3D Variability Analysis: Directly resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM images

Ali Punjani[1,2,3] David J. Fleet[1,2]

[1]University of Toronto [2]Vector Institute [3]Structura Biotechnology Inc.

April 2020

Abstract

Proteins are dynamic molecules that constitute the molecular machinery of living cells. Single particle cryo-EM excels as an imaging technique to solve static structures of proteins, but existing computational 3D reconstruction methods for cryo-EM data have not been effective in practice for solving structures or modelling motion of flexible proteins. We introduce 3D variability analysis (3DVA), an algorithm that models the conformational landscape of a molecule as a linear subspace, and is able to fit the model to experimental cryo-EM data at high resolution. 3DVA makes it possible, for the first time, to resolve and visualize detailed molecular motions of both large and small proteins. Through extensive experimental results on cryo-EM data, we demonstrate the ability of 3DVA to resolve multiple flexible motions of α-helices in the sub-50 kDa transmembrane domain of a GPCR complex, bending modes of a sodium ion channel, five types of symmetric and symmetry-breaking flexibility in a proteasome, and large motions in a spliceosome complex. We also show that 3DVA can uncover discrete conformational states of heterogenous mixtures of ribosome particles from a single sample. The results demonstrate that 3DVA enables new biological insight to be extracted from single particle cryo-EM data. 3DVA is implemented in the *cryoSPARC* software package, and has already been used successfully in several notable structural studies.

1 Introduction

Protein dynamics hold the key to understanding function in many molecular machines that underlie life. Single particle cryo-EM collects thousands of static 2D protein particle images that in aggregate span the target protein's 3D conformational space. Cryo-EM therefore holds great promise for studying protein dynamics [21]. However, there are significant computational challenges in resolving dynamics from static image data, particularly in the presence of continuous heterogeneity. These stem from the need

to simultaneously estimate the canonical structure of the molecule, the one or more ways in which the structure can change (due to motion, dissociation, etc), and the specific state of the molecule in each single particle image. These problems are exacerbated by the high levels of noise in the imaging process and the need to also estimate, for each particle image, the CTF corruption and 3D coordinate transform between the 3D model the 3D pose of the particle.

A heterogeneous target molecule will exist in a mixture of conformational states within the sample. This distribution of 3D structures lies on a manifold in the abstract space of all possible 3D structures. Each observed single particle image X provides a CTF corrupted, 2D projection of the 3D structure $\mathcal{V}$ at a single point on the manifold. The true manifold in the general case may have a complex, non-linear geometry. However, for many molecules of interest, a linear subspace model provides a simple but effective way to represent the 3D structures present in a cryo-EM sample. To this end, a K-dimensional linear subspace represents all possible 3D structures in the sample as the sum of a base structure V_0 and a weighted combination of K components, $V_{1:K} \equiv \{V_1 \ldots V_K\}$; i.e.,

$$\mathcal{V}(z) \;=\; V_0 + \sum_{k=1}^{K} z_k V_k \tag{1}$$

The weight vector, $z = (z_1, ..., z_K)$, is a per-particle latent variable representing the conformational state of the molecule.

This class of subspace models, including Principal Component Analysis (PCA), Multivariate Statistical Analysis (MSA), or eigenanalysis, has been discussed extensively in the literature (e.g., [1, 28, 42, 44]). They are attractive for their simplicity, their analytical properties, and the existence of relatively stable numerical methods. However, methods for fitting linear subspace models are not widely used in practice compared to K-way discrete 3D classification methods [12, 36, 38, 39]. This is due to the inability of existing methods to estimate the components and weights in Equation (1) at moderate or high resolutions or to do so for several components simultaneously. This makes it difficult to resolve detailed conformational changes that are most important at high resolutions.

This paper describes an effective new algorithm for fitting accurate high-resolution 3D linear subspace models to single particle cryo-EM data, and shows that it yields remarkably useful information about flexible dynamics and heterogeneity of the target molecule. We refer to this specific algorithm for cryo-EM as *3D Variability Analysis* (3DVA). Through experimental results we show that 3DVA allows for directly resolving molecular motion, flexibility, changes in occupancy, and even discrete heterogeneity of a wide range of protein molecules. In particular, it resolves high resolution flexibility, with motions of individual α-helices as small as a few Å, for both large and small proteins. To our knowledge this is the first method to work on small proteins; for example, we are able to visualize three different types of flexible motion of the small sub-50 kDa transmembrane region of a GPCR complex, at a resolution of $\sim$3Å- 4Å.

We further show that 3DVA identifies several symmetric and assymetric motion modes of high resolution motion of a symmetric T20S proteasome, as well as large motions of a ribosome and a pre-catalytic spliceosome complex. For a bacterial large ribosomal subunit assembly, we show that 3DVA resolves

several discrete conformational states. Recently proposed deep generative models [48, 49] have also been experimentally demonstrated to resolve continuous and discrete heterogeneity, but for large motions at coarse resolution, and only for large complexes (spliceosome and ribosomes, all in the multiple-MDa size range). 3DVA provides similar insights on these same datasets, but it is 20-50x faster to train and requires no manual parameter tuning or hyperparameter optimization for individual datasets.

An efficient GPU implementation of 3DVA has been made available in the *cryoSPARC* [31] software package since v2.9+, and it has already been used in several notable structural studies (see Section 4). Among them is the recent structure of the spike protein of the SARS-CoV-2 virus, responsible for the global COVID19 pandemic [46]. Generally, 3DVA enables new biological insights to be gleaned from single particle cryo-EM data.

Related methods for manifold fitting

The linear subspace model (Eqn 1) and related forms, also known as Principle Component Analysis (PCA), Multivariate Statistical Analysis (MSA), eigenanalysis, and other names, has been extensively studied in single particle cryo-EM literature. Early methods involved a bootstrap technique using 3D reconstructions from random subsets of particle images to coarsely simulate sampling from the true conformational distribution of a target molecule [28]. Another family of methods have attempted to construct a low resolution complete covariance matrix of the 3D density map distribution underlying a sample, extracting a linear subspace using eigenanalysis [1]. A direct method has also been proposed to solve K linear subspace components via optimization [42]. These methods for linear subspace models are most closely related to 3DVA, as discussed in detail in Section 2.

Non-linear manifold embedding techniques are also in development. Diffusion maps have been used in 2D, with multiple 2D manifold embeddings combined to resolve a 3D manifold of continuous conformational change [6, 10]. These methods have been demonstrated on experimental data of large proteins, and recent work aims to improve the difficult process of combining 2D manifolds into 3D [22]. Non-linear methods employing the graph Laplacian of similar 2D particle images have also been developed [25] and demonstrated on simulated synthetic data. Most recently, another class of non-linear models, deep generative models, have been proposed and demonstrated to resolve large-scale conformational changes from experimental cryo-EM data of large complexes [48, 49]. Methods attempting to directly model protein motion via motion models (rather than generic heterogeneity models) have also been proposed, and demonstrated on simplified synthetic data [11, 19].

Relationship to other SPA heterogeneity techniques

The predominant method currently used to resolve heterogeneity in cryo-EM data is K-way 3D classification [12, 36, 38, 39]. Unlike classification, a linear subspace model does not constrain particles to exist in only a (small) finite number of discrete conformational states. Rather, the subspace components, $V_{1:K}$, span a K dimensional continuous vector space of possible states. Empirically, it is well known that clustering methods cannot readily resolve continuous flexibility.

Local refinement methods (e.g., [12, 38]) and multi-body refinement [26] methods allow high resolution refinement of some regions of flexible molecules, by assuming that the molecule is composed of a small number of rigid parts. For successful alignment during 3D refinement, such methods require that each part has sufficient mass that it can be (rigidly) aligned independent of the other parts, once other parts have been subtracted from the molecule. A linear subspace model does not make this assumption, as it allows generic continuous flexibility within the subspace.

Techniques like normal-modes analysis [4] make assumptions about the energy landscape and dynamics of a protein molecule to predict possible flexibility. Methods have been proposed to exploit these fixed prior models to recover information from cryo-EM data of flexible molecules [40]. In contrast, the linear subspace model does not presuppose knowledge about the energy landscape, bending or flexing of the molecule. Instead, the model effectively learns this from the cryo-EM image data.

Finally, linear subspace models have already been useful in the related field of cryo-electron tomography (cryo-ET), specifically in sub-tomogram averaging where multiple 3D reconstructions of individual molecules, are obtained. These can then be averaged to improve the resolution of reconstruction. In this case, Principle Component Analysis (PCA) has been directly applied [15] since tomography provides complete 3D observations for each individual particle.

2 Linear Subspaces, PCA, and 3D Variability Analysis

Under the standard cryo-EM image formation model [9, 13, 30, 37], each 2D experimental image, X_i, is generated as a contrast transfer function (CTF) -corrupted projection of an unobserved 3D density V from an unobserved pose ϕ_i (i.e., the 3D coordinate transformation between the density V and the microscope projection direction), plus additive noise:

$$X_i \;=\; C_i \, P(\phi_i) \, V + \eta \tag{2}$$

The unknown 3D density, V, is typically represented in real-space as a regular 3D grid of N^3 density voxels, or in the Fourier domain as a corresponding 3D grid of N^3 Fourier coefficients. Here, $P(\phi_i)$ is the projection operator for pose ϕ_i, and C_i is the CTF operator for image i. The noise, η, is modelled as colored Gaussian noise; the additive noise in the Fourier coefficients is mean zero and independent, and the variance is constant in spherical shells centered at the origin. More formally, $\eta_l \sim \mathcal{N}(0, \sigma_l^2)$ for the Fourier shell with radius l.

