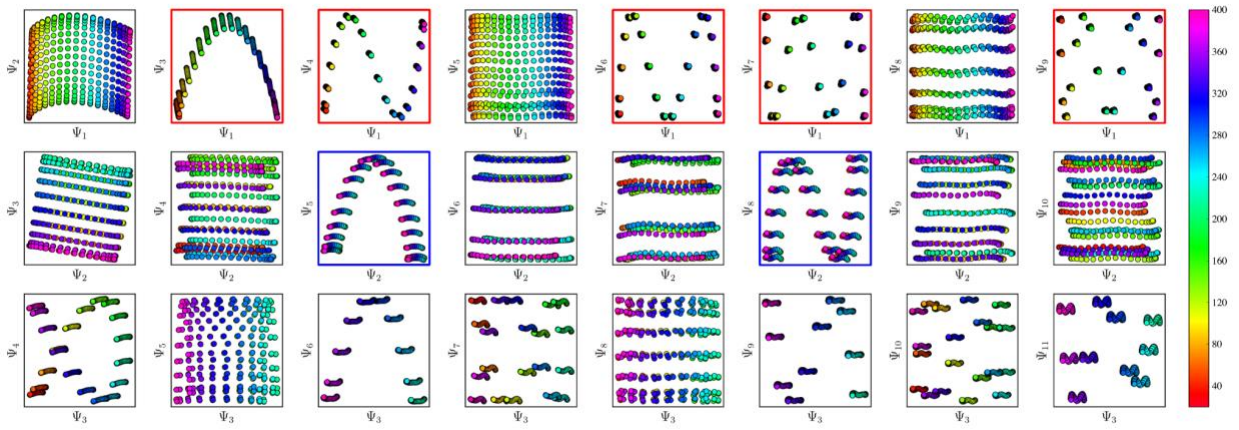


Figure 34: A subset of the space of 2D subspaces for PD_1 in SS_2 from data-type I. As demarcated in red and blue boxes, a set of conformational modes exists for both $\mathtt{CM}_1$ (red boxes, $\{\Psi_1 \times \Psi_i\}$) and $\mathtt{CM}_2$ (blue boxes, $\{\Psi_2 \times \Psi_j\}$); where $i > 1$ and $j > 2$), interspersed throughout each row. The indices for points in each set of polynomials can be visualized here via the corresponding color mapping, where $\mathtt{CM}_1$ points follow along the full spectrum of colors (i.e., a rainbow with indices $1 - 400$) while $\mathtt{CM}_2$ 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 $\mathtt{CM}_1$ located at $\{\Psi_3 \times \Psi_6\}$. Similar plots for the remaining four PDs are provided in Figure G1 in Appendix G.

Indeed, each Chebyshev polynomial mapping $\mathtt{CM}_1$ information in Figure 34 (visualized with subplots enclosed by blue boxes) corresponds to the subset of sinusoidal eigenfunctions which emerged in the reference frame of $\mathtt{CM}_1$ in Figure 33-A; with similar relations holding for $\mathtt{CM}_2$ in Figure 33-B. For convenience, a set of Chebyshev polynomials corresponding to a given CM will be referred to as the *conformational modes*). Thus, even though the knowledge required to view these CM sinusoids in unavailable outside of ground-truth studies, our analysis confirms that these CM relationships are ever-present, and further, that we can rely on their existence—via the composites of carefully chosen eigenvectors—to elucidate conformational type and order.

Combining this empirically-obtained knowledge with our *a priori* understanding of the eigenfunctions of the LBO on known domains (Chapter 9), we were able to intuit the analytical form of these Ω_{PD} eigenfunctions. In close approximation, we found that the leading Ω_{PD} eigenfunctions appear in the form

$$\Psi_i = \cos(\theta)\cos(v\pi x) + \sin(\theta)\cos(w\pi y) = A\psi_v + B\psi_w \tag{13}$$

such that a given eigenvector Ψ_i may contain some linear combination of n canonical eigenfunctions $\{\cos(k\pi x_q) \mid k \in \mathbb{Z}^+\}$ corresponding to each degree of freedom $x_q \subset \mathbb{R}^n$. In Figure 35, we use this explicit expression to near-perfectly emulate the heuristic results obtained in Figure 33 and Figure 34.

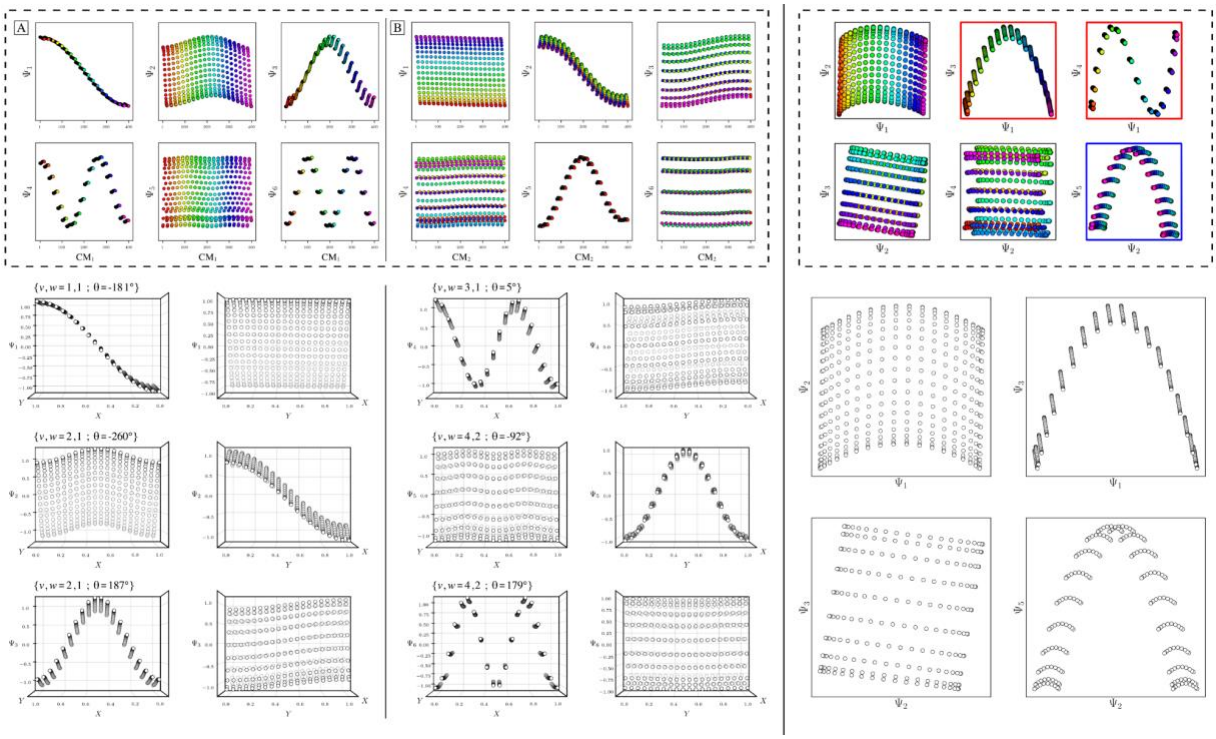


Figure 35: Comparison of analytically-generated functions with heuristic results previously obtained for PD_1 (Figure 33 and Figure 34). For each pair of subplots, values for θ were approximated by eye. Our approximations share a remarkable similarity with heuristic results, and are able to account for geometric minutiae previously unaccounted for, as well as larger-scale rotations seen in the composite of eigenfunctions. Discrepancies can be seen in the slightly tilted appearance of parabolas in Ψ_3 and Ψ_5 of Figure 33, as well as the clumping of points as observed in the CM_2 reference frame of Ψ_6. These differences can be understood as additional, small-scale perturbations which are currently unaccounted for in our general expression.

As also demonstrated in our analysis of eigenfunction misalignments in Chapter 9, the sum of these squared coefficients is conserved across pairs of eigenvectors, such that the base functions $\Psi_i' = \psi_v$ and $\Psi_j' = \psi_w$ can be expressed as a rotation $\mathbf{\Psi} = \mathbf{R}^T\mathbf{\Psi}'$, having the form

$$\begin{bmatrix}\Psi_i(\theta)\\ \Psi_j(\theta)\end{bmatrix} = \begin{bmatrix}\cos(\theta)\psi_v + \sin(\theta)\psi_w\\ -\sin(\theta)\psi_w + \cos(\theta)\psi_v\end{bmatrix} \quad (14)$$

From our analytical expression, it is clear that depending on the PD, conformational information—pertaining to each of the system's degrees of freedom—will lie on some linear combination of the embedded manifold's orthogonal eigenvectors. This feature is seen most strikingly in $\{\Psi_3 \times \Psi_4\}$ of PD_3 in Figure G1, where the parabolic surface described by the Chebyshev polynomial is significantly out of alignment with the plane of the 2D subspace containing it. Similar instances, albeit in more subtle form, also arise for surfaces in the remaining three PDs of Figure G1. In Chapter 9, we demonstrated that the need for eigenfunction realignment is due to the change in apparent interatomic distances dependent on projection direction (Figure 30). This disparity among PDs is inevitable, and poses a fundamental problem that must be addressed.

State Space 3. We next investigated the 1000 states making up SS_3. For each conformational motion present in a given PD data set (this time for CM_1, CM_2 and CM_3), a set of unique Lissajous curves were again found spanning specific 2D subspaces of the embedded manifold, with the Chebyshev subset describing the corresponding CM along a 2D trajectory explicitly. As an example, Figure G2 shows the set of 2D subspaces where these modes exist for PD_5. To note, due to the increased complexity of SS_3, these patterns were much more interspersed throughout the embedding, but still followed a similarly predictable ordering. In addition, due to the relatively small range of motion exhibited by the third conformational domain (as seen from these PDs and as designed in the ground-truth structures), all CM_3 modes were found in higher-order eigenvectors; e.g., Ψ_5 and higher for these five PDs. As similar patterns were identified in SS_3 as in previous accounts, for the remainder of our study, focus is honed onto mapping data sets generated specifically for SS_2.

10.3 Embeddings for Data-type II

In the following section, an analysis of data-type II using SS_1 and SS_2 is performed. As finite SNR is an important attribute of any experimental data set, we next sought to understand how the structure of these manifolds change with varying SNR and state space coverage. First, we sought to better understand the differences between embeddings obtained via linear and nonlinear dimensionality reduction methods within varying SNR regimes. To this end, we compared the manifolds from principal component analysis (PCA) and diffusion maps for PD_1 with additive Gaussian noise, such that the images in each data set had unique SNR $\in \{0.01, 0.1\}$ consistently applied to all images in a set. For the PCA approach, instead of defining the Gaussian kernel as previously used in DM for 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 data set $\mathbf{Z}$ of dimension $P \times N$). Before embedding, we standardized the images in each data set by removing the mean and scaling to unit variance, and generated eigenfunctions of the resultant $N \times N$ matrix $\mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}^T$ is the transpose of $\mathbf{Z}$. To note an important comparison between PCA and DM, the matrix $\mathbf{Z}^T\mathbf{Z}$ is symmetric and positive semi-definite (i.e., all eigenvalues are non-negative; Lay et al., 2016), which is also the case for the Markov transition matrix used in the DM method.

A set of different projections of this embedding as obtained from selected eigenvectors (i.e., principal components, PC_i) can be seen in the first three columns of Figure G3, with results from DM similarly presented for comparison in the remaining columns. As demonstrated, the eigenvalue spectra and eigenvectors obtained from performing PCA and DM are almost identical, except for subtle differences in the spacing between states and boundaries for the pristine case (data-type I, shown in the first and fourth column of Figure G3). These similarities align with our

a priori knowledge of the existence of quadric surfaces for positive semidefinite matrices. The similarity between these manifolds holds for all subspaces explored, with distinctions diminished in the presence of noise. Overall, the corresponding spectral geometry obtained from each method became increasingly similar as the SNR was decreased (Figure G4). The results of PCA versus DM on SS_1 and SS_3 show similar behavior.

We next investigated the effects of varying state space coverage across several SNR regimes, and its effects on the robustness of the corresponding manifolds produced by PCA. As the choice of PCA or DM proved irrelevant in these low-SNR regimes, PCA was chosen here so as to bypass uncertainties introduced by the need of tuning in DM via the Gaussian bandwidth. For this study, we used the 20 images in PD_1 representing SS_1 (i.e., the full range of CM_1), and varied both the number of times (τ) these $M = 20$ ground-truth states were duplicated as a group—with each instance having a different realization of additive Gaussian noise—and the SNR of each image therein. Here, Gaussian noise of constant variance was applied for each SNR regime and uniquely added to each of the $\tau M = N$ images independently. An excerpt from the results of our analysis is shown in Figure 36, where a highly structured pattern emerged. Specifically, when increasing levels of noise was added to each image (decreasing SNR), increasingly larger values of τ were required to reestablish coherent structure in the spectral geometry; i.e., the set of Lissajous curves and corresponding Chebyshev polynomials.

To quantify these relations, each member of the set of PCA-embedded manifolds in a τ series was fitted with a set of leading Chebyshev polynomials, as seen in Figure G5 for SNR of 0.1. The coefficient of determination ($\mathcal{R}^2$), which can be interpreted as the proportion of variance in one variable accounted for by another (Schober et al., 2018), was then computed for each mode therein. The resultant trends across several SNR regimes are plotted in Figure G6. Our findings

show that as τ is increased, the rate at which the geometry of each subspace reaches its most stable regime is dependent on SNR. A critical τ_c value was determined both visually and analytically by assessment of the asymptotes for each SNR regime, beyond which larger values of τ provided no further improvement to the spectral geometry.

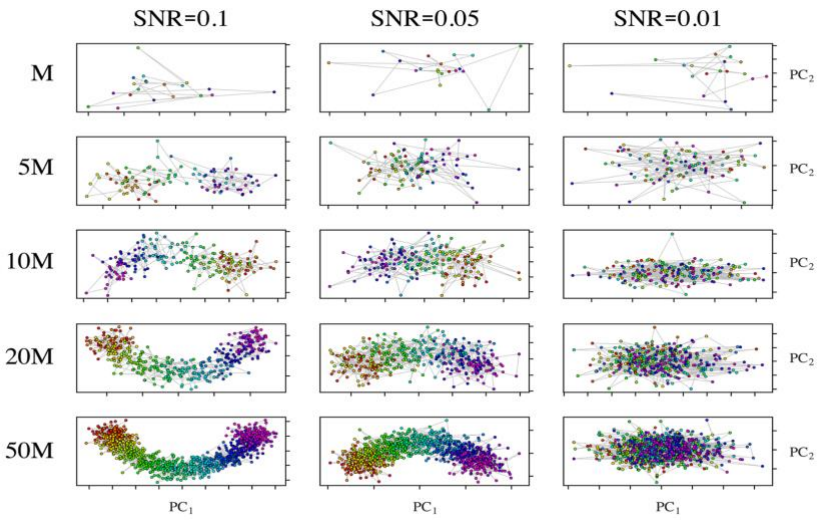


Figure 36: Set of PCA subspaces over a range of SNR values and state coverage. Set of $\{PC_1, PC_2\}$ subspaces correspond to PD_1 images in SS_1. As can be seen in the columns, the fidelity of the point-cloud distribution in each subspace to the parabolic form increasingly deteriorated due to decreasing SNR regimes. However, as the $M = 20$ state space was populated by increasing values of τ in each of these SNR regimes, the intrinsic parabolic structure of the embedding reemerged. To be precise, $5M$ represents five exact copies of the 20 SS_1 images ($\tau M =$ 100 images), with unique Gaussian noise added to each image independently as prescribed by its SNR regime. It can be seen here that all values of τ shown (up to 50) in the SNR of 0.01 regime are too low for recapitulating the intrinsic parabolic structure of the embeddings, and, as further illustrated by the color mapping of their points, no sensible ordering of snapshots can be ascertained within these subspaces.

Further, across all of these regimes, each subsequently higher-order Chebyshev polynomial required a larger value of τ_c to be properly resolved (Figure G6-A), which is a consequence of higher-frequency patterns requiring more points to resolve when the number 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 data set is infeasible. Parameters influencing τ_c include not only unknowns such as the number of ground-truth states M and SNR regime, but also the intrinsic dimensionality n of the data set and the free energy of the system.

We next describe specific characteristics of CM subspaces obtained from data sets generated with these noisy-duplicate images. Specifically, we examine the parabolas generated via PCA from SS_2 PDs with SNR of 0.1 and $\tau = 10$, which will guide several choices made for our methodology in the following chapter. Figure 37 shows the composite parabolic trajectory and corresponding sinusoidal form of each eigenfunction for CM_1 and CM_2 of PD_1, as well as a collection of similar CM subspaces from randomly-selected PDs. Each subplot has been assigned a color map matching the ground-truth sequence of states of the CM to which it corresponds, with this sequence partitioned into 20 equally-occupied bins (i.e., CM states). As can be seen, while each of the two underlying point clouds corresponding to a unique eigenfunction maintains well-defined structure after introduction of noise, CM state partitioning becomes increasingly disordered in their composite parabolic point cloud.

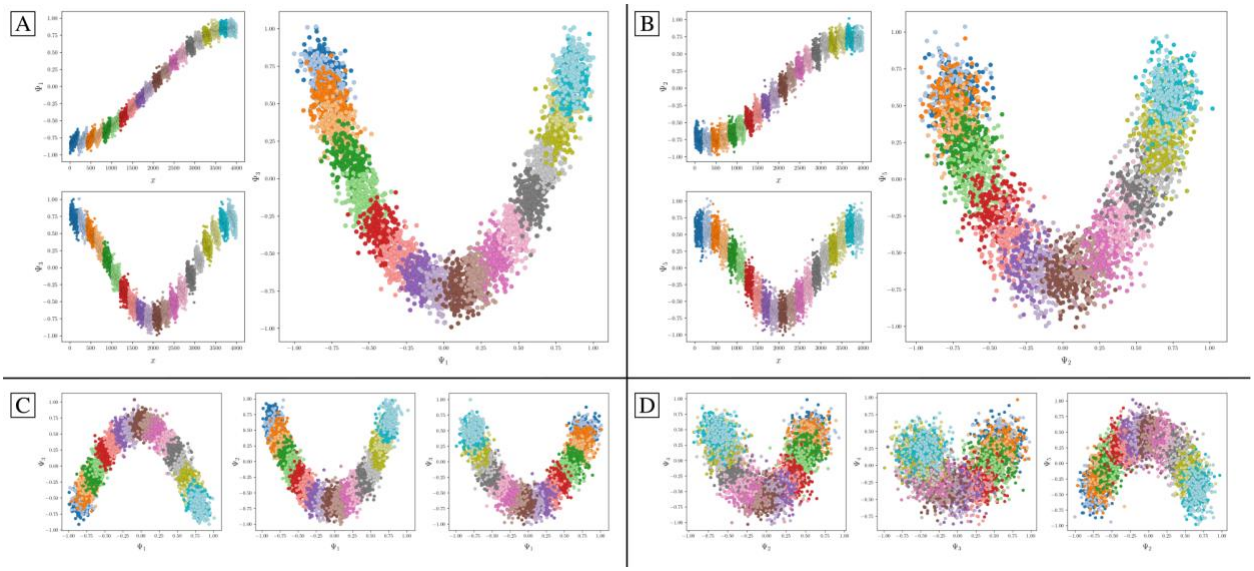


Figure 37: Comparison of CM subspaces for five PDs generated from data-type II. Here, SNR of 0.1 and $\tau = 10$ is used, with embeddings achieved via PCA. The coordinates within each

point cloud are colored to indicate their ground-truth CM state assignment, such that each point belongs to one of the 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 embeddings obtained from noisy images.

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 density maps along a conformational trajectory, certain clusters of PDs on 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.

10.4 Embeddings for Data-type III

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 TEM. For these trials, we first generated and applied a CTF to each image as previously described (Figure 17). Specifically, using images from PD_2 of SS_2 with $\tau = 10$, we assigned to each image a random defocus value from the interval $[5000, 15000]$ Å. Similar intervals are typically used in modern cryo-EM experiments. Likewise for each image, constant values were used for voltage (300 kV), spherical aberration

coefficient (2.7 mm), and amplitude contrast ratio (0.1) to emulate typical TEM conditions. These parameters were jointly used to construct a CTF for each image, which was applied via multiplication to the image's Fourier transform. With the collection of images modified by unique CTFs, additive Gaussian noise was next applied such that the SNR of each image in the resultant ensemble was approximately 0.1.

We next set out to measure the extent of interference of the CTF on the corresponding manifold for an example PD. As described in Chapter 2, images with different CTFs—as constructed in our study in emulation of the experimental situation—are no longer directly comparable using a standard distance metric. Instead, an adjustment to the kernel must first be made to account for our introduction of CTF. We show here the results of applying the previously established double-filtering kernel, which ensures a zero Euclidean distance between any two images that differ in defocus only (Dashti et al., 2014). The corresponding manifold embedding is shown in Figure 38, which juxtaposes these results with the same data set generated without CTF and using a standard Gaussian kernel.

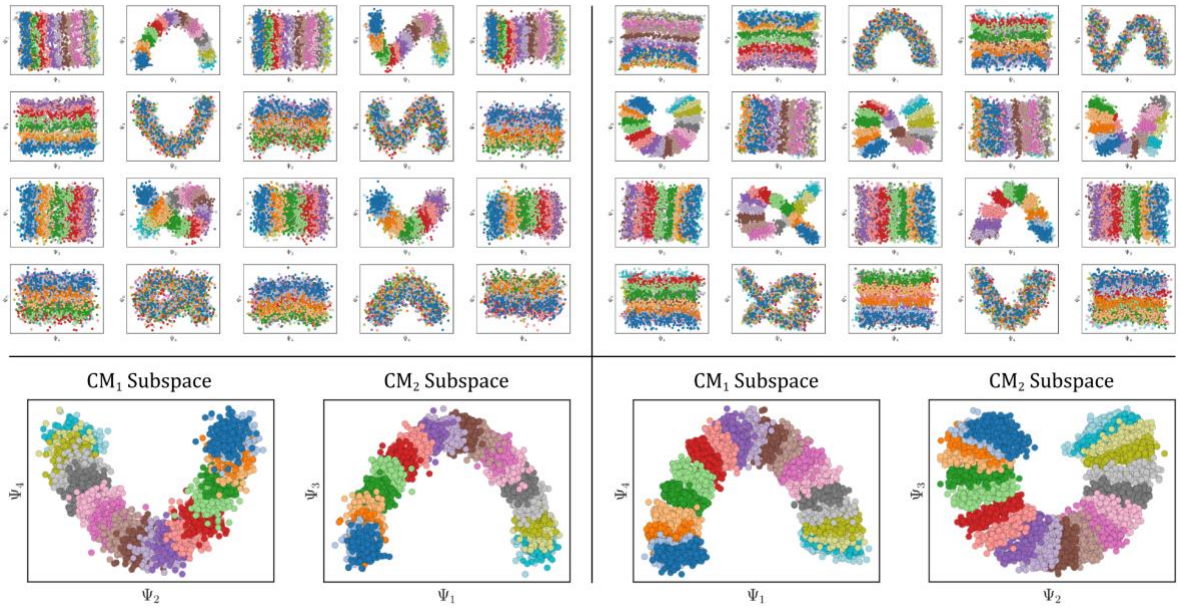


Figure 38: Embeddings obtained with and without double-filtering kernel, as shown on the right and left-hand side, respectively. Specifics for these embeddings are: SS_2, PD_2, $SNR = 0.1$ and $\tau = 10$. For the case of the embedding obtained from images without defocus, protocols for synthetic generation follow those established in Figure 17 (A, B). Likewise, on the right, protocols follow synthetic generation of images with microscopy parameters as shown in Figure 17 (D, E). The non-CTF manifold embedding was generated via DM with the standard Gaussian kernel, while the CTF-manifold embedding was obtained via DM with the double-filtering kernel. On the top insets, colors displayed represent the ground-truth CM_2 bins, while for the bottom insets, for both sides, colors represent CM_1 bins (left) and CM_2 bins (right).

As seen on the right-hand side of Figure 38, there is a noticeable inward-curling at the ends of the CM subspace parabolas generated using the double-filtering kernel. Notwithstanding this artifact, we found that the double-filter kernel was successful in preserving the most important aspects of the manifold. This approach also proved superior to alternative techniques explored, such as embedding using a standard kernel from sets of CTF-corrected images. It should be noted, however, that perfect defocus assignments were used here for CTF correction, when in reality these values would be estimated first using established algorithms (Tang et al., 2007; Scheres, 2012; Rohou and Grigorieff, 2015).

10.5 Closing Remarks

In the following chapter, several considerations are provided pertaining to the relevancy and breadth of this heuristic analysis. Our analysis is focused on data models originating from molecules undergoing collective rigid-body motions, which we believe are sufficient for most molecular machines, but may fall short of addressing instances involving more complex situations. 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 its binding partners. For such a situation, a similar heuristic analysis could be conducted using synthetic models occupying two or more state spaces.

For completeness, we further tested the ability of PCA and DM to correctly embed PD manifolds formed from models exercising more complex domain motions. For this purpose, an ensemble of projections of the mouth-wings toy model (Figure 1) was generated, modeled using a collection of 2-spheres (representing atoms; as described in Figure G7) and rendered using professional 3D modeling software (Cinema 4D; Maxon Computer GmbH, 1989). Notably, the spheres making up the "mouth" domain were uniquely positioned by hand (keyframed) for each state so as to gradually clump together, ultimately presenting a higher density towards the fully-open state (Movie G1). In a similar vein, the "wings" were programmed in unison to both open at constant angle about their hinge, while also curling inwards.

Compared to the synthetic framework used to generate the Hsp90 data set, this workflow provides a radically different approach, and accounts for complex interactions of particles within each domain motion. Nonetheless, the embedding of these mouth-wings images still manifested all essential geometric characteristics previously detailed for Hsp90: presenting SS_2 across a parabolic sheet (Figure G8), as expected. Although the procurement of the mouth-wings model is

nowhere near an exhaustive search, we believe the correspondence between its outputs and those of the independently-designed Hsp90 data set strongly establishes the universality of our discoveries.

This example illustrates the broad scope of our heuristic analysis, which not only provides insights for cryo-EM data, but also for any method in general dealing with projection data. Further, as demonstrated in Chapter 9, several portions of our analysis have been directly extended to other experimental techniques dealing with alternative manifold inputs, such as the use of atomic models in molecular dynamics and 3D density maps in cryo-electron tomography. As such, we believe that there is a potential for the application of these insights to a wide range of experimental data sets beyond cryo-EM, and particularly those obtained from systems exercising multiple degrees of freedom in a continuous manner.

