A path collective variable for extracting free energy profiles from cryo-electron microscopy

by

Julian David Giraldo Barreto

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Physical Sciences

Advisor: Pilar Cossio Tejada, Ph.D.
co-advisor: Johans Restrepo Cárdenas,Ph.D

Biophysics of Tropical Diseases - Max Planck Tandem Group
Faculty of Exact and Natural Sciences
November, 2021

© Julian David Giraldo Barreto
all rights reserved, 2021

Acknowledgments

I give special thanks to Dr.Pilar Cossio, Director of Max Planck Tandem Group: Biophysics of Tropical Diseases. Without her leadership, this project would not have been possible.

To the University of Antioquia, Colciencias (now MinCiencias) and the Max Planck Society for funding and supporting this work, which is part of the FP44842-292-2017 project.

I would like to thank the members of Biophysics of Tropical Diseases group for their constant comments and helpful discussions on all of the group's projects.

I would like to thank Professor Johans Restrepo for being part of the co-direction of this work and for his important contribution to my learning process.

I also specially thank to my mom and my brothers for their constant support, even from a distance.

Finally, I thank Alex H. Barnett, Bob Carpenter, and Erik H. Thiede from CCM group at Flatiron Institute for their great contributions to this work.

To my mother, my brothers and my friends

Abstract

Cryo-electron microscopy (or cryo-EM) is an experimental technique to obtain structures of biomolecules. Although single-particle cryo-EM is widely used for 3D reconstruction, it has the potential to provide information about a biomolecule's conformational variability, which leads to the underlying free-energy landscape of the system. However, cryo-EM as a single-molecule technique uses the 2D projections of individual particles with low signal-to-noise ratio (SNR), making it difficult to work directly with the raw cryo-EM data. Even though there are some methods that overcome the SNR issue, those normally need a big image data bank or are difficult to reproduce.

This work proposes a new method called cryo-BIFE (cryo-EM Bayesian Inference of Free-Energy profiles), which uses a path collective variable to extract free-energy profiles and their uncertainties from cryo-EM images. The method is tested for different realistic experimental conditions, leading to a very good performance of extracting free energy profiles for different benchmark systems using different sets of synthetic images. The results show that, to recover the underlying free energy, the SNR in the cryo-EM images and the accuracy in estimating the orientation of biomolecule's projection, are crucial factors. Then, the method is used to study the conformational transitions of a calcium-activated channel with real cryo-EM particles. Interestingly, we recover not only the most probable conformation (used to generate a high-resolution reconstruction of the calcium-bound state) but also a metastable state that corresponds to the calcium-unbound conformation. As expected for turnover transitions within the same sample, the activation barriers are on the order of k_BT.

We expect our tool for extracting free-energy profiles from cryo-EM images will enable more complete characterization of the thermodynamic ensemble of biomolecules.

Contents

Acknowledgments **3**

Dedication **4**

Abstract **5**

1 Introduction **11**
- 1.1 Problem statement . . . 13
  - 1.1.1 Objectives . . . 13

2 Theory and methods **15**
- 2.1 Cryo-electron microscopy . . . 15
  - 2.1.1 Preparing the sample . . . 16
  - 2.1.2 Irradiating the sample . . . 16
  - 2.1.3 Image acquisition and processing . . . 16
- 2.2 Free-energy landscape . . . 17
- 2.3 Collective Variables . . . 17
- 2.4 Theory: A path collective variable for extracting free energy profiles from cryo-EM images 17
  - 2.4.1 A path collective variable . . . 17
  - 2.4.2 The free-energy profile along the path . . . 19
  - 2.4.3 cryo-BIFE: a Bayesian approach for extracting the free-energy profile using cryo-EM images . . . 20
- 2.5 Methods . . . 22
  - 2.5.1 BioEM analysis . . . 22
  - 2.5.2 Markov chain Monte Carlo . . . 23
  - 2.5.3 Synthetic particles . . . 24
  - 2.5.4 Benchmark systems . . . 24
  - 2.5.5 TMEM16F: experimental cryo-EM data . . . 27

3 Results **29**
3.1 Free energy profile recovery over controlled datasets 29
3.1.1 Hsp90 chaperone 29
3.1.2 Real cryo-EM data: TMEM16F ion channel 35

4 Conclusions **39**

5 Perspectives **41**

Appendix A Appendix **42**
A.1 The BioEM likelihood 42
A.2 Input files for the BioEM analysis 42

References **44**

List of figures

2.1 Schematic representation of the path collective variable and Bayesian formalism for cryo-BIFE. 18

3.1 1D analysis of Hsp90. 30
3.2 Free-energy profile recovery for different cryo-EM conditions. 31
3.3 **Cryo-BIFE versus supervised particle-classification for 1D Hsp90 using images with SNR [0.001,0.1].** Free-energy profile recovery using cryo-BIFE (same as in Fig 3.1C blue line) and from particle classification (brown line) by using directly the BioEM likelihood (round 2), assigning each particle to the closest node, calculating a histogram for all particles, and using Boltzman's factor to extract the free energy. Cryo-BIFE outperforms standard classification because individual particle contributions are weighted by the posterior and are not assigned to a single node. 32
3.4 **HSP90 system. Comparing image sets with a wide range of SNR values.** A shoulder-shape can be identified in figures 3.1, 3.2 and 3.3. The shoulder around 0.55 is an artifact that appears when there are very low SNR images in the data set. 33
3.5 2D analysis of Hsp90. 34
3.6 Free-energy profiles from 2D images (cryo-BIFE) or 3D conformations of the VGVAPG hexapeptide. 36
3.7 Free-energy profiles from 2D images (cryo-BIFE) or 3D conformations for the semiSWEET transporter. 37
3.8 Real cryo-EM data for studying the TMEM16F Ca^{+2}- bound/unbound transition with cryo-BIFE 38

List of Tables

A1 Parameters, integration ranges and prior information for calculating the first BioEM round1. 43

List of Abbreviations

cryo-BIFE	cryo-EM Bayesian Inference of Free Energy profiles
cryo-EM	cryogenic elecron microscopy
BioEM	Bayesian inference of electron microscopy
FEP	Free Energy Profile
SNR	Signal-to-noise ratio

1 Introduction

Cryo-electron microscopy (cryo-EM) is a method that enables obtaining 3D density maps from 2D projection-images extracted from a sample immersed in vitrified ice, normally at liquid nitrogen temperature, which was irradiated with an electron dose. The main difference between cryo-EM and X-ray crystallography is that, in cryo-EM, a vitreous ice solution contains the sample in diverse configurational states, whereas in X-ray crystallography the biomolecule lies within an ordered crystal where each vertex is the same configuration.

Cryo-EM began as a low-resolution technique, however, with advances in the direct electron detection cameras [1] and improvements in the image analysis algorithms [2, 3], cryo-EM now enables resolving density maps at near-atomic resolution (< 4 Å [4]), with the highest reported resolution close to 1.22 Å[5, 6]. In this way, cryo-EM has become a widely used technique in structural biology, playing a central role in understanding biological systems of a wide range of sizes (from a few KDa to a hundreds of MDa) [7].

After obtaining the 2D projections, the common cryo-EM task is to build a 3D density map. These maps give important clues not only about the structure of the biomolecule, but also about its binding partners (such as ions or ligands). This allows for the understanding of, for example, how mutations work and, in some cases, about the interactions with new drugs [8].

For the 3D map reconstruction, Bayesian methods that analyze the cryo-EM images emerge as a popular option. These approaches can avoid some overfitting in the determination of the maps. RELION [3] and FREALIGN [9] are two examples of Bayesian 3D reconstruction methods that improve the refinement and obtain information of the molecular assemblies. These methods have been successful to determine the

3D maps of many relevant biomolecules. However, they have the disadvantage of requiring the images in the same conformation to extract a 3D map, and valuable information about the biomolecule's dynamics could be lost because of this assumption.

On the other hand, Bayesian methods can be used for conformational-ensemble determination. The absence of rigid crystal-structures in cryo-EM is of great advantage to obtain dynamical information of the sample using the 2D projections [10]. The 2D images extracted from the sample show different random particle states and orientations, giving information about the diverse configurations, and thus, making it possible to map the configurational space [11, 7]. BioEM [12], for example, takes into account a spectrum of nuisance parameters of the images, in order to assign probability values to a 3D ensemble of models with respect to the 2D experimental cryo-EM projections for discovering which of the models replicates with largest accuracy the experimental images. Therefore, using the raw data directly becomes a priority for studying heterogeneous ensembles. As a consequence, the paradigmatic vision about a rigid single structure (as seen in X-ray crystallography) must be broken towards a dynamical visualization of the biological macromolecules [13].

A way to overcome the information-loss is by using the raw experimental data with diffusion maps. This method together with machine learning algorithms was proposed by Dashti and collaborators [14]. In essence, the method selects the images belonging to the same projection direction, then it organizes the images in a "time evolving" way, then, it uses a diffusion-map algorithms with machine-learning processes to project the images over a low-dimensional manifold and classifies the conformations to obtain the free-energy landscape, this is, a map where each point identifies a configurational state of the system with a free energy value. Unfortunately, replicating this process is difficult, and the bank of images required is very large (in ref. [14] researchers use almost a million experimental images). Another limitation is that the low-dimensional space where the particles are projected is difficult to interpret, *i.e.*, it is not easy to understand what a point in this space represents.

For these reasons, some recent studies have returned to particle-classification schemes for extracting free energies using an increased number of 3D conformations in the classification. Haselbach and co-workers [15] studied the dynamics of the Human Spliceosomal B^{act} Complex by performing PCA on the reconstructed 3D volumes. The population of each sub-state along the first two PCA eigenvectors was used to extract the free-energy landscape using the Boltzmann factor. A different study assessed the motion of unbound glutamate dehydrogenase [16] through a hybrid approach that combined PCA over a molecular dynamics (MD) trajectory (to define the low-dimensional space) with the populations of four cryo-EM maps. The weights of the MD conformations and the relative occupancy of the particles were combined to produce a hybrid free-energy landscape. These methods have the advantage of mapping the free energy onto an easy-to-interpret low-dimensional space. However, PCA assumes that the motions can be modeled in a linear regime, which might not be the case for large conformational changes. Moreover, for highly flexible molecules, generating 3D maps may be challenging.

Free-energy profiling by means of reaction coordinates or collective variables (CV) has been widely used to understand biomolecular processes. CVs reduce the dimensionality of the system by projecting

the molecular coordinates onto a low-dimensional, continuous variable (note that PCA is a particular method for constructing CVs). CVs provide a simple and continuous low-dimensional projection of the free-energy landscape of complex multidimensional systems. A good CV should be able to discriminate between key regions of the underlying multidimensional free energy, such as metastable states and transition states. By constructing a free energy profile over the CV and examining features such as barrier heights, practitioners can gain insight into how a reaction takes place and how relevant conformational changes occur. Free energies are commonly extracted by evaluating the CV for each conformation, taking a histogram of the values, and relating the population of each bin to the free energy using the Boltzmann factor. However, approaches based on Bayesian methods also exist [17]. CVs have also been used with enhanced sampling techniques, such as umbrella sampling [18] or metadynamics [19], which bias the simulation along the CVs to more efficiently explore the conformational space for extracting the free-energy landscape. Along these lines, several methods [20, 21] have been proposed to extract free energies from MD simulations with CVs that use 3D maps instead of directly using the individual particles.

Inspired by the CVs from the MD community [22] and motivated to extract free energies from cryo-EM particles [14], we propose the cryo-BIFE method (cryo-EM Bayesian Inference of Free-Energy profiles), a Bayesian formalism for extracting free-energy profiles and their uncertainties from an ensemble of cryo-EM images. We apply the method to several datasets representing a diverse set of biomolecular systems, using controlled parameters and comparing with known underlying free-energy profiles. We show that under several realistic cryo-EM conditions it is possible to recover the free-energy profile using our methodology. We then apply it with real cryo-EM data to study the transition between the calcium bound/unbound states of a membrane channel. We expect that free-energy profiles from cryo-EM particles will bring new information about the metastable states, barriers, and transition states to help practitioners obtain a more complete thermodynamic characterization of the biomolecular system.

1.1 Problem statement

Extracting free-energy profiles of macromolecular systems is an important task in a wide range of disciplines. In principle, this is possible using 2D cryo-EM projections, but the current methods [14] are not easy to reproduce, and some features are difficult to understand. Therefore, the development of new, original methods that use the raw cryo-EM data to extracting free-energy profiles of macromolecular systems are still needed.