In the standard homogeneous model (Eq. 2), there is a single 3D density V from which all particle images are generated. This assumes a homogeneous population of molecules in the sample, all having the same identity and existing in the same single conformation, differing only by a rigid coordinate transformation. Extending this basic model, K-way 3D classification methods [12, 36, 38, 39] model each particle as having been generated from one of K different, independent, 3D densities. This assumes that the population of molecules in the sample contains a small number of distinct identities or distinct conformations.

Many protein molecules have a mechanistic function, exhibiting flexible motion across a continuous landscape of energetically favourable conformations. This flexibility can result in a corresponding continuum of conformational states present in the frozen EM sample. In this case, the assumptions of the standard homogeneous or K-way classification models are incorrect.

Instead, here we consider a class of models that capture continuous flexibility, namely, the K-dimensional linear subspace model. This model assumes that each experimental particle image is generated from a **mean density** V_0, plus weighted contributions from each of several different **variability components** $V_{1:K}$. Each component is itself a 3D density, capturing a particular type of change to V_0. This model is designed to explicitly account for continuous variability in the molecule. When $K > 1$, each component ideally corresponds to a different, simultaneously occurring type of change in the molecule, e.g. bending vs. twisting.

The basic linear subspace model for 3DVA can be written as:

$$X_i = \alpha_i C_i P(\phi_i) \mathcal{V}(z_i) + \eta \tag{3}$$

$$= \alpha_i C_i P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) + \eta \tag{4}$$

$$\text{where} \quad V_i^{\mathrm{T}} V_j = [i = j] \tag{5}$$

Here, $\mathcal{V}(z)$ is a generic 3D density generator that takes in a per-particle latent variable z and produces the 3D density that is present in that particular 2D particle image. The generator is simply a linear combination of $V_{0:K}$ densities, with weights $z_i = (z_{i,1}, ..., z_{i,K})$, expanded in the second line. The latent variable z is a vector of numbers, also known as the **reaction coordinates**, for the particular particle image. α is a per-particle scale factor that accounts for varying ice-thickness and contrast level of each particle image. In 3DVA we assume that the CTF C and pose ϕ for each particle are known, along with the mean density V_0. These quantities would be available, for example, after processing the particle dataset using standard homogeneous refinement. Finally, like PCA and related subspace methods, in 3DVA we also require that the K variability components be orthonormal (Equation 5).

The linear subspace model can be constructed and understood in another way as well. Rather than explicitly writing a linear manifold as a model for 3D densities that are present in the sample, consider a model where the specific 3D density in each given particle image is drawn from a multivariate distribution, with unknown mean V_0 and unknown covariance matrix Σ. The mean is a single 3D density. The covariance is a very large matrix, of dimensions $N^3 \times N^3$ that captures the variance of each 3D density voxels, as well as the covariance between every pair of 3D density voxels. Principal Component Analysis decomposes the covariance matrix into its eigenvectors and eigenvalues to reveal the orthogonal patterns of how groups of voxels vary together. The leading eigenvectors (those with largest eigenvalues) capture 3D patterns of density variation that occur with the largest magnitude across particle images.

The connection between eigenvectors of a covariance matrix and linear subspace fitting is the basis for the PCA algorithm used extensively in many data analysis contexts [2]. In PCA, it is assumed that all data points in a sample are generated from the same underlying distribution. The mean and covariance Σ are

constructed explicitly as the sample mean and sample covariance of the data, and the eigenvectors of Σ are computed numerically. The K leading eigenvectors, named **principle components** or eigen-volumes (since each one is a 3D density) are equal to the 3DVA variability components $V_{1:K}$. They are retained by PCA, and the projection of each data point onto each principle component returns the latent representation of the data point, equivalent to reaction coordinates in 3DVA.

Although widely studied and understood, the linear subspace model has proven difficult to use in practice due to several features of single particle cryo-EM data that make it impossible to use the standard PCA algorithm directly. First, each single particle 2D image is only a **partial observation** from a single pose of the entire 3D density, and therefore the covariance matrix Σ of 3D densities that we require cannot be constructed explicitly from particle images. Second, the high dimensionality of Σ means that it is often **not possible to store** it in memory or to compute its eigenvectors unless the model is restricted to low resolution volumes. Third, each 2D particle image is corrupted by a potentially **different CTF** C, breaking the assumption required for PCA that all data are generated from the same distribution.

Given these challenges, many methods have been proposed to solve the linear subspace fitting problem for single particle data. Restricting the problem to 2D only (for example by only computing the 2D covariance of particles within a single 2D class or from a single viewing direction) removes the partial observation problem, and reduces dimensionality of Σ but does not capture 3D variability or account for the CTF [44]. In 3D, the statistical technique of bootstrapping [28] can be employed, where multiple random subsets of particle images are used to reconstruct multiple 3D structures under the homogeneous model (Eq. 2). Each of these "bootstrapped" 3D densities is then taken to be a single sample from the underlying distribution of 3D densities. This avoids the missing information problem in 3D, but can only operate at low resolutions where the dimensionality of Σ is still tractable, and it further suffers from statistical inefficiency due to the random noise injected into each bootstrap sample. Together these drawbacks limit the method to resolving only very coarse resolution eigenvectors of large complexes [28]. Some techniques have been proposed that make further simplifying assumptions about the form of Σ, reducing the effective dimensionality so that the entire matrix can be estimated. These techniques have also been demonstrated only for resolving coarse resolution variability components [1].

The 3DVA algorithm builds on the formulation of probabilistic principal component analysis (PPCA), a generative model in which one assumes the data are drawn from a Gaussian distribution [35, 43]. In addition to model formulated above in Equation (4), 3DVA, like PPCA, assumes a Gaussian prior over the latent reaction coordinates z. A maximum likelihood (ML) estimate of the linear variability components $V_{1:K}$ can then be obtained using the Expectation-Maximization algorithm [8], allowing one to marginalize out the latent variables [43]. The algorithm also provides ML estimates for the covariance of the additive noise in Equation (4) and the reaction coordinates. Importantly, the Expectation-Maximization algorithm for PPCA works with partial observations, and with extremely high-dimensional data without the need to explicitly store or approximate the covariance matrix, and it allows direct estimation of only the top K components, as desired [35]. In the limiting case, as the variation of the additive noise in Equation (4) approaches zero, the PPCA model reduces to standard PCA. As such, the Expectation-Maximization

algorithm also applies directly to this case, providing maximum likelihood estimation of the variability components $V_{1:K}$ and the reaction coordinates z for each image. This is the formulation adopted for 3DVA. It naturally allows for 2D observations of 3D volumes and the inclusion of particle-specific CTF corruption.

In the 3DVA method, the Expectation-Maximization algorithm entails an iterative solution in which each iteration comprises an E-step and an M-step. These are given as follows:

$$\text{E-step:} \qquad z_i^* = \arg\min_{z_i} \frac{1}{2} \left\| X_i - \alpha_i C_i P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) \right\|_{\sigma_l^2}^2 \tag{6}$$

$$\text{M-step:} \qquad V_j^* = \arg\min_{V_j} \frac{1}{2} \sum_i \left\| X_i - \alpha_i C_i P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) \right\|_{\sigma_l^2}^2 \tag{7}$$

$$\text{such that} \quad V_i^{\mathrm{T}} V_j = [i = j] \tag{8}$$

where $[i = 1]$ is defined to be 1 when $i = j$, and 0 otherwise.

These update equations correspond to "hard" Expectation-Maximization, where the E-step computes a point estimate for z for each particle, and the M-step computes improved estimates for $V_{1:K}$ at each iteration. An earlier work by Tagare *et al.* [42] constructed a very similar linear subspace model and also applied maximum-likelihood optimization using the Expectation-Maximization algorithm. In [42], the M-step of Expectation-Maximization is solved via an approximate iterative algorithm (conjugate gradient, CG), and each V_k is solved sequentially, requiring K complete rounds of optimization. Furthermore, the CTF is treated approximately using Wiener filtering, noise in experimental images is assumed to be white, and particles are binned into a finite number of 3D pose bins. These choices and approximations limit the method to resolving coarse low-resolution motion and variations. Results on experimental data are only demonstrated at 15Å resolution, and only for $K = 2$ variability components. As such, the method explored in [42] has not been widely used.

In contrast, 3DVA follows standard techniques [35] to compute the M-step (Equation 7) exactly via matrix operations, without approximate or iterative methods. All K variability components are solved simultaneously in one round of optimization. We explicitly include the exact CTF for each particle, use all 2D particle images from all viewing directions simultaneously without binning poses, and optimize the variability components $V_{1:K}$ at the full resolution of the raw particle images. In addition, we optimize per-particle scale factors α_i and a joint colored noise model σ_l^2. The Gram-Schmidt orthogonalization operation is used to satisfy the orthogonality constraint (Equation 8) for the variability components at each iteration.