Finally for consideration, we have only dealt here with molecular machines that specifically exhibit each of their domain motions along an independent and mutually unrestricted sequence of quasi-continuous states. All n-wise combinations of these bounded intervals (one for each conformational motion) produce an n-dimensional shape with a rectangular boundary. In Chapter 9, we have shown that the corresponding Laplacian eigenfunctions are well defined for this domain. However, in general, analytically solving the Laplacian for any arbitrary boundary is impossible. Eigenfunctions can change drastically depending on the boundary, and are analytically only known for certain elementary shapes, such as rectangles, discs, ellipses and special triangles (McCartin, 2008; Grebenkov and Nguyen, 2013). On the other hand, geometric machine learning approaches can obtain solutions numerically, in principle for any boundary. However, such geometric machine learning methods still require the boundary to be known *a priori*. For systems with unknown boundaries, the problem is intractable.

As the set of all possible molecular machines is unfathomably complex, it is unlikely that one single algorithm could ever be so versatile as to anticipate every possible instance. Instead, we are interested in casting a wide enough net so as to capture the dynamics of a large portion of these systems, which we surmise operate within rectangular boundaries of an n-dimensional latent space of relatively-rigid multi-body motions. However, one can still imagine all sorts of other situations, such as a system where one domain blocks—via *steric hindrance*—another domain from its full range of motion in a specific region of the state space. This topic will be returned to after introduction of the ESPER method in the following chapter, with the potential impact on that methodology thoroughly addressed.

Chapter 11: The ESPER Method

11.1 Overview of ESPER

Having explored all three data-types, we now lay out our informed strategy—ESPER: *Embedded subspace partitioning and eigenfunction realignment*—for the recovery of conformational motions in the form of 3D movies and a corresponding free-energy landscape. The following is a continuation of the previously described analysis on manifold embeddings, with identical contributions, and follows the comprehensive description available via Seitz et al. (2021a). The ESPER methodology requires several steps that will be introduced in turn, with the entire process schematized in Figure 39.

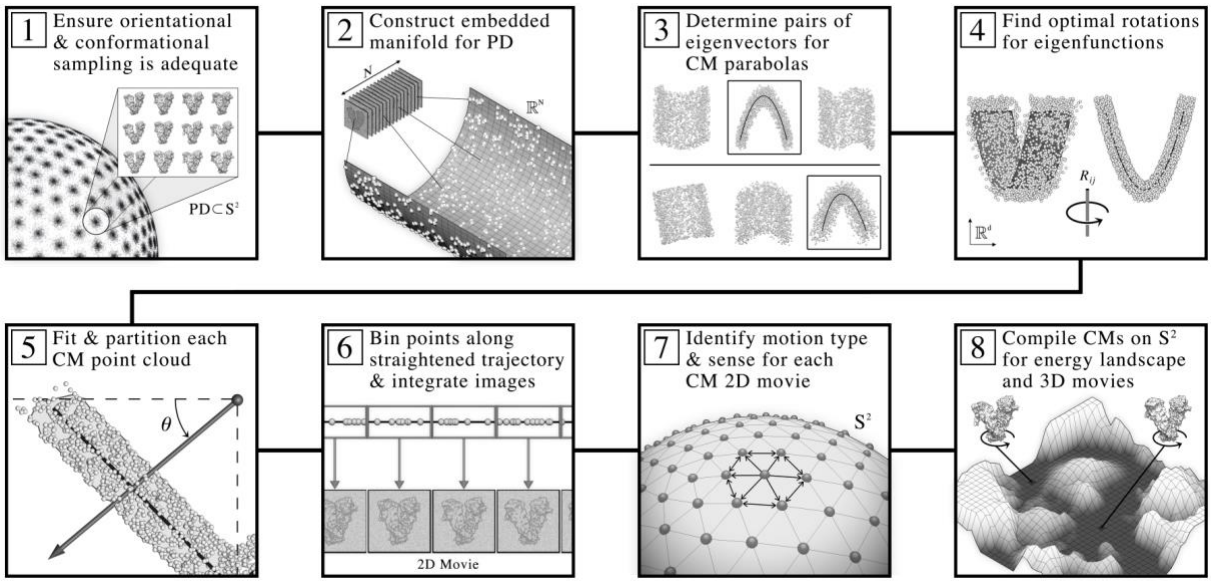


Figure 39: Schematic of the ESPER workflow for recovery of conformational continuum as informed by heuristic analysis. Through this methodology, 3D movies and corresponding free-energy landscapes are obtained for the set of conformational motions in a given data set. Note that the previous ManifoldEM workflow branches off after completion of step 2 above, and after performing a series of alternative steps required by NLSA, then enters again with our pipeline at step 7, before splitting off again to form final outputs—as we achieve independently via ESPER—in step 8.

The general intuition for our approach is as follows. Ideally, for each Ω_{PD} embedding, we first wish to translate the n conformational-variation signals residing along a high-dimensional parabolic surface into a rectilinear n-dimensional state space. To this end, one can imaging 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 with each bin's boundary. This procedure should then be repeated for each Ω_{PD} independently. To reconcile the contents of these PD manifolds on S^2, which may contain conformational information along different coordinates due to PD disparity, the orientation of each n-dimensional grid (and thus ordering of bins therein) must be aligned so as to match across all PD manifolds. Next, the set of images belonging to each compiled bin must be combined to reconstruct a 3D density map of the molecule, with the total image count used to define a state occupancy. As a result of this construction, an n-dimensional occupancy map (and thus, via the Boltzmann relationship, free-energy landscape) can be formed, along with a set of corresponding 3D density maps representing every state.

In application, however, there are many complications to this procedure. For one, the desired high-dimension parabolic surface presents difficulties in both discovery and direct mapping. Since there are many such potential subspaces housing parabolic surfaces (even for two degrees of freedom) with the embedding of a given Ω_{PD}, there exists ambiguity as to which one contains the desired information, further exacerbated in the presence of harmonics and experimental artifacts. In addition, due to the complex nature of these hypersurfaces—which can vary in features ranging from elliptical, to parabolic, to hyperbolic, and each with boundary aberrations—it is much easier to instead fit and partition the set of its orthogonal components; i.e.,

the parabolas residing in easily identifiable 2D subspaces. Taking 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 ESPER approach first aims to find the set of parabola-housing subspaces (via least-square conic fits) required for elucidating higher-dimensional CM information, while accounting for eigenvector rotations and carefully eliminating harmonics (steps 3 and 4 in Figure 39). Each parabola-housing subspace is next transformed (along with its conic fit) to account for nonuniform rates of change, with points partitioned into contiguous equal-area bins (step 5). Each bin is then filled with a set of image indices corresponding to all points falling within its geometric bounds. Images belonging to each bin are next averaged to form the frame of a 2D movie (step 6), which is used to identify both the type of CM and the directionality of its motions (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 higher-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 density maps in all bins are thus obtained (step 8). In the following sections, we provide a more detailed description of these steps.

11.2 Eigenfunction Realignment

We describe here our methodology for calculating the rotations required for eigenfunction realignment of each embedded Ω_{PD}. We consider this calculation to be the first step in the ESPER method (i.e., step 3 in Figure 39). To note, after generation of each embedding from PD-images as previously described (steps 1 and 2), our methodology deviates here from the existing ManifoldEM framework (Dashti et al., 2014; Mashayekhi, 2020; Seitz et al., 2021b), which would next move on to NLSA without accounting for eigenfunction realignment of Ω_{PD} subspaces. First, recall that, depending on the PD, the observed conformational eigenfunctions may be misaligned with respect to the ideal eigenfunctions of the LBO.

As a remedy to this problem, since each CM is represented by a set of orthogonal sinusoids (one per degree of freedom), we thus aim to isolate these sinusoids in their complete form within each PD-manifold eigenbasis. As detailed in Chapter 9, by use of appropriate rotation operators $R_{i,j}$, the summands within each eigenfunction pair can be maximally separated among a set of eigenvectors (e.g., $\Psi_i = \psi_v$ and $\Psi_j = \psi_w$) such that an ideal (i.e., canonical) eigenbasis is recovered. As a result of this decoupling of eigenfunctions onto a set of appropriate eigenvectors, each corresponding parabolic surface becomes manually aligned within its 2D subspace, such that the projected structure is again that of the 2D Chebyshev parabola carrying information about a single CM along its curve. In this projected view, states differing in coordinates that are orthogonal to the projection plane—and thus describe ulterior CM information embedded on a higher-dimension surface—overlap; a feature we will take full advantage of later when generating 2D conformational movies. Thus, as long as each parabolic trajectory corresponding to a given CM is aligned with the plane of an independent 2D subspace, we can restrict our study to an analysis of only a few essential subspaces.

To provide rationalization for this technique, Figure 40 shows the eigenvectors for the highly-misaligned PD_3 eigenbasis (SS_2 from data-type I), ordered along CM_1. As seen in the first column of Figure 40-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 heavily deformed. As a direct consequence, any Lissajous curve that inherits one of these deformed sinusoids (e.g., any subspace composed in combination with Ψ_2 or Ψ_3) will be misaligned with respect to its ideal form (Figure 40-B). Given this insight, we introduce a method for correcting these misalignments using orthogonal transformations. Specifically, we apply a d-dimensional rotation operator of sufficiently large dimensions to single-handedly reorient all aberrant surfaces in their respective 2D subspaces. The results of this operation on the embedding associated with PD_3 can be seen in Figure 40-B and Figure 40-C; before and after applying a 5D rotation matrix, respectively.

Mathematically, this d-dimensional rotation is a subgroup of the orthogonal transformation in d dimensions with determinant one. These orthogonal transformations are linear and represented by a $d \times d$ matrix $\mathbf{O}$ with the property $\mathbf{O} \times \mathbf{O}^T = \mathbf{I}$. As a consequence, orthogonal transformations leave lengths and angles between vectors unchanged. Each such matrix $\mathbf{O}$ can further be represented by $d(d-1)/2$ rotation sub-matrices $R_{i,j}$, with each sub-matrix parameterized by a unique angle θ and operating on a specific plane.

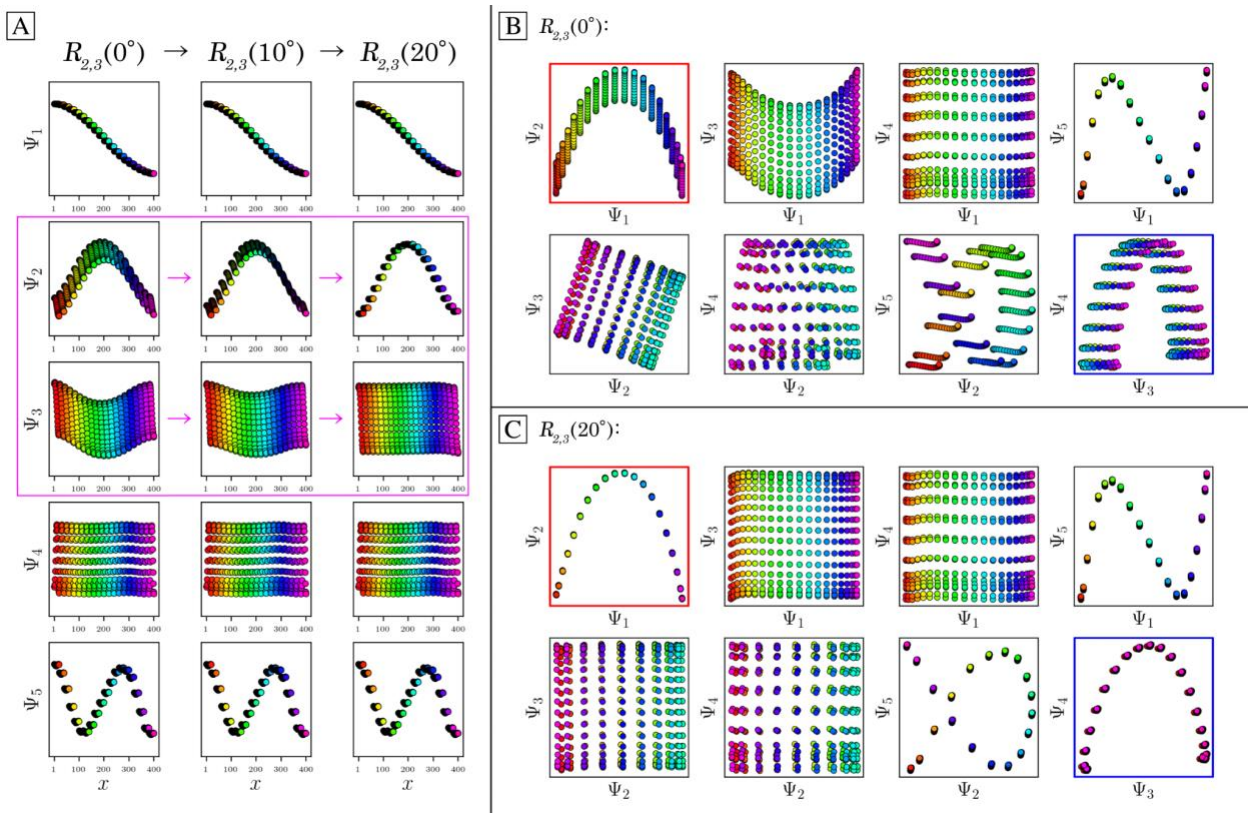


Figure 40: Application of a 5D rotation matrix on a SS_2 embedding generated for PD_3 from data-type I. The three columns in [A] display the individual eigenfunctions (as plotted by indices corresponding to the CM_1 frame of reference) before the $R_{2,3}(\theta)$ rotation is applied, at $R_{2,3}(10°)$, and finally at $R_{2,3}(20°)$, respectively. Note that $R_{2,3}(20°)$ 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 $R_{2,3}(20°)$ 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.

For the specific case of the 5D rotation matrix used in Figure 40, there exists 10 rotation sub-matrices in total, with each corresponding to a specific planar rotation on the eigenbasis. Of these 10 matrices, we found that only one had to be altered to achieve the results shown, having the general form

$$R_{2,3}(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ 0 & \cos(\theta) & -\sin(\theta) & 0 & \cdots \\ 0 & \sin(\theta) & \cos(\theta) & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{15}$$

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 Figure 40-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). As seen in Figure 40-A, along the way in reaching $R_{2,3}(\theta)$, Ψ_2 and Ψ_3 have effectively exchanged between each other an equal share of their initial content via a series of deformations, with each initial eigenvector thus sharing some combination of the other's initial sinusoidal form (as was analytically demonstrated in Chapter 9). After this exchange, the initially overlapping sinusoidal information contained in part between Ψ_2 and Ψ_3 is maximally separated between both eigenvectors, resulting in the alignment of all corresponding Lissajous surfaces with their 2D subspaces (Figure 40-C), as desired.

An additional example is provided in Movie H1, showing the effects of applying a 4D rotation to the 4D subspace for data-type II, with SNR of 0.1 and $\tau = 10$, corresponding to PD_2 in SS_2. Here, only one of the six required rotation matrices was 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. To generalize this solution for any Ω_{PD} embedding, there are thus three unknowns: (i) the dimensionality d of the matrix $\mathbf{O}$; (ii) the required rotation sub-matrices $R_{i,j}$; and, for each of these $R_{i,j}$, (iii) the rotation angle θ. After careful observation of all PDs across numerous data sets, we have determined that the dimensionality d and rotation

operators $R_{i,j}$ required are linked to the indices of eigenvectors housing each CM parabola. As a consequence, we need to first determine these CM subspaces, which can be identified by a systematic comparison of least-squares fits, while eliminating subspaces housing parabolic harmonics. The pseudocode of the eigenfunction determination procedure is given in Algorithm H1 in Appendix H. As a result of Algorithm H1, eigenvectors housing CM subspaces $\wp$ are identified (line 11), while excluding possibility of parabolic harmonics (line 13), with $2(\tilde{n}!)/(\tilde{n}-2)!$ essential d-dimensional $R_{i,j}$ operators defined (line 1.18). The removal of parabolic harmonics can be easily understood, since any 2D subspace formed in part by an eigenfunction corresponding to a known CM parabola cannot combine to form some other orthogonal CM parabola.

Once these CM subspaces are known, we approximate the third unknown—the rotation angle—using 2D histograms. In the case of noisy data, as each 2D subspace is rotated by a given $R_{i,j}(\theta)$, 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 θ. When we plot the number of nonzero bins in the corresponding 2D histogram as a function of $R_{i,j}(\theta)$, the minimum in this distribution corresponds to the angle required to properly counter-rotate each 2D subspace by the current operator (Movie H2). The pseudocode for the eigenfunction realignment procedure is given in Algorithm H2 in Appendix H. To good approximation, the d-dimensional rotations performed for each $R_{i,j}$ operator in Algorithm H2 realign the essential eigenfunctions of each Ω_{PD} CM subspace. An example visualization of this entire workflow, demonstrating Algorithm H1 and Algorithm H2 applied on a SS_2 embedding from data-type II, is provided in Movie H3.

We additionally perform a final least-squares fit $\widehat{\Psi}_{fit}$ on each rotated CM subspace. For data-type III, we found an implicit equation of a general conic section to be most flexible, defined by a polynomial of degree two

$$ax^2 + bxy + cy^2 + dx + ey + f = 0 \tag{16}$$

which allows for the possibility of parabolic-like trajectories with elliptic or hyperbolic features. This flexibility is essential for fitting parabolic-like point clouds with inwards curling near the boundaries, as was observed for manifolds formed by images modified by the CTF (Figure H1).

11.3 Subspace Partitioning

Once the required Ω_{PD} eigenfunctions are correctly rotated into a common eigenbasis as defined by the desired CMs, and each 2D subspace housing CM information is identified for each PD (steps 3 and 4 in Figure 39), we next partition these 2D subspaces into contiguous equal-area bins (step 5), representing collectively a quasi-continuum of conformational states. To note, here the ESPER method differs decisively from the NLSA approach. The motivation for the ESPER subspace-partitioning approach stems from our analysis of PD disparity in the presence of noise (as shown in Figure 37), where it is observed that the ground-truth bins and overall area of each point cloud manifest in a variety of sizes depending on PD viewing angle. These observations inspired an area-based point-cloud fitting approach able to correctly chart spatial discrepancies while remaining unencumbered by changing densities (i.e., occupancies) along each trajectory. Figure 41 provides an overview of our strategy for splitting up each CM subspace into a sequence of equal-area bins, with subplots detailing recovery of CM_1 states and corresponding occupancies for a single PD using data-type II. As well, the pseudocode for this subspace partitioning procedure is given in Algorithm H3.

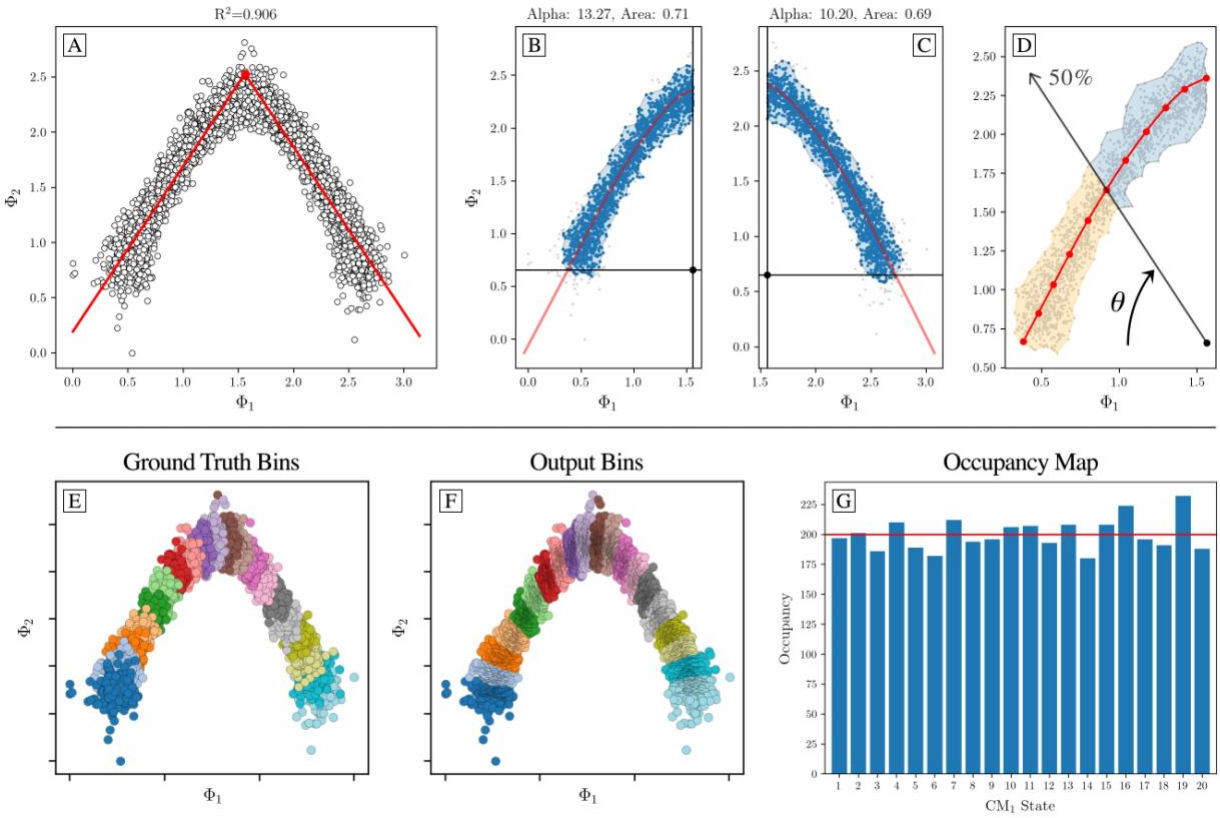


Figure 41: Overview of the ESPER subspace-partitioning workflow for extracting sequential conformational information from a 2D subspace. Subplots [A] through [G] display our algorithm's outputs on the CM_1 subspace of an arbitrary PD from data-type II. First, [A] shows the inverse-cosine transformation 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 halved subspace, with an 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 produced by the ESPER method. 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.

To initiate this procedure, we must first correct for the nonuniform rates of change in each CM subspace which arise innately as a result of taking the Cartesian product of sinusoids. As a remedy, we investigated the use of an inverse-cosine mapping on each eigenfunction. Figure H2 provides the results of this transformation on both (i) the analytically-derived cosine functions $k \in \{1, 2\}$ and (ii) the SS_1 eigenfunctions obtained by applying DM on images in PD_1 using data-

type I. The first two subplots in Figure H2 further highlight the remarkable fidelity of the DM eigenfunctions of the graph Laplacian to the analytical form of the LBO, while the third subplot illustrates the results of inverse-cosine transformation. As can be seen, this mapping presents the coordinates of each eigenfunction in a space with uniform rates of change, consistent with the ground-truth relationships between atomic-coordinate structures. We will indicate any eigenvectors Ψ_i under this transformation with the insignia Φ_i.

To mitigate the overall complexity of operations, the axis of symmetry of the conic equation is then used to split the subspace into two halves, such that each half can be operated on individually. We next apply a balltree algorithm (Omohundro, 1989) to temporarily prune outlier points for heightened accuracy during subsequent steps. Specifically, the balltree 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, the alpha shapes algorithm (Edelsbrunner et al., 1983) is used to define the overall area of each CM point cloud with a polygon—a generalization of the convex hull—representing the key features of its geometric shape (Figure 41-B and Figure 41-C). To achieve this, the alpha shapes algorithm defines the boundaries of the point cloud by a series of α-discs (1-spheres of radius $1/\alpha$), such that an edge of the alpha shape (polygon) is drawn between two members of the point cloud whenever there exists an α-disc containing no members of the point cloud, and carrying the property that the two points lie on its boundary. A family of alpha shapes can thus be defined for each halved subspace via the α parameter, ranging from coarse (a convex hull) to increasingly finer fits around the point cloud. Within this class of polygons, there exists a member providing an optimal level of refinement, in which the alpha-shape area and point-cloud area are equal (Gardiner et al., 2018). For our purposes, the determination of a suitable value for this parameter

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

Next, using rays emanating from a point opposite the point cloud's apex (Figure 41-D), we divide this polygon into a collection of contiguous sub-polygons of equal area (Figure 41-F). Each of these sub-polygons, in sequence, corresponds to one of the CM's unique states, with the total number of points contained within each sub-polygon defining the corresponding state's occupancy (Figure 4-G). Importantly, we store the image indices belonging to each bin along the given CM for subsequent use in forming an $n > 1$ occupancy map. Since the points in any CM subspace that are aligned orthogonal to the respective projection plane describe ulterior CM information, averaging points together in that subspace only reveals the conformational information corresponding to the current CM. Hence, cryo-EM images assigned to each state 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 CM parabola resides, and so on for higher degrees of freedom. The 2D movies produced by this procedure for SS_2 from data-type II can be found in Movie H4, with a similar construction provided for data-type IV in Movie H5.

11.4 Conformation Compilation

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. For these decisions, recall that a comprehensive automated strategy has been developed using optical flow and belief propagation algorithms (Maji et al., 2020). Once CM

types and senses are 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 of that same CM across all PDs. We next generate an n-dimensional occupancy map by taking the intersection (overlap) or image indices corresponding to each combination of bins in the CM trajectories per PD. Since the CM coordinates are intrinsically linked by the independent occurrence of image indices from the same PD image stack, the operation effectively reconstructs the n-dimensional hypersurface on which the images jointly reside. (If only one degree of freedom is desired, naturally no intersection is required).

Next, image stacks—one for each state—are generated and paired with an alignment file that carries the input alignment and microscopy information for each image therein. This file can then be used as input for the 3D reconstruction (e.g., as can be performed by RELION; Scheres, 2012) of the molecule in each state in the compiled state space. The ESPER method is thus concluded when these occupancies are transformed into a free-energy landscape. For any information left out in these preceding steps, a more detailed description is additionally available (Seitz et al., 2021a), including comprehensive Python code (Seitz, 2021).

11.5 Analysis of ESPER Results with Synthetic Data

The results of applying the entire ESPER method on the 126 PDs (recall Section 6.7) from SS_2 in data-type IV (with experimentally-relevant SNR and CTF) are shown in Figure 42 and Movie H6. The former demonstrates the accuracy of occupancy assignments for comparison with Figure 18, with the 2D occupancy map 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. As a note, considering there were only 126 PDs to decipher, we opted to determine the type of CM and its

sense with perfect accuracy by visual inspection of the 2D movies. Overall, the results are very accurate, with only a subtle difference in occupancies near the boundaries of each CM, which manifest on the four corners of the 2D occupancy map. These discrepancies are mainly due to a combination of PD disparity, CTF-induced inward curling, and the vanishing derivatives of the DM eigenfunctions at the boundaries (Coifman et al., 2005) arising in each Ω_{PD} embedding.