1.1.1 Objectives

General objective

Develop the mathematical formalism to calculate free-energy profiles from cryo-EM experiments.

SPECIFIC OBJECTIVES

- Understand the main aspects of the cryo-EM technique.
- Learn the BioEM method and how to use the BioEM software.
- Develop a mathematical formulation to calculate free-energy profiles using a path collective variable for cryo-EM
- Test and validate the novel cryo-BIFE method over a set of benchmark systems.
- Apply the cryo-BIFE method to obtain a free energy profile related to an experimental system.

2
Theory and methods

This chapter contains the following sections. First, an introductory section, describing the concepts of cryo-EM, free energy landscape, and collective variables.

Then, in the Theory section a path collective variable (path-CV) is defined, and then the explicit relationship between this path-CV and the free energy profile is described. After that, using the theoretical framework of Bayesian statistics, the cryo-BIFE method is presented and developed.

In the methods section, we describe BioEM and the Markov Chain Monte Carlo as tools to calculate the posterior probability $p(G|w)$ and the free energy profile, G.

Finally, in the benchmark systems section, the experimental system, the particularities of how the configurations for each structural path were chosen, the number of synthetic or real cryo-EM images used, and other features, are stated.

2.1 Cryo-electron microscopy

cryo-EM is an experimental method that irradiates with an electron beam a sample immersed in vitrified ice, normally at liquid nitrogen temperature [10]. Cryo-electron tomography and single-particle cryo-EM are two examples of the method used to analyse different samples which range from a few KDa to hundreds of MDa [7]. In the following, we describe the single-particle cryo-EM method.

2.1.1 Preparing the sample

The sample consists of a vitrified-ice aqueous solution where the system under study is immersed. Aspects such as the thickness of the ice, a support structure where the solution stays, and the density of the particles within the solution, are critical in order to obtain good enough resolution of the 2D projections. The quality of the sample is, typically, checked by a negative-stain method, obtaining images of macromolecules with high contrast (a deeper discussion of this method can be seen in [23]). In essence, it provides information about the presence of contaminants, the size and shape of the sample, and potential conformational variability, among others [24]. The layer normally consists of a carbon film put on a supported structure, such as a copper grid. The thin layer of the vitrified ice shouldn't exceed $50\ nm$, and the main hope of the cryo-EM reconstruction methods is that the sample particles are randomly orientated inside the vitrified ice. Nevertheless, this sometimes is not true, since it is possible that some systems show preferred orientations on the grid relative to the incident electron beam [10].

2.1.2 Irradiating the sample

After preparing the sample, with the microscope in vacuum conditions, it is irradiated with an electron dose, which depends on the voltage of the microscope. Typically, the electron dose is kept below $20\ e/\text{Å}$ [4], if higher doses are used there is high chance of generating radiation damage. For tomography measurements, the tilt angle of the sample is changed in order to obtain different views of the structure and win some information at the moment of the 3D reconstruction [25]. However, for single-particle cryo-EM this is not possible because of radiation damage.

2.1.3 Image acquisition and processing

The main goal of cryo-EM is to obtain images with high contrast and improved resolution, in such a way that the reconstruction of the 3D density maps has as much information as possible–typically measured with the resolution. The EM parameters depend, to some extent, on the equipment factory features. The other aspects can be set in the experiment: the contrast responds to the thickness of the ice, the defocus depends on the focus of the magnetic lens, and some lenses might have issues like spherical aberration. Also, the resolution strongly depends on the electron dose and the defocus [4].

After irradiating the sample, a 2D image micrograph with the information of the sample is recorded. The number of micrographs per day can range from 1000 to 5000 [25], and with randomly orientated particles, the information is enough to build a 3D model from the micrographs by means of the Fourier slice theorem. But, taking the micrographs is only the first part of the process; particle picking has an important role such because bad particles samples can be a problem at the moment of the 3D model reconstruction. Although fully automated and semi automated methods for particle picking exist in the literature, when there is little prior information about the structure, manual picking is preferred [4].

2.2 Free-energy landscape

Macromolecular systems are not static over the time, but they are dynamic machinery constantly changing their conformational state. However, the way they change is not completely random: the underlying nature of these conformational changes follows the free-energy landscape. Here, using the definition in ref. [26], we define the free energy landscape (also known as potential mean force, PMF) *to be that energy whose Boltzmann factor gives the probability distribution on the coordinates of interest.*

In cryo-EM, the free-energy landscape determines the structural configurations that are projected into the images. So by sampling the variety of states from the 2D cryo-EM projections, in principle, it is possible to obtain the free energy, which in practice is the population of each state. However, extracting free-energy landscapes is difficult because the systems conformations are embedded in a high dimensional space.

2.3 Collective Variables

Widely speaking, a collective variable is a function that projects a large dimensional degree of freedom space into a low dimensional space. The objective of this reduction of dimensions is to convert something abstract into human understandable and easy to visualize. For macromolecular systems, there can be as many degrees of freedom, as atoms in the structure; in general, mapping the free-energy hypersurface is an impossible task.

One of the main objectives in biophysics, or chemistry, is to find adequate collective variables (also known as reaction coordinates), to project the free-energy hypersurface, in a low dimensional map, extracting the most important thermodynamical information of the system. In the work of Branduardi *et.al* [22], for example, a path collective variable (path CV) is defined. The main idea of the Branduardi's path CV is generate a discrete path where each node of the path is a different structural configuration of the system. Then, the set of all configurational states of the system one can access (3D models obtained from a molecular dynamics, for example) are projected over the path, and from this projection, one is capable to obtain the 1D free energy profile associated to the system.

Optimal CVs should be able to discriminate the most frequented configurations of the system, whereas infrequent configurations can be related with free-energy barriers.

2.4 Theory: A path collective variable for extracting free energy profiles from cryo-EM images

2.4.1 A path collective variable

Consider a biomolecule of N atoms. Inspired by ref. [22], we will define a collective variable by projecting every possible molecular configuration onto a path in the biomolecule's configuration space. We will use $x \in \mathbb{R}^{3N}$ to denote a particular configuration (conformation). We define the CV in a manner that allows for the extraction of a 1D free-energy profile.

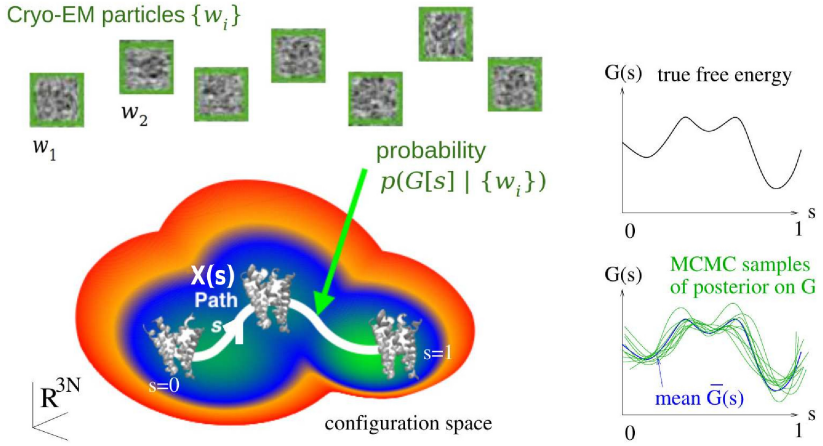

Figure 2.1: Schematic representation of the path collective variable and Bayesian formalism for cryo-BIFE. The main goal of our methodology is to determine the posterior probability distribution of free-energy profiles $G(s)$ over a given configuration space path $X(s)$, given a set of noisy cryo-EM particle (projection) images $w = \{w_i\}$ from $i = 1, ..., I$. The green graphs on the right show independent samples were drawn from this posterior, and the blue curve their mean. The black curve represents the true free-energy profile. Variation between sampled free energy surfaces arises from a detailed Bayesian model of imaging noise. The path $0 \leq s \leq 1$ is discretized using M nodes.

Let a predetermined smooth 1D path X in configuration space be parameterized by $0 \leq s \leq 1$, so that $x = X(s)$ is a particular configuration chosen to be on the path. This path should span the relevant conformational changes of the system, and thermal motion should be relatively small in all directions transverse to the path. In Figure 2.1, we show a schematic representation of the path X (white curve) that connects the relevant metastable states (basins) in the conformational space. At each configuration $x = X(s)$ one sets up transverse coordinates $z \in \mathbb{R}^{3N-1}$, so that any configuration x in a tubular neighborhood of the path may be written uniquely via a map $x = \mathcal{X}(s, z)$, where $X(s) = \mathcal{X}(s, 0)$. This means that inverse functions $S(x)$ and $Z(x)$ exist such that $\mathcal{X}(S(x), Z(x)) = x$ for all x in this neighborhood. Our CV is defined by $S(x)$, *i.e.* the parameter value s of the unique point on the path nearest to a given thermally-accessible configuration x. For all points $X(s)$ on the path, $S(X(s)) = s$ extracts their CV parameter.

In practice, one must discretize integrals (*e.g.*, for the Bayesian analysis presented below) over the parameter $0 \leq s \leq 1$. For this we use a simple M-node equispaced rule,

$$\int_0^1 f(s)ds \approx \frac{1}{M}\sum_{m=1}^{M} f(s_m), \tag{2.1}$$

which applies to smooth functions f, the parameter nodes being $s_m := (m-1)/(M-1)$. This defines a discrete set of 3D conformations (which we refer to as nodes) $x_m := X(s_m)$, that take the system from

a starting conformation x_1 to a final one x_M. Note that M is a numerical convergence parameter (the results are expected to converge as $M \to \infty$), and should be chosen large enough so that conformational changes are small between adjacent nodes. Ideally, the parameterization of the path should also have roughly uniform "speed" $|X'(s)|$, so that discrete conformations x_m are approximately evenly spaced in $\mathbb{R}^{3N}$, although satisfying this condition may be challenging in many applications. If the path is well chosen, then the assumption that the cryo-EM images come from conformations near the path is justified by the Laplace approximation in the low-temperature limit, as in path-based algorithms for MD simulations [27, 22].

The CV defined in reference [22] compares 3D conformations (*e.g.* from an MD trajectory) to the set of nodes belonging to the path X. Inspired by this, we develop the cryo-BIFE method, a Bayesian formalism to infer the free-energy profile along the predetermined path, given an ensemble of raw cryo-EM images from the same biomolecule.

2.4.2 The free-energy profile along the path

Here, we consider the biomolecule at thermal equilibrium. From Boltzmann statistics, the probability density at configuration $x \in \mathbb{R}^{3N}$ is given by

$$\rho(x) = \frac{1}{Z_0} e^{-\beta H(x)}, \tag{2.2}$$

where $H(x)$ is the system's Hamiltonian (potential energy of conformation x), and $Z_0 = \int e^{-\beta H(x)} dx$ is the full partition function. We now project this down to the CV. One may choose the map $\mathcal{X}(s,z)$ so that, at each point on the path, $\frac{\partial x}{\partial z_j}$ for each of the transverse coordinates $z_j, j = 1, \ldots, 3N-1$, are mutually orthonormal, and orthogonal to the path tangent vector $X'(s)$. Then, near to the path, the Jacobean of the map is the "speed" $|X'(s)|$ (note that $|z|^2$ then matches the squared-distance variable preferred in ref. [22]).

A change of variables gives the marginalized probability density as

$$\rho(s) = \int \delta\left(S(x) - s\right) \rho(x) dx = \frac{1}{Z_0} |X'(s)| \int e^{-\beta H(\mathcal{X}(s,z))} dz, \qquad 0 \le s \le 1, \tag{2.3}$$

where δ is the 1D Dirac delta distribution, and in the last step we used equation (2.2) and the Jacobean. Since only conformations near to the path are assumed relevant, for simplicity the Jacobean here was approximated as constant with respect to z. Note that the final integral in equation (2.3) is a partition function restricted to the "slice" transverse to X at s. It is then standard to interpret this $\rho(s)$ as the equilibrium density due to an effective 1D free-energy profile (or potential of mean force) $G(s)$ defined by

$$\rho_G(s) = \frac{1}{Z_1} e^{-\beta G(s)}, \qquad 0 \le s \le 1, \tag{2.4}$$

a 1D analog of equation (2.2) with $Z_1 = \int_0^1 e^{-\beta G(s)} ds$. Our goal is to infer the function G from a large set of 2D cryo-EM images in a statistically rigorous fashion, up to an additive offset. Note that, by equation (2.4), this is equivalent to inferring the population density ρ_G.