Together, accounting for these details yields a computationally demanding method, but a careful GPU implementation in *cryoSPARC* allows us to scale up to large experimental datasets with 10^6 2D particle images. During optimization, the variability components $V_{1:K}$ are regularized by both a high-pass filter and low-pass filter. The high-pass filter is an 8th order butterworth filter, and the low-pass filter is a 2nd order butterworth filter, both with user-specified cutoff resolution. The high pass filter is intended to

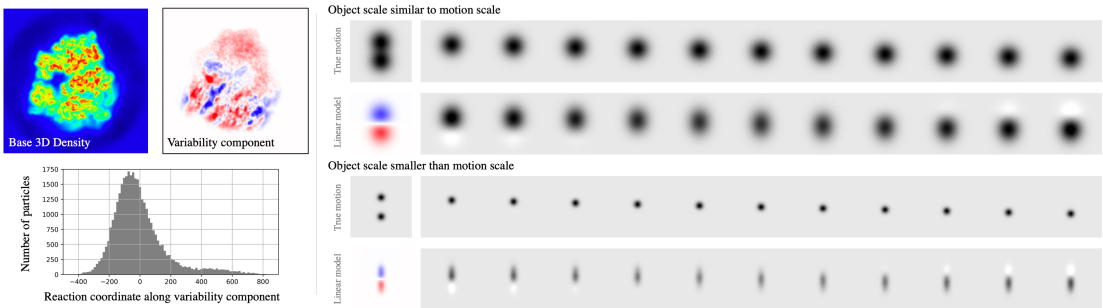

Figure 1: **Left:** An example of a linear subspace model variability component for a Ribosome structure. The base 3D density and variability component are shown as 2D slices through the 3D density. The variability component is colored in red and blue to indicate negative and positive values respectively. The variability component captures the change to be made to the base density to explain heterogeneity in the dataset; density should be decreased in red regions and increased in blue regions. A histogram (bottom) shows the distribution of the amount of the variability component present in each particle image, also known as a reaction coordinate. **Right:** Synthetic examples of how a linear subspace model like 3DVA resolves motion of objects in the data. In both cases, the top row depicts 11 positions of the object observed in the data. The bottom row depicts the resulting variability components and the image generated by the linear subspace model at the same 11 positions. When the object scale is similar to the motion scale, the linear model is a good approximation to motion. When the object scale is smaller than the motion scale, the linear model is less accurate.

remove the contribution of low-resolution variability caused by contaminants in the particle images (e.g. denatured protein mass at the air-water interfaces). In our implementation, it is optional, and is not used in the experiments in this work. The low-pass filter is intended to filter out high-frequency noise and prevent over-fitting.

The 3DVA method can be used to resolve many variability components simultaneously, limited only by GPU memory and signal present in the data. The resulting variability components allow us to resolve and detect detailed heterogeneity in single particle EM samples at resolutions as high as 3Å-4Å, far more effectively than previous approaches for linear subspace models.

Interpretation and properties of 3D Variability Analysis

When 3DVA is applied to single particle EM data, it yields the variability components $V_{1:K}$ and reaction coordinates z. These can reveal significant biological insights. Variability components are those directions in the space of 3D density that, starting from the mean structure V_0, define a linear manifold that best fits the data. As such, each variability component is a 3D volume, with density values that are positive or negative at each voxel. These values indicate where density should be removed and where density should be added, relative to V_0, to explain variability amongst the particles.

Orthogonality of the K variability components ensures that they each explain a different mode of possible change that the true molecule exhibits. In should be noted that orthogonality does not mean that the

components identified are statistically independent, nor that they correspond to physically disconnected flexibility or heterogeneity. Instead, each particle image is explained by a combination of the components, and in general, several variability components may together represent the physically independent motions and changes in the molecule. Figure 1 shows an example of a variability component on a simple ribosome dataset. Variability components solved by 3DVA will often capture flexible motion of the underlying molecule in a sample. Figure 1 (right) shows qualitatively how a linear subspace model will capture motion of an object. The model is a good approximation for an object that is moving a distance similar to its own size. Small objects moving large distances are less well modelled. In other words, 3DVA can represent large motion in low resolution regions of a 3D density map, and can model detailed small motions in high resolution regions.

3DVA also yields reaction coordinates z. For each particle, there are K reaction coordinates, one for each variability component. These values indicate the level to which a particular variability component is present in a particular 2D image. If a particle image has a large positive value for its kth reaction coordinate, this means that the image is best described by adding a large amount of V_k to V_0. Figure 1 shows an example of reaction coordinates along a single variability components for a ribosome dataset, as a 1D histogram. Across a population of particle images, the mean value of each reaction coordinate will be zero, since V_0 itself best explains the average particle. The variance of each reaction coordinate across the population (known as the eigenvalue) indicates the importance of that component in describing the data. Components with large variance explain the most variability in the particle images. When a population of particle images contains well-separated clusters of different conformations, 3DVA components will identify the differences between clusters, and the reaction coordinates of particles will also be clustered providing direct insight into discrete heterogeneity in the dataset. In fact, under mild assumptions, one can prove that when there are discrete clusters present, variability components (or more generally, principal components) will span the subspace that maximally separates the clusters [5].

3DVA also has useful mathematical properties. When the 3DVA model is optimized using Expectation-Maximization, the resulting algorithm is a Krylov subspace method [35]. Krylov subspace methods are a general class of iterative optimization algorithms from computer science for solving the eigenvectors of matrices, used in popular linear algebra techniques like SVD. Krylov subspace methods like 3DVA operate without explicitly constructing the large matrix Σ, and have the beneficial property that when used to solve K eigenvectors, will return the top-K eigenvectors with the largest eigenvalues first. This means that running 3DVA with $K = 3$ will yield the top three components, and running with $K = 6$ will yield the same top three plus the next three important components.

Another related useful property of 3DVA is that it is not sensitive to initialization bias. Generally, iterative algorithms that solve non-convex optimization problems like Eq. 7 can give incorrect results if initialized from a poor starting point. It has been proven, however, that for PPCA models (like 3DVA) the only stable local optimum is also a global optimum of the objective function [43]. This means that no matter how $V_{1:K}$ are initialized at the start of optimizing 3DVA, the results will be equivalent. This is in contrast to existing methods for K-way 3D classification, where initialization is critical and multiple runs

can often yield significantly different results.

3DVA requires as input the mean density V_0, and requires that particle images already be aligned to this mean density. This means that other methods that produce high quality homogeneous structures and alignments of particles in the presence of flexibility and heterogeneity can improve the results of 3DVA. An example of such a method is non-uniform refinement [32].

3 Results and Discussion

The major contribution of this work is to show empirically that the linear subspace model, when appropriately fit to single particle cryo-EM data using 3DVA, can yield useful, interpretable, and detailed information about both continuous and discrete heterogeneity present in real cryo-EM data. 3DVA is fast and does not require any tuning or parameter changes between datasets, other than resolution limits and the number of components.

We demonstrate 3DVA results on six real cryo-EM datasets. In all cases, the data are first processed in *cryoSPARC* to create a consensus refinement of the entire available particle set. In the case of membrane proteins, non-uniform refinement [32] is used to improve image alignments and overall resolution. The resulting particle images and 3D pose alignments are then used to compute variability components and reaction coordinates via 3DVA. No prior information is provided about the type or form of heterogeneity in each dataset. 3DVA is run in all cases with a mask that excludes solvent, and for membrane proteins the mask also excludes micelle/nanodisc. In all cases, 3DVA executes 20 iterations of Expectation-Maximization over the dataset, starting from random initialization. All experiments are run on a standalone workstation using a single NVIDIA V100 GPU.

For results displaying continuous heterogeneity, we display renderings of the 3D density generated by the linear subspace model at negative (red) and positive (blue) values of the reaction coordinate for individual variability components, along with a superimposed rendering of both densities to help clearly indicate the difference between them. Guide lines (solid and dashed) and feature markers (+) are provided as visual reference points. For each dataset, videos displaying flexible conformational changes are available at `http://3dva.cryosparc.com` and as supplementary movies.