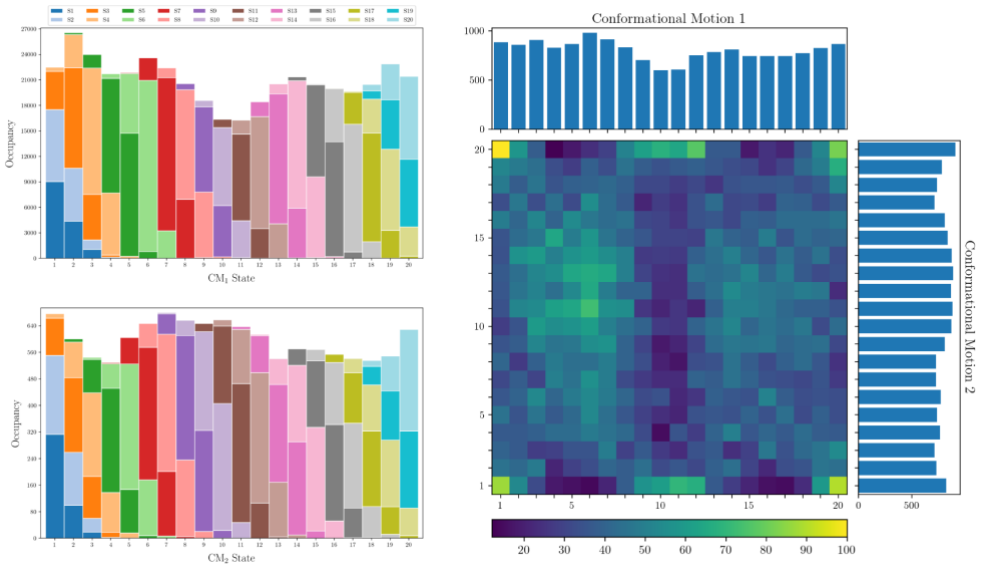


Figure 42: Final occupancy maps produced by the ESPER method for Hsp90. On the left, the occupancy maps for the 20 states in CM_1 (top) are shown alongside an equivalent representation for the 20 states in CM_2 (bottom). Each plot was obtained by integration of the corresponding 20 bins (corrected for sense) in each of the 126 PDs. The total number of images as assigned to each state via our subspace fitting procedure is shown by the height of the 20 bars. Within each bar, the different colors represent how many of the assignments therein belonged to which ground-truth states (as seen in the legend), allowing an assessment of the true positive rate. On the right, the final 2D occupancy map for the 400 states formed by CM_1 and CM_2 is shown.

To circumvent issues stemming from inclusion of CM subspaces with poor geometric structure due to PD disparity, while all images are used for subsequent 3D reconstructions, only those occupancy assignments for CM subspaces above an $\mathcal{R}^2$ threshold value (0.7) were integrated during this analysis. In Movie H6, the occupancy map without $\mathcal{R}^2$ thresholding is shown, along with the corresponding final 3D density maps for an example trajectory (3D movie) from the

compiled 2D state space. As can be seen, the ESPER method upholds the spatial relationships in the ground-truth CMs with striking accuracy.

We additionally validated our results by calculating the Fourier shell correlation (FSC; Frank, 2006) between 3D density maps recovered by the ESPER method and their ground-truth counterparts (Figure H3, as generated via Warshamanage and Murshudov, 2019). In general, the FSC curve provides a global measure of how well one 3D density map matches another, for which we found a good agreement of all states up to a resolution near 3 Å; i.e., the ground-truth value. We also calculated Q-scores (Pintilie et al., 2020) as a local quantitative validation of the structural fidelity of the ESPER outputs. Using this approach on the ground-truth atomic-coordinate structures and their corresponding ESPER-recovered 3D density maps, we found highly favorable agreement across all residues in each state. On average, the Q-scores obtained were approximately 1.3 times that of the expected value (i.e., the average Q-score at a resolution of 3 Å) as calculated based on a data bank of reported resolutions of 3D density maps (Figure H4).

A comparison of the outputs of ESPER and NLSA for three example PDs from data-type IV are provided in Movie H7, with these PDs selected based on both the visual appearance of their images and their embedded geometries. It should be noted that the same preliminary steps were performed for both methods (i.e., steps 1 and 2 in Figure 39) before the branch in the workflow. During this branch, recall that ESPER includes the use of a unique 2D subspace per CM, while avoiding parabolic harmonics and applying eigenfunction realignments, ultimately resulting in 2D and 3D movies that retain the raw cryo-EM images. In contrast, the NLSA approach operates on only one of the initial DM eigenvectors per CM, and performs no steps for avoiding eigenvectors or realigning subspaces. In the process, the raw cryo-EM images are unavoidably discarded during

the NLSA procedure, ultimately resulting in final 2D and 3D movies formed from interpolated NLSA images.

Immediately apparent for all three PDs in Movie H7 is the difference in quality of the Hsp90 domains under motion corresponding to the given CM. For ESPER, these domains are highly resolved across all frames produced, while for NLSA these regions are much less resolved and noticeably smeared out. Second, while the visual differences between frames of the ESPER movies appear to evolve at an even pace, differences in frames appear less emphasized near the beginning and end of the NLSA movies, as if the movie was decelerating near these regions. In addition, the NLSA occupancies share little resemblance to our ground truth, with errors accentuated near the boundaries. Similar boundary problems do exist but are significantly less pronounced in the ESPER occupancy maps, with each map showing reasonable agreement with ground truth (i.e., bimodal for CM_1 and unimodal for CM_2).

Differences in outputs due to methodology are most pronounced for the example PD_{33}, which is a representative from the class of embeddings with appreciably unaligned eigenfunctions from the ideal eigenbasis, with the subspace of CM_2 here requiring a larger counter-rotation than CM_1. As can be seen, the overall range of motion for CM_1 is noticeably reduced compared to outputs from ESPER. For CM_2, matters are much worse. While our procedure using ESPER correctly charted a rotated, properly-aligned set of eigenfunctions, ManifoldEM employing NLSA used the existing embedding without accounting for realignment. As a result, the 2D movie produced by NLSA having closest resemblance to CM_2 (i.e., Ψ_4) demonstrated a physically-impossible sequence of motions: the splitting of the CM_2 domain into two separate domains. At the end of Movie H7, the 2D NLSA movies obtained for the leading four eigenvectors are shown for comparison. Here, both (i) a physically-impossible splitting of the CM_1 domain, and (ii) a

subdued CM_2 motion can be seen in the 2D movies obtained for both Ψ_2 and Ψ_3. A more detailed account of this comparison is also available (Seitz et al., 2021a).

In summary, while NLSA and ESPER have been provided the exact same data—even up to generation of identical manifold embeddings—only our method is able to fully leverage the geometric structure present to consistently recapitulate ground-truth CMs and occupancies from a variety of PD manifolds. Further, while the ESPER method offers strategies to procedurally avoid introduction of nonsensical contextual output, NLSA can generate 2D movies with a wide range of defects, with each having the potential of appearing as a likely CM candidate to the naïve eye. It is also important to examine the total computation time for performing these two techniques on the same CM-eigenvector (Ψ_1) from PD_2, with final output a single 2D movie (as seen in Movie H7). While the application of the ESPER method to retrieve a 2D movie required approximately three minutes, the total computation time for NLSA for this same endeavor was over 90 times longer, with both methods having been run using a single-processor on the same workstation (3.8 GHz 8-Core Intel Core i7; 8 GB 2667 MHz DDR4). In the current release of the ManifoldEM framework (Mashayekhi, 2020; Seitz et al., 2021b), it is required that this time-expensive NLSA computation is then repeated in its entirety for every Ω_{PD} eigenvector chosen during final computation of the free-energy landscape. Meanwhile, by applying our intersection of image-indices approach, as afforded by retainment of the raw cryo-EM images, the ESPER method compiles CM content for all PDs and generates the free-energy landscape within minutes. All in all, our method has the potential to push the total computation time for a typical data set of approximately 500,000 images down from weeks or months to only a few days.

These high computational demands were rationalized for NLSA as a way to handle the unknown manifold structures (Dashti et al., 2014). In contrast, our heuristic analysis directly

informs us of anticipated characteristics of the spectral geometry, enabling us to circumvent these previous unknowns, and perform the necessary operations required to accurately retrieve high-resolution images and a corresponding occupancy map for all CM states. Based on this knowledge, the ESPER method is able to produce appreciably more accurate outputs than the previous technique in a fraction of the time.

As a general note, it is important to keep in mind that results from synthetic data will always be superior to experimentally-obtained data, since even the most sophisticated simulations will be unable to capture all complexities existent in an experimental system. These complexities can be considered as introducing higher-order terms in our parameter space, which has been designed to emulate all lower-order terms up to a threshold deemed satisfactory. Any limitations or uncertainties that do emerge using synthetic data should be anticipated to arise in real-world data, and potentially in exacerbated form. To gain a better understanding of these influences, in the next section, the results obtained by ESPER with two experimentally-obtained data sets are explored.

11.6 Analysis of ESPER Results with Experimental Data

To assess the performance of ESPER—and the capacity of our heuristic knowledge—on real, experimentally-obtained data, we deployed our method on two data sets: the 80S ribosome (Dashti et al., 2014) and the ryanodine receptor type 1 (RyR1, ligand-free; Zalk et al., 2015). This work[12] was conducted in 2021 after our latest bioRxiv release (Seitz et al., 2021a), and is currently unpublished material that has been submitted for review. Recall that both of these were previously studied using ManifoldEM with NLSA. As these data sets are used primarily to compare method

[12] Contributions are as follows. E. Seitz: Conceptualization; formal analysis; methodology; software; validation; manuscript draft, review and editing. F. Acosta-Reyes: Data preprocessing and manuscript review. S. Maji: Data preprocessing and manuscript review. P. Schwander: Manuscript review and editing. J. Frank: Direction and project administration; manuscript review and editing.

outputs between ESPER and NLSA, minimal conclusions will be supplied pertaining to the deeper biological context of results.

Motivated by our analysis of synthetic data, we first searched through these data sets for Ω_{PD} embeddings with distinct geometric features matching those encountered during our ground-truth studies. This search was enabled by the interactive tools in the ManifoldEM Python GUI (Seitz et al., 2021b), which provided a flexible means to view the distribution of images and occupancy of each PD as the angular width of each PD was uniformly altered. For different PD angular widths, up to 10° on S^2, we then embedded a set of highest-occupancy PDs and analyzed the results. Overall, the structure of all Ω_{PD} embeddings observed across these data sets fell into three broad categories, with leading subspaces exhibiting either a (i) *robust*; (ii) *marginal*; or (iii) featureless, *globular* geometric form.

For the majority of manifolds analyzed, embeddings with globular form were the most frequently encountered, followed by marginal, then robust. We found that the presence of each could be reliably predicted based on two parameters: the angular width and occupancy of the respective PD. Specifically, for PDs with small angular widths (approximately 3°) and a relatively high occupancy (typically greater than 1000 images), robust parabolic features emerged in the corresponding embedded subspaces.

Of the two experimental data sets, only the 80S ribosome was able to meet this criterion (and consistently for numerous PDs), which was statistically favored given the sheer number of images available: nearly 850,000 total. For the approximately 350,000 RyR1 images, PDs formed with 3° angular widths typically contained less than 400 images each, resulting in globular-shaped embeddings as expected from the results in Figure 36. In the case of the RyR1 data set, as the angular width was increased to include a sufficient number of images in each PD, embeddings

with marginal parabolic features emerged. Even still, these were intermixed among other PD embeddings exhibiting no apparent geometric features, which we assume is indicative of compounding factors, including: alignment mis-estimations (e.g., recall Figure 21), aberrant particles, and dispersed arrangement of angular assignments in a given PD. Given this initial assessment, we next provide the outputs of ESPER on example subspaces exhibiting each of the three properties observed.

80S Ribosome. The results of applying subspace partitioning (Algorithm H3) via ESPER on an embedded subspace with robust parabolic features is shown in Figure 43. Since robust parabolic features were present only in the leading subspace $\{\Psi_1 \times \Psi_2\}$, and given appropriate use of the minimum cutoff $\mathcal{R}^2_{min}$ for the coefficient of determination in Algorithm H1, eigenfunction realignment was not applied. To note, after defining CM_1 at $\{\Psi_1 \times \Psi_2\}$, Algorithm H1 additionally defines all eigenvectors Ψ_j in composite with Ψ_2 (i.e., $\{\Psi_2 \times \Psi_j\}$) as CM_1 harmonics. The 2D movie results for the first four eigenvectors are provided in Movie H8, with outputs using ESPER compared directly with those obtained from the same Ω_{PD} embedding using NLSA. At the end of Movie H8, a schematic of the 80S ribosome is shown as viewed from this PD, with its most important domains labelled.

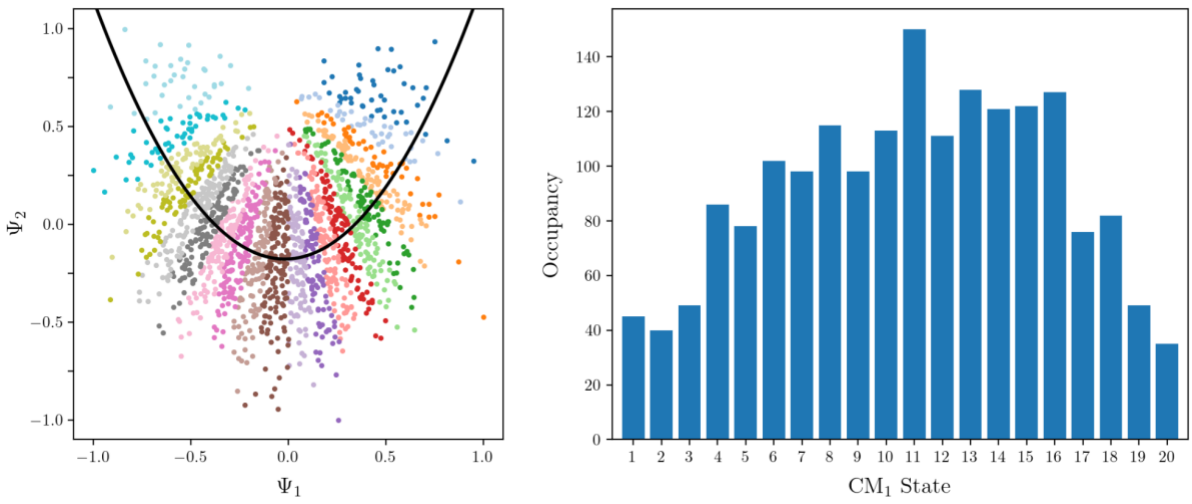


Figure 43: Results of applying ESPER on the experimental ribosomal data set. The embedding was performed on a PD formed with a 2° angular width containing a total of 1825 images, forming a robust parabolic point cloud in the embedded space.

The 2D movies produced by NLSA align well with the findings described in the original analysis (Dashti et al., 2014), which were obtained from a suitable great circle in that analysis, where "typical" embeddings showed a parabolic form. The 2D NLSA movie we obtained corresponding to Ψ_1 appears to exhibit the previously-described CM_1 as seen from the current PD, including a ratchet-like intersubunit motion, a closing of the L1 stalk towards the intersubunit space, and a rotation of small subunit head along its long axis (Dashti et al., 2014). The 2D movie corresponding to Ψ_2 likewise appears to describe CM_2: a so-called nodding motion of the head which is needed for selection of tRNA during decoding (Dashti et al., 2014). However, this motion is not isolated, and is also accompanied by similar—yet more subtle—domain motions as seen in the Ψ_1 NLSA movie. The third NLSA movie (Ψ_3) exhibits a collection of subtle domain motions as seen in the first two NLSA movies, while in the fourth NLSA movie (Ψ_4), there appears to be a previously-undescribed shift in intensity within the intersubunit space. Through NLSA, the original work (Dashti et al., 2014) defines the $\{\Psi_1 \times \Psi_2\}$ subspace as the basis for constructing the 2D energy landscape, corresponding to the motions observed in both the Ψ_1 and Ψ_2 NLSA movies.

As seen in Movie H8, the ESPER method provides nearly identical outputs as achieved by NLSA for Ψ_2, Ψ_3 and Ψ_4. (To note, we had to force the calculation of Ψ_2 in the ESPER workflow, as it was initially removed as a harmonic in Algorithm H1). A striking difference appears, however, in the sequence of states describing the leading CM, and specifically as it pertains to the trajectory of the head subunit. Instead of a simple left-to-right head motion as seen via NLSA, the motion charted by our method shares a likeness to a combination of the motions observed in both the first and second NLSA movies. Specifically, in the ESPER movie, the head subunit is first seen moving up and to the left, followed by a downwards motion (nodding) into the intersubunit space. Meanwhile, all other domains charted for CM_1 by the ESPER method move in a consistent fashion to those seen in the Ψ_1 NLSA movie.

As understood by our approach, and in contrast to the analysis performed using NLSA, the $\{\Psi_1 \times \Psi_2\}$ subspace is not a 2D state space $\{\mathrm{CM}_1, \mathrm{CM}_2\}$ laid out conspicuously like a parabola, but actually a parabolic point-cloud $\{\psi_1 \times \psi_2\}$, where both ψ_1 and ψ_2 represent the same degree of freedom (CM_1). This noticeable difference arises strictly due to the different treatments of subspace geometry by the two methods. While NLSA projects onto one eigenvector before organizing images into supervectors to form interpolated NLSA movies, the ESPER method carefully charts the geometric structure of the 2D subspace, and creates each frame of the respective movie by selectively averaging subsets of the available cryo-EM images.

Ryanodine receptor (RyR1). We next describe the results of ESPER on embeddings with marginal and globular geometric structure. In Movie H9, we show the 2D movies obtained from an RyR1 PD containing 976 images within an angular diameter of 6°. Here, the $\{\Psi_1 \times \Psi_3\}$ point cloud corresponding to CM_1 has a significantly more marginal parabolic appearance compared to those observed for the ribosome, while the best-fit point cloud for CM_2 at $\{\Psi_2 \times \Psi_4\}$ appeared

devoid of parabolic features altogether. For this embedding, the conformational motions output by using ESPER for each 2D subspace were highly similar—albeit noisier—to those output by NLSA for each eigenvector individually. In either case, the leading CM corresponds to the entire assembly of cytoplasmic shell, activation core and pore (appearing like an opening of the central channel pore), while CM_2 corresponds to movement of the cytoplasmic shell resulting in an apparent lowering of the handle and clamp domains.

Overall, these results are comparable to those described in the original study (Dashti et al., 2020), with no major deviations observed, other than noise, between the performance of NLSA and ESPER. For this latter discrepancy, we found that by noise-filtering each 2D movie produced by ESPER using singular value decomposition (as denoted in Movie H9 with the initials "SVD"), we were able to very closely reproduce the appearance of the NLSA outputs. This likeness increases with the occupancy of the PD, which corresponds to the presence of more pronounced signal in the initial 2D movie used for decomposition during SVD. As a loose estimate, below 900 images we begin to see a significant drop in the consistency of these SVD outputs.

11.7 ESPER Limitations and Uncertainties

Despite the remarkable performance of the ESPER method on our 126-PD data set with experimentally-relevant SNR and CTF, we believe there is still room for improvements for each of its techniques. In the following section, several challenges and potential improvements to the ESPER method will be addressed.

Eigenfunction Determination. At the beginning of the ESPER workflow, an ability to accurately detect and fit CM parabolas encountered in both favorable and unfavorable conditions is imperative to the final accuracy of ESPER outputs. While we found that the general conic least-

squares fit sufficiently handled robust geometric features (including inward curling due to CTF) encountered in embeddings from synthetic data, its performance was much less predictable—and required supervision—when applied on less-structured embeddings. In the current repository (Seitz, 2021), alternative fitting strategies are also provided, such as a standard parabolic least-squares fit and absolute value function. However, this choice currently requires user intervention, with each of the fitting algorithms likewise varying in performance based on data quality. As a remedy, more robust procedures should be employed to identify and fit both highly-structured and less-structured regimes, such as a constrained least squares method (Harker et al., 2004) or a generalized Hough transform (Duda et al., 1972). Ideally, for a given subspace, the choice of one fitting strategy over all others should be fully automated.

Furthermore, the presence of complex physical constraints—which change the boundary shape of the manifold's domain—can alter the required rotation operators currently discovered by Algorithm H1. Importantly, our requirement of adequate coverage (step 1 in Figure 39) excludes the case of obtaining poorly sampled data from a rectangular domain, which would allow any number of arbitrary shapes to emerge. This exclusion also holds for state spaces with "holes" (i.e., interior boundaries; Grebenkov and Nguyen, 2013), where the occurrence of certain states is forbidden due to energetic restraints. Instead, as described in Section 10.5, of most significance are those constraints due to steric hindrance between moving domains, resulting in manifolds with arbitrarily-shaped domains granting unknown eigenfunction form.

To better understand the effects of these boundary challenges within the scope of the ESPER method, we have created a 2D state space with an octagonal domain (noting the Laplacian eigenfunctions of the octagon is an analytically unsolved problem), which was achieved by eliminating states at the four corners of our standard rectangular domain (Figure H5-A). To

circumvent the occurrence of eigenfunction misalignments due to PD disparity—which may complicate the interpretation of the boundary influence—we opted to embed the 3D electron density maps instead of 2D projections for this analysis. The corresponding manifold embedding obtained from this octagonal state space is shown in Figure H5-C, which features a number of deviations from the canonical rectangular eigenbasis (Figure 28).

Manually, we attempt to find a transformation from the octagonally-derived eigenbasis (Figure H5-C) to the rectangular form (Figure 28) by intuiting a collection of suitable rotation operators. Indeed, we are able to show that such a transformation is possible, up to some level of uncertainty (Figure H5-D). Thus, it is not that the eigenfunctions are dramatically changed by the imposed boundaries, but instead that the eigenvectors can now contain multiple cosine terms. We note that both the indices and number of rotation operators required for this transformation deviated from our PD-manifold findings on eigenfunction realignment performed on rectangular state spaces, creating a more complex tree of decisions. We believe this instance motivates the need for a future comprehensive method for estimating the preferred eigenbasis rotations, with ESPER outputs possibly refined by a maximum-likelihood approach or by machine learning using a convoluted neural network.

Eigenfunction Realignment. For noisier, less-structured embeddings, the 2D histogram method can provide suboptimal counter-rotations. To supplement this method, it may prove beneficial to gather other metrics alongside 2D histogram values for each θ in Algorithm H2, such as an $\mathcal{R}^2$ value for each instance that the subspace is rotated. As well, while it is true that finite rotations about different axes do not commute (Townsend, 2000, chap. 3), we found only minor deviations in the final orientation of each eigenbasis, given different permutations in the sequence of the four operators obtained for two degrees of freedom. Future studies should explore more

degrees of freedom, and alternative boundary conditions, to ensure that this simplification is performed within reason.

Subspace Partitioning. Overall, the area-based procedures employed—including the balltree and alpha shape methods—are strongly affected by large changes in parameters, and thus may require initial supervision on externally-obtained data sets. For the purposes of our study, we investigated over 500 subspaces, with each of our parameters broadly tuned for robust, high-quality performance on the SS_2 subspaces from data-type II, III and IV. Detailed notes on these procedures have been provided in our published repository (Seitz, 2021), which includes comments describing less-significant decisions not explicitly stated here. In general, even in ideal settings, 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. As well, while it is very easy to split the CM subspace crudely into only tow or a handful of states, as the number of states requested δ is increased, the degree of sophistication in mapping and segmenting each point cloud must also increase in turn.

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. Ultimately, 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 virtually identical for means of our final outputs. Naturally, the use of this optimal value should maximize the amount of information ascertainable in our system. In theory, we desire a minimum number of snapshots in the lowest-occupancy bin, such that every frame of the subsequent movie has significant content; e.g., as possibly defined via both the average SNR of each image and the number of images in the lowest-occupancy bin.

Additionally, as stated in our analysis of RyR1, noise-filtered 2D movies can be generated by ESPER using singular value decomposition, for use strictly by belief propagation. In noisier circumstances, SVD results can be enhanced by decreasing δ so as to boost signal; it is also worth investigating alternative methods for these means.

With the use of synthetic data, we also show that final occupancy assignments can have slight inaccuracies, which are most emphasized near the boundaries of each CM. Although not pursued here, since our method retains the raw cryo-EM images, these misassignments could be further corrected to improve 3D density maps and corresponding occupancy distributions. Specifically, an optimization approach could be designed to compare images within each bin and reassign erroneously-assigned snapshots into neighboring bins in which they most likely belong. To note, a maximum-likelihood approach does already exist that aims to extract such granular conformational heterogeneity (Giraldo-Barreto et al., 2021), as does a method based on neural networks (Zhong et al., 2021).

Chapter 12: ManifoldEM with NLSA and ESPER

12.1 Motivation

Given our previous analysis of results from synthetic and experimental data sets, it is clear that, for embeddings of sufficient quality, the ESPER method offers many benefits over the founding ManifoldEM approach using NLSA. Of most importance, ESPER outputs are significantly more accurate in this regime. However, in the regime of manifold embeddings completely lacking discernible form, only the NLSA route should be used, while still potentially incurring its known limitations and uncertainties (Chapter 7). Further complicating matters, we anticipate that these two regimes will typically coexist within the same experimentally-obtained data set. Thus, it is clear that the availability of both ESPER and NLSA within the overarching ManifoldEM framework is essential, with the application of one over the other demarcated by an assessment of geometric conditions in each PD embedding. In this chapter, such a strategy is laid out, along with considerations for future work.

12.2 Overview of Approach

By applying the ManifoldEM framework on our synthetic data, we have demonstrated that serious problems can emerge in the analysis, including presence of physically-impossible, stunted, or hybrid conformational motions, as well as erroneous occupancy maps. These issues mainly arise during one critical step, where the geometry of each embedding must be correctly charted to render a set of CMs and corresponding occupancies. This task is originally performed in most part by the application of NLSA, where each eigenvector of the DM embedding is treated as an independent coordinate for a conformational change. By concatenating the set of cryo-EM images along a given

eigenvector, interpolated images are produced via NLSA, and re-embedded to form a new space from which a 2D NLSA movie is extracted representing the deduced CM.

However, our heuristic analysis shows that the observed problems can arise when each eigenvector is treated as an independent source for a CM, while in actuality, a single eigenvector can correspond to some combination of CM eigenfunctions, as well as to eigenfunction harmonics. We additionally demonstrate how each CM is better mapped by a parabolic trajectory in a 2D subspace defined by two eigenvectors (corrected for misalignments), for which the projection of that trajectory onto a single (uncorrected) eigenvector is naturally problematic. Our analysis found that it was essential to correct for these properties in order to accurately map each CM. Depending on the PD and data characteristics, these issues can combine to create several systematic errors, limitations and uncertainties that were previously pointed out in Chapter 7.