2.4.3 cryo-BIFE: a Bayesian approach for extracting the free-energy profile using cryo-EM images

In general, the underlying free energy for a system is unknown. However, in cryo-EM, we have access to a collection of (noisy) raw images $w := \{w_i\}_{i=1}^I$. The model for each image w_i is a noisy unknown projection of the biomolecule with an unknown configuration x taken to be independently distributed following equation (2.2). In the CV approach sketched above we restrict this to the 1D configuration path $x = X(s)$, where s is a Boltzmann-distributed random variable as in equation (2.4).

For simplicity of notation, we use the symbol G to represent the profile, *i.e.*, function $G(s)$ over $0 \leq s \leq 1$, keeping in mind that in all numerical computations it will be represented by its vector of values at the nodes, $\{G(s_m)\}_{m=1}^M$ (see the Methods). In the Bayesian approach, uncertainty about G is encoded by a *posterior* density over the space of functions. Then, by Bayes' rule,

$$p(G|w) = \frac{p(w|G)\, p(G)}{p(w)}, \tag{2.5}$$

where $p(G|w)$ is the desired posterior density over free-energy profiles induced by the observed data. $p(w|G)$ is the sampling density (or *likelihood*) of the set of all observed images w, assuming a specific free-energy profile function G. The term $p(G)$ encodes any prior knowledge about the free-energy profile. In this work, we will impose only a weak-smoothness prior, whose functional form is given in the Methods section. The normalizing constant $p(w)$, also known as the evidence, will be ignored since it is not needed for inference of G. Note that in equation (2.5), and many subsequent formulae, each term is of course conditioned on the path X, and thus one could write $p(G|w, X)$, etc. However, since X is fixed, for notational simplicity we leave this dependence implied.

We assume that the cryo-EM images are conditionally independent given G, giving

$$p(w|G) = \prod_i p(w_i|G)\,, \tag{2.6}$$

where $p(w_i|G)$ is the sampling density (likelihood) of the single image w_i given G.

Our imaging model, encoded by $p(w_i|G)$, may be interpreted as having two steps: first we draw s randomly according to ρ_G in equation (2.4), then we draw a noisy image of the 3D molecular configuration $x = X(s)$ according to the full random set of imaging parameters (orientation, translation, noise, etc). Because s is an unobserved (a.k.a. latent) variable, the likelihood of an image can be computed by

marginalizing over s, thus

$$p(w_i|G) \;=\; \int p(w_i|X(s))p(s|G)\,\mathrm{d}s \;\approx\; \frac{1}{M}\sum_m p(w_i|x_m)p(s_m|G)\,, \tag{2.7}$$

where the second step applies the quadrature, equation (2.1), and our assumption that images come from conformations near the path. The second factor in this sum is, under the Boltzmann assumption, the normalized equilibrium density (2.4) evaluated at the mth parameter node, that is,

$$p(s_m|G) = \rho_G(s_m) = \frac{1}{Z_1}e^{-\beta G(s_m)}\,. \tag{2.8}$$

The first factor $p(w_i|x_m)$ in the sum (2.7) is interpreted as the likelihood density of image w_i conditioned on a known conformation x_m. The cryo-EM imaging process is quite well understood, and considerable work has gone into evaluating such likelihoods [3, 9, 28]. Here, we will use the BioEM formalism from ref. [12], which uses a set of numerical marginalizations over all imaging parameters, analogous to (but much larger in scale than) the above one over s. See the Methods, and refs. [12, 29], for details about the BioEM calculations. We note that the present method is not limited to the use of BioEM: any other likelihood formalism (*e.g.*, those used for 3D reconstruction [3]) could be inserted.

Plugging equations (2.6), (2.7) and (2.8) into Bayes's rule $p(G|w) \;\propto\; p(G)p(w|G)$, and dropping irrelevant normalization factors, the posterior becomes

$$p(G|w) \;\propto\; p(G)\prod_i \Big[\sum_m p(w_i|x_m)\,\frac{e^{-\beta G(s_m)}}{Z_1}\Big]. \tag{2.9}$$

Here, the reader might worry that since the normalization factor Z_1 depends on G it should be retained; however, it may in fact be dropped because G is defined only up to an additive constant. In other words, only differences in $G(s)$ are meaningful.

Given a set of particles, the cryo-BIFE algorithm consists of three main steps: *i)* define a path X and discretize it with M nodes $x_m = X(s_m)$, *ii)* pre-calculate the BioEM likelihoods $p(w_i|x_m)$ for all nodes $m = 1, \ldots, M$, for every image w_i, then *iii)* use a Markov chain Monte Carlo (MCMC) method to *sample* from the posterior, equation (2.9), and from these samples—each a possible profile $G(s)$—estimate the expected value of the free-energy profile, $\bar{G}(s)$, and also its uncertainty. Steps *ii)* and *iii)* are described in the Methods. Step *i)*, defining the path, is challenging because it depends on the particular system of interest. In practice, we select a set of conformations x_m that go from one relevant state of the system to another, as is done with the CV from ref. [22]. In future work, we hope to adapt algorithms from the molecular-simulation community, such as the String method[27, 30] and Nudged Elastic Band[31], to let us determine optimal path-CVs directly from the cryo-EM data.

In the following, we validate and test cryo-BIFE over a diverse set of systems, from a conformational

change along one dimension, using synthetic images, to a membrane channel's calcium bound/unbound transition, using real cryo-EM data.

2.5 Methods

2.5.1 BioEM analysis

The likelihoods $p(w_i|x_m)$ in equation (2.9) were calculated using the BioEM algorithm [12], as follows. Given an image w_i and a 3D conformation (from a density map or atomic model) x_m, BioEM computes the probability density $p(w_i|x_m)$ that w_i is a projection of x_m. This probability was calculated by integrating the likelihood function $L(w_i|\Theta, x_m)$ (see the Appendix), weighted by prior probabilities $p(\Theta)$, over all relevant physical parameters Θ for image formation (rotation angles, displacements, CTF parameters, noise variance, normalization factor and offset [12, 29]):

$$p(w_i|x_m) \propto \int L(w_i|\Theta, x_m)p(\Theta)d\Theta. \tag{2.10}$$

The integrals over the noise variance, offset and normalization were performed analytically, and all others were computed numerically, as described in ref. [29]. The prior probabilities of the orientation angles and the displacements were taken to be uniform over the integration interval. The prior for the CTF defocus parameter was a Gaussian distribution whose center and width depended on the BioEM rounds described below. The normalization constant in equation (2.10) requires some care,since for Bayes' rule, hence equation (2.9), to be correct, the likelihood $p(w_i|x_m)$ must be normalized over the space of 2D images w_i. Instead, we used a weaker but also correct criterion that the normalization factor is merely independent of configuration x_m.

The BioEM orientational integral was divided into two stages referred to as Round 1 and Round 2, respectively. In BioEM round 1, $p(w_i|x_m)$ was calculated by integrating over a uniform orientation grid of 36864 quaternions, which was constructed following the method described in ref. [32]. The BioEM integration ranges and number of grid points for round 1 are presented in the Appendix for each system. In BioEM round 2, a finer quaternion grid of 125 points was created around the ten best orientations (*i.e.*, with the highest probability) selected from BioEM round 1. In total, a 1250 quaternion grid were used for the second BioEM orientation round. For this round, the Gaussian prior for the defocus was centered at the synthetic/experimental value of each particle and its width was 0.3 μm. This procedure is similar to that described in refs. [33, 34]; however, here we calculated BioEM rounds 1 and 2 independently for each node of the path. We used the BioEM code from ref. [29] with CPU and GPU acceleration. For one node along with the path and 10000 particles of 128×128 size, BioEM round 1 takes $\sim$ 6 hours on 24 CPU cores + 2 GPUs, and BioEM round 2 takes $\sim$ 3 hours on 24 CPU cores.

Recalling equation (2.9), one needs to evaluate equation (2.10) for every image-node pair, i.e., MI distinct evaluations. Then, to estimate the free-energy profile, we performed the MCMC algorithm de-

scribed below to draw samples from its posterior, equation (2.9).

2.5.2 Markov chain Monte Carlo

We used a Markov chain Monte Carlo (MCMC) method to draw independent samples of the free-energy profile $G(s)$ from the posterior defined in equation (2.9). Such a set of samples captures the full posterior in a much more practical fashion than trying to represent it as a function in the high-dimensional space $\mathbb{R}^M$. We found that a standard Metropolis-Hastings algorithm, sampling the unknown vector of values $\{G(s_m)\}_{m=1}^M$ at the discrete quadrature nodes, was adequate for our needs. Initial values $G^0(s_m)$ were chosen independently and uniformly at random in $[-2, 2]$, for each $m = 1, \ldots, M$. Then, each MCMC step $i = 1, 2, \ldots, N_{MC}$ comprised the following sub-steps:

- We randomly selected a node $m \in [1, M]$ with uniform probability.
- We randomly displaced the free-energy profile at the selected node $G^i(s_m) = G^{i-1}(s_m) + \delta g$ where δg was uniformly randomly chosen in $[-0.5, 0.5]k_BT$.
- We shifted the free-energy profile so that $\sum_m G^i(s_m) = 0$. (Note that the particular choice of shift here is irrelevant.)
- We evaluated the posterior in equation (2.9) using the samples $G^i(s_m)$ of this free energy, and the pre-calculated values of $\log(p(w_i|x_m))$ (described above by equation (2.10)) for all images and all nodes $m = 1, \ldots, M$. For the prior in equation (2.9), we used $p(G) = \int \lambda e^{-\lambda \mathcal{G}} d\lambda = 1/\mathcal{G}^2$, where $\mathcal{G} = \sum_{m=1}^{M-1} (G(s_{m+1}) - G(s_m))^2$, which is a standard ℓ^2 smoothness prior on the discrete differences, marginalized over the precision parameter λ.
- From this, the log-acceptance probability of the proposal was computed (here we omit s for notational simplicity, so that G may be thought of as a vector in $\mathbb{R}^M$):

$$A(G^i, G^{i-1}) \;:=\; \log\big(p(G^i|w)\big) - \log\big(p(G^{i-1}|w)\big), \tag{2.11}$$

- We chose a uniform random number $u \in [0, 1]$. Then, if $\log(u) \leq A(G^i, G^{i-1})$, the move was accepted, otherwise it was rejected (in which case $G^i = G^{i-1}$).

This procedure was iterated well beyond the time by which the distribution over samples has converged. For the systems analyzed in this work, we ran $R = 8$ independent MCMC chains each with a total of $N_{MC} = 200000$ steps. The expected value of the free energy at each node was calculated using all samples $i = 1, \ldots, R N_{MC}$, that is,

$$\bar{G}(s_m) = \frac{1}{R N_{MC}} \sum_i G^i(s_m). \tag{2.12}$$

Finally, since it is assumed that the nodes adequately discretize a continuous path, to recover a continuous function $\bar{G}(s)$, we fitted a cubic spline through the values $\{\bar{G}(s_m)\}_{m=1}^{M}$ with knots being the nodes s_m. Because only free-energy differences are relevant, we shifted $\bar{G}$ such that its minimum was zero. The credible interval for each node was calculated at 5% and 95% of the resulting empirical distribution. We performed the R-hat diagnostic test[35], which compares the inter-chain variance to the variance within each chain to monitor convergence of the MCMC using the arviz package [36]. R-hat values ≤ 1.1 indicate convergence of the sampling.

The MCMC code was written in python3.5. It was optimized with the numba compiler, taking approximately 2 hours on 24 CPU cores for $I = 13000$ particles, $M = 20$ nodes, and $R = 8$ replicas each with $N_{MC} = 200000$ MC steps.