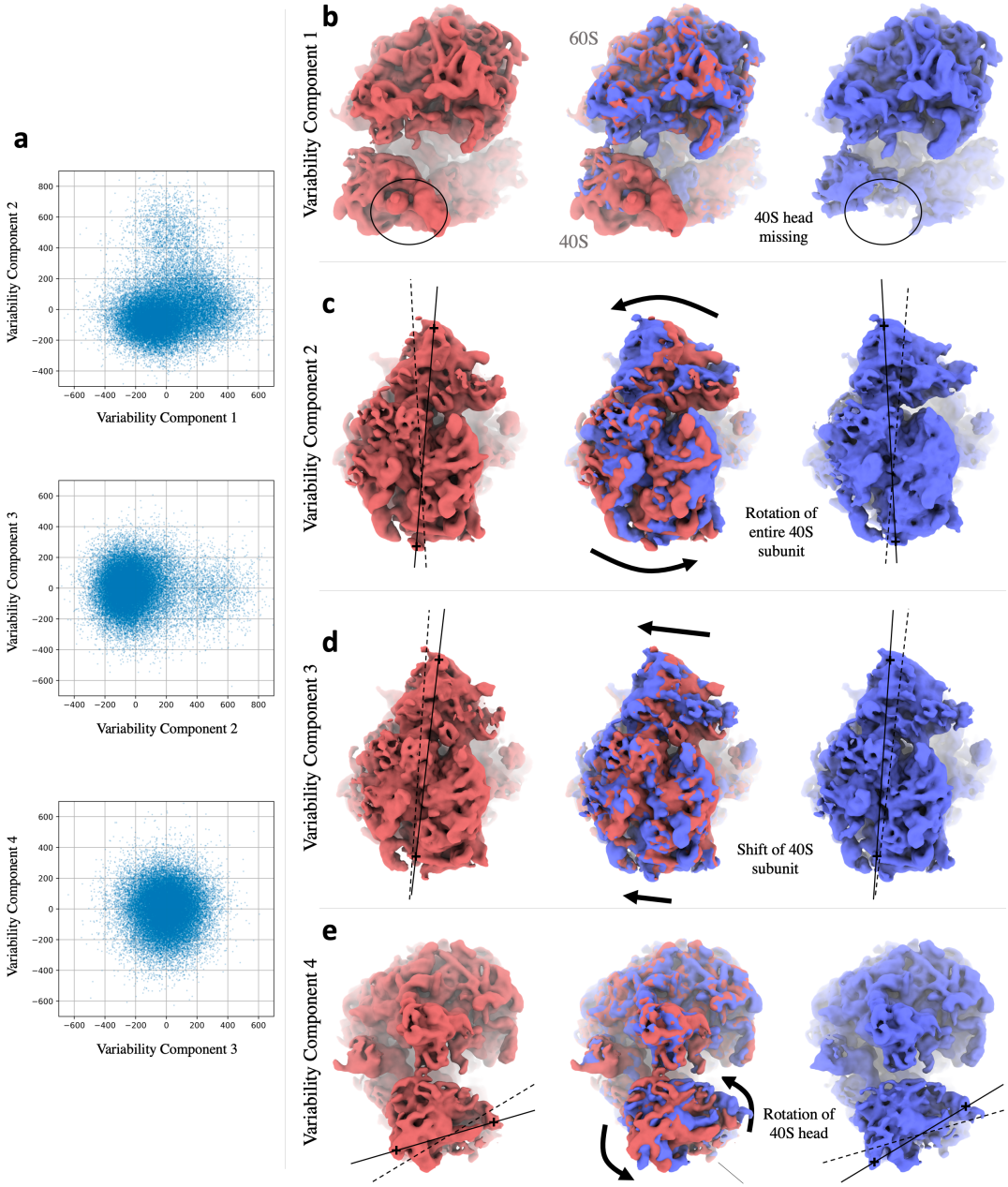

Figure 2: Results of 3DVA with four variability components on 105,247 particle images of *Pf* 80S Ribosome. A video version of these results are available at http://3dva.cryosparc.com and as Movie S1. **a**: Four-dimensional reaction coordinates of individual particles are solved by 3DVA, plotted as scatter plots along sequential pairs of components. The scatter plots indicate the extent of variability along each dimension. **b-e**: Renderings of 3D density maps generated by 3DVA at negative (red) and positive (blue) positions along each variability component. **b**: (front view) Component 1 identifies the presence and absence of the 40S head region. **c**: (bottom view) Component 2 resolves the rotation of the entire 40S subunit. **d**: (bottom view) Component 3 resolves a lateral shifting of the 40S subunit. **e**: (side view) Component 4 resolves the transverse rotation of the intact 40S head region.

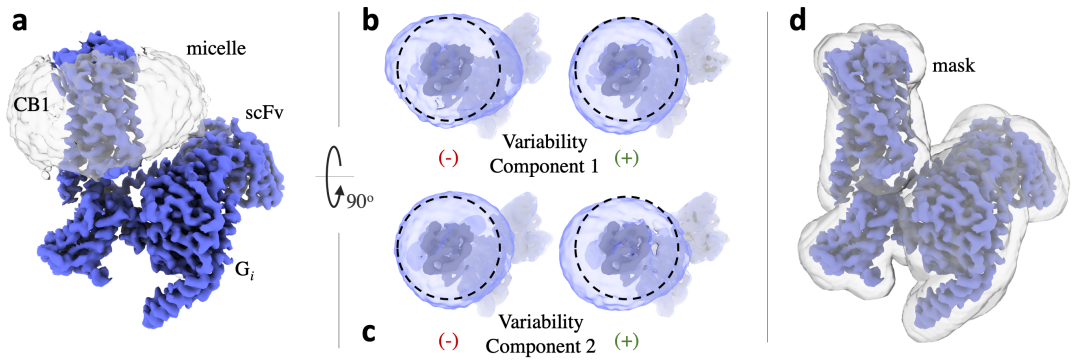

Figure 3: **a**: Consensus 3D refinement density of 250,649 particle images of Cannabinoid Receptor 1-G GPCR complex. The micelle is transparently shaded. **b-c**: Results of 3DVA with two components, with a mask that only excludes solvent. Both components (shown at negative and positive values of reaction coordinate) resolve different types of change and motion of the micelle, rather than protein. **d**: Mask that is use for subsequent 3DVA processing that excludes solvent and micelle.

80S Ribosome: Solving multiple dimensions of heterogeneity

On a dataset of 105,247 *Pf*80S ribosome particles (EMPIAR-10028 [45]), 3DVA with four variability components resolves four types of heterogeneity (Figure 2). The first component identifies the presence or absence of a portion of the "head" region of the 40S small subunit. The second component resolves rotational motion of the entire 40S subunit along an axis that connects the 40S and 60S subunits. The third component resolves lateral shifting motion of the 40S subunit. The fourth component resolves transverse rotational motion of the "head" region of the small subunit, perpendicular to the rotation in the second component.

In this case, 3DVA was run with low-pass filter corner resolution set to 8Å. The four orthogonal types of heterogeneity in the molecule were solved in a single run of 3DVA, using the full resolution 360 pixel box size particle images, taking a total of 71 minutes on a single NVIDIA V100 GPU. As a comparison, on this dataset the recently proposed cryoDRGN deep generative model [49] was reported to resolve conformational changes similar to variability components 1, 2, and 4 [48], but required model training for 150 epochs taking over 3600 minutes (i.e., 50x more than 3DVA) on the same hardware.

Cannabinoid GPCR complex: high resolution flexible motion of small proteins

GPCRs are an important family of small membrane proteins responsible for cell signalling and transmission [14]. 3DVA is applied to a dataset of Cannabinoid Receptor 1-G GPCR complex particles [18]. The complex contains the CB1 GPCR, G protein, and scFv. The raw microscope images (EMPIAR-10288) are processed in *cryoSPARC v2* to obtain a consensus refinement of the entire structure (Figure 3a) with 250,649 particle images. When 3DVA is first run on the entire 3D density, the variability components are dominated by changes in the shape and position of the micelle (Figure 3b and 3c). To inspect the heterogeneity of the protein itself, a mask is created that excludes solvent and micelle (Figure 3d). 3DVA is then run with this mask using three variability components and a low-pass filter resolution of 3Å.