We have developed the ESPER method as a means to circumvent these problems and refine the existing framework. The operations introduced in this study offer several enhancements, including our procedure for isolating CM subspaces, removing CM harmonics, correcting for eigenfunction misalignments, and directly retrieving each CM from the raw cryo-EM snapshots as arranged within the initial (corrected) embedding. In the last case, the use of the raw images is shown to improve both the accuracy of occupancy determination and final resolution of 3D structures, while providing a vastly simplified workflow for determining multidimensional free-energy landscapes. We have further shown that our implementation of these enhancements drastically decreases the overall computation cost.

Our findings on both synthetic and experimental data sets establish a minimum requirement for PD-manifold studies of conformational continuum. Specifically, we have found that for maximal fidelity of final outputs with ground truth, a data set must contain well-structured

geometry when embedded. The performance of ESPER hinges on the presence of such geometric information, and as the quality of the embedded geometry increases, more of our method's benefits become available. As seen in our two examples, the ability to sensibly avoid harmonics or apply eigenvector rotations is only applicable up to the number of CMs present having pronounced geometric structure that is viable for reliable parabolic fits. If no geometric form can be deciphered in a PD embedding, it is effectively impossible to solve for these unknowns.

This limitation presents a problem when dealing with typical experimental data sets, where it is most realistic to anticipate the presence of only a subset of PD embeddings which have adequate geometric information as required by the ESPER approach. Even given just one such high-quality PD, our analysis shows that the ESPER method is able to provide essential information on the molecular machine's conformational spectrum. If such a PD-embedding was both highly-populated and available from a viewing angle where all CMs were well-visualized, all information pertaining to the machine's total number and approximate types of degrees of freedom—as well as corresponding occupancies—could be accurately calculated from the 2D images alone.

We next expand this idea to the more likely presence of not just one, but a subpopulation of such informative PDs, with the ESPER approach individually applied on each. To reconstruct adequate 3D density maps, alternative methods would then need to be devised to effectively fill in conformational information for the remaining PDs lacking geometric form. Indeed, the reliability of our approach decreases rapidly when approaching the regime of globular embeddings, since there is no geometric information to leverage. For these remaining PDs, the NLSA approach is better suited, since it can at least retrieve reasonable 2D movies from low-occupancy embeddings. However, since the apparent absence of geometric features in these embeddings does not discount

their latent presence and potential impact, NLSA outputs may still incur the known limitations and uncertainties. To mitigate these unavoidable issues, we recommend an altered use of NLSA, which is directly informed by the conformational spectra obtained by the ESPER method in the highest-quality PDs. Under such a scheme, the ESPER outputs would serve as a high-quality template on which NLSA outputs can be mapped.

In this tradeoff, there exists some gray area where it is difficult to make out which technique should take precedence. Certainly, low-occupancy globular embeddings should be handled by NLSA, and although the ESPER outputs on high-occupancy globular embeddings are similarly convincing and highly consistent with NLSA, it is our belief that the decision to run ESPER over NLSA on a given PD should be governed by a sensible coefficient of determination threshold $\mathcal{R}^2_{\min}$. Specifically, for each Ψ_i, the fit score $\mathcal{R}^2$ corresponding to the 2D subspace with the highest fit score among all other $\{\Psi_i, \Psi_j\}$ subspaces should exceed the value of $\mathcal{R}^2_{\min}$. For embeddings with fit scores above this threshold, the ESPER method can leverage a number of benefits over NLSA, with this number increasing as the quality of the geometry improves. Since the appearance of robust geometry is also dependent on high PD occupancy (Figure 36), and high PD occupancy boosts signal, the ESPER method is additionally qualified to furnish high-quality SVD movies in this regime while retaining the raw cryo-EM images.

As we have shown for RyR1, these noise-filtered 2D movies have a quality almost identical to the respective NLSA outputs, and serve the single purpose for use by belief propagation across S^2 during CM assignments. In noisier circumstances, SVD results can be enhanced by binning the movie frames to boost signal. Meanwhile, the raw cryo-EM images are retained for use during 3D reconstruction, and, aside from improving fidelity of those outputs, allow use of our efficient approach using intersection of image-indices in generating occupancy maps and energy

landscapes. Since our proposed strategy leaves the PD-embeddings without discernable geometry to be analyzed using the preexisting NLSA approach, final outputs would next need to be combined between these two methods. Notably, for each PD analyzed by either ESPER or NLSA, the corresponding free-energy landscape and projections (i.e., raw cryo-EM images or NLSA images, respectively) must be combined to form a consolidated free-energy landscape and corresponding 3D density maps. If necessary, SVD could then be applied on the final sequence of reconstructed 3D density maps, as has previously been done in ManifoldEM (Mashayekhi, 2020; Seitz et al., 2021b). We anticipate that the next public distribution of the ManifoldEM Python suite (Seitz et al., 2021b) will include these advancements as a refinement to its workflow.

11.3 Future Advancements

As it stands, the current Python release of ManifoldEM (Seitz et al., 2021b) provides users with only the 1D implementation of the founding ManifoldEM approach using NLSA. As previously described, while many of the most important frontend features are already coded for 2D use (Section 2.5), it is a work in progress to update the backend Python code to match the Matlab repository (Mashayekhi, 2020). Upon doing so, the ManifoldEM Python suite will be superior for researchers in all regards. To further advance this frontier, the ESPER methodology must also be incorporated along with NLSA 1D and 2D functionality.

In Section 11.2, an automated strategy for deciding between use of NLSA and ESPER on each Ω_{PD} embedding was described, with the decision based on the fit scores ($\mathcal{R}^2$) of corresponding subspace geometries. As these $\mathcal{R}^2$ values are currently calculated based on fitting methods designed specifically for well-defined (robust) embeddings, a recommendation was made in Section 11.7 to enhance their performance for embeddings with suboptimal (marginal) features.

As well, several other enhancements were described there—specifically for the ESPER method—recommending, for example, incorporation of more robust automation schemes for defining essential rotation operators $R_{i,j}(\theta)$ and final occupancy assignments. The ESPER outputs obtained from these informed decisions on the highest-quality PDs should next be used as a template on which NLSA outputs can be mapped. Finally, Section 11.2 described an approach for consolidating the outputs obtained exclusively by either ESPER or NLSA from each Ω_{PD} into a final free-energy landscape and corresponding set of 3D density maps.

Here, several ulterior enhancements to the ManifoldEM Python suite will be described that serve to further bolster the use of the ESPER method alongside NLSA. The first of these recommendations involve one of the first steps in the original ManifoldEM framework: tessellating S^2 into a collection of PDs (i.e., "Division into PDs" in Section 2.3). Currently, this automated approach defines the angular width of each PD uniformly based on several input values, and is inadequate for several reasons. First, and most importantly, it is impractical. In each of the studies shown here, this automated value was consistently changed to explore better results, with the initial value seldom—if ever—the best choice available. In the ManifoldEM Python user manual (Seitz et al., 2021b), we go so far to recommend a comprehensive method for users to alter this parameter themselves based on the resulting distribution of images across all PDs. It could even be argued that this expression incorporates arbitrary values, such as the resolution of a 3D refinement generated via an upstream method. As previously described (i.e., "Assessment of Reconstructed Volume" in Section 8.2), the resolution of a reconstructed volume has little relevance in the study of a heterogeneous data set.

Besides these issues, even if tuned "appropriately", there is a deeper problem with the uniform assignment of the angular width. In our analysis of the ribosomal data set, we found that

the quality of geometric features in an embedding was closely tied to the angular width and occupancy of the corresponding PD. However, these conditions must be met on a case-by-case basis, since satisfying them only for the PDs requiring the smallest angular width will, as a result of uniformity, define all other PD with suboptimal boundary and image counts. Such a globally-restrictive assignment only subverts the intentions underlying the design of the ESPER method, which aims to operate on optimally-formed PDs, while leaving all other PDs to NLSA. Instead, this S^2 tessellation procedure needs to be a dynamic one, able to locally adjust the angular width to capture optimal particle density per PD. A schematic of the intended output of this procedure is provided in Figure 44-B.

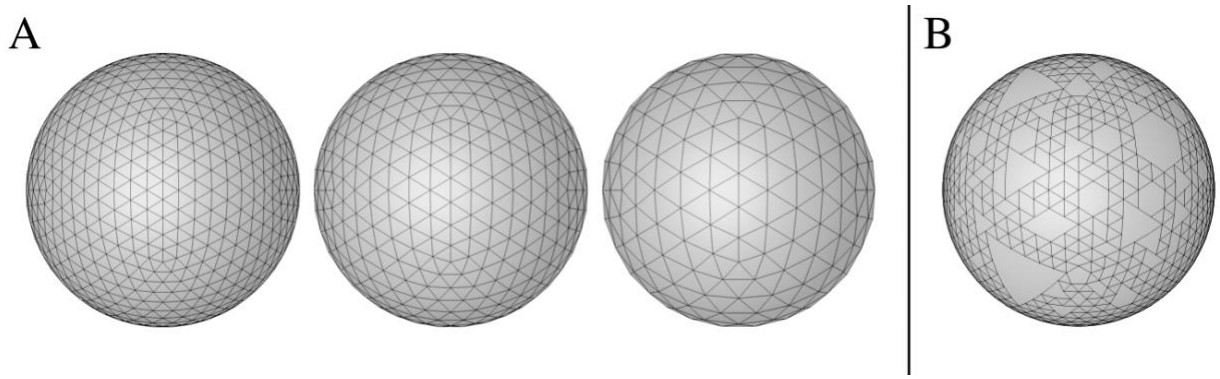


Figure 44: Schematic for a dynamic tessellation strategy for ManifoldEM with ESPER. In [A], three icosahedrons generated with 45, 35 and 25 segments (from left to right) are displayed, each representing a standard tessellation of S^2 with a uniform angular width. The hybrid pattern in [B] is an example of a desired output from an (undeveloped) algorithm for dynamically assigning nonuniform PD boundaries as defined by the local density of images on S^2.

Briefly, the decision for the creation of these adaptive tessellation bins could be defined locally by the requirement of each bin to enclose a minimum number of images within its bounds. One potential scheme is a hierarchal scaling procedure (similar to POLARIS) that, starting with an initially coarse mosaic of repeated shapes, recursively subdivides each shape into finer sub-tessellations until this criterion is globally met. Given the importance of these decisions, the

development of such an algorithm is deserving of the scale delivered for the other ManifoldEM subproblem, belief propagation (Maji et al., 2020).

Next, a number of smaller "quality of life" improvements are suggested for future implementation within the ManifoldEM framework. First, the GUI must stay up to date with upstream refinement methods, and be programmed to allow inputs from the latest versions as they are continuously released, such as from RELION and cryoSPARC. (Currently, the GUI has been set up for use with outputs from RELION-3.1 and earlier). Following this logic, the ManifoldEM backend code should also be updated to incorporate increasingly-detailed parameters from these upstream assignments, such as use of elliptical defocus. During analysis of PD manifolds by ESPER, heightened user supervision must be granted, such as GUI windows for examining best least-squares fits and alpha-shape boundaries, which allow users to alter the corresponding parameters on the spot as needed. Similar instances could be designed for intermediate NLSA outputs, such as an ability to redefine the SVD eigenvectors chosen based on a user analysis of the chronos distributions. In either case, users should be able to view the distribution of images using polar coordinates within each PD (i.e., on the "Eigenvectors" tab) and prune outliers.

After compilation of ESPER and NLSA results, the POLARIS method should also be made available within the ManifoldEM backend (and not external to it), allowing for direct assessment of minimum-energy paths on the 2D free-energy landscape. Along these same lines, the GUI should be expanded for the possibility of multi-manifold mapping; such as for data sets incorporating ligands, mutations, or drugs (Figure 2). As well, the possibility of generating 3D free-energy landscapes should also be provided, which require additional backend code for NLSA, as well as a frontend changes, such as a Mayavi viewer for analyzing free-energy slices on the cube. Finally, the post-processing procedures—as currently run via a series of external scripts—

should be fully implemented into the GUI. A new tab therein could allow users to both clean (e.g., via Gaussian filters or SVD) and analyze their final ManifoldEM outputs. For the latter, users could view 3D animations of final volumes animating along different energy landscape trajectories, while rotating the camera fluidly in 3D space to view different domains during playback. As is often done, these post-processing features could likewise be incorporated within a future Chimera plugin, and take advantage of the extensive toolset which it already provides. A detailed list of other, less-essential recommendations is also available in the ManifoldEM Python user manual (Seitz et al., 2021b).

11.5 Closing Remarks

For a number of the ESPER-related enhancements to be made, testing can be done using the synthetic data set (i.e., data-type IV) already provided. However, for experimental use of the ManifoldEM framework with ESPER and NLSA, it is of utmost importance that testing is performed across several experimental data sets for further refinement of these algorithms for public release. Alongside the ribosomal data set previously described, an additional data should be obtained: having a sufficient number of particles and S^2 distribution to form a significant number of high-quality PD manifolds, alongside several readily-identifiable conformational motions.

Epilogue

In this work, a comprehensive analysis of the conformational continuum and corresponding free-energy landscapes of molecular machines has been performed, as primarily informed by manifold embedding of single-particle cryo-EM ensembles using ManifoldEM. The findings presented here on experimental data sets—including the ribosome, ryanodine receptor, vacuolar ATPase, and SARS-CoV-2 spike protein—provide a number of new advancements, such as discovery of novel conformational changes and a successful cross-validation with molecular dynamics simulations. As well, new strategies were introduced to analyze the free-energy landscapes obtained from these systems, including a minimum-energy pathfinding algorithm, POLARIS, as well as an application of molecular dynamics flexible fitting to rapidly estimate the atomic-coordinate structures for each state along a given minimum-energy path. Likewise, an advanced user interface—available to the public via a Python distribution—was also described, offering users a complex, multi-tab experience enabling intuitive and dynamic navigation across each module in the ManifoldEM framework.

During this exposition, for each data set, several critical ManifoldEM limitations and uncertainties were documented and categorized. The occurrence of these motivated an extensive heuristic analysis using simulated, controlled data sets of the Hsp90 protein undergoing several quasi-continuous conformational changes, with projections formed under experimentally-relevant noise conditions and microscopy aberrations. Through this heuristic analysis, a foundation was created for identifying the way sets of images originating from a varying atomic structure are represented in low-dimensional embeddings obtained by prominent dimensionality-reduction techniques, and further, how this spectral geometry can be navigated to recover the machine's ground-truth conformational continuum. These findings on synthetic noisy data sets provide a

number of new insights unaccounted for in the founding ManifoldEM framework, and emphasize the need for a refined workflow when analyzing the eigenvectors from embeddings of single-particle cryo-EM data sets of molecules exercising numerous degrees of freedom.

Based on this knowledge, a novel, unsupervised workflow—ESPER—was introduced that, for data sets of sufficient quality, offers several enhancements to substantially improve the ability of ManifoldEM to recover continuous conformations. These include essential eigenvector rotations to consistently align the spectral geometry of all embeddings, an informed subspace fitting procedure using specific combinations of eigenfunctions while excluding parabolic harmonics, and a novel method for direct retrieval of each conformational motion using the raw cryo-EM snapshots, as arranged within the initial embedding, to form high-quality 2D and 3D movies. In the last case, the use of the raw images improves both the accuracy of occupancy determination and final resolution of 3D structures, while providing a vastly simplified workflow for generation of multidimensional free-energy landscapes. Further, it was shown that corrections for high-dimensional rotations are essential, and can lead to serious systematic errors downstream when unaccounted for.

A strategy was next laid out for combining the ESPER method within the founding ManifoldEM framework using NLSA. Specifically, for each PD manifold, the decision to use either ESPER or NLSA exclusively must be made based on the quality of the respective spectral geometry. Under this strategy, for the highest-quality embeddings, outputs provided by ESPER are used as a template for informing NLSA decisions, with final outputs from both methods then recompiled to form a consolidated free-energy landscape and corresponding 3D movies. Alongside this re-envisioning of the ManifoldEM framework, the need for several important enhancements to the next public release were also described, such as the use of nonuniform—instead of

uniform—tessellations in angular space, as well as quality-of-life additions to the Python GUI to optimize user experience.

As a few final notes beyond ManifoldEM, first, it is my opinion that the synthetic data protocol introduced here—which has already seen use in two external publications—is a necessary benchmark in validating and assessing the limitations of other approaches involving the recovery of conformational continuum. An example of such an exposition was provided in this work for both RELION and cryo-SPARC maximum-likelihood methods, which assessed both the accuracy of the methods while providing intuition for outputs in the scope of ground-truth knowledge. Here, the analysis of the ManifoldEM methodology has been grueling and relentless, but vital for scientific advancement; and no less should be expected from other published techniques. Finally, I note that this document also described several insights and strategies for general analysis of heterogeneous data types beyond cryo-EM projection data, such as the indexing of ground-truth information to elucidate eigenfunction form, and an analysis of the changes to the induced metric achieved by conversion of ground-truth content through a series of unique data representations. Along the way, several challenges, fundamental limitations and uncertainties are documented that emerge in geometric machine learning of heterogeneous data, and when possible, how these challenges can be remedied. Ultimately, it is my hope that the insights gained during the course of this work will be useful not only in the cryo-EM field, but more broadly to other fields aimed at untangling complex systems exercising multiple degrees of freedom.

References

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 9th ed., 1972.

X. Agirrezabala, J. Lei, J. Brunelle, R. Ortiz-Meoz, R. Green and J. Frank. Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Mol Cell*, 32(2):190—97, 2009.

X. Agirrezabala, H. Liao, E. Schreiner, J. Fu, R. Ortiz-Meoz, K. Schulten, R. Green and J. Frank. Structural characterization of mRNA-tRNA translocation intermediates. *Proc Natl Acad Sci*, 109(16):6094—99, 2012.

M. Ali, S. Roe, C. Vaughan, P. Meyer, B. Panaretou, P. Piper, C. Prodromou and L. Pearl. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, 440(7087):1013—17, 2006.

C. Anklin. Spotlight on nuclear magnetic resonance: a timeless technique. *Spectroscopy Europe*, 30(4):1—4, 2018.

A. Arkhipov, P. Freddolino and K. Schulten. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, 14:1767—1777, 2006.

N. Baker, G. Hamilton, J. Wilkes, S. Hutchinson, M. Barrett and D. Horn. Vacuolar ATPase depletion affects mitochondrial ATPase function, kinetoplast dependency, and drug sensitivity in trypanosomes. *Proc Natl Acad Sci USA*, 112(29):9112—7, 2015.

W. Baxter, R. 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(2):126—32, 2009.

D. Benton, A. Wrobel, P. Xu, C. Roustan, S. Martin, P. Rosenthal, J. Skehel and S. Gamblin. Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. *Nature*, 588(7837):327—30, 2020.

H. Berman, K. Henrick and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, 10:980, 2003.

A. Botea, M. Muller and J. Schaeffer. Near optimal hierarchical path-finding. *Journal of Game Development*, 1:7—28, 2004.

P. Buser. *Geometry and Spectra of Compact Riemann Surfaces.* Springer, 1992. ISBN 0817634061.

L. Casalino, Z. Gaieb, J. Goldsmith, C. Hjorth, A. Dommer, A. Harbison, C. Fogarty, E. Barros, B. Taylor, J. McLellan, E. Fadda and R. Amaro. Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent Sci*, 6(10):1722—34, 2020.

J. Chan, S. Yuan, K. Kok, K. To, H. Chu, J. Yang, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395(10223):514—23, 2020.

R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc Natl Acad Sci USA*, 102(21):7426—31, 2005.

R. Coifman and S. Lafon. Diffusion maps. *Appl Comput Harmon A*, 21(1):5—30, 2006.

R. Coifman, Y. Shkolnisky, F. Sigworth and A. Singer. Graph Laplacian tomography from unknown random projections. *IEEE Trans Image Process*, 17(10):1891—1899, 2008.

M. Craioveanu, M. Peta and T. Rassias. *Old and New Aspects in Spectral Geometry*. Springer Science, 2001. ISBN 978-94-017-2475-3.

E. Cruz-Chú, A. Hosseinizadeh, G. Mashayekhi, R. Fung, A. Ourmazd and P. Schwander. Selecting XFEL single-particle snapshots by geometric machine learning. *Struct Dyn*, 8(1):014701, 2021.

H. Cundy and A. Rollet. *Mathematical Models*. Tarquin Publications, 3rd ed., 1989.

A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseinizadeh, H. Liao, J. Pallesen, G. Sharma, V. Stupina, A. Simon, J. Dinman, J. Frank and A. Ourmazd. Trajectories of the ribosome as a brownian nanomachine. *Proc Natl Acad Sci*, 111(49):17492—97, 2014.

A. Dashti, G. Mashayekhi, M. Shekhar, D. 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(1):4734, 2020.

E. Dijkstra. *Communication with an Automatic Computer*. Doctoral thesis, 1959.

R. Duda and P. Hart. Use of the Hough Transformation to Detect Lines and Curves in Pictures. *Comm ACM*, 15(1):11—5, 1972.

J. Durrant and J. McCammon. Molecular dynamics simulations and drug discovery. *BMC Biol*, 9: 71, 2001.

H. Edelsbrunner, D. Kirkpatrick and R. Seidel. One the shape of a set of points in the plane. *IEEE T Inform Theory*, 29(4):551—9, 1983.

D. Ermolenko, Z. Majumdar, R. Hickerson, P. Spiegel, R. Clegg and H. Noller. Observation of intersubunit movement of the ribosome in solution using FRET. *J Mol Biol*, 370(3):530—40, 2007.

A. Ferguson, A. Panagiotopoulos, P. Debenedetti and I.G. Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc Natl Acad Sci USA*, 107(31):13597—602, 2010.

R. Feynman, R. Leighton and M. Sands. *The Feynman Lectures on Physics*. Basic Books, New York, NY, 2010 (originally published: 1965).

J. Fiaux, E. Bertelsen, A. Horwich and K. Wuthrich. NMR analysis of a 900K GroEL-GroES complex. *Nature*, 418:207—11, 2002.

N. Fischer, A. Konevega, W. Wintermeye, M. Rodnina and H. Stark. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature*, 466(7304):329—33, 2010.

J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules In Their Native State*. Oxford University Press, Oxford, New York, 2006.

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

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.

J. Frank. Generalized single-particle cryo-EM: a historical perspective. *Microscopy (Oxf)*, 65(1):3—8, 2016.

J. Frank. *Nobel Lecture: Single-Particle Reconstruction—Story in a Sample*. 2017.

J. Frank. New opportunities created by single-particle cryo-EM: the mapping of conformational space. *ACS: Biochemistry*, 57(6):888, 2018.

J. Gardiner, J. Behnsen and C. Brassey. Alpha shapes: Determining 3D shape complexity across morphologically diverse structures. *BMC Evol Biol*, 18(1):184, 2018.

A. des Georges, O. Clarke, R. Zalk, Q. Yuan, K. Condon, R. Grassucci, W. Hendrickson, A. Marks, J. Frank. Structural basis for gating and activation of RyR1. *Cell*, 167(1):145—57, 2016.

D. Giannakis and A. Majda. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc Natl Acad Sci*, 109(7):2222—27, 2012a.

D. Giannakis, P. Schwander and A. Ourmazd. The symmetries of image formation by scattering. I. Theoretical framework. *Opt Express*, 20(12):12799—826, 2012b.

P. 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(1066):89—102, 1972.

J. Giraldo-Barreto, S. Ortiz, E. Thiede, K. Palacio-Rodriguez, B. Carpenter, A. Barnett and P. Cossio. A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci Rep*. 11:13657, 2021.

D. Goodsell. Molecule of the month: Vacuolar ATPase. PDB-101, 2018. doi:10.2210/rcsb_pdb/mom_2018_3

P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D—Nonlinear Phenomena*, 9(1):189—208, 1983.

N. Grigorieff. Frealign: An exploratory tool for single particle cryo-EM. *Methods Enzymol*. 579:191—226, 2016.

H. Guo, G. Courbon, S. Bueler, J. Mai, J. Liu and J. Rubinstein. Structure of mycobacterial ATP synthase bound to the tuberculosis drug bedaquiline. *Nature*, 589:143—7, 2021.

H. Gupta, T. Phan and J. Yoo. *Multi-CryoGAN: Reconstruction of Continuous Conformations in Cryo-EM Using Generative Adversarial Networks*. Computer Vision — ECCV 2020 Workshops, LNCS. Springer, 12535: 429—44, 2020.

M. Harker, P. O'Leary and P. Zsombor-Murray. Direct and Specific Fitting of Conics to Scattered Data. *Proc BMVA*, 1—10, 2004.

A. Hosseinizadeh, G. Mashayekhi, J. Copperman, P. Schwander, A. Dashti, R. Sepehr, R. Fung, M. Schmidt, C. Yoon, B. Hogue, G. Williams, A. Aquila and A. Ourmazd. Conformational landscape of a virus by single-particle X-ray scattering. *Nature Methods*, 14: 877—81, 2017.

A. Hospital, J. Goni, M. Orozco and J. Gelpi. Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem*, 8: 37—7, 2015.

G. Huber and S. Kim. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys J*, 70(1):97—110, 1996.

W. Humphrey, A. Dalke and K. Schulten. VMD — visual molecular dynamics. *J Molec Graphics*, 14:33—38, 1996.

M. Jaskolki, Z. Dauter and A. Wlodawer. A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *Febs J, Special Issue: Celebrating the International Year of Crystallography*, 281(18):3985—4009, 2014.

J. Jost. *Geometry and Physics*. Springer-Verlag Berlin Heidelberg, 2009. ISBN 978-3-642-00541-1. doi:10.1007/978-3-642-00541-1.

N. Kolimi, A. Pabbathi, N. Saikia, F. Ding, H. Sanabria and J. Alper. Out-of-equilibrium biophysical chemistry: The case for multidimensional, integrated single-molecule approaches. *J Phys Chem B*, 125:10404—18, 2021.

Y. Kuzmin. Bresenham's line generation algorithm with built-in clipping. *Comput Graph Forum*, 14:275—80, 1995.

D. Lay, S. Lay and J. McDonald. *Linear Algebra and its Applications*. Pearson, 5th ed., 2016.

Z. Li, Q. Ding and W. Zhang. *A Comparative Study of Different Distances for Similarity Estimation*. Intelligent Computing and Information Science, Springer Berlin Heidelberg, 483—8, 2011. ISBN 978-3-642-18129-0.

X. Li, P. Mooney, S. Zheng, C. Booth, M. Braunfeld, S. Gubbens, D. Agard and Y. Cheng. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods*, 10(6):584—90, 2013.