2.5.3 Synthetic particles

We used a modification of the BioEM program [29] to generate the synthetic cryo-EM particles following similar ideas to those described in ref. [37]. Each image was created by coarse-graining the molecular configuration (*e.g.* one taken from an MD simulation) on the residue level. Each residue was represented as a sphere with a corresponding radius and number of electrons [12]. The contrast transfer function (CTF) was modeled on top of the ideal image given a defocus, amplitude and B-factor (for details see the SI of ref. [12]). For the synthetic particles, the amplitude was 0.1 and the B-factor was 1Å. Gaussian noise was added on top of the CTF convoluted image. The standard deviation of the noise was determined (as in ref. [37]) using the SNR and variance of the image without noise (calculated within a circle of radius 40 pixels centered at the box center). All synthetic images were of box size 128×128 pixels, however, the pixel size varied for each system.

2.5.4 Benchmark systems

Hsp90 system

The Hsp90 chaperone is a flexible protein involved in several biological processes related to protein folding [38]. When bound to certain ligands, its conformational landscape can be approximated by two relative motions of its chains (A and B) [37]. The Hsp90 dynamics was reduced to a 2D dimensional phase space, where both chains are rotated in mutual normal directions and perpendicular to the axis of symmetry. In this work, we first assessed conformations from just one degree of freedom (1D analysis), and then we assessed images from conformations belonging to the 2D conformational space (2D analysis).

To generate the conformations for the first degree of freedom (1D case), we started from the closed state (PDB ID 2cg9[39]), removed the ATP ligand and residues 1-11 to avoid overlapping crashes. Chain B was fixed and chain A was rotated at 1° steps around the center of mass of residues LEU674-ASN677, up to 20° from the starting position, generating 20 conformations along this degree of freedom (denominated CMA motion [37]). These 20 conformations were used to define the path for the 1D analysis

(Figure 3.1A). Along this reaction coordinate, we proposed a synthetic free energy (which determines the population occupancy) given by

$$e^{(-\beta G_{true}(s))} = e^{(\frac{-(19s-6)^2}{8})} + \frac{1}{3}e^{(\frac{-(19s-15)^2}{18})} \text{ for } 0 \leq s \leq 1 \tag{2.13}$$

This ground truth-free energy is shown as a black solid line in Figure 3.1C. Using this synthetic population for the conformations along the path, we generated 13333 synthetic images of pixel size 2.2Å with uniformly distributed random orientations in $SO(3)$, SNR in $\log_{10}[0.001, 0.1]$ and defocus in [0.5,3] μm.

For the 2D conformational landscape, we add a new rotation. Starting from each rotated chain A from the 1D case, residues ILE12-LEU442 of chain B were rotated in 2° steps around the center of mass of residues LEU442-LEU443, in the normal direction to the plane generated by the 1D movement of chain A and the axis of symmetry. This normal motion mode was referred to as CMB [37]. In total, 400 models were generated corresponding to 20×20 rotations. We proposed a 2D synthetic free energy given by

$$e^{(-\beta G_{true}(u,v))} = e^{(\frac{-(u-6)^2}{18} - \frac{(v-6)^2}{10})} + e^{(\frac{-(u-15)^2}{18} - \frac{(v-15)^2}{10})}, \tag{2.14}$$

where (u, v) are the CMA, CMB models respectively. This distribution was characterized by two minima localized at models $(6, 6)$ and $(15, 15)$ separated by a barrier of around $2k_BT$. We generated 6800 synthetic images of pixel size 2.2Å with uniformly distributed random orientations in $SO(3)$, SNR in $\log_{10}[0.01, 0.1]$ and defocus in [0.5,3] μm. For this case, we defined three paths: CV1 is a good reaction coordinate that passes through the minima and transition state following the function $CMB = CMA$ (black dashed line Figure 3.5B), CV2 has model $CMA = 10$ fixed and CMB varying (orange dashed line Figure 3.5B) and CV3 has CMA varying and model $CMB = 10$ fixed (green dashed line Figure 3.5B).

3D ensemble of the hexapeptide VGVAPG

We used the conformational ensemble of the hexapeptide VGVAPG from a long all-atom MD simulation in explicit solvent. GROMACS [40] was used to perform a 230 ns MD simulation. The initial conformation was extracted from the crystal structure of the Ca6 site mutant of Pro-SA-subtilisin [41] with PBD code 3VHQ (residues 171 to 176) [42]. The peptide was solvated with a cubic water box, centered at the geometric center of the complex with at least 2.0 nm between any two periodic images. The AMBER99SB-ILDN [43] force field and TIP3P water model were used [44]. Minimization was done with the steepest descent algorithm and stopped when the maximum force was $\leq$ 1000 kJ/mol·nm. Periodic boundary conditions were used. We performed a 100 ps equilibration in an NVT ensemble using the velocity rescaling thermostat [45] followed by a 100 ps equilibration in an NPT ensemble using Parrinello-Rahman barostat[46]. The MD production run was performed without restraints, with a time step of 2 fs in an NPT ensemble at 300.15 K and 1 atm. We extracted MD snapshots (or frames) every

40 ps, obtaining 5688 conformations.

We selected ten conformations to create the path such that the nodes covered the relevant conformational changes of the system. To do so, we use the end-to-end distance of the peptide, *i.e.*, the distance between the nitrogen atom of the N-terminus, and the carboxyl carbon of the C-terminus [42].The path was created by selecting ten conformations from the MD with equally spaced end-to-end distances between successive nodes of 1.8 Å. The path is shown at the bottom of Figure 3.6A, and it was used both with the path-CV[22] and cryo-BIFE. The path-CV was calculated using the RMSD between all the MD frames and the ten nodes belonging to the path with parameter $\lambda = 50\text{Å}^{-2}$ (using equation 8 of ref. [22]). To calculate the free-energy profile, we computed the value of each CV for all MD conformations, took the histogram (with a number of bins equal to the number of nodes along the path), and then estimated the free energy using the Boltzmann factor and the histogram bin populations.

From each MD conformation, we generated a synthetic image with pixel size of 0.3 Å and with uniformly distributed random orientations in $SO(3)$, SNR in $\log_{10}[0.01, 0.1]$ and defocus in [0.1,1.0] μm. Using the 5688 synthetic images and the same ten nodes of the path, we performed the cryo-BIFE analysis.

Path-CV for semiSWEET

We used the conformations of the SemiSWEET membrane protein from eight unbiased MD trajectories that present a conformational change from the outward-open state to the inward-open state [47]. 1761 snapshots of the trajectories were taken every 1ns. We used these conformations as a reference ensemble. The path-CV of ref.[22] was used to calculate the reference free-energy profile with the same parameters as described for the VGVAPG hexapeptide.

To select the nodes of the path, we performed a conformational clustering of the 1761 MD snapshots. The GROMACS [40] *g_cluster* tool was used with the RMSD defining the distance between conformations. The clustering was performed using the single-linkage algorithm with an RMSD cutoff of 1.3Å, resulting in 39 clusters. The cluster-center is the conformation with the smallest average distance to all conformations belonging to the cluster. We compared each cluster center to the outward-open and inward-open crystal structures using the RMSD. 12 cluster centers with the quasi-equidistant differences of RMS, between successive nodes, to the outward-open state, were selected.

Synthetic semiSWEET images

To generate the synthetic images, we used the SemiSWEET conformation together with a nanodisk (*i.e.*, lipid) belt of 25Å centered at the center of mass of the protein and extracted from the MD simulation. To coarse-grain the nanodisk and imitate the effects of averaging, the heavy atoms of lipids were modeled with a 3Å radius and 10 electrons. We generated three images from each MD snapshot (protein and nanodisk) of pixel size 0.7Å with uniformly distributed random orientations in $SO(3)$, random SNR $\in \log_{10}[0.01, 0.1]$ and random defocus $\in [0.5, 1.5]\mu m$. A nanodisk belt was also included for each

node of the path. We note that for real datasets the lipid nanodisks will have larger uncertainties because of their variable shape and size.

2.5.5 TMEM16F: EXPERIMENTAL CRYO-EM DATA

CRYO-EM PARTICLES

The cryo-EM particles of the TMEM16F membrane channel used to generate the calcium bound state [48] from the EMPIAR dataset [49] with code EMPIAR-10278 were used. See ref. [48], for information about the experimental conditions. The images were recorded with a pixel size of 1.059Å box size of 256×256 pixels, with defocus values within the interval $[0.5, 2.7]$ μm. For this work, we randomly selected 15000 images from this Ca^{+2}-bound (Digitonin_Ca) set. Note that these images represent the entire set and not only those used for the final reconstruction. Since only 13% of the particles from the EMPIAR-10278 set are used to create the Ca^{+2}-bound reconstruction[48], our hypothesis is that not all imaged particles belong to this state. Our aim was to extract a free-energy profile from the Ca^{+2}-bound to the Ca^{+2}-unbound states using only the cryo-EM particles from the Ca^{+2}-added set.

STEERED MD FOR CREATING THE TMEM16F PATH

To generate the path, we used steered MD simulations from the Ca^{+2}-bound to the Ca^{+2}-unbound state. The simulations were performed as follows. We started from the Ca^{+2}-bound structure (PDB ID 6p46). Since the structure has atoms missing, we added these using the Swiss model webserver [50]. We note that because some residues have to accommodate to fit the missing residues the full atom structure was not identical to the PDB. Starting from the full atom model of 6p46, we added the membrane using CHARMM-GUI [51], in a 3:1:1 ratio of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC), 1-palmitoyl-2-oleoylsn-glycero-3-phosphoethanolamine (POPE), and 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (POPS), respectively. A box size of $16.8076 \times 16.8076 \times 17.2012nm$ was used with periodic boundary conditions and 122923 TIP3P water molecules were inserted. We used the GROMACS program [40] with the CHARMM36M force field [52]. The temperature was controlled in the simulation through Berendsen thermostat at 300 K, whereas the pressure was controlled with the Berendsen barostat at 1.0 atm [53]. The energy was then minimized using the steepest descent algorithm and stopped when the maximum force was $\leq$ 1000 kJ/mol·nm. We used the leapfrog algorithm to propagate the equations of motion. The long-range electrostatic interactions are calculated using a PME scheme with a 1.2 nm cutoff. We performed two consecutive equilibrations, of 125 ps each, in an NVT ensemble with a time step of 1 fs. Then, we performed two equilibrations in an NPT ensemble, where the first was of 125 ps and time step of 1 fs, and the last was of 1.5ns, with a time step of 2 fs. For the equilibration in the NPT ensemble, the pressure coupling was of semi-isotropic type. The backbone atoms of the protein were restrained throughout the equilibration runs.

After the MD equilibration, we performed steered MD simulations [54] using the GROMACS program [40] patched with the PLUMED 2.5 library [55]. The first target structure for the steered MD was the Ca^{+2}-unbound state (PDB ID 6p47). We used the RMSD of the C_α atoms to steer the dynamics between the initial structure and the target structure. The steering harmonic potential had an initial force constant of 5000 and ending at $260000\,kJ/mol/nm^2$. We noticed that a threshold of 0.2Å in RMSD to the Ca^{+2}-unbound reference was reached very quickly, in less than $1ns$. A second steered MD simulation was needed to go from the initial system (all-atom system) to the 6p46 PDB structure. This steered MD used the same parameters mentioned before. We also ran two short ($1ns$) unbiased MD simulations starting from each state (*i.e.*, closest conformation to PDB 6p47 and 6p46). These trajectories allowed us to build a path from the Ca^{+2}-bound to the Ca^{+2}-unbound states. We used the C_α-RMSD to the Ca^{+2}-bound state to select 19 nodes, where successive nodes are as equidistant as possible (see Figure 3.8B). To mimic the detergent in the cryo-EM images, we included a membrane nanodisk surrounding each node. It was taken from the lipids from the MD simulations, centered at the center of mass of the protein and of 50Å radius. The nanodisk was modeled in a coarse-grained manner, similarly to the SemiSWEET transporter.

3 Results

To understand the effects of the physical parameters (*e.g.*, those involved in the image formation process) for recovering free-energy profiles with cryo-BIFE, we designed several control systems where the projections are generated synthetically following the ideas of ref.[37]. The first system consists of conformations of the Hsp90 chaperone representing a low-dimensional (1D-2D) conformational space. The analysis is then extended to more realistic ensembles from MD simulations. Lastly, we apply cryo-BIFE to experimental cryo-EM data. To this end, we chose raw images of TMEM16F, a membrane channel and lipid scramblase [48] available at the EMPIAR databank [49].