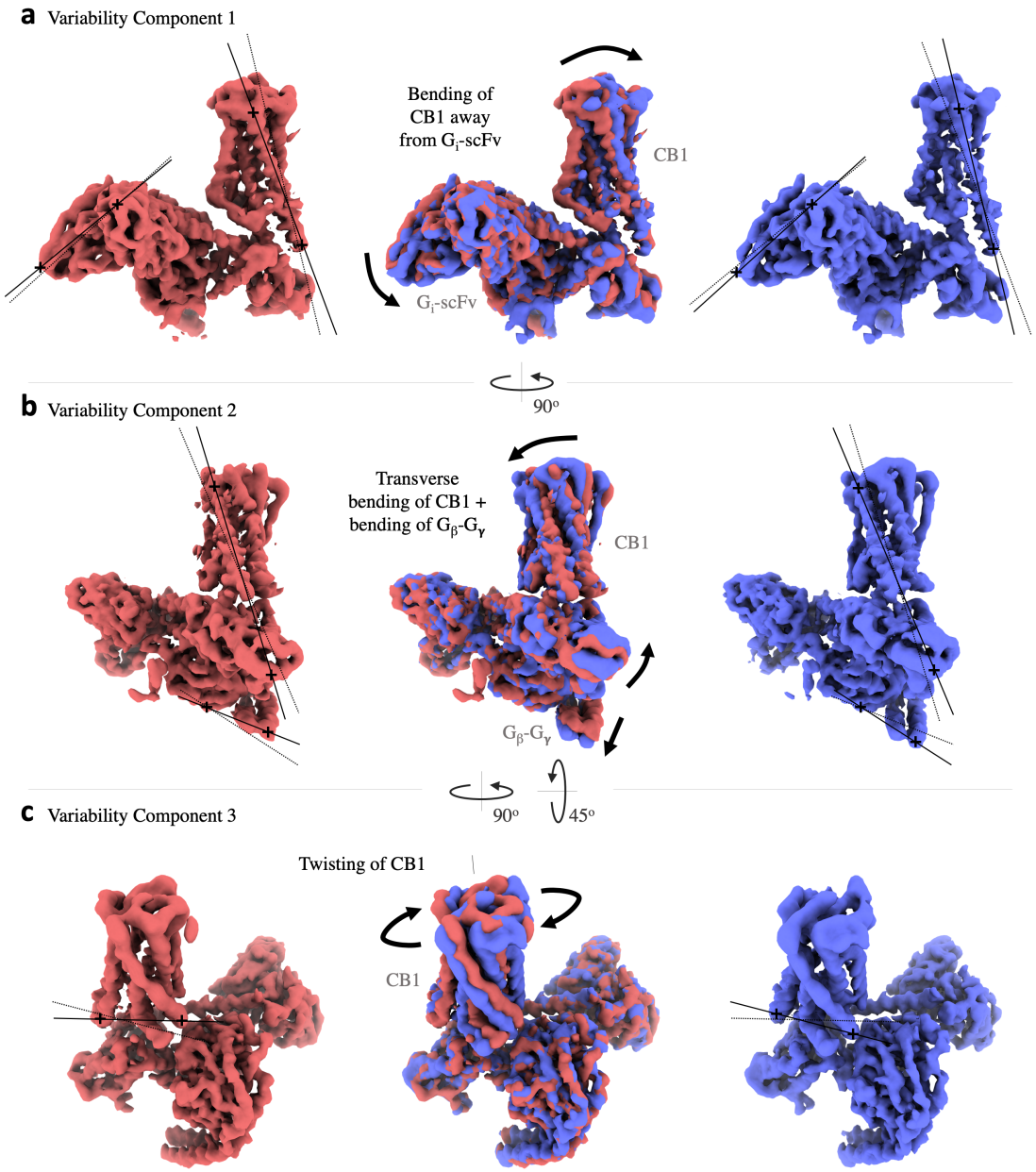

Figure 4: Results of 3DVA with three variability components on 250,649 particle images of a Cannabinoid Receptor 1-G GPCR complex. This result demonstrates the capacity for 3DVA to resolve detailed high resolution motion of small proteins and small subregions of proteins. A video version of these results are available at http://3dva.cryosparc.com and as Movie S2. **a**: Component 1 resolves bending of the CB1 transmembrane region away/towards the G-protein. Bending also affects the pose of the scFv subunit. **b**: Component 2 resolves a perpendicular (compared to component 1) bending of the CB1 region, with simultaneous motion of the G_β-G_γ pair of helices. **c**: Component 3 resolves twisting of the CB1 transmembrane domain around an axis perpendicular to the membrane.

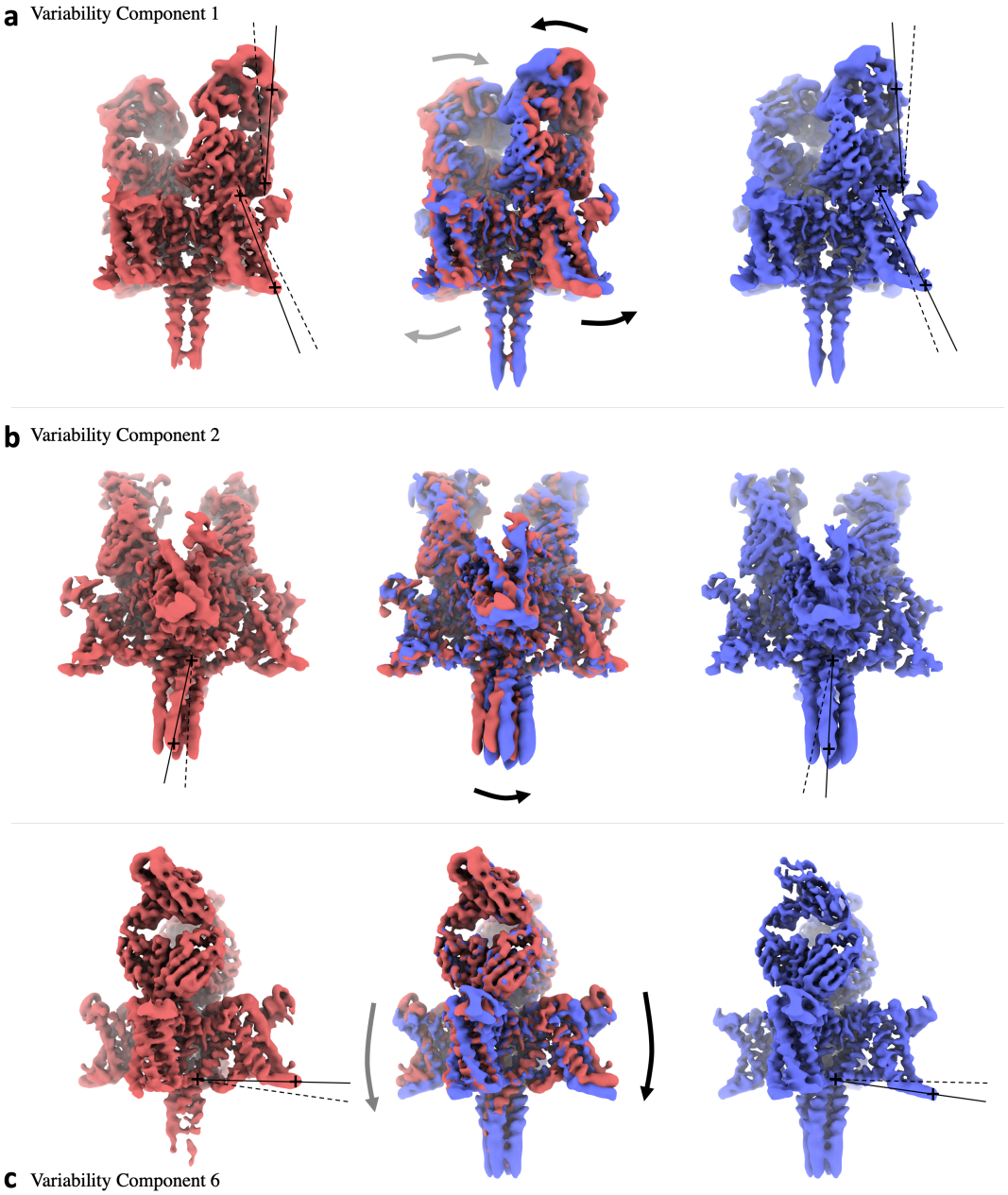

Figure 5: Results of 3DVA with six variability components (three shown) on 300,759 particle images of the $Na_V1.7$ channel membrane protein. Particles are subject to symmetry expansion around a C2 symmetry. A video version of these results are available at http://3dva.cryosparc.com and as Movie S3. **a**: Component 1 resolves simultaneous bending of the two protein subunits that are bound to Fabs, and the motion of the Fabs together/apart. **b**: Component 2 resolves lateral bending of the 4-helix bundle. **c**: Component 6 resolves up/down bending of the two subunits that are not bound to Fabs.

The results of 3DVA (Figure 4) provide, to our knowledge, the first high resolution continuous flexibility resolved for a protein as small as a GPCR complex. The first component resolves bending of the CB1 transmembrane domain towards and away from the G-protein. The second component resolves a perpendicular bending of the CB1 domain, with simultaneous motion of the G_β-G_γ domain. The third component resolves twisting of the CB1 transmembrane region around a vertical axis, perpendicular to the membrane. This result is notable, as 3DVA is able to resolve the detailed motion of a very small subregion of the complex. The entire CB1 protein is only 53 kDa [16] and the CB1 transmembrane domain is even smaller.

Nav17: multiple high resolution modes of bending in a sodium ion channel

The $\mathrm{Na_V}1.7$ channel [47] is a voltage-gated sodium channel found in the human nervous system. $\mathrm{Na_V}$ channels are fundamental to the generation and conduction of action potentials in neurons. They are mutated in various diseases, and are targeted by toxins and therapeutic drugs (e.g., for pain relief). A dataset of 431,741 particle images of a $\mathrm{Na_V}$-Fab complex is created by complete processing in *cryoSPARC v2* from raw data (EMPIAR-10261). This particle set is processed through standard 3D classification methods to separate the discrete conformational states of active and inactive channels. The active class contains 300,759 particles. The overall protein-Fab complex is C2 symmetric, and so the particle images are duplicated with their 3D poses rotated 180° around the symmetric axis (i.e. symmetry expansion). The resulting 601,518 particles are then provided as input to 3DVA with six variability components and a low-pass filter resolution of 3Å.

Of the six components solved, we display three in Figure 5. The first (component 1), resolves bending of two of the transmembrane subunits of the tetrameric protein, along with motion of the Fabs that are bound. The outer transmembrane helices move left and right while the Fabs move closer and further apart. The second component (component 2), resolves the lateral bending of the 4-helix bundle. The third component (component 6) resolves bending of the two subunits that are not bound to Fabs, in an up-down direction. For this dataset, refinement resolution of the peripheral transmembrane helices is limited [32], and 3DVA provides insight into the flexibility that causes this limitation.