H. Liao and J. Frank. Classification by bootstrapping in single particle methods. *Proc IEEE Int Symp Biomed Imaging*, 2010:169—172, 2010.

D. Liebschner, P. Afonine, M. Baker, G. Bunkoczi, V. Chen, T. Croll, B. Hintze, L. Hung, S. Jain, A. McCoy, N. Moriarty, R. Oeffner, B. Poon, M. Prisant, R. Read, J. Richardson, D. Richardson, M. Sammito, O. Sobolev, D. Stockwell, T. Terwilliger, A. Urzhumtsev, L. Videau, C. Williams and P. Adams. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst*, D75:861—77, 2019.

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(12):2628—35, 1995.

L. Lovisolo and E. da Silva. Uniform distribution of points on a hyper-sphere with applications to vector bit-plane encoding. *IEEE Proceedings - Vision, Image and Signal Processing*, 148(3):187—93, 2001.

R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, 395(10224):565—74, 2020.

L. Maaten, E. Postma and J. Herik. *Dimensionality reduction: A comparative review*. Tilburg University Technical Report, 2009.

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(5):2484—2491, 2020.

S. Maji and J. Frank. What's in the black box?—A perspective on software in cryo-electron microscopy. *Biophys J*, 120(20):4307—11, 2021.

I. Marcos-Alcalde, J. Setoain, J. Mendieta-Moreno, J. Mendieta and P. Gomez-Puertas. MEPSA: Minimum energy pathway analysis for energy landscapes. *Bioinformatics*, 31(23):3853—55, 2015.

G. Mashayekhi. ManifoldEM Matlab repository. GitHub, 2020 (accessed Aug. 17, 2020). https://github.com/GMashayekhi/ManifoldEM_Matlab/

Maxon Computer GmbH. Cinema 4D. Version 24, 1989-2021.

B. McCartin. On polygonal domains with trigonometric eigenfunctions of the Laplacian under Dirichlet or Neumann boundary conditions. Appl Math Sci, 2(58):2891—901, 2008.

M. Merkulova, T. Paunescu, A. Azroyan, V. Marshansky, S. Breton and D. Brown. Mapping the H^+ (V)-ATPase interactome: identification of proteins involved in trafficking, folding, assembly and phosphorylation. *Sci Rep*, 14827, 2015.

A. Moscovich, A. Halevi, J. Anden and A. Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *Inverse Probl*, 36(2), 2020.

J. Munkres. *Topology*. Prentice Hall, Incorporated, 2nd ed., 2000.

J. Munro, K. Sanbonmatsu, C. Spahn and S. Blanchard. Navigating the ribosome's metastable energy landscape. *Trends Biochem Sci*, 34(8):390—400, 2009.

K. Murata and M. Wolf. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim Biophys Acta—Gen Subjects*, 1862(2):324—34, 2018.

T. Nakane, D. Kimanius, E. Lindahl and S. Scheres. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *Elife*, 7(36861), 2018.

H. Noller, L. Lancaster, J. Zhou and S. Mohan. The ribosome moves: RNA mechanics and translocation. *Nat Struct Mol Biol*, 24(12):1021—27, 2017.

S. Omohundro. *Five balltree construction algorithms*. ICSI Technical Report, TR-89-063, 1989.

R. Owens (Ed.) et al. *Structural Proteomics: High-Throughput Methods*. Methods in Molecular Biology Series, vol. 2305, 3rd ed., 2021.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos Mag*, 2:559—72, 1901.

P. Penczek. Variance in three-dimensional reconstructions from projections. *Proc IEEE Int Symp on Biomed Imaging*, 749—52, 2002. doi:10.1109/ISBI.2002.1029366.

P. Penczek, J. Frank and C. 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(2):184—194, 2006a.

P. Penczek, C. Yang, J. Frank and C. Spahn. Estimation of variance in single-particle reconstruction using the bootstrap technique. *J Struct Biol*, 154(2):168-183, 2006b.

P. Penczek, M. Kimmel and C. Spahn. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. *Structure*, 19(11):1582—1590, 2011.

M. Peplow. The next big hit in molecule Hollywood. *Nature*, 544:408—10, 2017.

E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng and T. Ferrin. USCF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605—1612, 2004.

R. Phillips, J. Kondev and J. Theriot. *Physical Biology of the Cell*. Garland Science, Taylor and Francis Group, New York, 2008.

J. Phillips, D. Hardy, J. Maia, J. Stone, J. Ribeiro, R. Bernardi, R. Buch, G. Fiorin, J. Henin, W. Jiang, R. McGreevy, M. Melo, B. Radak, R. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. Kale, K. Schulten, C. Chipot and E. Tajkhorshid. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys*, 153:044130, 2020.

I. Poudyal, M. Schmidt and P. Schwander. Single-particle imaging by x-ray free-electron lasers—How many snapshots are needed? *Structural Dynamics*, 7:024102, 2020.

G. Pintilie, K. Zhang, Z. Su, S. Li, M. Schmid and W. Chiu. Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat Methods*, 17(3):328—34, 2020.

A. Punjani, J. Rubinstein, D. Fleet and M. Brubaker. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods*, 14(3):290—96, 2017.

A. Punjani and D. Fleet. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J Struct Biol*. 213(2):107702, 2021.

M. Rapp, L. Shapiro and J. Frank. Contributions of single-particle cryo-electron microscopy toward fighting COVID-19. Trends Biochem Sci, 2021.

L. Reimer and H. Kohl. *Transmission Electron Microscopy: Physics of Image Formation*. Springer Series in Optical Sciences, vol. 36, 5th ed., 2008.

A. Rohou and N. Grigorieff. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol*. 192(2):216—21, 2015.

B. Roux. *Molecular Machines*. World Scientific, 1st ed., 2011.

S. Scheres, R. Nunez-Ramirez, Y. Gomez-Llorente, C. San Martin, P. Eggermont and J. Carazo, Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure*, 15(10):1167—77, 2007.

S. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol*. 180(3):519—30, 2012.

S. Scheres. Processing of structurally heterogeneous cryo-EM data in RELION. *Methods Enzymol*, 579:125—57, 2016.

P. Schober, C. Boer and L. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesth Analg*, 126(5):1763—8, 2018.

F. Schopf, M. Biebl and J. Buchner. The HSP90 chaperone machinery. *Nat Rev Mol Cell Biol*, 18:345—60, 2017.

Schrödinger, LLC. The PyMOL molecular graphics system. Version 1.8, 2015.

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(1647):20130567, 2014.

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. doi: 10.1101/864116.

E. Seitz. Cryo-EM synthetic continua Python repository. Zenodo, 2019. https://zenodo.org/record/3561105#.YUPFiJ5KgUY

E. Seitz and J. Frank. POLARIS: Path of least action analysis on energy landscapes. *ACS: J Chem Inf Model*, 60(5):2581—90, 2020.

E. Seitz. POLARIS Python repository. Zenodo, 2020 (accessed Feb. 4, 2020). https://zenodo.org/record/3612185#.XjnV9WhKiUk/

E. Seitz, F. Acosta-Reyes, S. Maji, P. Schwander and J. Frank. Geometric machine learning informed by ground truth: Recovery of conformational continuum from single-particle cryo-EM data of biomolecules. *bioRxiv*, 2021a. doi:10.1101/2021.06.18.449029.

E. Seitz. ESPER Python repository. Zenodo, 2021. https://zenodo.org/record/5362645#.YUPG2p5KgUY

E. Seitz, H. Liao, S. Maji, J. Frank, et al. ManifoldEM Beta Python repository. Zenodo, 2021b. https://zenodo.org/record/5578874#.YXrw69bMIUY

J. Spence, U. Weierstall and H. Chapman. X-ray lasers for structural and dynamic biology. *Rep Prog Phys*, 75(10):102601, 2012.

A. Spirin. Ribosome as a molecular machine. *Febs Lett*, 514:2—10, 2002.

T. Sztain, S. Ahn, A. Bogetti, L. Casalino, J. Goldsmith, E. Seitz, R. McCool, F. Kearns, F. Acosta-Reyes, S. Maji, G. Mashayekhi, J. McCammon, A. Ourmazd, J. Frank, J. McLellan, L. Chong and R. Amaro. A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat Chem*. 2021.

G. Tang, L. Peng, P. Baldwin, D. Mann, W. Jiang, I. Rees and S. Ludtke. EMAN2: An extensible image processing suite for electron microscopy. *J Struct Biol*, 157(1):38—46, 2007.

J. Townsend. *A Modern Approach to Quantum Mechanics*. University Science Books, 6th ed., 2000.

L. Trabuco, E. Villa, K. Mitra, J. Frank and K. Schulten. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure*, 16(5):673—83, 2008.

M. Vulovic, L. Voortman, L. Vliet and B. Riegar. When to use the projection assumption and the weak-phase object approximation in phase contrast cryo-EM. *Ultramicroscopy*, 136:61—6, 2014.

R. Wade. A brief look at imaging and contrast transfer. *Ultramicroscopy*. 46(1-4):145—56, 1992.

D. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

A. Walls, Y. Park, M. Tortorici, A. Wall, A. McGuire and D. Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike protein. *Cell*, 181(2):281—92, 2020.

R. Warshamanage and G. Murshudov. EMDA: Electron microscopy difference and average map calculation, 2019. Available Online: https://www2.mrc-lmb.cam.ac.uk/groups/murshudov/content/emda/emda.html

P. Whitford, R. Altman, P. Geggier, D. Terry, J. Munro, J. Onuchic, C. Spahn, K. Sanbonmatsu and S. Blanchard. *Dynamic Views of Ribosome Function: Energy Landscapes and Ensembles*. Springer, Vienna, 2011.

D. Wrapp, N. Wang, K. Corbett, J. Goldsmith, C. Hsieh, O. Abiona, B. Graham, and J. McLellan. Cryo-EM structure of the 2019-nCov spike in the prefusion conformation. *Science*, 367(6483):1260—3, 2020.

R. Zalk, O. Clarke, A. des Georges, R. Grassucci, S. Reiken, F. Mancia, W. Hendrickson, J. Frank, and A. Marks. Structure of a mammalian ryanodine receptor. *Nature*, 517(7532):44—9, 2015.

B. Zhang, D. Jasnow and D. Zuckerman. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *J Chem Phys*, 132(5):054107, 2010.

J. Zhao, S. Benlekbir and J. Rubinstein. Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. *Nature*, 521:241—5, 2015.

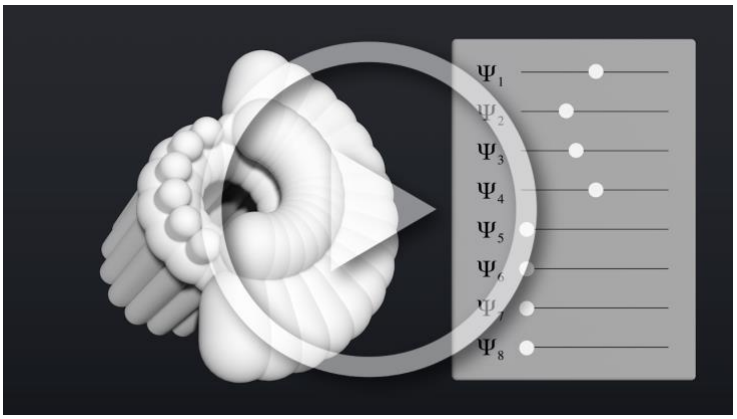
E. Zhong, T. Bepler, B. Berger and J. Davis. CryoDRGN: Reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat Methods*. 2021.

J. Zivanov, T. Nakane, B. Forsberg, D. Kimanius, W. Hagen, E. Lindahl and S. Scheres. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife*, 7(42166), 2018.

X. Zou. *What is Electron Crystallography*. Electron Crystallography. NATO Science Series II: Mathematics, Physics and Chemistry, vol. 211, Springer, Dordrecht, 2006.

Appendix A

The following is supplementary content for Chapter 1: Function from Structure.



Movie A1: Animated schematic of a toy model shown exercising $n = 5$ degrees of freedom (of eight in total). Each degree of freedom represents a unique sequence of conformational changes (i.e., a conformational motion, CM) represented by an interval slider on the right. As the slider is altered, the machine transforms between neighboring states belonging to that CM coordinate. Orthogonality of these CMs ensures that all of the machine's allowable states are accessible by some unique permutation of slider values.

Movie A2: Animated schematic of conformational diffusion. Here, the molecular machine—the toy model seen in Figure 1—is portrayed as a marble in a bowl-like basin, undergoing random perturbations. These are portrayed by the large force arrows, which appear intermittently (with magnitude ideally drawn from a Boltzmann distribution) and alter the marble's position (i.e., its current state) along the available 2D coordinates ("mouth", "wings"). Much like an actual marble in a bowl, unfavorable energetics impede the machine's progress "up the hill" (i.e., opening its wings), such that several consecutive, unidirectional thermal forces are required for it to overcome this barrier and reach a new region in the free-energy landscape.

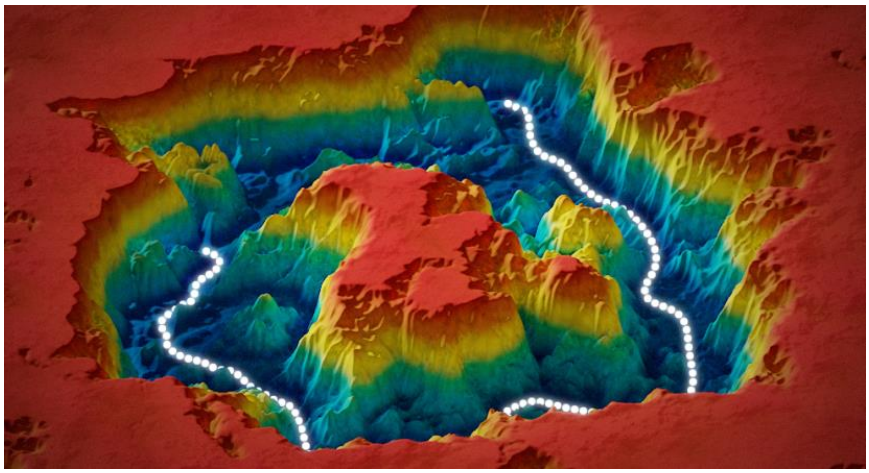


Figure A1: Schematic of a free-energy landscape, with topographic features describing probabilities of transition between states in a 2D state space. The white pathway demonstrates one possible low-energy route along which a molecular machine may transform between two conformations, where each white dot in between represents a unique conformational state.

Appendix B

The following is supplementary content for Chapter 2: The ManifoldEM Framework.

Movie B1: A video overview of the ManifoldEM (Beta) Python GUI. This demonstration was created by E. Seitz and presented by J. Frank at the 2019 Computation Cryo-EM workshop (Flatiron Institute). The molecular machine shown is V-ATPase, as provided by Zhao et al. (2015).

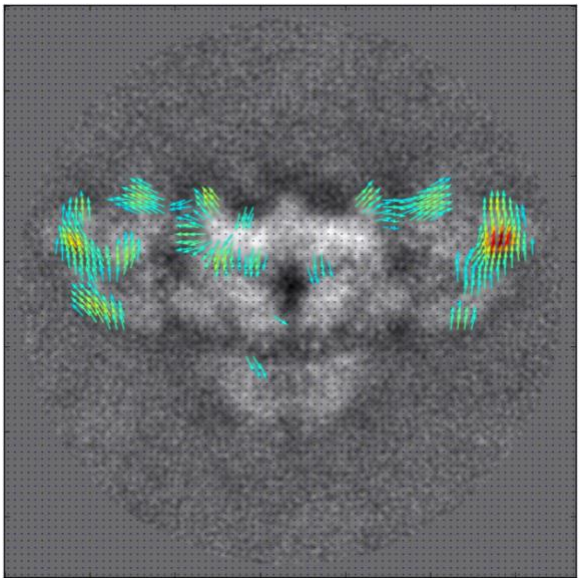


Figure B1: Example of "flow vectors" obtained from optical flow on a 2D NLSA movie (generated via ManifoldEM with the RyR1 data set). Image motion is expressed by the direction of the vectors, while the magnitude of motions is expressed by the relative color of each vector. As seen here in the wing-like domains, red denotes the maximum magnitude of motion in the image. (Figure by S. Maji).

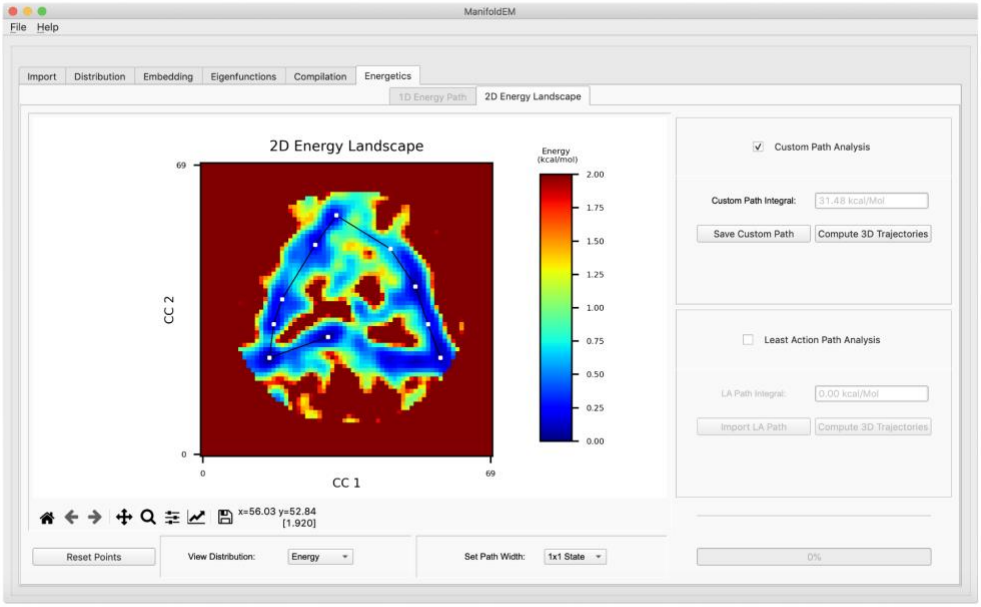
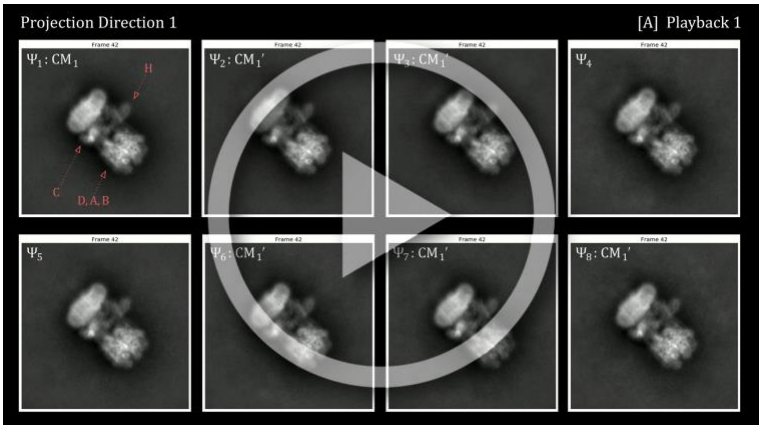


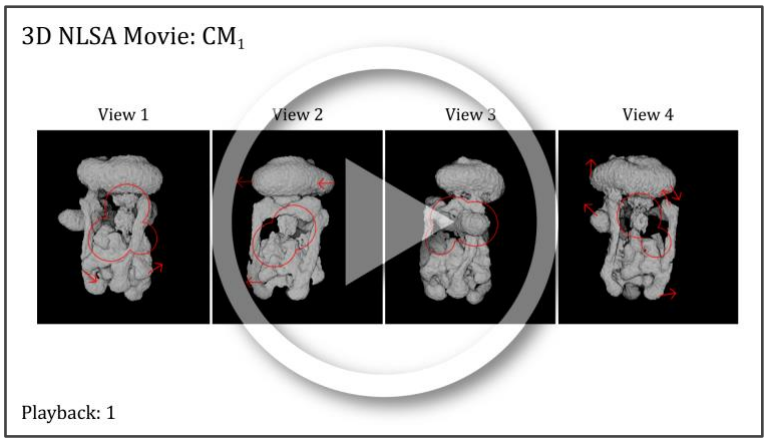
Figure B2: The "2D Energy Landscape" tab showing outputs from the ribosomal data set (Dashti et al., 2014). Although not active in the Beta release, as shown here, a set of user-defined coordinates can be selected using the GUI interface. Alternatively, the path of least action can be imported (as generated via external programs; e.g., POLARIS) and similarly used to produce a final sequence of NLSA volumes.

Appendix C

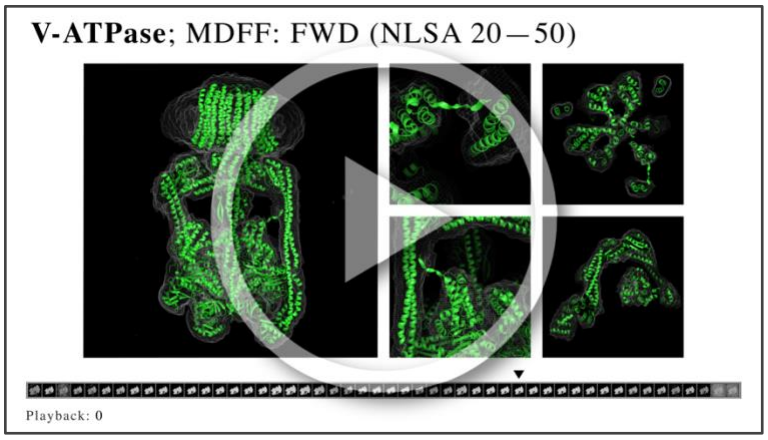
The following is supplementary content for Chapter 4: ManifoldEM Analysis of V-ATPase.



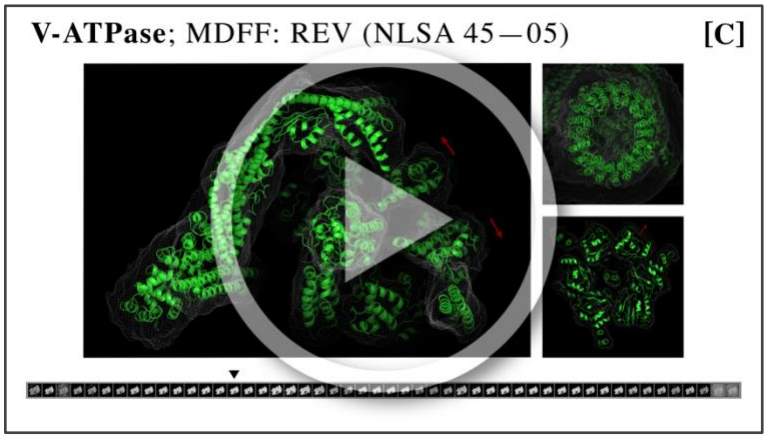
Movie C1: Collection of 2D NLSA movies as obtained from an assortment of PDs and corresponding eigenvectors. In [A], the 2D NLSA movies for the leading eight eigenvectors for the same PD are shown. Here, Ψ_1 can be seen exercising the previously-described CM_1. Harmonics of CM_1 can be found occupying Ψ_2, Ψ_3, Ψ_6, Ψ_7 and Ψ_8 (uncorrected for sense). Ψ_4 demonstrates a conformational coordinate corresponding to an enlargement of the molecule, possibly due to the images occupying varying depth in the ice. Similarly, Ψ_5 demonstrates an apparent rotation of the molecule, which can arise due to dispersed S^2 angular distributions of images in the PD. In [B], the 2D NLSA movies corresponding to CM_1 are shown as expressed in eight highly-occupied PDs residing along the S^2 equator (Figure 9). Finally, in [C], a second "spring"-like motion is shown (CM_2), which was only observed in a small number of PDs.



Movie C2: Four views showing the sequence of V-ATPase NLSA volumes reconstructed along CM_1, as displayed using Chimera. In each view, red arrows indicate the directions of motion of the respective domains.



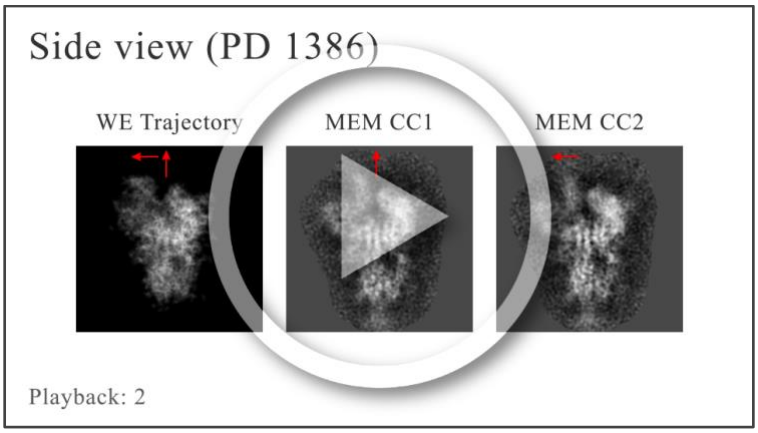
Movie C3: V-ATPase atomic-coordinate structures fit to NLSA volumes using MDFF, and displayed using PyMOL. Here, the results of the initial forward MDFF simulation are shown, which include NLSA states 20 through 50. See Section 4.3 for a description of the results shown.



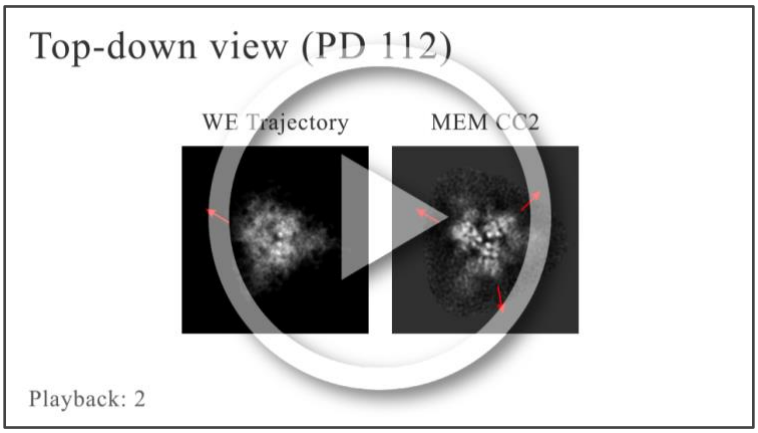
Movie C4: V-ATPase atomic-coordinate structures fit to NLSA volumes using MDFF, and displayed using PyMOL. Here, the results of the reverse simulation are shown, including NLSA states 45 through 5. See Section 4.3 for a description of the results shown.