3.1 Free energy profile recovery over controlled datasets

3.1.1 Hsp90 chaperone

Hsp90 (a heat shock protein) is a chaperone involved in the folding process of several kinases, transcription factors, and steroid hormone receptors [38]. This protein consists of two chains (A and B, containing 677 residues each) forming a V-like shape. Although Hsp90 is flexible, in the presence of certain ligands (*e.g.*, ATP) its conformational space can be reduced to a few degrees of freedom that go from an open to a closed state of the chains. Following the ideas described in ref. [37], we reduced the open-closed dynamics of the Hsp90 into a one (1D) and two (2D) dimensional phase space where both chains are rotated in mutual, normal directions and perpendicular to the axis of symmetry (see the Methods).

Free-energy profile recovery for a 1D conformational change

In Figure 3.1A, we show a 1D conformational change of Hsp90, where chain B is fixed and chain A is rotated from the closed state to the open state (denoted by CMA). We define the path using twenty conformations, equally spaced by 1° in the rotation angle. The underlying synthetic free-energy profile (*i.e.* constructed reference) along the path is shown as a black line in Figure 3.1C. We generated around 13300 synthetic images from the predetermined population of the twenty conformations (given by the Boltzmann factor of the constructed reference free energy). The synthetic images have a uniform random signal-to-noise-ratio (SNR) $\log_{10}([0.001, 0.1])$, defocus [0.5,3] μm and orientation angles (see the Methods). Examples of the synthetic particles are shown in Figure 3.1B.

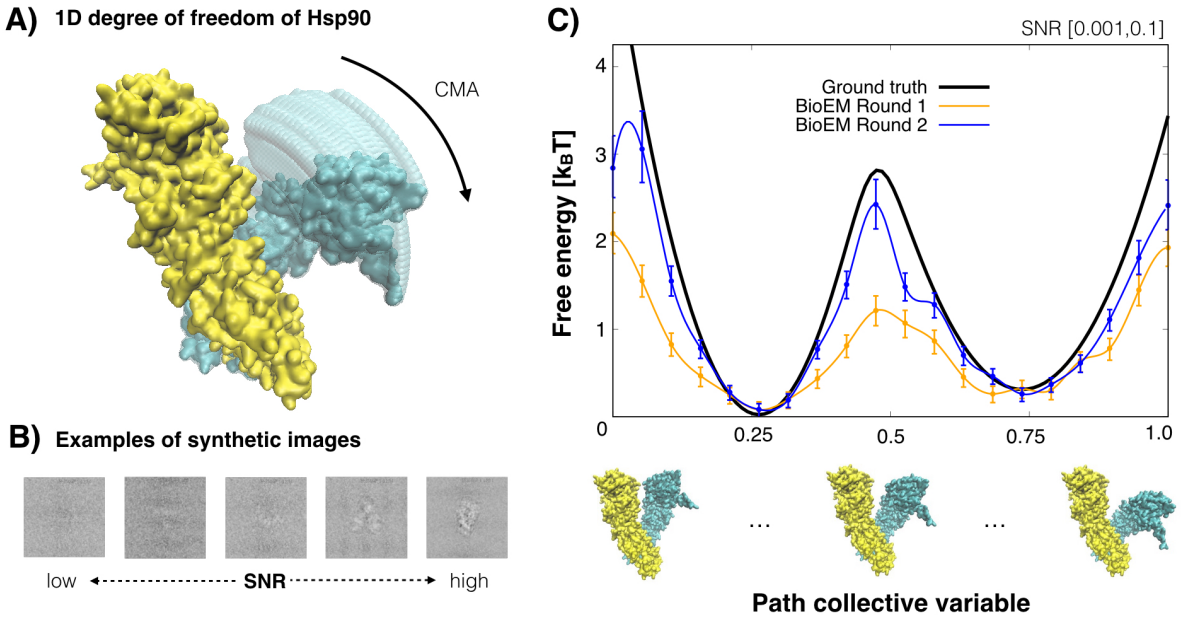

Figure 3.1: 1D analysis of Hsp90. A) Movement of Hsp90 along the single degree of freedom (CMA). The rotation of chain A relative to a fixed chain B. B) Examples of the synthetic images with varying SNR between [0.001,0.1]. C) Free-energy profiles along the path for the entire set of images recovered from cryo-BIFE. The constructed reference free-energy profile is shown in black. The expected free energy profile using cryo-BIFE is shown for BioEM orientation rounds 1 and 2 in orange and blue, respectively. The R-hat test for the MCMC yielded 1.000 and 1.001 for BioEM round 1 and 2, respectively. The bars show the credible interval at 5% and 95% of the empirical quantile at each node. A cubic spline is used to fit the expected free-energy profile, providing a smooth profile.

To apply cryo-BIFE, we first precalculated the BioEM probabilities for the nodes along the path and all synthetic images for two BioEM rounds of orientation estimation (see the Methods). The MCMC sampling strategy described in the Methods was applied to extract the expected $\bar{G}(s)$ and the credible

interval at 5% and 95% of the empirical quantile at each node. Figure 3.1C, shows the results of $\bar{G}(s)$ using all particles for the first and second BioEM rounds of orientation estimation. Note that the second round was more accurate than the first. This was also reflected in the recovery of the free-energy profile $\bar{G}(s)$, where the second round had a much better performance. This suggests that the pose accuracy of the particles is crucial for extracting an adequate free-energy estimate. The results from BioEM round 2 show that cryo-BIFE was able to recover the free-energy profile for a wide range of SNRs and defocus. Interestingly, the credible intervals widen for higher free-energy values, *i.e.,* near the barrier, where there are fewer particles and the error is expected to be larger.

The performance of the method for different cryo-EM conditions was then studied. In Figure 3.2A, the particle set was divided in two: high SNRs from [0.01,0.1] and low SNRs from [0.001,0.01], each with an equal number of particles ($\sim$6600 each). The expected free energy calculated from cryo-BIFE is shown for the high and low SNRs sets (light blue and green, respectively) for the second BioEM orientation round. The expected free energy was also compared to $\bar{G}(s)$ using the entire set (blue line). We observed a poor recovery for the low SNR set [0.001,0.01] and large errors, whereas the high SNR set behaved well. Interestingly, the free-energy estimate for the entire particle set (SNR [0.001,0.1]) was slightly worse than for the high SNR set but much better than the low SNR set. The reason for this is that the Bayesian posterior (equation (2.9)) naturally weighs the contribution of each particle and particles with high SNR contribute much more weight to the posterior. If particles with even higher SNR are added, the free-energy profile recovery is better, and for example, artifacts like the shoulder around s =0.55 vanish.

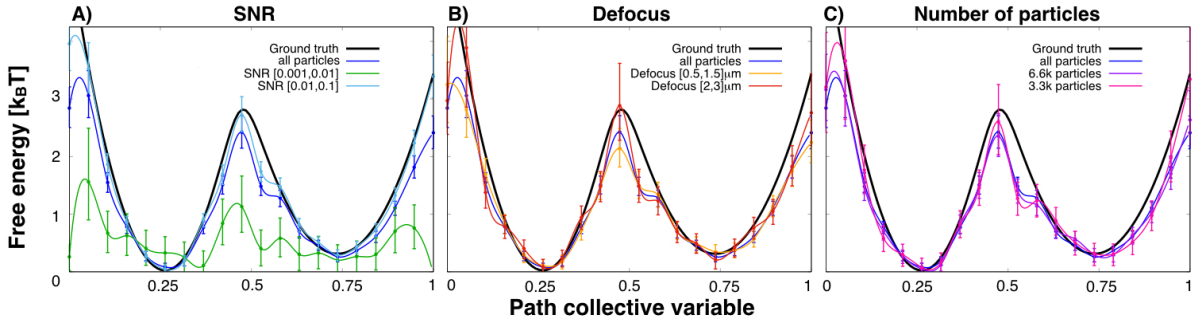

Figure 3.2: Free-energy profile recovery for different cryo-EM conditions. A) Particles grouped by SNR from [0.01,0.1] (cyan) and from [0.001,0.01] (green). Each subset contained around 6600 particles. B) Particles grouped by defocus. Sets with small defocus [0.5,1.5]μm (orange) and large defocus [2,3] μm (red). Each subset contained around 5300 particles. C) Particle subsets with a different number of particles: 3300 (pink) and 6600 (purple). For reference, the constructed reference and expected free-energy profiles using all particles are shown in black and blue, respectively. The R-hat test for the MCMC yielded values < 1.01 for all cases. The bars show the credible interval at 5% and 95% of the empirical quantile at each node. The results are for the second BioEM round of orientation estimate.

In Figure 3.2B, the effects of the defocus by grouping the particles with small defocus [0.5,1.5]μm (orange line) and large defocus [2,3]μm (red line) were analyzed. The results for the large defocus were

slightly better, but these have large errors around the barrier. The number of particles needed to recover the free-energy profile was also studied. In Figure 3.2C, the results are shown for sets with 3300 (pink line) and 6600 (purple line) particles. In agreement with previous results for 3D map validation [34], just a small set of particles ($\geq$ 3000) randomly picked from the entire set is able to reproduce the underlying statistics. Contrary to 3D refinement, where large numbers of particles are required, our results indicate that conformational variability can be captured from a small set of particles.

Cryo-BIFE has several advantages over standard particle-classification methods for calculating the populations (or equivalently the free-energy profile). These classification methods treat each particle equally, whereas cryo-BIFE weighs them differently (*e.g.*, depending on their SNR). Moreover, most methods assign each particle to a single node along the path and calculate a histogram over all particles to extract the populations. In Figure 3.3, this analysis (using the BioEM likelihood) was compared to the cryo-BIFE results for the 1D Hsp90 data with a wide range of SNR [0.001,0.1]. These results show that cryo-BIFE outperforms standard classification because individual particle-contributions are weighted by the posterior and are not assigned to a single node.

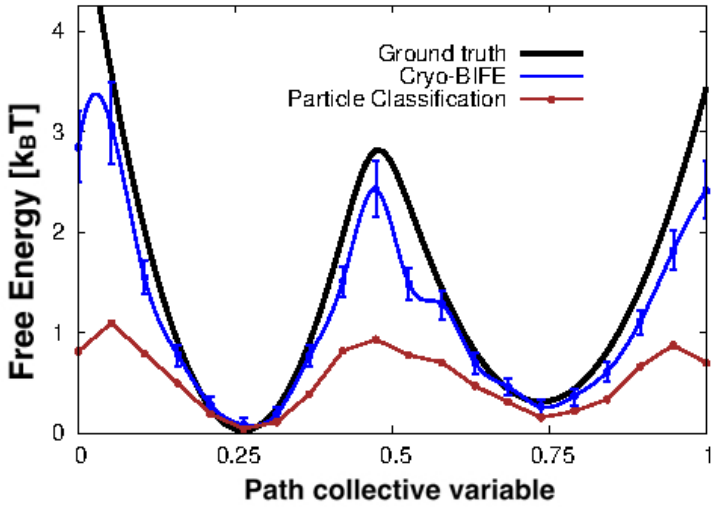

Figure 3.3: Cryo-BIFE versus supervised particle-classification for 1D Hsp90 using images with SNR [0.001,0.1]. Free-energy profile recovery using cryo-BIFE (same as in Fig 3.1C blue line) and from particle classification (brown line) by using directly the BioEM likelihood (round 2), assigning each particle to the closest node, calculating a histogram for all particles, and using Boltzman's factor to extract the free energy. Cryo-BIFE outperforms standard classification because individual particle contributions are weighted by the posterior and are not assigned to a single node.

The shoulder-shape artifact around 0.55 for the HSP90 system is due to the particles with low SNR that have a similar magnitude for many nodes along the path. In Figure3.4, we show the cryo-BIFE results for two-particle sets: with SNR [0.001,0.1] (blue) and a new set with SNR [0.01,1] (red). The shoulder

vanishes for the set that has particles with very high SNRs, suggesting that it is an artifact of the low SNR images.