T20S proteasome: symmetric and asymmetric flexible motion at high resolution

The T20S proteasome is a D7-symmetric large, stable protein that is commonly used as a test specimen for cryo-EM microscopes [3]. A dataset of 84,605 particle images of T20S is created by complete processing in *cryoSPARC v2* from raw data (EMPIAR-10025). The particle images are duplicated around the 14-fold D7 symmetry (i.e. symmetry expansion), and the resulting 1,184,470 particles are then used for 3DVA with five variability components and a low-pass filter resolution of 5Å.

Despite the generally stable nature of the T20S protein, 3DVA is able to detect five types of continuous bending flexibility in the molecule (Figure 6). This is an interesting case due to the high symmetry; three of the solved variability components (Figure 6a, left) are symmetric, corresponding to symmetric motion of all 14 subunits simultaneously. The first component resolves a stretching-compression motion of the

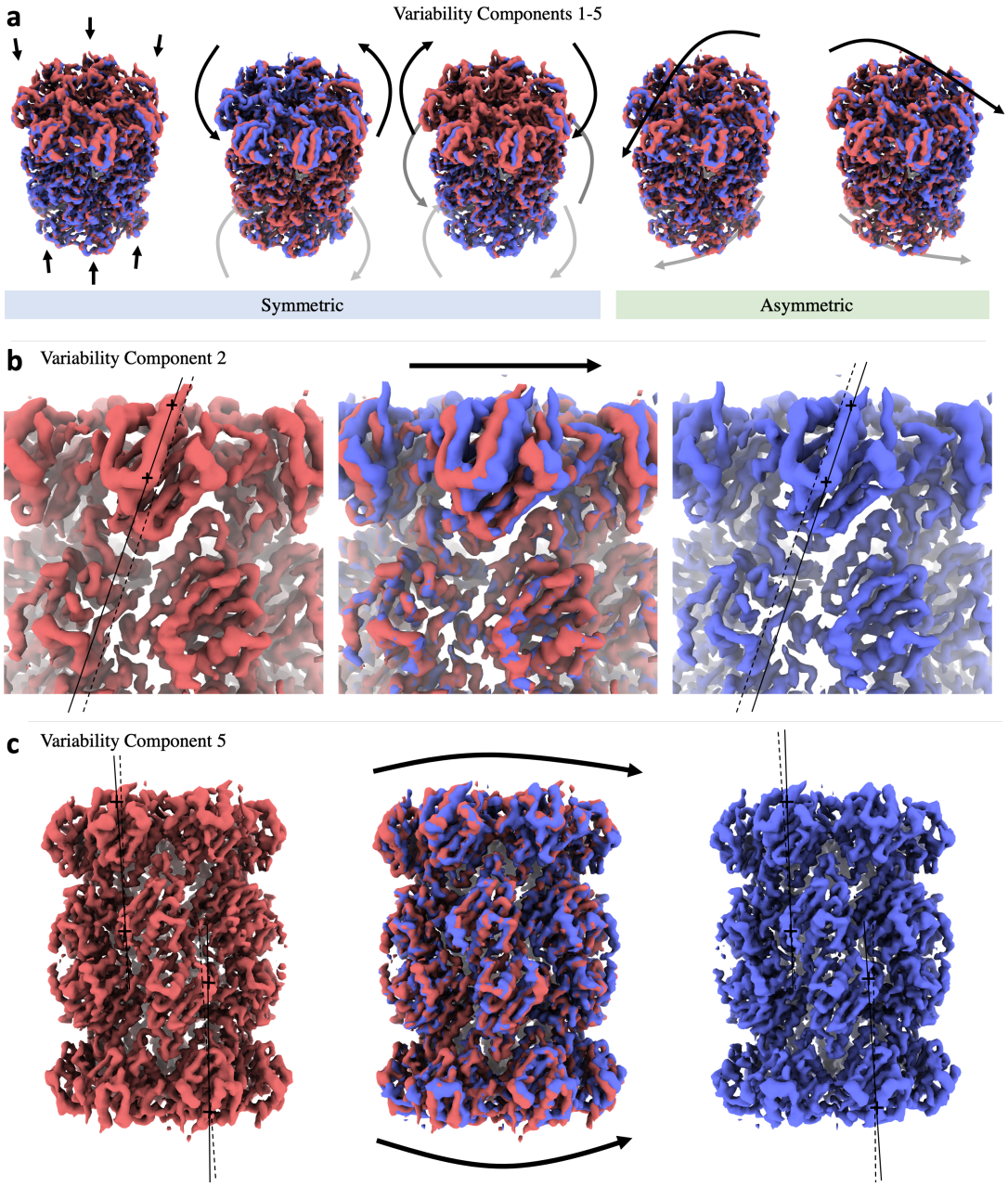

Figure 6: Results of 3DVA with five variability components on 84,604 particle images of the T20S Proteasome. Particles are subject to symmetry expansion around a D7 symmetry. A video version of these results are available at http://3dva.cryosparc.com and as Movie S4. **a**: Overlayed renderings of 3D density maps generated by 3DVA at negative (red) and positive (blue) reaction coordinate values along each of five variability components. The first three components resolve symmetric flexible motion of all subunits simultaneously. The next two components resolve asymmetric bending of the entire molecule where the flexing of each subunit is different. **b**: Detail view of variability component 2, showing rotational motion of the top region of the barrel. **c**: Detail view of variability component 5, showing bending of the entire molecule, breaking D7 symmetry.

top and bottom regions of the barrel. The second component (detailed in Figure 6b) shows rotational motion of the top and bottom of the barrel in opposite directions around the symmetry axis. The third component resolves a twisting motion where the middle and ends of the barrel both rotate around the symmetric axis. The final two components are asymmetric (Figure 6a, right), resolving bending of the entire barrel in two different directions (detail of component 4 in Figure 6c). This result is notable as it shows that 3DVA can resolve several orthogonal modes of detailed high resolution flexible motion of larger molecules. Moreover, it can help to detect pseudo-symmetries created by asymmetric flexibility of symmetric molecules.

Spliceosome: large flexible motions of large complexes

On a dataset of 327,490 particle images of a pre-catalytic spliceosome complex (EMPIAR-10180 [29]), 3DVA was run with two components and a low-pass filter of 8Å and it resolved two large motions of multiple parts of the complex (Figure 7). The first component (Figure 7b) resolves large rotational motion of both the helicase and SF3b regions towards the front/back of the complex, while the foot region also bends slightly forward and back. The second component (Figure 7c) resolves side-to-side rotation of SF3b, diagonal rotation of the helicase, and slight side-to-side bending of the foot region.

In this case, the linear subspace model captures both types of large motion, and the reaction coordinates of individual particles provide an estimate for the position of the helicase and SF3b regions in each image (Figure 7a, top). Nevertheless, the linear subspace model underlying 3DVA is limited in its ability to faithfully represent large motions (Figure 1), and so a simple local weighting strategy can be used to create intermediate 3D reconstructions along each variability component. Particles are weighted based on their position along each reaction coordinate (Figure 7a, bottom) and 3D reconstructions are created using those weighted particles. This type of local neighborhood weighting is commonly used in manifold embedding applications generally, and has been used previously in other methods for manifold embedding of cryo-EM data (e.g. [6]). The resulting 3D density maps depict the molecule at different positions along its continuous flexibility (Figure 7d and 7e).

The pre-catalytic spliceosome dataset was processed in 3DVA in 176 minutes. By comparison, on this dataset the recently proposed cryoDRGN deep generative model [49] was reported to resolve conformational changes similar to the two variability components [48], but it required model training for 50 epochs taking over 1800 minutes, for less than half the number of particles (i.e., 20x slower than 3DVA) on the same hardware.

Bacterial Large Ribosomal Subunit: directly resolving discrete heterogeneity

The bacterial large ribosomal subunit (LSU) is a large macromolecular assembly of several parts, and a cryo-EM dataset has been collected that contains a mixture of intermediate partially assembled states (EMPIAR-10076 [7]). For this dataset of 131,899 particles, 3DVA is run with four variability components and a low-pass filter of 5Å. In this case, the solved 3DVA components lie in a subspace that spans several discrete 3D states. Within that subspace, as shown in Figure 8a, the reaction coordinates of particle