Appendix D

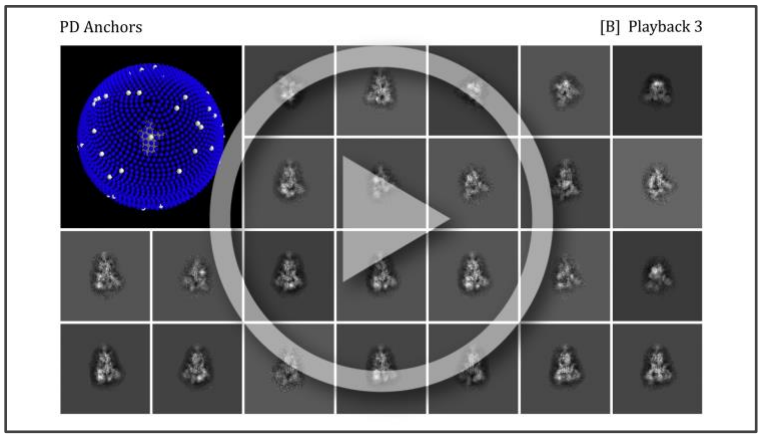
The following is supplementary content for Chapter 5: ManifoldEM Analysis of SARS-CoV-2.



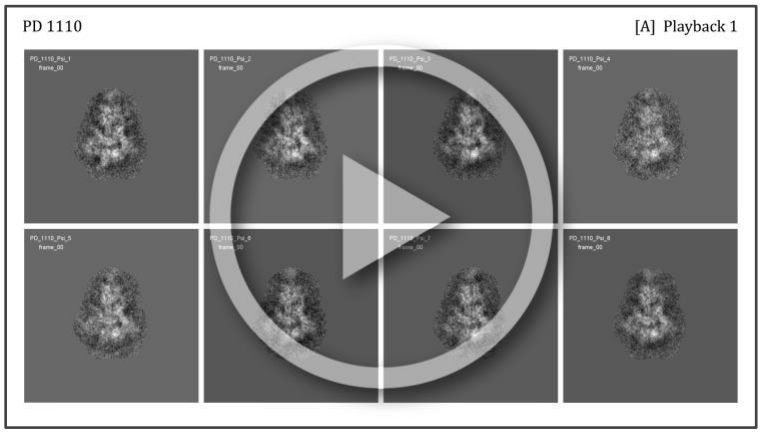
Movie D1: A comparison of the WE trajectory and ManifoldEM CC_1 and CC_2 (i.e., CM_1 and CM_2) for a side view. It can be seen that there is a strong agreement between the full WE trajectory and the sequential, piecewise combination of both conformational motions. Red arrows indicate the direction of motion.



Movie D2: A comparison of the WE trajectory and ManifoldEM CC_2 (i.e., CM_2) for a top-down view. A strong agreement can be seen between the outputs of these two frameworks. To note, the first conformational motion was not readily achievable from this view via manifold embedding, since the RBD-down to RBD-up trajectory from this view is orthogonal to the plane of the projection. Red arrows indicate directions of subunit motions.



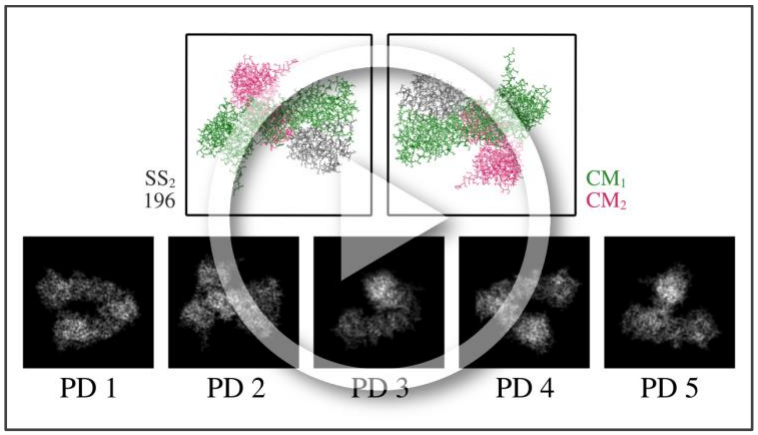
Movie D3: In [A], three (contrast corrected) 2D NLSA movies are shown, each for a different projection direction. PD_{1326} displays CM_1, while PD_{1074} and PD_{1664} appear to display a hybrid of CM_1 and CM_2. Red arrows indicate direction of subunit motions. In [B], a set of 2D NLSA movies is shown, with each one corresponding to one of the CM_1 anchor PDs chosen for initiation of belief propagation. The location of these anchor nodes on S^2 are shown in the top-left insert.



Movie D4: Based on the appearance of decoupled CMs in the 2D NLSA movies in [A], PD_{1110} is an example of one of the few gold standard embeddings. The RBD can be seen transitioning from the outward position to the inward position (CM_1) in Ψ_1, while the claw motion (CM_2) can be seen clearly in Ψ_5. Motions in the remaining NLSA movies are harder to untangle, and likely represent either harmonics or hybrid motions. Likewise, in [B], PD_{1387} shows CM_1 in Ψ_3 and CM_2 in Ψ_2 respectively, with several harmonics and hybrids mixed in. In [C], PD_{1222} demonstrates the increase in difficulty for CM decisions, which were encountered in the majority of PDs. All of these NLSA movies have the appearance of hybrid motions, with Ψ_1 or Ψ_3 possibly expressing a marginal preference for CM_1. Finally, PD_{150} [D] represents the vast majority of the 1678 PD embeddings retrieved, where corresponding NLSA frames are noisy and display jittery or stalling movies. Here, CMs are very difficult to consistently assign, with these PDs typically having a much lower image count (e.g., PD_{150} encompassed 158 cryo-EM images).

Appendix E

The following is supplementary content for Chapter 6: The Synthetic Continuum Framework.



Movie E1: An animation cycling through the 400 SS_2 states as represented by the Hsp90 atomic structures (top) and subset of PDs (bottom). The ordering of these states chronologically follows a clockwork pattern, with each progression from $CM_2\{1\} \rightarrow CM_2\{20\}$ iterating CM_1 (akin to the minute hand) forward once. Thus, of the 400 states in SS_2, the first index corresponds to state $CM_1\{1\}\_CM_2\{1\}$, the second to state $CM_1\{1\}\_CM_2\{2\}$, the 21st to state $CM_1\{2\}\_CM_2\{1\}$, and the 400th to state $CM_1\{20\}\_CM_2\{20\}$.

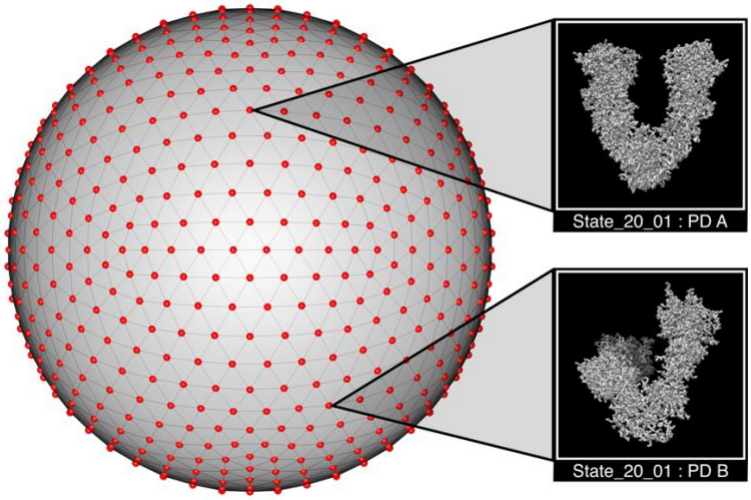


Figure E1: Distribution of PDs (shown as red points) as arranged across the space of all possible viewing angles (S^2). This tessellation was achieved by forming an icosahedron with 48 segments in Cinema4D (Maxon Computer GmbH, 1989) to create 812 evenly spaced vertices (with edges seen in black) representing a set of available PDs. Each vertex (PD) is approximately 6° from its nearest neighbor.

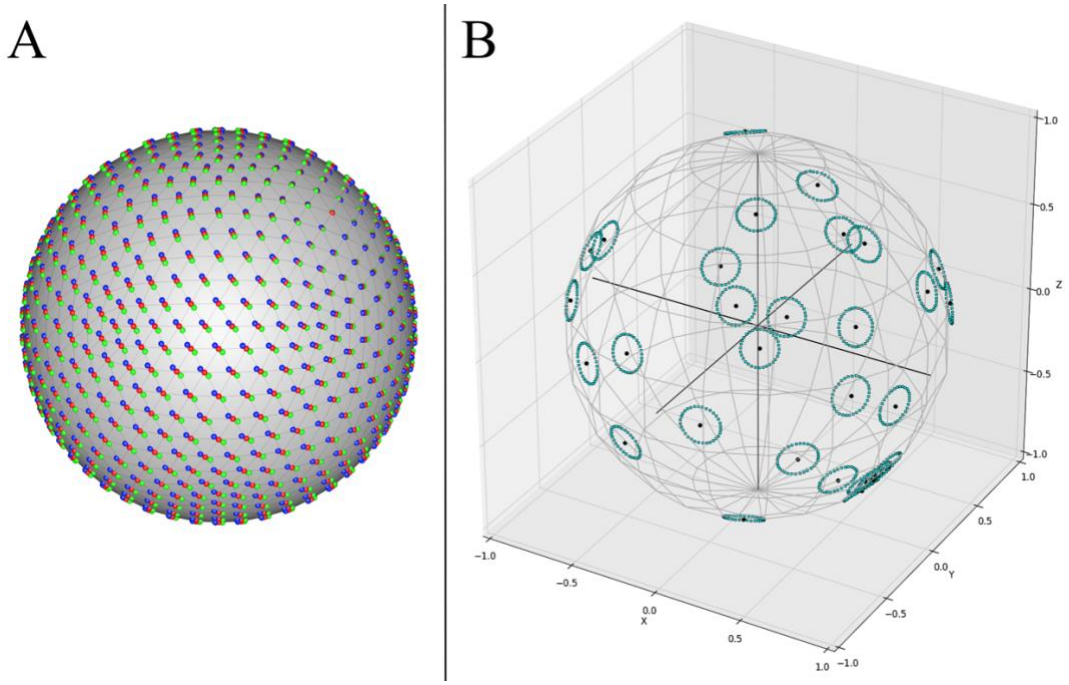
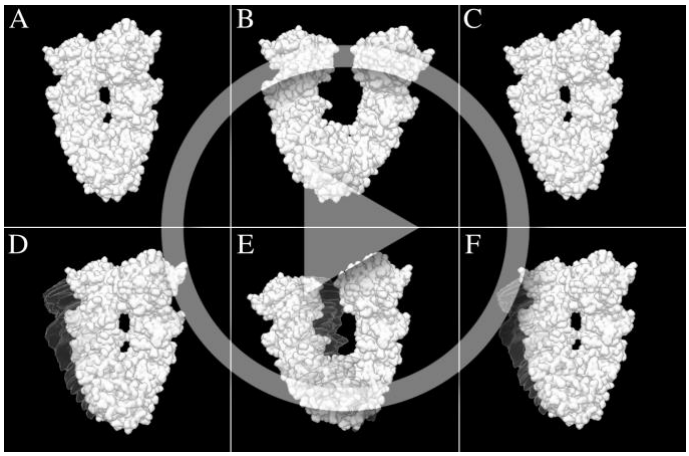


Figure E2: In [A], the problem of applying global rotations to all PDs is illustrated, which results in a non-uniform angular distribution on S^2. The original PD locations are represented by red vertices, while two oppositely-directed global rotations are shown as blue and green vertices. As can be seen, the distribution of changes of blue and green vertices about each corresponding red vertex are non-uniform, and vanishes near the poles formed by the axis of rotation. In [B], a schematic is presented showing the effects of the back-to-back quaternion transformations, resulting in a uniform distribution on S^2. In this example, the first quaternion transformation rotates each point radially away from its PD center by an equal amount ($\theta_1 = \pi/15$), while the second quaternion transformation incrementally rotates those points ($\theta_2 = \pm\pi/8$) along a circle enclosing the original PD with angular radius θ_1.

Appendix F

The following is supplementary content for Chapter 8: Heuristic Analysis of Upstream Methods.



Movie F1: Comparison of the same CMs as observed from two different frames of reference (due to the choice of alignments). The top row [A, B, C] presents states as viewed in the original alignment (i.e., our synthetic workflow presentation), while the bottom row [D, E, F] presents states as viewed via the RELION alignment. The first column [A, D] represents $\mathtt{CM}_1$, column [B, E] represents $\mathtt{CM}_2$, and the third column represents the diagonal across the $\{\mathtt{CM}_1, \mathtt{CM}_2\}$ state space. The semi-transparent overlay in the second row represents the actual volume obtained from the RELION "3D Auto-Refine" procedure. This volume was then used to re-align the states in the original alignment (top row) to the RELION alignment, which is the end result seen in the animation. Alignments of one volume into another were done with the Chimera "fitmap" command. Importantly, while our original synthetic presentation chose the immobile region between the arm and elbow as the reference frame, RELION has instead chosen chain A, such that it is presented as immobile while $\mathtt{CM}_1$ and $\mathtt{CM}_2$ now occur entirely on chain B.

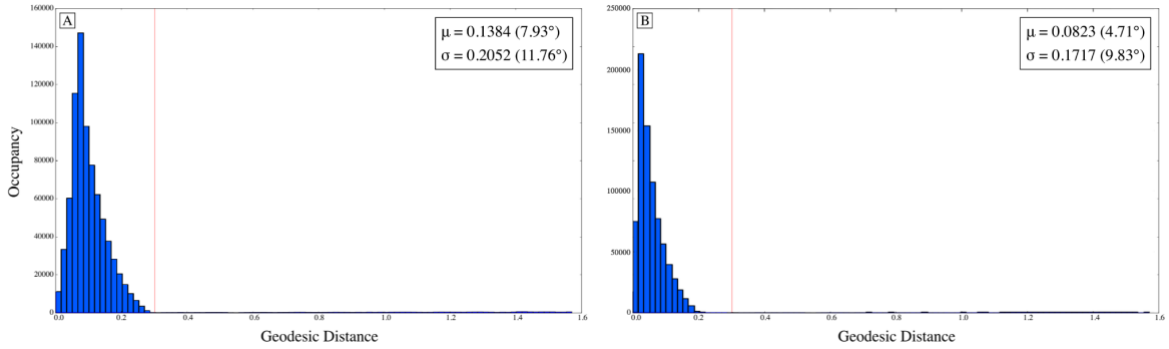


Figure F1: Geodesic distance between each image's assigned ground-truth location on S^2 (i.e., the set of Euler angles from which it was projected in the synthetic workflow) and the placement predicted by [A] RELION 3D auto-refinement and [B] cryo-SPARC non-uniform

refinement. In both cases, statistics of erroneous angular assignments appear to follow a chi-squared distribution, with an elongated tail between 0.3 and 1.6 radians (17.19° and 91.67°, respectively) as demarcated past the red line. In [A], this tail comprised 4.15% of the images (33,733 in total), while in [B], the same region comprised 2.70% of all images (21,901 in total). If this region is removed from both distributions (all values to the right of the red line), in [A], the mean distance μ decreases from 7.93° to 5.72° and the corresponding standard deviation σ decreases from 11.76° to 2.92°. Likewise, in [B], μ decreases from 4.71° to 2.86° and σ decreases from 9.83° to 2.22°.

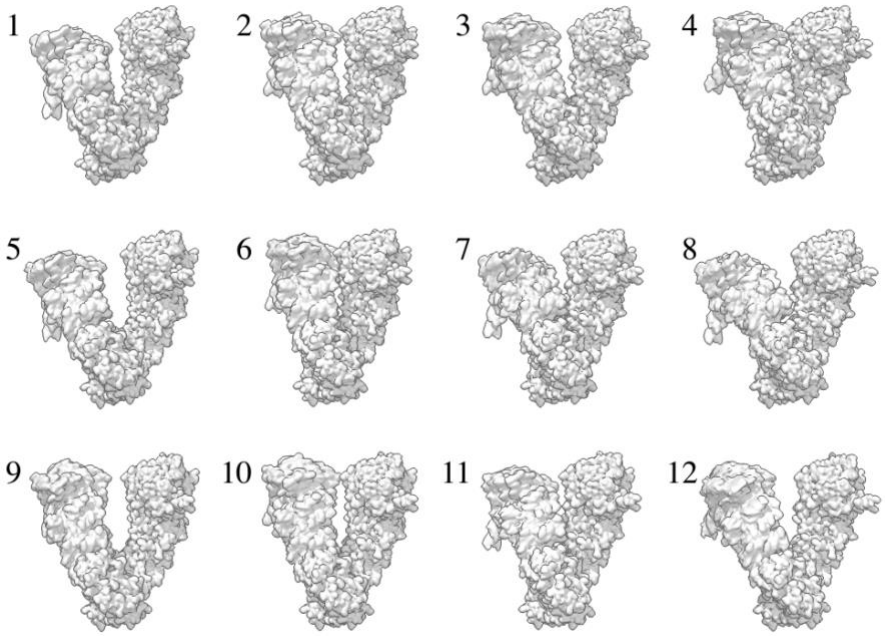


Figure F2: The first 12 RELION classes, as ranked by the number of particles from the ensemble assigned to each. Each class represents a local average of states in the conformational state space (Figure 23). In total, these classes explain 78.21% of all particles in the data set.

Appendix G

The following is supplementary content for Chapter 10: Heuristic Analysis of PD Manifolds.

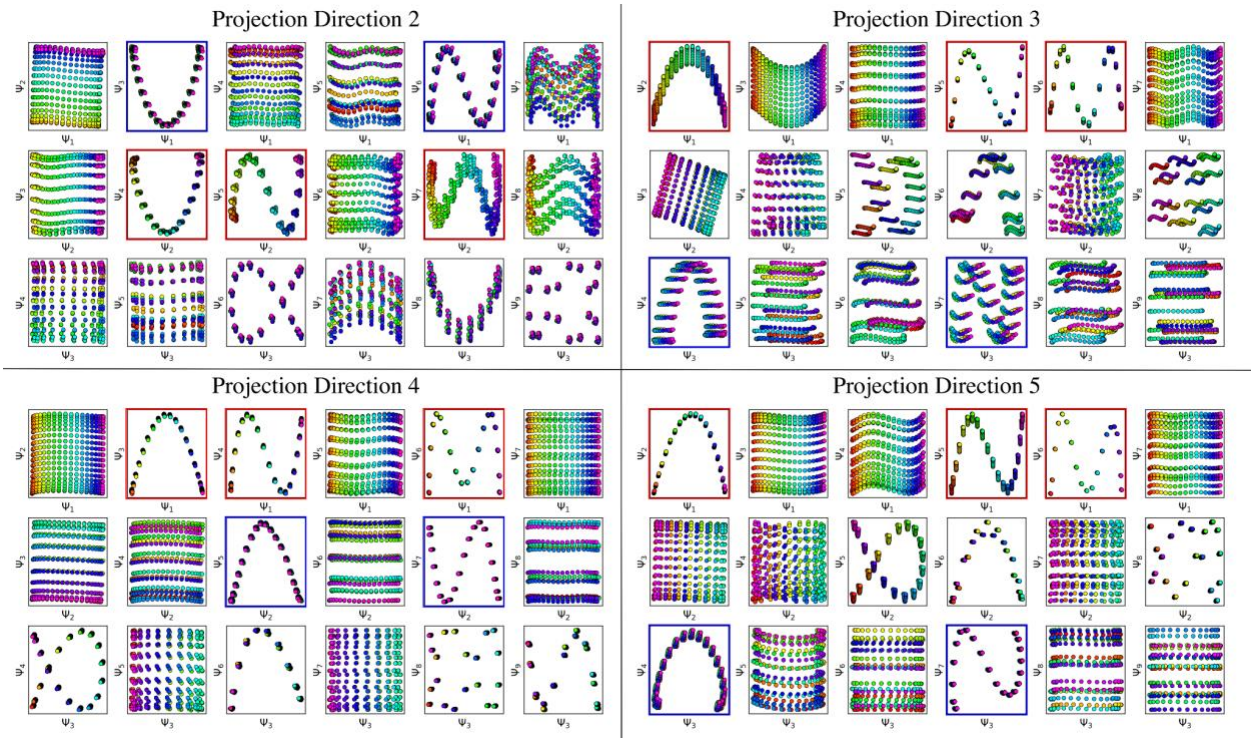


Figure G1: A similar presentation as is shown in Figure 34 for the remaining four PDs. Here, subspaces requiring eigenvector rotations (e.g., both parabolas in PD_3) and housing subtle boundary problems (e.g., the curling inwards of the point-cloud trajectory in $\{\Psi_3 \times \Psi_4\}$ of PD_5) can also be seen in certain 2D subspaces. Note that in PD_2, due to PD disparity, the hierarchy of CM information is actually reversed from those seen in the other four PDs. Here, the CM_2 Chebyshev polynomials are instead present along $\{\Psi_1 \times \Psi_i\}$ combinations (in the first row), while CM_1 Chebyshev polynomials are present along $\{\Psi_2 \times \Psi_j\}$ combinations (in the second row).

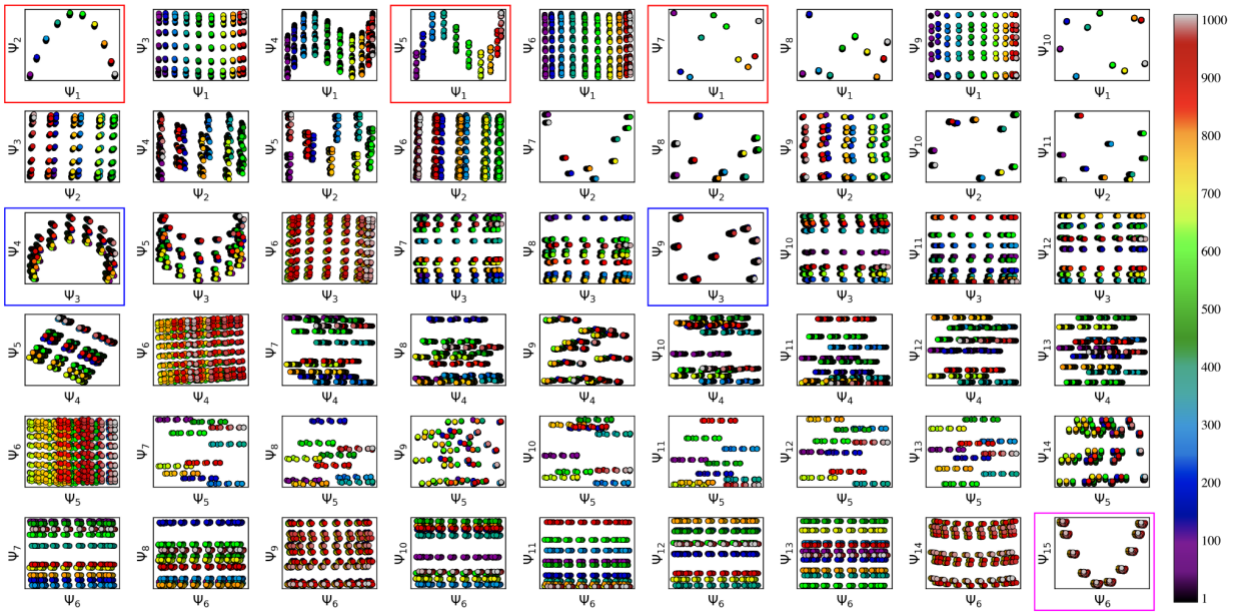


Figure G2: A set of 2D subspaces projected from the embedding obtained from PD_5 in SS_3. The set of conformational modes corresponding to CM_1 is demarcated by the red boxes around interspersed Ψ_1 plots, and occupy specific $\{\Psi_1 \times \Psi_i\}$ combinations (where $i > 1$). Likewise, CM_2 and CM_3 are both separately represented by a set of their own conformational modes; demarcated by blue boxes around interspersed Ψ_3 plots and magenta boxes around interspersed Ψ_6 plots (of which only the first is displayed), respectively. As expected, points on the trajectories defined by CM_1 modes follow along the full spectrum of colors (i.e., indices 1 through 1000), while CM_2 and CM_3 points cover a span of 100 and 10 colors, respectively. Note the presence of harmonics in neighboring rows (for each CM), which are characterized by the presence of α-shaped trajectories, as seen in $\{\Psi_2 \times \Psi_5\}$.

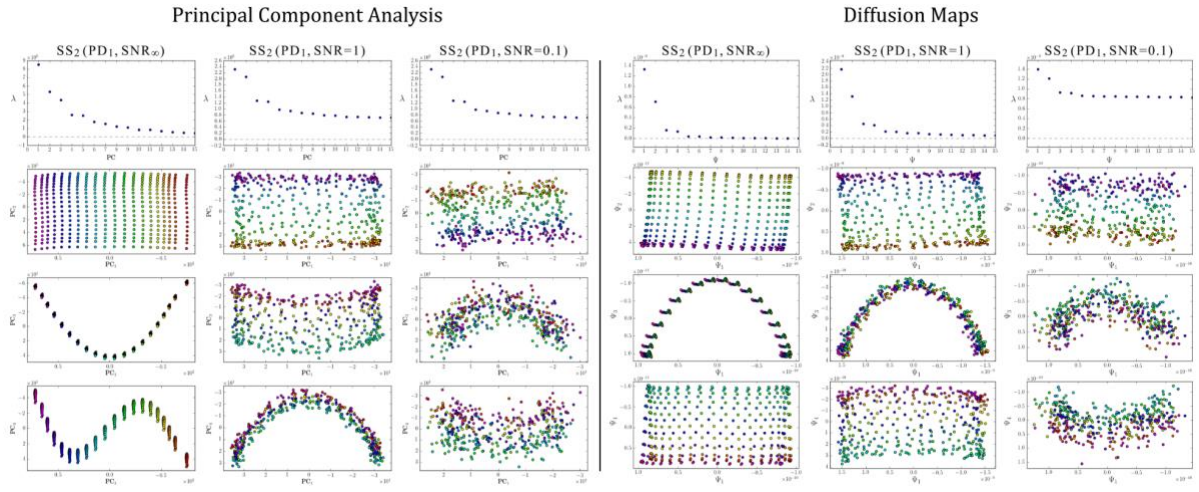


Figure G3: Comparison of 2D subspaces and eigenvalue spectra obtained via PCA (left) and DM (right) for PD_1 in SS_2 across three SNR regimes (one SNR regime per column; with

uniform occupancy $\tau = 1$). As can be seen for both linear and nonlinear dimensionality reduction methods, the well-defined structure of these subspaces deteriorates rapidly as increasing amounts of additive Gaussian noise is introduced on each image. Overall, the outputs of PCA on these data sets revealed a striking resemblance to those produced by DM (with the latter generated within its optimal range of Gaussian bandwidth). Importantly, the parabolic mode is conserved for both methods even within experimental regimes.