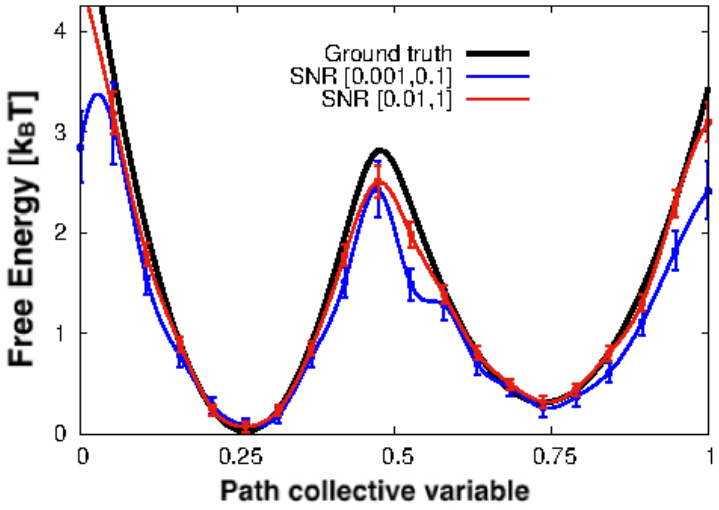

Figure 3.4: HSP90 system. Comparing image sets with a wide range of SNR values. A shoulder-shape can be identified in figures 3.1, 3.2 and 3.3. The shoulder around 0.55 is an artifact that appears when there are very low SNR images in the data set.

2D conformational change of Hsp90

As described in ref. [37], Hsp90 is also characterized by a second degree of freedom; the rotation of chain B relative to the 1D rotation of chain A (see Figure 3.5A, and the Methods). A synthetic 2D underlying free-energy surface was generated, shown in Figure 3.5B, with an energy barrier of around $2k_BT$. Given the imagining conditions in cryo-EM experiments, free-energy barriers around this range are expected. We generated 6800 synthetic particles, using the population given by the Boltzmann factor of constructed reference free energy, with SNR [0.01,0.1], defocus [0.5,3]μm and random orientations in $SO(3)$ (see the Methods).

To study the effects of the path-CV, we defined three paths. The black dashed line (CV1) in Figure 3.5B shows a good path-CV that passes along the relevant basins and the transition state of the system. In contrast, the orange and green dashed lines in Figure 3.5B (CV2 and CV3, respectively) are able to discriminate between the states (*i.e.*, good order parameters) but are not ideal reaction coordinates because they underestimate the barrier. In Figure 3.5C, we compare the expected free-energy profile extracted with cryo-BIFE to the constructed reference (given by equation (2.4)) along each path. Relatively good agreement between the underlying profile and the extracted free energy using the cryo-EM images along the three paths was observed. However, using only CV1, the metastable states of the system, the transition

state, and true barrier height were recovered. Conversely, using non-ideal CVs, *e.g.*, CV2 and CV3, the barrier can be underestimated. In extreme cases, the identification of the metastable states could also be lost. We note that these are artifacts caused by choosing a poor projection direction, and are not the result of using 2D images. This highlights the importance of choosing an adequate path-CV.

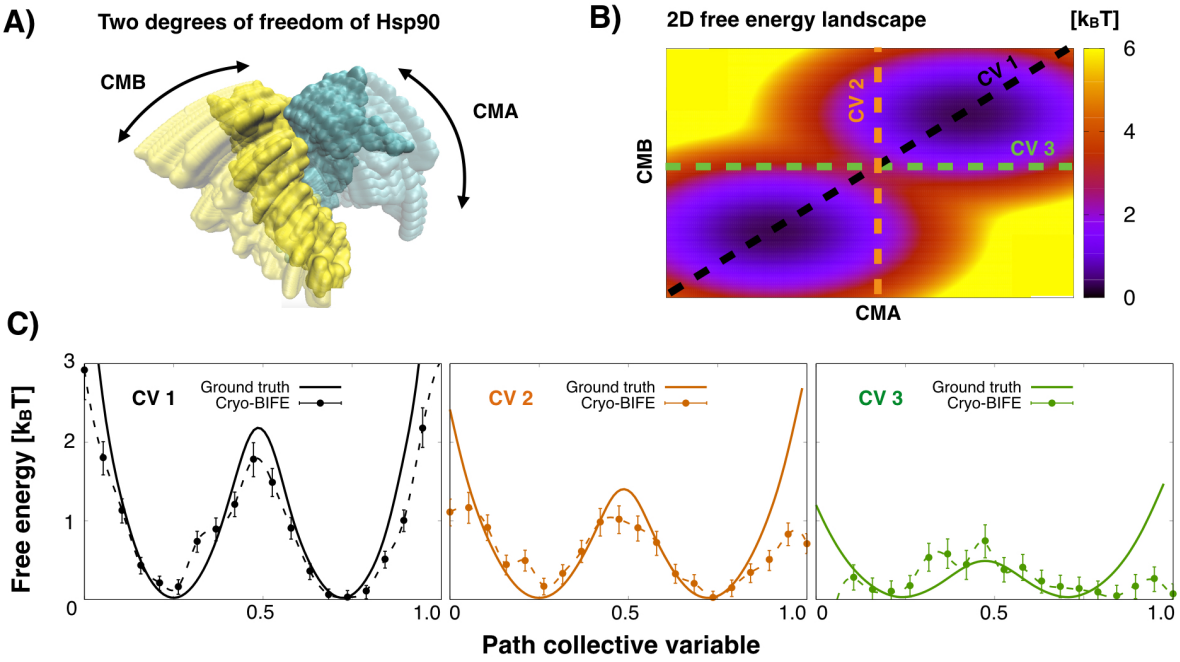

Figure 3.5: 2D analysis of Hsp90. A) Two degrees of freedom of Hsp90 along the CMA and CMB rotation directions (see the Methods). B) Constructed reference free-energy surface along CMA and CMB directions. Black (CV1), orange (CV2) and green (CV3) dashed lines show three paths used for the cryo-BIFE analysis. C) The free-energy profiles along these three path CVs, extracted with cryo-BIFE using synthetic particle images (dashed lines), are compared to the constructed reference projected profiles (solid lines). The R-hat test for the MCMC yielded values < 1.003 for all cases. The bars show the credible interval at 5% and 95% of the empirical quantile at each node. The results are for the second BioEM round of orientation estimate.

Cryo-BIFE over conformational ensembles

MD simulations of the VGVAPG hexapeptide have been extensively used to test methods, such as Girsanov reweighting [42]. The peptide has opposite charges at its extremes and exhibits a conformational change between an open state and a closed state (see the Methods). Here, we will compare the free energy extracted from the 3D ensemble to one estimated by cryo-BIFE using 2D particles with the same path (Figure 3.6A). The path was created by selecting ten conformations from the MD with equally spaced end-to-end distances between successive nodes (see the Methods). To calculate the free energy from the

3D conformations, we used the path-CV proposed by Branduardi *et. al.* [22] with the RMSD as a metric. This path-CV was evaluated for each MD conformation, then a histogram was taken and the free energy was calculated via Boltzmann's factor and the population of each histogram bin. For cryo-BIFE, we used a set of 5688 synthetic images generated from the MD ensemble. The synthetic images had uniformly distributed random SNR, defocus and orientations (see the Methods). Cryo-BIFE was applied to extract the expected $\bar{G}(s)$ along the same path used for the 3D conformations. In Figure 3.6B, the free-energy profiles from cryo-BIFE and the path-CV [22] were compared. The end-to-end distance curve is added as a reference. The difference is that cryo-BIFE extracts the FE profile from 2D cryo-EM images, whereas the path-CV uses 3D conformations (Figure 3.6A).

To investigate whether cryo-BIFE is able to resolve the free-energy profile of membrane proteins with nanodisk belts (as in the cryo-EM experiment), and small conformational changes (< 4Å), we attempted to recover a free-energy profile from synthetic images of the semiSWEET transporter generated from MD configurations (see next section). In conjunction with our results on the VGVAPG hexapeptide, they demonstrate that cryo-BIFE is able to recover the free-energy profile from 2D cryo-EM projections for a realistic ensemble.

Cryo-BIFE analysis for MD conformations of the semiSWEET transporter

The semiSWEET transporter is a membrane protein that transports sugar between cell membranes. Several unbiased MD simulations [47] of this transporter were perform starting from the outward-open conformation. Eight trajectories showed a conformational change to the inward-open conformation. We thought it interesting to investigate if cryo-BIFE could resolve the free-energy profile of membrane proteins with nanodisk belts (as in the cryo-EM experiment), and small conformational changes (< 4Å). We used these MD snapshots to generate a synthetic ensemble of semiSWEET conformations, which we used as a reference. We generated 5280 synthetic images from this ensemble having each a nanodisk belt (see chapter 2). To define the path, we clustered the MD conformations, and selected successive nodes that had a quasi-equidistant difference in RMSD to the outward-opened state (see chapter 2). In Figure 3.7, we compare the free-energy profile from the 2D images to that from the 3D ensemble using Branduardi's path-CV [22]. A relatively good agreement between the profiles around the minimum was found, however, for the high free-energy regions, the agreement was not as good.

3.1.2 Real cryo-EM data: TMEM16F ion channel

TMEM16F is a membrane channel and lipid scramblase that is activated by calcium binding. In ref. [48], cryo-EM experiments using different Ca^{+2} conditions and membrane/detergent compositions were performed to resolve TMEM16F's Ca^{+2} bound and unbound states. The cryo-EM particles under different conditions are available at the EMPIAR [49]. In this work, we focus on the EMPIAR dataset with around 1.2 million particles that was used to generate the Ca^{+2}-bound state in digitonin (EMPIAR code 10278). Since around 13% of these particles are used to generate the final reconstruction (all other particles are

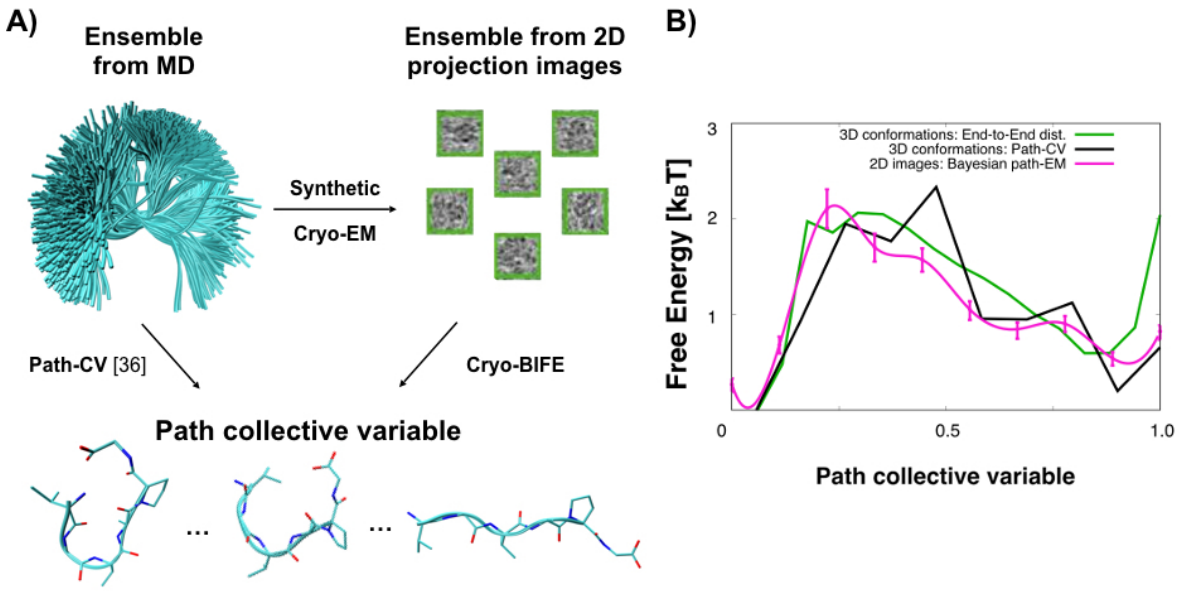

Figure 3.6: Free-energy profiles from 2D images (cryo-BIFE) or 3D conformations of the VGVAPG hexapeptide. A) The conformational ensemble of the VGVAPG hexapeptide from MD simulations is used to generate synthetic images. The nodes belonging to the path (bottom) are selected with equally spaced end-to-end distances between successive nodes (see the Methods). The path-CV [22] method compares 3D conformations to the path nodes, whereas cryo-BIFE compares 2D particle images to the same nodes. B) Free-energy profile calculated over the 3D ensemble using the path-CV with RMSD metric (equation 8 in ref.[22]) with $\lambda = 50\text{Å}^{-2}$ (black), and the expected free energy $\bar{G}(s)$ extracted using cryo-BIFE with synthetic cryo-EM particles (pink line). The R-hat test for the MCMC yielded values < 1.01 for all cases. The bars show the credible interval at 5% and 95% of the empirical quantile at each node. See the Methods for details about the path and set of images for each system.