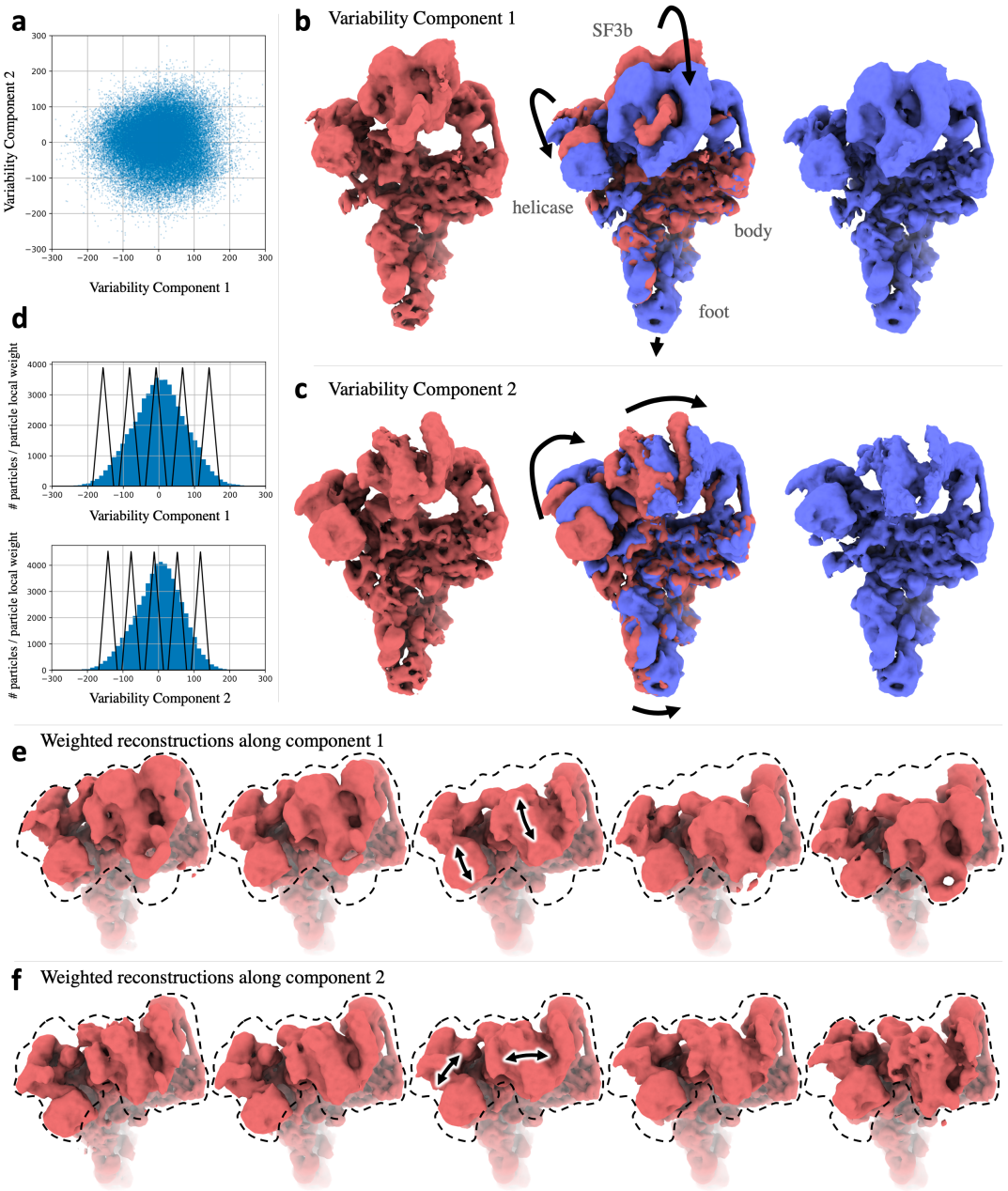

Figure 7: Results of 3DVA with two variabiltiy components on 327,490 particle images of a pre-catalytic spliceosome complex. A video version of these results are available at http://3dva.cryosparc.com and as Movie S5. **a**: Particle images are spread over two reaction coordinates. **b-c**: Renderings of 3D density maps generated by 3DVA at negative (red) and positive (blue) reaction coordinate values. Component 1 resolves forward/back rotation of SF3b and helicase regions, and slight forward/back bending of the food region. Component 2 resolves side-to-side rotation of SF3b and diagonal rotation of helicase. **d**: 1D histograms of both variability components showing local weightings for each of five intermediate positions that are used for weighted reconstructions. **e-f**: Weighted local reconstructions along both variability components, showing five intermediate positions along both types of motion.

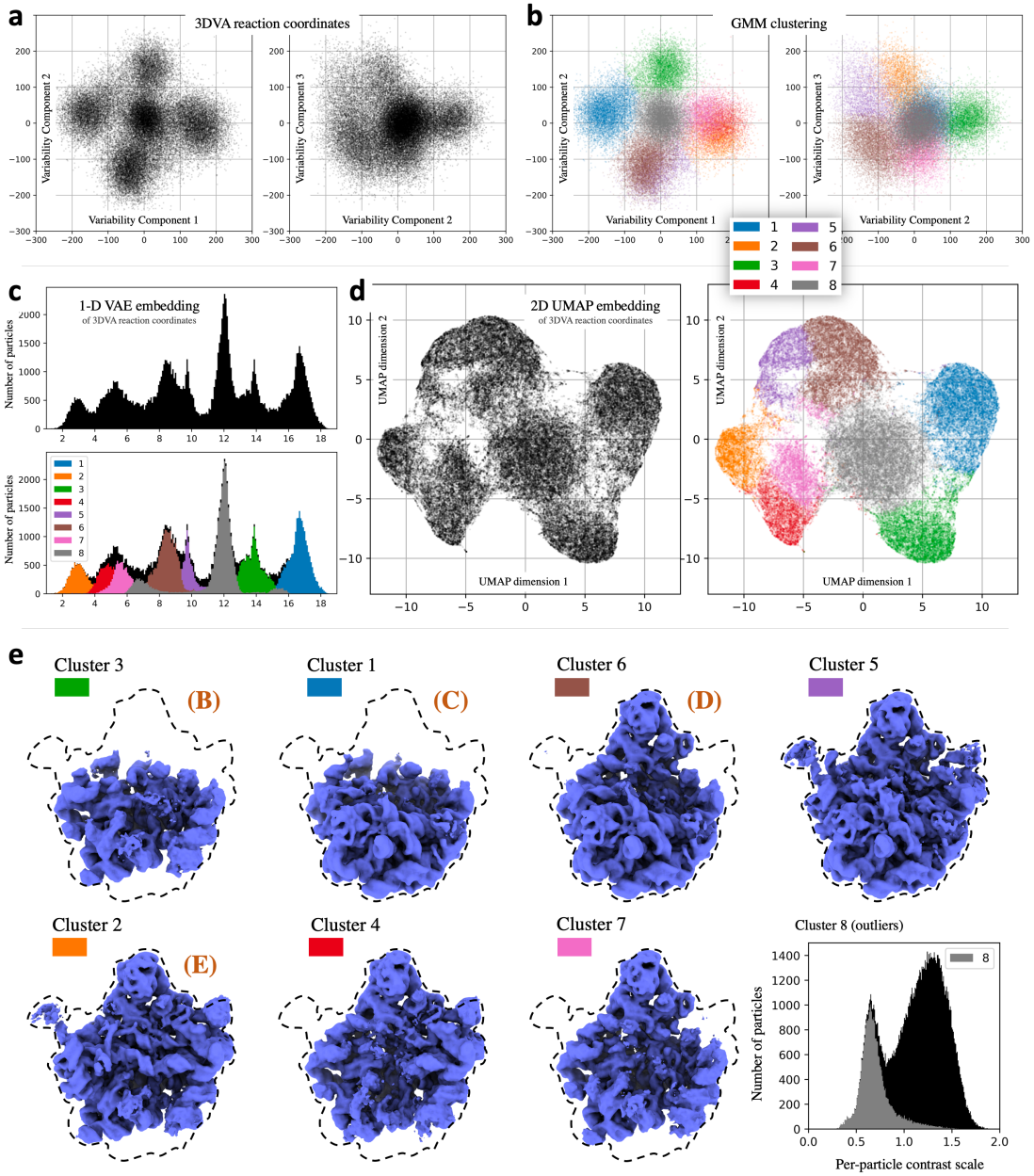

Figure 8: Results of 3DVA with four components on 131,899 particle images of a bacterial ribosomal large subunit. 3DVA identifies clusters corresponding to discrete conformational states. **a**: 3DVA particle reaction coordinates, as 2D scatter plots of sequential pairs of dimensions. **b**: Reaction coordinates colored by 8-way Gaussian Mixture Model (GMM) clustering. **c**: 1D histogram of Variational Autoencoder (VAE) embedding of reaction coordinates. The histogram shows seven major peaks. Overlayed histograms (bottom) of each of the GMM clusters from (b) show that peaks correspond to clusters. **d**: 2D UMAP embedding of reaction coordinates, showing overall geometry of clusters. Points are colored (right) by GMM clusters from (b). **e**: 3D reconstructions from particles in each GMM cluster. Cluster 3 contains the least assembled compled, and other clusters have more subunits assembled. The 8th cluster contains outliers for which 3DVA estimates lower per-particle scale factor α (histogram).

images are well clustered in multiple dimensions. We find that each cluster corresponds to a different partial assembly state of the complete LSU complex.

In order to analyze and understand the clustering of particles in the 3DVA reaction coordinate space, we first fit an 8-component Gaussian Mixture Model (GMM) to the reaction coordinates (Figure 8b). This operation fits 8 clusters to the major groupings of particles. The particles from each of these clusters are then used to perform a 3D reconstruction, and the resulting 3D densities are displayed in Figure 8e. Cluster 3 contains the least density, and each other cluster adds on a part to the assembly. Cluster 5 displays the entire LSU complex. Clusters 1, 2, 3 and 6 correspond to the four large 3D class subpopulations that were reported in the original study (C, E, B, D respectively) [7]. Cluster 8 appears to contain mainly outlier particles (contaminants and junk particles). The presence of outliers was also noted in the original study [7]. The identity of this cluster as outliers can be seen in the histogram of estimated per-particle contrast scales α in Figure 8e, where particles in cluster 8 have a low estimated contrast (indicating a low level of agreement between the particles and the consensus 3D density) compared to all other clusters.