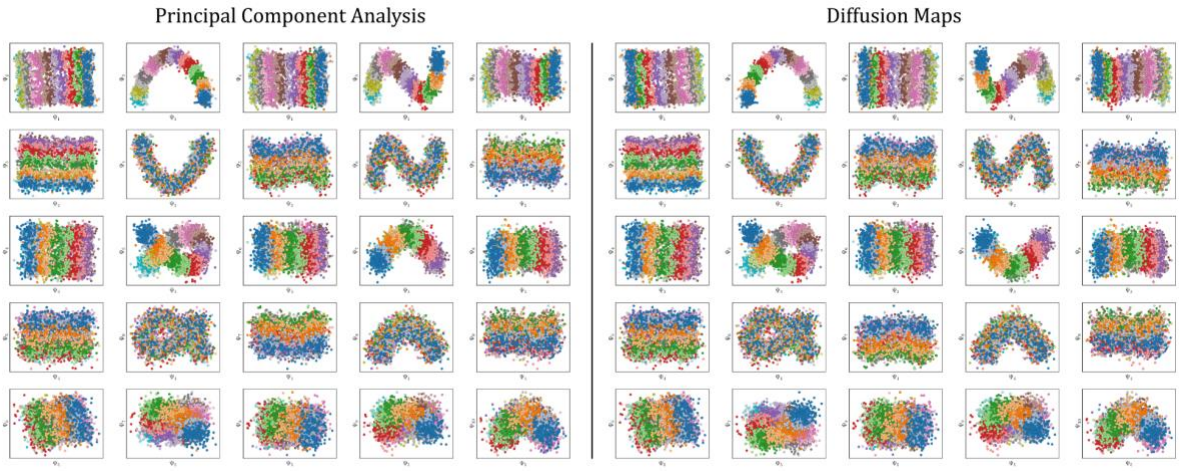


Figure G4: Comparison of 2D subspaces obtained via PCA (left) and DM (right) for PD_2 in SS_2, with image sets generated with SNR of 0.1 and $\tau = 10$. Colors have been assigned to data points so as to match the ground-truth indices of states covering CM_2, which is the most visually pronounced motion as viewed from this PD. Here, the CM_2 subspaces can be seen in the first row, the CM_1 subspaces in the second row, CM_2 first-harmonics in the third row, CM_1 first-harmonics in the fourth row, and CM_2 second-harmonics in the fifth row. Overall, the similarity in outputs between these two methods is undeniable, with the only visual difference appearing in the arbitrary directionality (sense) of each coordinate; as expected.

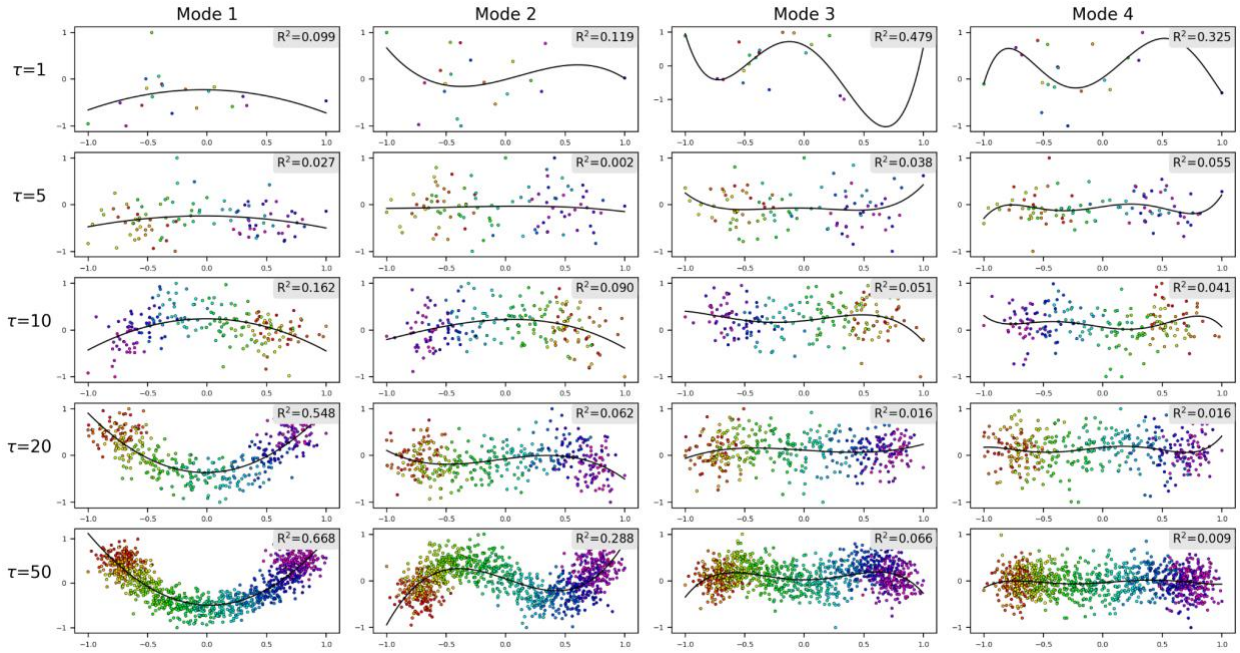


Figure G5: Each of the above rows show a set of 2D subspaces from a τM-dimensional SS_1 embedding with $M = 20$, as obtained from an ensemble of images created with that row's given τ value and SNR of 0.1. Each 2D subspace within each row displays one of the CM's leading Chebyshev modes k via $\{PC_1 \times PC_{k+1}\}$; e.g., all 2D subspaces in the first column have x-axis and y-axis defined by PC_1 and PC_2, respectively. Note that each of these principal components has been scaled to have matching bounds $[-1, 1]$. Lines of best fit are then computed with coefficient of determination $\mathcal{R}^2$ values recorded, as displayed in each corner.

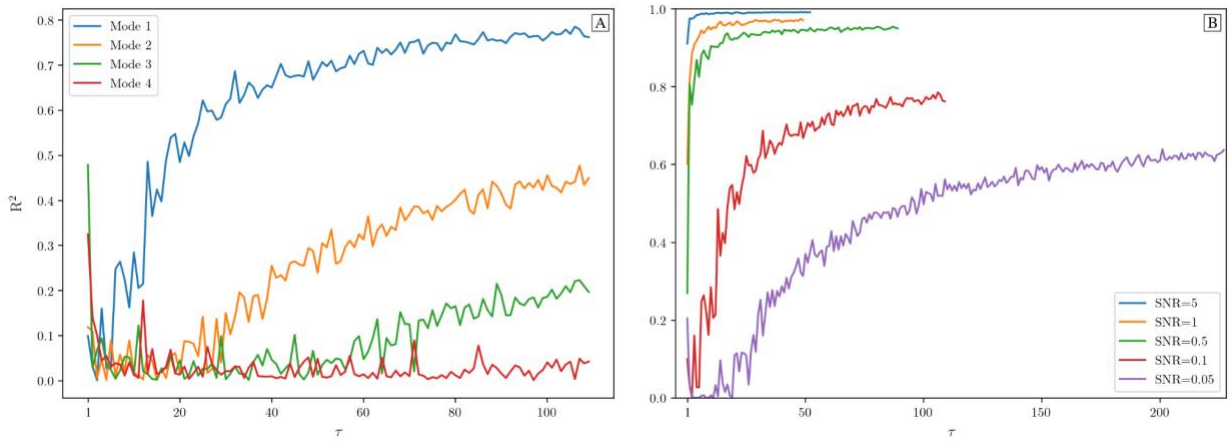


Figure G6: In [A], the coefficient of determination $\mathcal{R}^2$ values are plotted for leading modes with constant SNR. The first mode corresponds to the parabolic Chebyshev mode defined via the projection $\{PC_1 \times PC_2\}$, with the second Chebyshev mode defined via $\{PC_1 \times PC_3\}$, and so on. In [B], only the $\mathcal{R}^2$ values for the first mode are shown while SNR is altered.

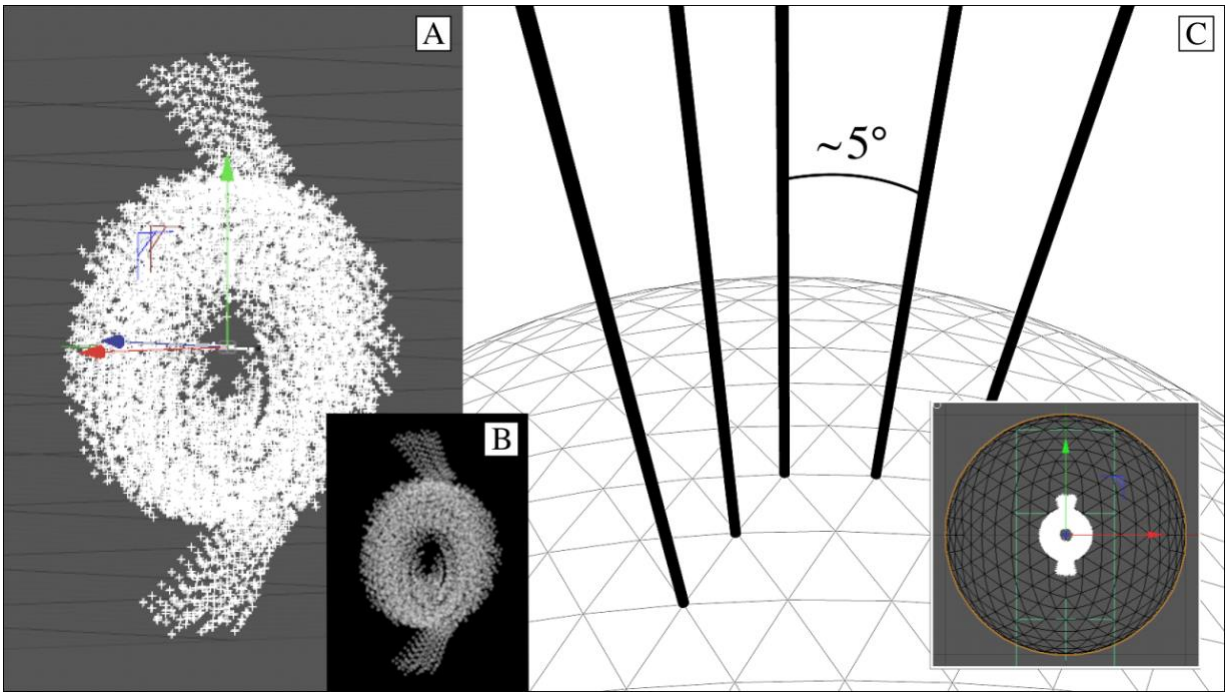
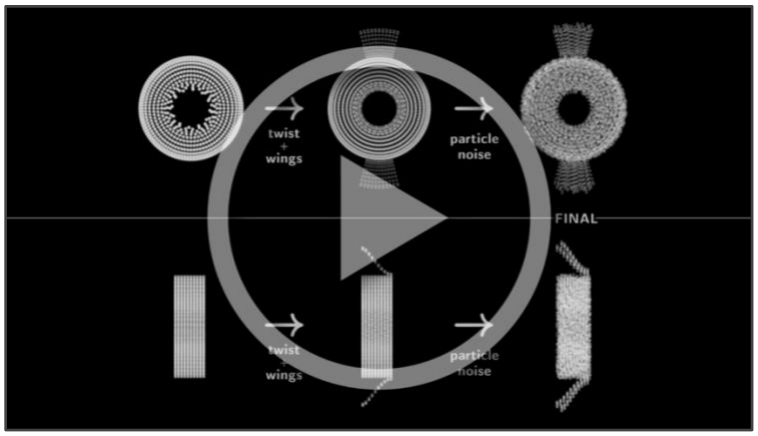


Figure G7: The mouth-wing toy model as generated using Cinema 4D. To uphold the conservation of mass across all states, an ensemble of sphere was generated in three arrays using the "cloner object": (i) a cylindrical array with adjustable inner radius (representing the main body of the object, with its central opening termed the "mouth"); and (ii, iii) two thin, rectangular arrays on either of its sides (termed the "wings"). As described in the main text, the spheres making up each section were keyframed to uniquely reposition across each range of their respective collective motion. From this, 20 "mouth" and 20 "wings" states were defined, using all combinations to generate the 20×20 state space (SS_2). Next, via the "PyroCluster" plugin, the spheres were transformed into "thinking particles" [A], whereby each particle can be envisioned as a "puff" that, when rendered, effectively stacks with integrated intensity with all other particles present along the camera's line of sight [B]. A tessellated sphere was then generated by segmenting an icosahedron, with an orthographic camera set to obtain projections of all states across 980 PDs spaced approximately 5° apart [C]. An animation of these motions as seen from different viewing angles can be found in Movie G1.



Movie G1: In the first scene, on the right is a set of animations cycling through the 400 SS_2 states of the mouth-wing toy model, as viewed from four example PDs. Of these, the northeast PD was used for the investigation in Figure G8. On the left is a static render of the toy model (state 400) as it is rotated 360°. The next scene of Movie G1 showcases the sequence of operations performed to generate this model in Cinema 4D. In the last scene, projections of the final model show it animating along a conformational trajectory representing the diagonal of the state space.

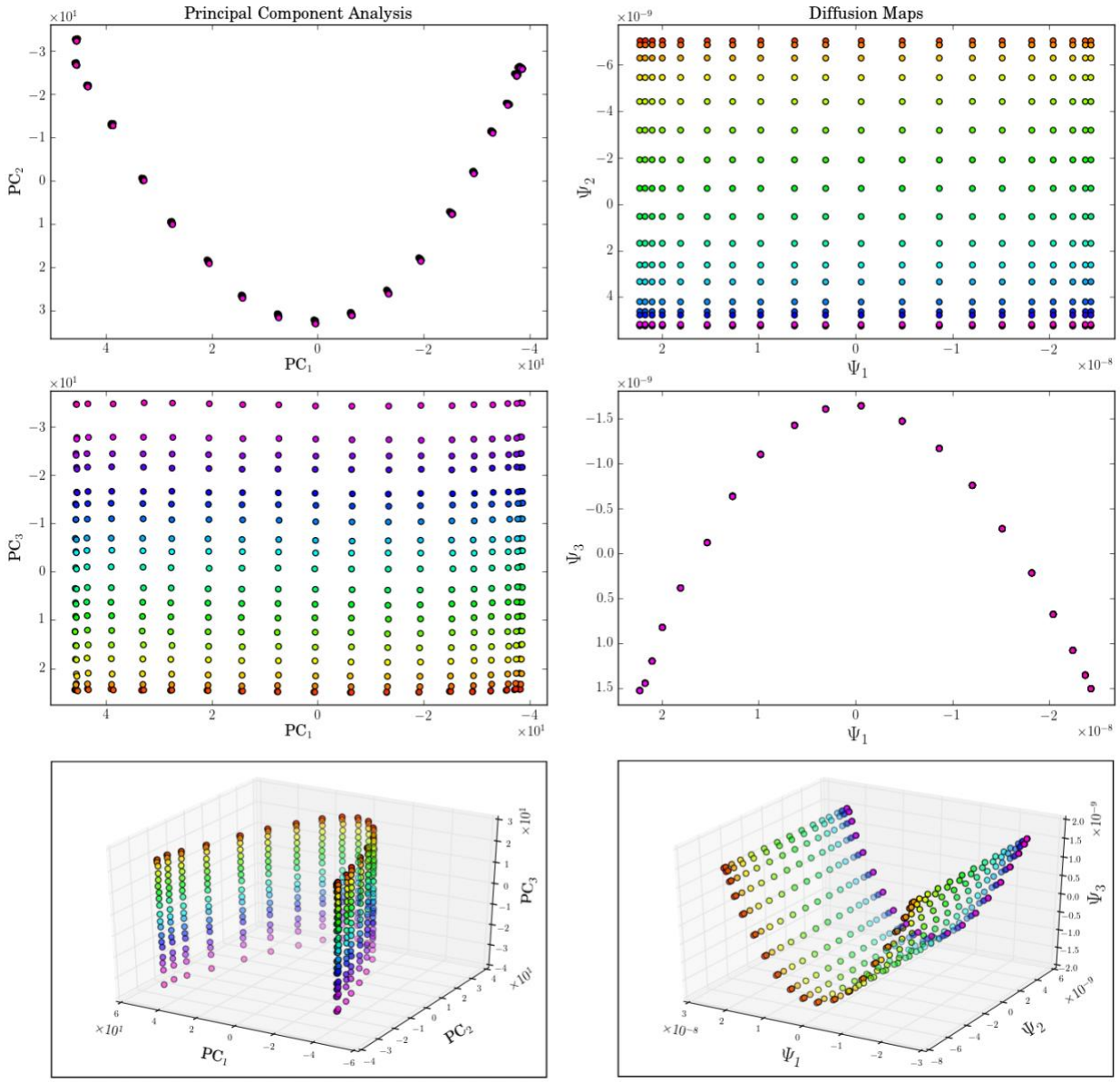
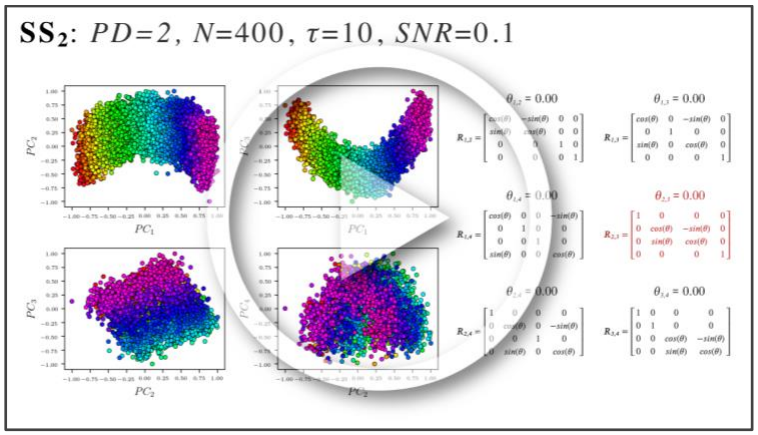


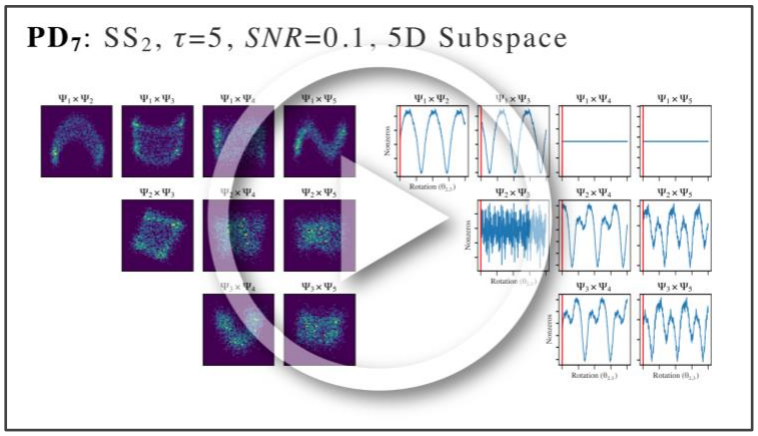
Figure G8: Comparison of PCA (left) and DM (right) embeddings of the 400 images of the mouth-wing toy model in SS_2 from a given PD (Figure G7). The anticipated 20×20 parabolic sheet of states is obtained by both linear or nonlinear dimensionality techniques. Of note, the points in the PCA embedding are less uniformly distributed than those in the DM embedding, suggesting that the DM method better approximates the data set's intrinsic relationships. Overall, these results closely match those obtained from application of PCA and DM on the Hsp90 synthetic continuum.

Appendix H

The following is supplementary content for Chapter 11: The ESPER Method.



Movie H1: Effects of applying a 4D orthogonal rotation to the 4D subspace (shown here using four projections of that subspace) obtained from PD_2 in SS_2 with $\tau = 10$ and SNR of 0.1. The six rotation matrices required for the rotation of a 4D subspace are shown on the right. By only applying rotation operator $R_{2,3}$ with 0.5 radians (28.65°), both CM_1 and CM_2 parabolic modes are corrected—preserving all distances between points—such that they reside completely in the plane of $\{\mathrm{PC}_1 \times \mathrm{PC}_3\}$ and $\{\mathrm{PC}_2 \times \mathrm{PC}_4\}$, respectively.



Movie H2: A movie displaying the 2D histogram approach for finding optimal angles and corresponding parabolic modes. Specifically, the effect of an incremental 4D rotation operator $R_{2,3}$ on each 2D subspace is shown. During these rotations, each 2D subspace exhibits a unique profile which can be characterized by the number of nonzero bins in the corresponding 2D histogram as a function of angle.

Algorithm H1: Eigenfunction Determination.

Input: $N \times N$ embedding ω of Ω_{PD} (N eigenvectors Ψ_i).

Output: Pairs of eigenvectors $\{\Psi_i \times \Psi_j\}$ for CM parabolic subspaces, $\wp$ (with harmonics eliminated); dimension of matrix $\mathbf{O}$, d; required rotation sub-matrices, $R_{i,j} \subset \widetilde{\mathbf{O}}$.

Parameters: Total number of Ψ_i to initially consider, $\widetilde{N}$; minimum cutoff for coefficient of determination ($\mathcal{R}^2$), $\mathcal{R}^2_{\min}$.

partition ω into $\frac{\widetilde{N}(\widetilde{N}-1)}{2}$ unique 2D subspaces $\{\Psi_i \times \Psi_j\}$
assign each subspace a (tuple) index $I_{i,j} \in I$; $\tilde{n} = 0$
for each $\{\Psi_i \times \Psi_j\}$ **do**
    compute best-fit parabola via least squares
    compute $\mathcal{R}^2$; indexed via $\mathcal{R}^2_{i,j}$
    if $\mathcal{R}^2_{i,j} < \mathcal{R}^2_{\min}$ **do** remove $I_{i,j}$ from I
for $i \in \{1, 2, \dots, \widetilde{N} - 1\}$ **do**
    for $j \in \{i + 1,\ i + 2, \dots, \widetilde{N}\}$ **do**
        if $I_{i,j} \in I$ **do**
            if $\mathcal{R}^2_{i,j}$ is max($\mathcal{R}^2_{i,j}$) **do**
                $\wp_i = \{\Psi_i \times \Psi_j\}$; $a_i = j$; $\tilde{n} \pm 1$
            else remove $I_{i,j}$ from I
    remove all $I_{a_i, j > a_i}$ from I
$d = \max(a_i)$
for $I_{i,j} \in I$ **do**:
    for $I_{i,j}{}' \in I$ **if** $I_{i,j} \neq I_{i,j}{}'$ **do**
        $((i,j)$ **for** i **in** $I_{i,j}$ **for** j **in** $I_{i,j}{}')$ $\rightarrow \{i,j\}$ of $R_{i,j}$
        form d-dimensional $R_{i,j}$ matrix; e.g., (5)
return $\wp$, d, $\widetilde{\mathbf{O}}$

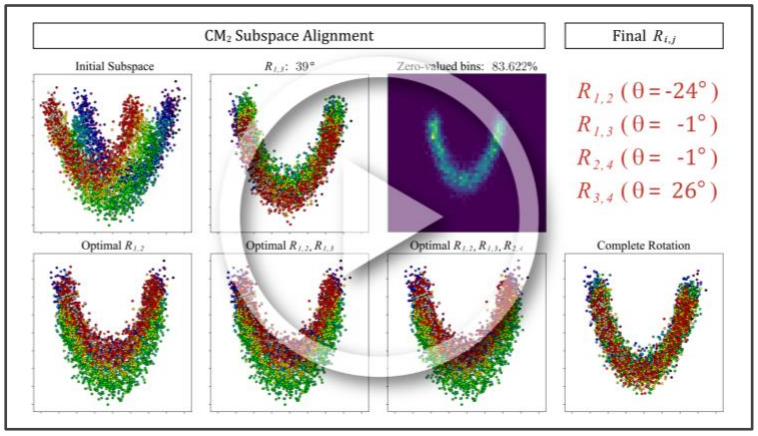
Algorithm H2: Eigenfunction Realignment.

Input: ω; $\wp$; d; $\widetilde{\mathbf{O}}$.

Output: Magnitude of each optimal rotation, $R_{i,j}(\theta_{\text{opt}})$.

Parameters: Number of 2D histogram bins, b; range of angles to explore, $[\theta_{\min}, \theta_{\max}]$ and step size, θ_{step}.

define $\widetilde{\omega}$ from first d eigenvectors of ω
$\theta_{\text{list}} = [\theta_{\min}, \theta_{\min} + \theta_{\text{step}}, \ldots, \theta_{\max} - \theta_{\text{step}}, \theta_{\max}]$
for $\{\widetilde{\Psi}_i \times \widetilde{\Psi}_j\} \subset \widetilde{\omega}$ **in** $\wp$ **do**
  for R_{ij} **in** $\widetilde{\mathbf{O}}$ **do**
    $\xi \coloneqq [\,]$
    for θ **in** θ_{list} **do**
      $\widehat{\omega} = R_{i,j}(\theta) \cdot \widetilde{\omega}$
      generate b-bin 2D histogram H of $\{\widehat{\Psi}_i \times \widehat{\Psi}_j\}$
      append number of zero entries in H to ξ
  define θ_{opt} for current $R_{i,j}$ by index of max(ξ)
return θ_{opt} for each $R_{i,j}$ per CM



Movie H3: An example PD from data-type II is chosen to demonstrate the inner workings of our eigenfunction realignment algorithm. Here, a $d = 5$ dimensional subspace is first isolated, with each 2D subspace therein assigned an $\mathcal{R}^2$ value based on least-square fits. Given presence of adequate fits, the parabola-housing subspace in each eigenvector row is determined via the best $\mathcal{R}^2$ value, with the corresponding eigenvector indices used to procure the four rotation operators (of 10, for $d = 5$) required to align each point cloud with the plane of its subspace. We next demonstrate the generation of 2D histograms as these operators are exercised to determine the optimal angles, as previously detailed in Movie H2. As can be visually assessed, slight inaccuracies may emerge during the histogram optimization (typically no more than 5°), but prove insignificant for downstream procedures.