classified out), we wanted to investigate *i)* if there could be a small population of the Ca^{+2}-unbound state in this set, and *ii)* if a free-energy profile from the Ca^{+2}-bound to the Ca^{+2}-unbound states can be extracted. Starting from the PDB structures (Figure 3.8A), steered MD simulations were used, which included a lipid membrane and explicit solvent (see the Methods), to generate a path connecting both states. The C_α-RMSD of the nodes to both states is shown in Figure 3.8B. We randomly selected around 15000 particles from the entire set, *i.e.*, not only those used for the final reconstruction. In Figure 3.8C, the free energy along the path using the same cryo-BIFE setup as for the previous systems is shown. It was observed that both the Ca^{+2}-bound and the Ca^{+2}-unbound states correspond to metastable basins of the system. Because the cryo-EM data set was prepared with Ca^{+2}, it is expected that the Ca^{+2}-bound state corresponds to the lowest free-energy minimum. However, it is interesting that not all the particles belong to this state, and that the Ca^{+2}-unbound state also has metastability. The highest barrier is around $2.2k_BT$, consistent with what is expected for turnover conditions in cryo-EM samples. The local mini-

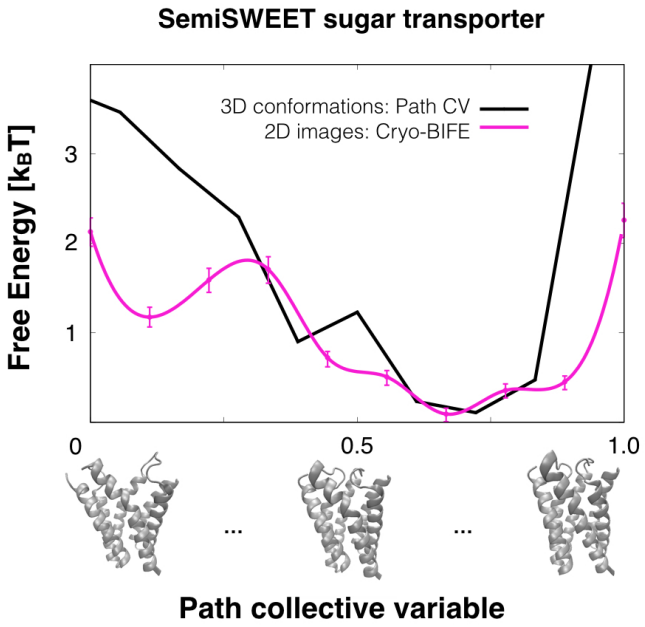

Figure 3.7: Free-energy profiles from 2D images (cryo-BIFE) or 3D conformations for the semiSWEET transporter. Free-energy profile calculated for the 3D ensemble from MD using the path-CV with the RMSD as metric (equation 8 in ref. [22]) and with $\lambda = 50\text{Å}^{-2}$ (black). The expected free energy $\bar{G}(s)$ extracted using cryo-BIFE from synthetic cryo-EM particles (pink line). The R-hat test for the MCMC yielded values < 1.01 for all cases. The bars show the credible interval at 5% and 95% of the empirical quantile at each node.

mum at 0.45 in the path-CV (see Figure 3.8) might represent a metastable configuration. However, we do not have enough experimental information to support it, due to the SNR levels in the bank of images and the number of images we used.

These results show that it is possible to extract a free-energy profile from real cryo-EM particles that agrees with the biophysical setup and expectations of the system.

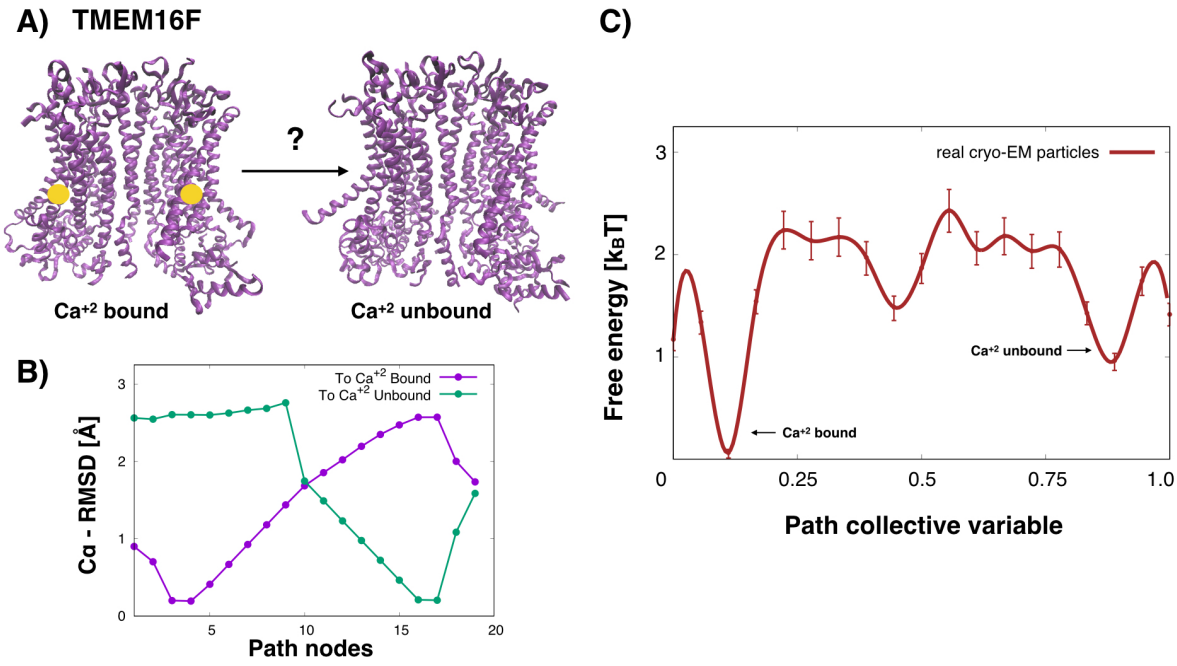

Figure 3.8: Real cryo-EM data for studying the TMEM16F Ca^{+2}- bound/unbound transition with cryo-BIFE. A) Ca^{+2}-bound to the Ca^{+2}-unbound states of TMEM16F (with PDB codes 6p46 and 6p47, respectively). B) C_α RMSD of the nodes along the path to the Ca^{+2}-bound and Ca^{+2}-unbound states (purple and green, respectively). C) Free-energy profile extracted along the path CV from real cryo-EM particles from the dataset used to generate the Ca^{+2}-bound reconstruction in digitonin[48] (EMPIAR code 10278). The R-hat test for the MCMC yielded 1.001. The bars show the credible interval at 5% and 95% of the empirical quantile at each node. Arrows point to the free-energy basins corresponding to the Ca^{+2}-bound/unbound states.

4
Conclusions

In this work, we have developed cryo-BIFE, a methodology for extracting free-energy profiles from cryo-EM experiments using a Bayesian approach with a path collective variable. The method was tested and validated over diverse systems ranging in complexity. Using controlled parameters, we found that the particle orientation accuracy and the SNR are important for adequately recovering the free-energy profile. This work is a proof of principle, demonstrating that under reasonable cryo-EM conditions it is possible to extract free-energy profiles using individual cryo-EM particles.

Primary focus has been given to extracting the *expectation* of the free-energy profile $G(s)$. However, this method produces (in the form of independent MCMC samples) the full posterior for such profiles, which contains much more information than just an average. In particular it quantifies the degree of certainty with which $G(s)$ can be extracted given the noise in particle images. Credible intervals can be placed on any function of G, such as downstream predictions (reaction rates, etc), simply by evaluating them for all G values in a set of MCMC samples.

The cryo-BIFE analysis should be performed on a raw, unbiased cryoEM-particle set. For cryo-BIFE, particles can be picked, polished, and motion-corrected. However, 3D-classification methods, which group particles with respect to conformational states, should not be performed before cryo-BIFE because these artificially modify the distribution of conformations. In other words, free-energy profiles extracted from classified-subsets of particles will be biased, and these will not represent the true thermodynamic ensemble.

Here, we have focused on developing, understanding and validating cryo-BIFE for a predetermined path. We have shown that under realistic cryo-EM-imaging conditions the extracted profile coincides

with the free-energy profile of the true conformational ensemble along that path. A demanding aspect is how to generate a conformational path for experimental cases. If the metastable states of the system have been resolved using standard cryo-EM 3D classification or from X-ray crystallography, then one could create a path by simply interpolating the maps (or structures) or by using steered MD (as done for the TMEM16F system). If metastable states are not available, then, one could generate conformational paths by directly analyzing the variability of the 2D images, for example, using the covariance matrix or spatial-VAE [56].

It is important to note that the temperature plays a crucial role in extracting free energies. In principle, the flash-cooling process [57] is done rapidly enough that the cryo-EM sample is trapped in the ensemble just before freezing. Consequently, the extracted free-energy profile should be a representation of the system at that temperature. However, freezing takes on the order of μs [58] to complete, so all relaxation processes faster than this timescale are lost. Since vitrification is not instantaneous, cooling might depopulate the barrier and cause the estimated barrier to be artificially large. Other experimental considerations, such as icesheet buckling during vitrification, can cause further perturbations to the observed structural ensemble. It remains to be fully assessed how much the freezing process affects the extracted free energy [59]. On the other hand, to obtain high-resolution reconstructions, it is common to set the system at temperatures below the ambient one for over-stabilizing a single state. We hope that these methods to extract free energies will motivate the field to measure more at ambient temperature, and moreover, use all particles (*i.e.,* without having to discard large percentages).

In summary, extracting free energies from cryo-EM experiments opens the field to the assessment of conformational dynamics from a biophysical perspective. By measuring the populations along relevant degrees of freedom, the results go beyond the discussion of discrete versus continuous, and the biophysical mechanisms are truly revealed. Additional clues to biomolecular function are unraveled by the information of the metastable states (*e.g.*, the size and shape of the free energy basins), of the activation barriers and of the location of the transition states of the system, as is common in single-molecule experiments.

5 Perspectives

A major challenge remains in determining if the path-CV is optimal. From a thermodynamic perspective, an optimal CV should separate the metastable states of the system, identify the transition states, and activation barriers, corresponding to those of the multidimensional landscape. The lowest free-energy path in the multidimensional space can be considered as an adequate CV. For simulations, several methods have been developed to measure the quality of a CV using transition state theory [60] or committor analysis [61], and algorithms exist to find optimal path-CVs [31, 27, 30] that can be shown to converge stably [62]. Recently, additional developments have standardized CV design [63, 64]. Nonetheless, a method to determine the optimal path-CV using cryo-EM images is still to be developed.

We propose two alternatives for solving the problem of choosing an optimal path-CV using cryo-EM images: using a gradient optimization method (in a string-method style) or performing MCMC moves on the biomolecule space, using the cryo-BIFE posterior probability for discriminating and optimizing the paths. Moreover, for some systems, a single degree of freedom may be insufficient and extending the CV to multiple dimensions would be advantageous.

A Appendix

A.1 The BioEM likelihood

The BioEM function $L(w_i|\Theta, x_m)$ calculates the likelihood of a 3D conformation x_m to have generated an image w_i given a set of parameters Θ, where these nuisance parameters are the projection direction, contrast transfer function (CTF) defocus, CTF amplitude, CTF b-factor, image center displacement, image normalization, offset and standard-deviation of the noise (λ). In essence, the likelihood measures the correlation of the experimental image and a calculated image from x_m using parameters Θ, assuming a Gaussian-noise model. It is defined by [12]

$$L(w_i|\Theta, x_m) = (2\pi\lambda)^{-N_{\text{pix}/2}} e^{-\sum_p (w_i(p)-I^{\text{cal}}(p))^2/(2\lambda^2)}\,, \tag{A.1}$$

where p indexes the pixel, and I^{cal} is a calculated image from conformation x_m with a given orientation direction, CTF defocus, CTF amplitud, b-factor, center displacement, image normalization, and offset (see ref. [12] for details about the image formation process). As mentioned in the Methods, the likelihood is multiplied by priors and integrated over the nuisance parameters.

A.2 Input files for the BioEM analysis

In the following table, the BioEM integration ranges, priors and input parameters (see details in https://github.com/bio-phys/BioEM) are presented for the first round and the systems studied.

BioEM parameter input for each system				
System	SemiSWEET	Hsp90	VGVAPG	TMEM16F
ORIENTATIONS	USE_QUATERNIONS	USE_QUATERNIONS	USE_QUATERNIONS	USE_QUATERNIONS
NUMBER_PIXELS	128×128	128×128	128×128	256×256
PIXEL_SIZE	0.7Å	2.2Å	0.3Å	1.059Å
CTF_DEFOCUS	[0.5,3.0] | 20gp	[0.5, 3.0] | 20gp	[0.1, 1.1] | 10gp	[1.0, 3.0] | 20 gp
CTF_B_ENV	[1, 1] | 1gp	[1, 1] | 1gp	[1, 1] | 1gp	[0.1, 0.1] | 1gp
CTF_AMPLITUDE	[0.1, 0.1] | 1gp	[0.1, 0.1] | 1gp	[0.1, 0.1] | 1gp	[0.1, 0.1] | 1gp
PRIOR_DEFOCUS_CENTER	1.75	1.75	0.6	1.75
SIGMA_PRIOR_DEFOCUS	4.0	4.0	1.7	3.0
SIGMA_PRIOR_B_CTF	1	1	1	3.0
DISPLACE_CENTER	[-2,2]	[-2,2]	[-1,1]	[-30,30]

Table A1: Parameters, integration ranges, and prior information for calculating the first BioEM round. The BioEM software was used to calculate probability $p(w_i|x_m)$ (see the Methods). The second BioEM round had a Gaussian prior for the CTF defocus centered around the experimental/synthetic defocus, and used a finer orientation grid around the best orientations from the first round. Abbreviation: gp= grid points. The center displacement was performed every pixel. See https://github.com/bio-phys/BioEM for details about the BioEM input keywords.

References

[1] G. McMullan, A. R. Faruqi, and R. Henderson. Direct Electron Detectors. In *Methods Enzymol.*, volume 587, pages 1–17. 2016.

[2] Pilar Cossio and Gerhard Hummer. Likelihood-based structural analysis of electron microscopy images. *Current Opinion in Structural Biology*, 49:162–168, apr 2018.

[3] Sjors H.W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.*, 180(3):519–530, 2012.

[4] Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy, 2015.

[5] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Breaking the next Cryo-EM resolution barrier Atomic resolution determination of proteins! *bioRxiv*, 2020.

[6] Takanori Nakane, Abhay Kotecha, Andrija Sente, Greg McMullan, Simonas Masiulis, Patricia M. G. E. Brown, Ioana T. Grigoras, Lina Malinauskaite, Tomas Malinauskas, Jonas Miehling, Tomasz Uchański, Lingbo Yu, Dimple Karia, Evgeniya V. Pechnikova, Erwin de Jong, Jeroen Keizer, Maarten Bischoff, Jamie McCormack, Peter Tiemeijer, Steven W. Hardwick, Dimitri Y. Chirgadze, Garib Murshudov, A. Radu Aricescu, and Sjors H. W. Scheres. Single-particle cryo-EM at atomic resolution. *Nature*, 587(7832):152–156, nov 2020.

[7] Kazuyoshi Murata and Matthias Wolf. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules, 2018.

[8] Karen Palacio-Rodríguez, Isaias Lans, Claudio N. Cavasotto, and Pilar Cossio. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Scientific Reports*, 2019.

[9] N. Grigorieff. Frealign: An Exploratory Tool for Single-Particle Cryo-EM. In *Methods Enzymol.*, volume 579, pages 191–226. 2016.

[10] Jacqueline L.S. Milne, Mario J. Borgnia, Alberto Bartesaghi, Erin E.H. Tran, Lesley A. Earl, David M. Schauder, Jeffrey Lengyel, Jason Pierson, Ardan Patwardhan, and Sriram Subramaniam. Cryo-electron microscopy - A primer for the non-microscopist, 2013.

[11] Roy R. Lederman, Joakim Andén, and Amit Singer. Hyper-molecules: On the representation and recovery of dynamical structures for applications in flexible macro-molecules in cryo-EM. *Inverse Probl.*, 36(4):044005, 2020.

[12] Pilar Cossio and Gerhard Hummer. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.*, 184:427–437, 2013.

[13] Joachim Frank. New Opportunities Created by Single-Particle Cryo-EM: The Mapping of Conformational Space, 2018.

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

[15] David Haselbach, Ilya Komarov, Dmitry E. Agafonov, Klaus Hartmuth, Benjamin Graf, Olexandr Dybkov, Henning Urlaub, Berthold Kastner, Reinhard Lührmann, and Holger Stark. Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell*, 172(3):454–464.e11, jan 2018.

[16] Mao Oide, Takayuki Kato, Tomotaka Oroguchi, and Masayoshi Nakasako. Energy landscape of domain motion in glutamate dehydrogenase deduced from cryo-electron microscopy. *FEBS J.*, 287(16):15224, feb 2020.

[17] Thomas Stecher, Noam Bernstein, and Gábor Csányi. Free Energy Surface Reconstruction from Umbrella Samples Using Gaussian Process Regression. *J. Chem. Theory Comput.*, 10(9):4079–4097, sep 2014.

[18] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23(2):187–199, feb 1977.

[19] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci.*, 99(20):12562–12566, oct 2002.

[20] Massimiliano Bonomi, Riccardo Pellarin, and Michele Vendruscolo. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. *Biophys. J.*, 114(7):1604–1613, apr 2018.

[21] John W. Vant, Daipayan Sarkar, Ellen Streitwieser, Giacomo Fiorin, Robert Skeel, Josh V. Vermaas, and Abhishek Singharoy. Data-guided Multi-Map variables for ensemble refinement of molecular movies. *J. Chem. Phys.*, 153(21):214102, dec 2020.

[22] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. From A to B in free energy space. *J. Chem. Phys.*, 126:054103, 2007.

[23] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, mar 2006.

[24] Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy. *Sup.Info. Cell*, 2015.

[25] Radostin Danev, Haruaki Yanagisawa, and Masahide Kikkawa. Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions, 2019.

[26] Daniel M. Zuckerman. *Statistical physics of biomolecules: An introduction*. 2010.

[27] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *Journal of Chemical Physics*, 125(2):024106, 2006.

[28] Sjors H W Scheres, Rafael Núñez-Ramírez, Carlos O S Sorzano, José María Carazo, and Roberto Marabini. Image processing for electron microscopy single-particle analysis using XMIPP. *Nat. Protoc.*, 3(6):977–990, jun 2008.

[29] Pilar Cossio, David Rohr, Fabio Baruffa, Markus Rampp, Volker Lindenstruth, and Gerhard Hummer. BioEM: GPU-accelerated computing of Bayesian inference of electron microscopy images. *Computer Physics Communications*, 210:163–171, 2017.

[30] Albert C Pan, Deniz Sezer, and Benoît Roux. Finding transition pathways using the string method with swarms of trajectories. *Journal of Physical Chemistry B*, 112(11):3432–3440, 2008.

[31] Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical And Quantum Dynamics In Condensed Phase Simulations*, pages 385–404. World Scientific, 1998.

[32] Anna Yershova, Swati Jain, Steven M. LaValle, and Julie C. Mitchell. Generating uniform incremental grids on SO(3) using the Hopf fibration. *Int. J. Robot. Res.*, 29(7):801–812, JUN 2010.

[33] Pilar Cossio, Matteo Allegretti, Florian Mayer, Volker Müller, Janet Vonck, and Gerhard Hummer. Bayesian inference of rotor ring stoichiometry from electron microscopy images of archaeal ATP synthase. *Microscopy*, 67(5):266–273, oct 2018.

[34] Sebastian Ortiz, Luka Stanisic, Boris A Rodriguez, Markus Rampp, Gerhard Hummer, and Pilar Cossio. Validation tests for cryo-em maps using an independent particle set. *Journal of Structural Biology: X*, 4:100032, 2020.

[35] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*, 16, 2021.

[36] Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo A. Martin. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *The Journal of Open Source Software*, 2019.

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

[38] Florian H. Schopf, Maximilian M. Biebl, and Johannes Buchner. The HSP90 chaperone machinery. *Nat. Rev. Mol. Cell Biol.*, 18(6):345–360, jun 2017.

[39] Maruf M. U. Ali, S. Mark Roe, Cara K. Vaughan, Phillipe Meyer, Barry Panaretou, Peter W. Piper, Chrisostomos Prodromou, and Laurence H. Pearl. Crystal structure of an Hsp90–nucleotide–p23/Sba1 closed chaperone complex. *Nature*, 440(7087):1013–1017, apr 2006.

[40] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, sep 2015.

[41] Ryo Uehara, Yuki Takeuchi, Shun Ichi Tanaka, Kazufumi Takano, Yuichi Koga, and Shigenori Kanaya. Requirement of Ca2+ ions for the hyperthermostability of Tk-subtilisin from Thermococcus kodakarensis. *Biochemistry*, 51(26):5369–5378, 2012.

[42] Luca Donati and Bettina G. Keller. Girsanov reweighting for metadynamics simulations. *J. Chem. Phys.*, 149(7):072335, aug 2018.

[43] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function and Bioinformatics*, 78(8):1950–1958, 2010.

[44] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[45] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[46] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[47] Naomi R. Latorraca, Nathan M. Fastman, A.J. Venkatakrishnan, Wolf B. Frommer, Ron O. Dror, and Liang Feng. Mechanism of Substrate Translocation in an Alternating Access Transporter. *Cell*, 169(1):96–107.e12, mar 2017.

[48] Shengjie Feng, Shangyu Dang, Tina Wei Han, Wenlei Ye, Peng Jin, Tong Cheng, Junrui Li, Yuh Nung Jan, Lily Yeh Jan, and Yifan Cheng. Cryo-EM Studies of TMEM16F Calcium-Activated Ion Channel Suggest Features Important for Lipid Scrambling. *Cell Rep.*, 28(2):567–579.e4, jul 2019.

[49] Andrii Iudin, Paul K. Korir, José Salavert-Torres, Gerard J. Kleywegt, and Ardan Patwardhan. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods*, 13(5):387–388, may 2016.

[50] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T. Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.*, 46(W1):W296–W303, jul 2018.

[51] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.*, 29(11):1859–1865, aug 2008.

[52] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D. MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, 14(1):71–73, jan 2017.

[53] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, oct 1984.

[54] H. Grubmuller, Berthold Heymann, and Paul Tavan. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science*, 271(5251):997–999, feb 1996.

[55] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, feb 2014.

[56] Tristan Bepler, Ellen D. Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from translation and rotation with spatial-vae. *arXiv*, 2019.

[57] Jacques Dubochet, Marc Adrian, Jiin-Ju Chang, Jean-Claude Homo, Jean Lepault, Alasdair W. McDowall, and Patrick Schultz. Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.*, 21(2):129–228, may 1988.

[58] Vanessa Cabra and Montserrat Samsó. Do's and Don'ts of Cryo-electron Microscopy: A Primer on Sample Preparation and High Quality Data Collection for Macromolecular 3D Reconstruction. *J. Vis. Exp.*, page 52311, jan 2015.

[59] Andrea Arsiccio, James McCarty, Roberto Pisano, and Joan-Emma Shea. Heightened Cold-Denaturation of Proteins at the Ice–Water Interface. *J. Am. Chem. Soc.*, 142(12):5722–5730, mar 2020.

[60] Gerhard Hummer. From transition paths to transition states and rate coefficients. *J. Chem. Phys.*, 120(2):516–523, jan 2004.

[61] John D. Chodera and Vijay S. Pande. Splitting Probabilities as a Test of Reaction Coordinate Choice in Single-Molecule Experiments. *Phys. Rev. Lett.*, 107(9):098102, aug 2011.

[62] Brian Van Koten and Mitchell Luskin. Stability and convergence of the string method for computing minimum energy paths. *Multiscale Modeling & Simulation*, 17(2):873–898, 2019.

[63] Mohammad M. Sultan and Vijay S. Pande. Automated design of collective variables using supervised machine learning. *J. Chem. Phys.*, 149(9):094106, sep 2018.

[64] Jutta Rogal, Elia Schneider, and Mark E. Tuckerman. Neural-Network-Based Path Collective Variables for Enhanced Sampling of Phase Transformations. *Phys. Rev. Lett.*, 123(24):245701, dec 2019.