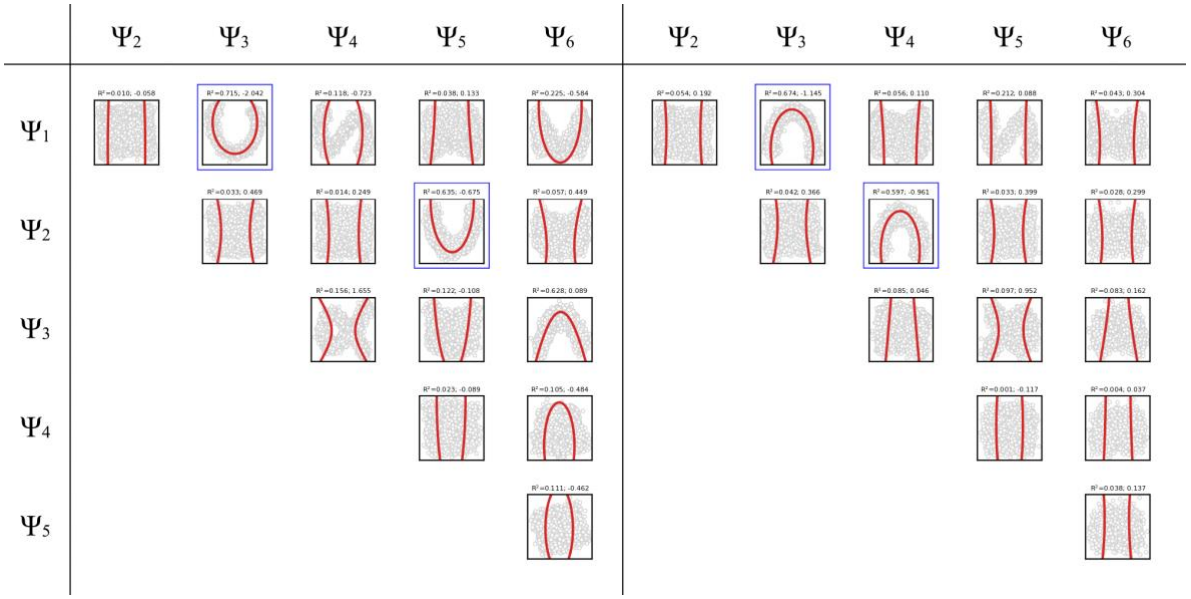


Figure H1: A collection of 2D subspaces for two PDs (left and right) from data-type IV is shown. As denoted in the blue boxes, the parabolic CM subspaces tend to curl inwards near their boundaries. This inward-curling effect varies depending on the CM subspace and thus on the type of motion as visualized from the corresponding PD. For all PDs explored, our use of the general conic least-squares fit proved highly robust to these changes. The coefficient of determination and discriminant of the implicit conic equation is provided above each subplot.

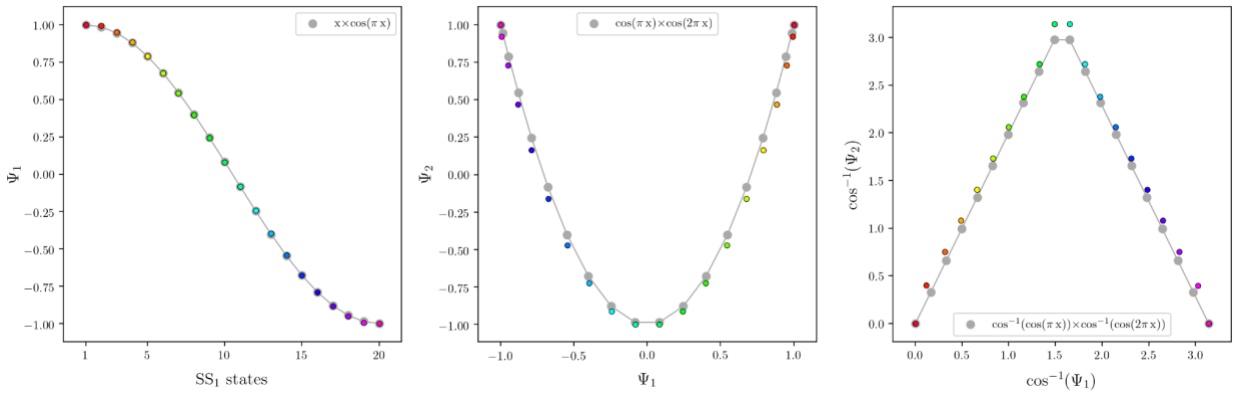


Figure H2: For each of the three subplots, (i) analytical cosine functions are colored in gray and overlaid with (ii) the coordinates of corresponding PD_1 eigenvectors, shown in different colors denoting the SS_1 sequence of states. To compare these two representations within the same coordinate system, the DM eigenvectors have been scaled to match the range of the analytical cosines. Specifically, for $x \in [0, 1]$, 20 x-values were used to generate each $\cos(k\pi x)$ function and subsequently scaled to match the number of SS_1 states; i.e., $x' \in [1, 20]$. Similarly, each eigenvector generated by DM was scaled to match the range of each cosine, such that $\Psi_k \in [-1, 1]$.

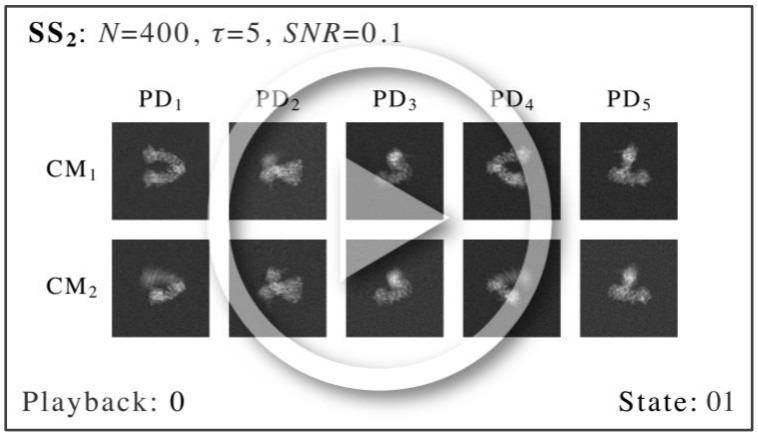
Algorithm H3: Subspace Partitioning.

Input: Rotated CM subspace, $\{\widehat{\Psi}_i \times \widehat{\Psi}_j\}$, and for each point, the corresponding image and CTF; least-squares fit, $\widehat{\Psi}_{\text{fit}}$.

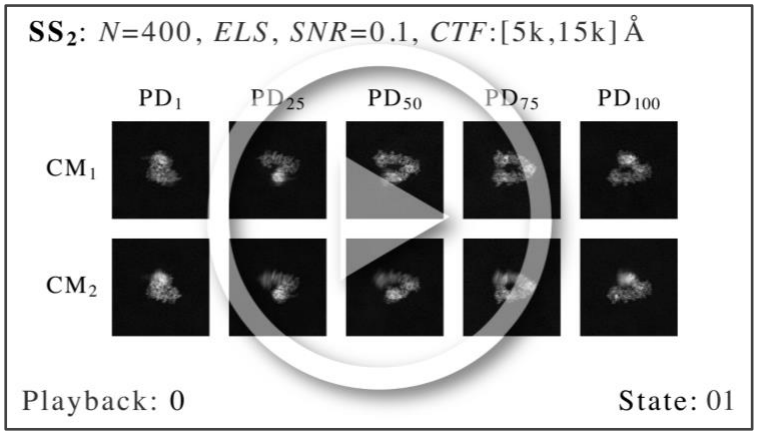
Output: CM state assignments for each point (i.e., image) in $\{\widehat{\Psi}_i \times \widehat{\Psi}_j\}$; 1D occupancy map; 2D conformational movie.

Parameters: Maximum ball-tree distance, $\mathcal{B}_r$, and minimum number of points for keeping a ball-tree cluster, $\mathcal{B}_p$ for ball tree algorithm $\mathcal{B}(\mathcal{B}_r, \mathcal{B}_p)$ [45]; the alpha parameter, α: as increased, the point cloud will be fit with a tighter bounding box by alpha shapes algorithm $\mathcal{A}(\alpha)$ [46]; even number of CM states, δ.

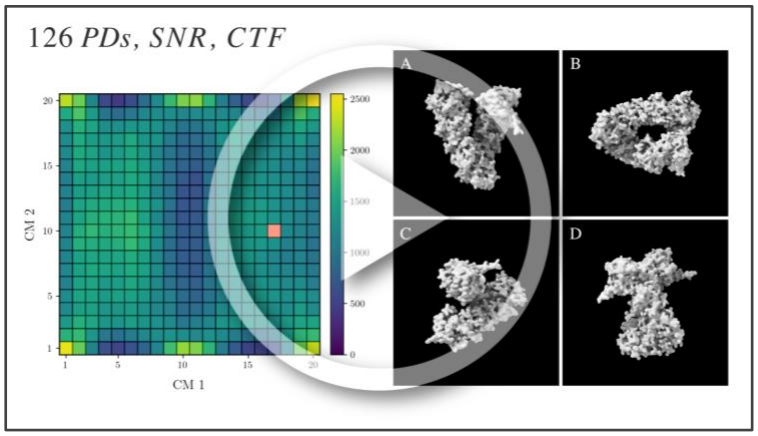
scale both eigenvectors $\{\widehat{\Psi}_i \times \widehat{\Psi}_j\}$ between $[-1, 1]$
apply $\cos^{-1}$: $\{\widehat{\Psi}_i \times \widehat{\Psi}_j\} \rightarrow \{\widehat{\Phi}_i \times \widehat{\Phi}_j\}$; $\widehat{\Psi}_{\text{fit}} \rightarrow \widehat{\Phi}_{\text{fit}}$
apply $\mathcal{B}(\mathcal{B}_r, \mathcal{B}_p)$ to temporarily prune outlier points
split $\widehat{\Phi}_{\text{fit}}$ into halves $\widehat{\Phi}^h_{\text{fit}}$ via v_x in vertex $v \coloneqq \{v_x, v_y\}$
split $\{\widehat{\Phi}_i \times \widehat{\Phi}_j\}$ into halves $\{h_1, h_2\} \in \mathcal{H}$ via v_x
$\beta_S \coloneqq [\,] \in \beta$, $1 \leq S \leq \delta$; one β_S bin per CM state S
for h **in** $\mathcal{H}$ **do**
    apply $\mathcal{A}(\alpha)$ to enclose most of h with a polygon $\mathcal{P}_h$
    reinclude previous $\mathcal{B}(\mathcal{B}_r, \mathcal{B}_p)$ outliers in h
    $\{x, y\} \coloneqq \mathcal{P}_h \cap \widehat{\Phi}^h_{\text{fit}}$, defined at outer boundary of $\mathcal{P}_h$
    $\kappa \coloneqq \{v_x, y\}$; anchor point for ray projections
    for each 2D point p **in** h **do**
        $r_p \coloneqq$ ray connecting κ and p
        project p onto $\widehat{\Phi}^h_{\text{fit}}$ via $r_p \cap \widehat{\Phi}^h_{\text{fit}}$; index via $\widehat{\Phi}^h_{\text{fit}}\{p\}$
    $\Sigma \coloneqq [\,]$
    for $\theta \in [0°, 90°]$ via 1° increments **do**
        $r_\theta \coloneqq$ ray emanating from κ
        $\mathcal{P}_h^\theta \coloneqq$ lower sub-polygon formed by $r_\theta \cap \mathcal{P}_h$
        append area of $\mathcal{P}_h^\theta$ to Σ
    if h_1 **do** $S_h \coloneqq \{1, 2, \ldots, \frac{\delta}{2}\}$ **else do** $S_h \coloneqq \{\frac{\delta}{2} + 1, \ldots, \delta\}$
    split $\widehat{\Phi}^h_{\text{fit}}$ into $\frac{\delta}{2}$ equal-area segments S_h via Σ
    for $\widehat{\Phi}^h_{\text{fit}}\{p\}$ **in** each S_h **do** $\widehat{\Phi}^h_{\text{fit}}\{p\} \rightarrow \beta_{S_h}\{p\}$
for β_S **in** β **do**
    tally $\beta_S\{p\}$ for occupancy of state S
    CTF correct images corresponding to points $\beta_S\{p\}$
    add CTF-corrected images for frame S of 2D movie
return β_S image indices, 1D occupancy map, 2D movie



Movie H4: Set of 2D movies captured along SS_2 subspaces, with each corresponding Ω_{PD} generated from images with SNR of 0.1 and $\tau = 5$. As seen in both the first and second row, as a result of the integration procedure on each corresponding CM parabola, there is a significant difference in resolution between the desired CM and the CM orthogonal to it. When 3D movies are eventually constructed from the full set of these 2D movies on S^2, pairwise information from both CMs is incorporated into each volume such that these anomalies resolve.



Movie H5: Set of 2D movies produced by ESPER from data-type IV in SS_2, corresponding to five PDs equispaced on a great circle. Unlike in Movie H4, which does not incorporate CTF, here the set of CTF-corrected and Wiener-filtered snapshots within each bin are integrated. Overall, the resolution of desired CM content in each row is superb across all states.



Movie H6: Output volumes from data-type IV: 126 PDs, SNR of 0.1 and CTF. A sequence of 69 3D density maps is shown, as seen from four orthographic views [A, B, C, D] animated along a chosen trajectory in the retrieved 2D state space. Here the 2D occupancy map before $\mathcal{R}^2$ thresholding has been supplied (in contrast to Figure 42), with all volumes reconstructed using RELION, without removal of any images in the original ensemble. Post-processing steps for display of each of these volumes included removal of dust (via the Chimera "hideDust" command with size 10) and application of Gaussian filter (via Chimera, using 1 Å standard deviations of the 3D isotropic Gaussian function).

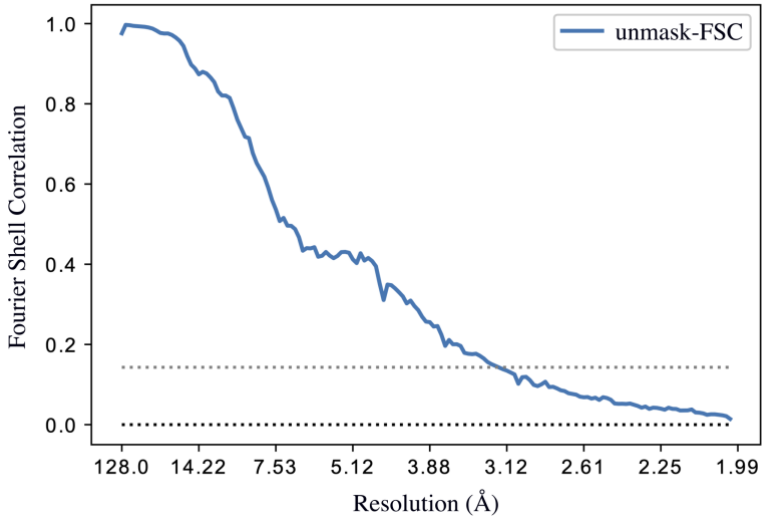


Figure H3: FSC curve comparing the state 05_10 input (ground-truth) and output (ESPER) 3D density maps. Specifically, the FSC measures the normalized cross-correlation between the two maps as measured over a series of shells in Fourier space, and is a global measure of how well the maps match.

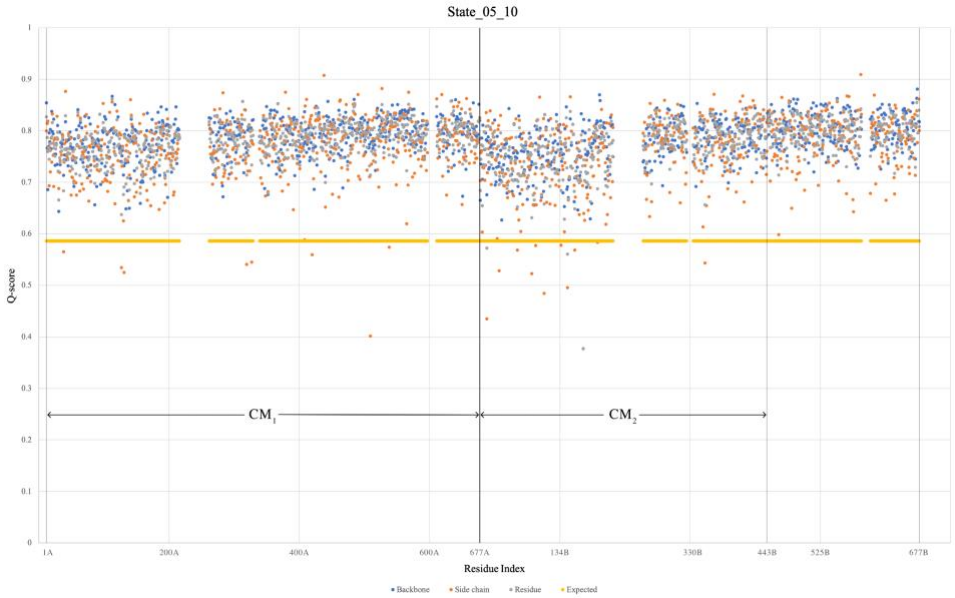
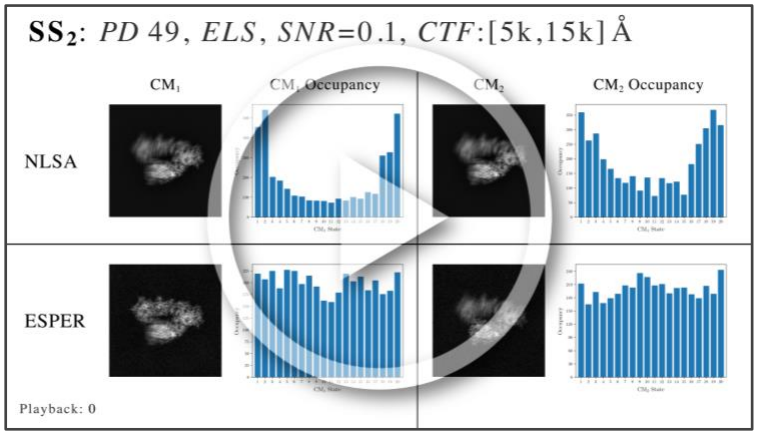
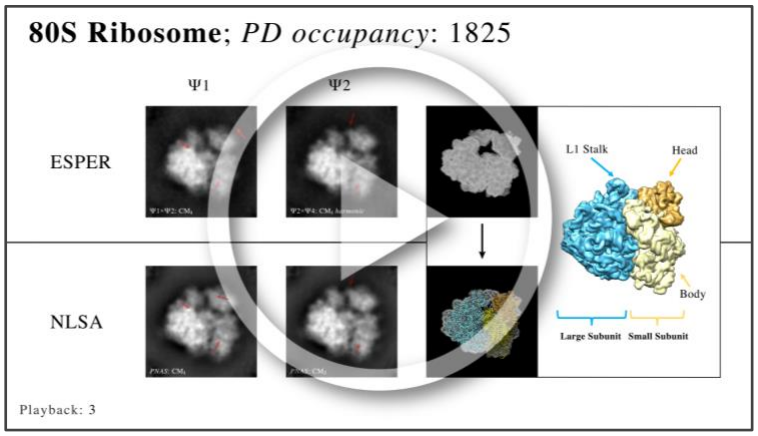


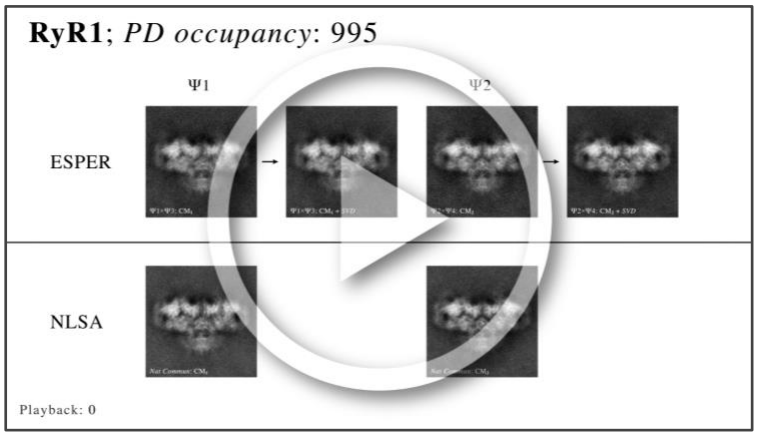
Figure H4: Q-scores for protein backbone, side chains and residues calculated using ESPER output map 05_10 with corresponding ground-truth atomic-coordinate structure. Q-scores were ascertained using the MapQ plugin for Chimera. The range of residues for each conformational motion (CM_1 and CM_2) are demarcated on chain A and B, respectively. Empty Q-scores correspond to those residue indices missing in the initial crystal structure (PDB 2CG9), whether due to insufficient resolution or electron density in the preceding study.



Movie H7: Three PDs chosen for direct comparison of 2D movies and occupancy maps produced by both ESPER and NLSA. In the first segment of Movie H7, PD_2 represents the class of extremely well-behaved embeddings, having near-perfect eigenfunction pre-alignment and irrelevant inward curling at the boundaries for both CM point clouds. Next, PD_{33} is a representative from the class of embeddings with eigenfunctions that are appreciably unaligned from the ideal eigenbasis. Finally, PD_{49} was similar to PD_2, but exhibited significant inward curling at the boundaries of its subspaces. The last segment of Movie H7 demonstrates the final 2D movies (one per eigenvector) output by NLSA for PD_{33} and PD_{49}.



Movie H8: A comparison of ESPER and NLSA as performed on the same Ω_{PD} embedding from the 80S ribosome data set. At the end of Movie H8, a schematic is provided detailing the location of ribosomal domains as observed from the current viewing direction. The caption "PNAS" is used only to match the CMs we obtained in this study using the Python ManifoldEM suite (Seitz et al., 2021b) with those described in the original study (Dashti et al., 2014). To note, from the original data set, each projection—originally or dimension 250×250 with a pixel size of 1.5 Å—was downsampled to 125×125 with a pixel size of 3 Å. The effect of this downsampling on the geometry of the manifold embeddings is negligible, and as validated by a separate ground-truth analysis, insignificant for our analysis. Finally, we aligned the particles and obtained their orientations using the "non-uniform refinement" module in cryoSPARC. Note that the caption "PNAS" is used only to match the CMs obtained in this study using the Python ManifoldEM suite with those obtained in the original study.



Movie H9: A comparison of ESPER and NLSA as performed on the same Ω_{PD} embedding from the RyR1 data set. Although only housing 976 images within a 6° angular width, this PD was chosen as an example due to the presence of a noticeable (albeit "marginal") parabolic appearance of a leading subspace (CM_1). Overall, the occupancy of PDs with smaller angular widths was insufficient for producing embeddings with robust geometric features. For the 2D movies, the ESPER method was run with $d = 10$ bins, with the first three eigenvectors of the SVD

decomposition retained for noise filtering. Note that the caption "Nat Commun" is used only to match the CMs we obtained with those described in the original study.

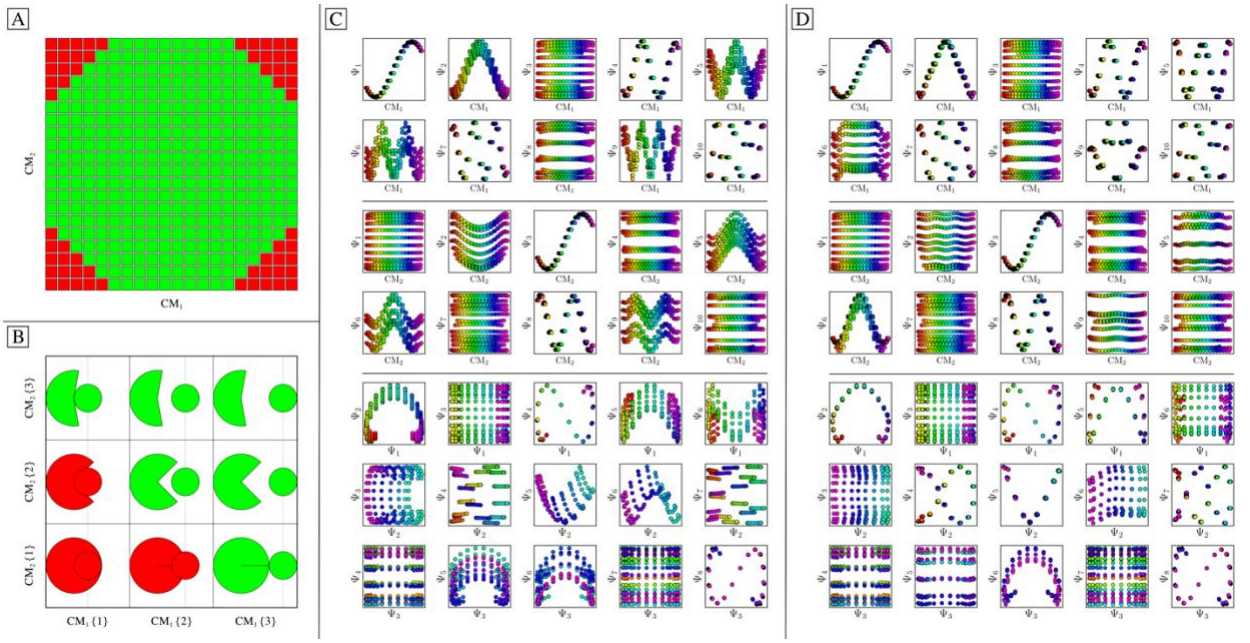


Figure H5: Analysis of the eigenfunctions [C] associated with the octagonal state space [A] of EDMs. The initial 20×20 rectangular state space is displayed in [A], where red boxes illustrate states that were removed to form an octagonal grid. The schematic in [B] provides some context for the possibility of a non-rectangular state space, which can be envisioned as a top-down view of (i) a large domain that opens and closes, and (ii) a small domain that translates left and right. Naturally, due to steric hindrance, while the larger domain is in a closed or half-closed state, the smaller domain is impeded from accessing a subset of its possible states, and vice versa. The eigenbasis obtained after application of a set of high-dimensional rotations (of dimension $d = 15$) is shown in [D]. The required operators were estimated by hand, and included several large and small transformations: $\{R_{5,6}(40°),\ R_{2,6}(-15°),\ R_{2,5}(3°),\ R_{2,9}(4°),\ R_{6,9}(20°),\ R_{6,12}(-25°),\ R_{2,11}(-6°),\ R_{9,11}(25°),\ R_{9,15}(5°),\ R_{6,11}(3°)\}$. We note these calculations are for completeness, and not a minimal set for a desired subspace. Still, we were unable to perfectly decouple the $\{\Psi_3, \Psi_6\}$ subspace. As Ψ_3 appeared well behaved in the CM_2 reference frame, it is possible that either the Ψ_6 eigenfunction was fundamentally altered due to the new boundaries, or our choice of an earlier operator trapped us in a suboptimal solution.