To further investigate and demonstrate the power of 3DVA components to capture discrete heterogeneity, two non-linear embedding methods are applied to the 4-dimensional particle reaction coordinate results. First, a simple Variational Autoencoder (VAE) [17] is trained to reduce the 4-dimensional reaction coordinates to a 1-dimensional embedding of particles where individual clusters can be directly visually identified. The VAE is constructed using *PyTorch* [27], with a single hidden layer of size 1024 in both encoder and decoder architectures. Training was completed in 2 minutes for 40 epochs. The result of training the 1-D VAE is shown in Figure 8c as a histogram of particle positions. The peaks correspond to the multiple discrete states and the outlier cluster detected by GMM fitting (shown using the same colors as Figure 8b). This result can be directly compared with the 1-D Spatial-VAE embedding carried out in processing of this same dataset using the cryoDRGN deep generative model [48], where a similar histogram is generated that resolves four major sub-states in the data, plus one outlier cluster. A second non-linear dimensionality reduction technique, UMAP [23] is applied to the 4-dimensional reaction coordinate results (Figure 8d) displaying the geometry of clusters in the 4-dimensional space, for example that clusters 2, 4, and 7 are sub-clusters of a larger cluster. This result can again be compared with the corresponding results in [48]. This result is notable as it shows the power of 3DVA to resolve nuanced discrete conformational heterogeneity, despite the apparent simplicity of the linear subspace model.

4 Conclusions

We introduce 3D Variability Analysis (3DVA), a new algorithm for fitting high resolution, multidimensional linear subspace models to single particle cryo-EM data. 3DVA uses a version of the Expectation-Maximization algorithm to solve K variability components simultaneously. The variability components and associated reaction coordinates reveal new biological insights from the data, including detailed flexible motion dynamics at the level of individual α-helices, even for small membrane proteins. 3DVA can also be used to identify discrete conformational states of proteins from heterogeneous samples. The

3DVA algorithm does not require tuning or hyperparameter optimization for individual datasets to which it is applied.

The 3DVA algorithm is based on a linear model of the manifold of 3D structures that make up the conformational landscape of a protein molecule. The true manifold can have non-linear and complex geometry, and further development of methods that can model these non-linear manifolds is required. For example, extensions of 3DVA using similar algorithmic techniques may be possible that allow for some forms of non-linearity.

The development of more accurate and powerful generic models of conformational manifolds is only the first step in truly solving flexible protein structures. In addition to modelling the manifold, techniques must be developed that can aggregate structural information across continuous conformational states along a manifold, in order to solve higher resolution 3D density maps of flexible molecules.

An efficient GPU-accelerated implementation of 3D Variability Analysis was previously released and in available in the *cryoSPARC* software package, version v2.9+. Indeed, it has already been used to study continuous flexible motion for a number of interesting new structures. Examples include mTORC1 docked on the lysosome [34], human oligosaccharyltransferase complexes OST-A and OST-B [33], a viral chaperonin [41], the *Acinetobacter baumannii* 70S Ribosome [24], Adrenomedullin Receptors AM1 and AM2 [20], and the prefusion conformation of the SARS-CoV-2 spike protein [46]. .

Acknowledgements

We thank the entire team at Structura Biotechnology Inc. that designs, develops, and maintains the *cryoSPARC* software system in which this project was implemented and tested. This work was financially supported in part by NSERC Canada.

References

[1] J. Andén and A. Singer. Structural Variability from Noisy Tomographic Projections. *SIAM journal on imaging sciences*, 11(2):1441–1492, 2018. 2, 3, 6

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. 5

[3] M. G. Campbell, D. Veesler, A. Cheng, C. S. Potter, and B. Carragher. 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *eLife*, 4:e06380, mar 2015. 15

[4] Q. Cui and I. Bahar. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Taylor and Francis, 2005. 4

[5] S. Dasgupta and L. Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007. 9

[6] A. Dashti, M. S. Shekhar, D. B. Hail, G. Mashayekhi, P. Schwander, A. des Georges, J. Frank, A. Singharoy, and A. Ourmazd. Functional Pathways of Biomolecules Retrieved from Single-particle Snapshots. *bioRxiv*, 2019. 3, 17

[7] J. H. Davis, Y. Z. Tan, B. Carragher, C. S. Potter, D. Lyumkis, and J. R. Williamson. Modular Assembly of the Bacterial Large Ribosomal Subunit. *Cell*, 167(6):1610–1622.e15, dec 2016. 17, 20

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 2020. 6

[9] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, 2007. 4

[10] J. Frank and A. Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data. *Methods*, 100:61–67, 2016. 3

[11] Y. Gong and P. C. Doerschuk. 3-d understanding of electron microscopy images of nano bio objects by computing generative mechanical models. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3161–3165, 2016. 3

[12] T. Grant, A. Rohou, and N. Grigorieff. *cis*TEM, user-friendly software for single-particle image processing. *eLife*, 7:e35383, 2018. 2, 3, 4

[13] N. Grigorieff. FREEALIGN: High resolution refinement of single particle structures. *Journal of Structural Biology*, 157:117–125, 2007. 4

[14] D. Hilger, M. Masureel, and B. K. Kobilka. Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology*, 25(1):4–12, 2018. 12

[15] B. A. Himes and P. Zhang. emClarity: Software for high-resolution cryo-electron tomography and subtomogram averaging. *Nature Methods*, 15:955–961, 2018. 4

[16] T. Hua, K. Vemuri, M. Pu, L. Qu, G. W. Han, Y. Wu, S. Zhao, W. Shui, S. Li, A. Korde, R. B. Laprairie, E. L. Stahl, J.-H. Ho, N. Zvonok, H. Zhou, I. Kufareva, B. Wu, Q. Zhao, M. A. Hanson, L. M. Bohn, A. Makriyannis, R. C. Stevens, and Z.-J. Liu. Crystal Structure of the Human Cannabinoid Receptor CB1. *Cell*, 167(3):750–762.e14, oct 2016. 15

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2013. 20

[18] K. Krishna Kumar, M. Shalev-Benami, M. J. Robertson, H. Hu, S. D. Banister, S. A. Hollingsworth, N. R. Latorraca, H. E. Kato, D. Hilger, S. Maeda, W. I. Weis, D. L. Farrens, R. O. Dror, S. V. Malhotra, B. K. Kobilka, and G. Skiniotis. Structure of a Signaling Cannabinoid Receptor 1-G Protein Complex. *Cell*, 176(3):448–458.e12, jan 2019. 12

[19] R. R. Lederman and A. Singer. Continuously heterogeneous hyper-objects in cryo-em and 3-d movies of many temporal dimensions, 2017. 3

[20] Y.-L. Liang, M. J. Belousoff, M. M. Fletcher, X. Zhang, M. Khoshouei, G. Deganutti, C. Koole, S. G. B. Furness, L. J. Miller, D. L. Hay, A. Christopoulos, C. A. Reynolds, R. Danev, D. Wootten, and P. M. Sexton. Structure and Dynamics of Adrenomedullin Receptors AM1 and AM2 Reveal

Key Mechanisms in the Control of Receptor Phenotype by Receptor Activity-Modifying Proteins. *ACS Pharmacology & Translational Science*, mar 2020. 21

[21] D. Lyumkis. Challenges and opportunities in cryo-EM single-particle analysis. *The Journal of biological chemistry*, 294(13):5181–5197, mar 2019. 1

[22] S. Maji, H. Liao, A. Dashti, G. Mashayekhi, A. Ourmazd, and J. Frank. Propagation of Conformational Coordinates Across Angular Space in Mapping the Continuum of States from Cryo-EM Data by Manifold Embedding. *Journal of Chemical Information and Modeling*, mar 2020. 3

[23] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. 20

[24] C. E. Morgan, W. Huang, S. D. Rudin, D. J. Taylor, J. E. Kirby, R. A. Bonomo, and E. W. Yu. Cryo-electron Microscopy Structure of the Acinetobacter baumannii 70S Ribosome and Implications for New Antibiotic Development. *mBio*, 11(1), 2020. 21

[25] A. Moscovich, A. Halevi, J. Anden, and A. Singer. Cryo-em reconstruction of continuous heterogeneity by laplacian spectral volumes. *Inverse Problems*, 36 024003, 2020. 3

[26] T. Nakane, D. Kimanius, E. Lindahl, and S. H. W. Scheres. Characterisation of molecular motions in cryo-em single-particle data by multi-body refinement in relion. *eLife*, 7:e36861, 2018. 4

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 20

[28] P. A. Penczek, M. Kimmel, and C. M. T. Spahn. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. *Structure (London, England : 1993)*, 19(11):1582–1590, nov 2011. 2, 3, 6

[29] C. Plaschka, P.-C. Lin, and K. Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, 2017. 17

[30] A. Punjani, M. A. Brubaker, and D. J. Fleet. Building proteins in a day: Efficient 3d molecular structure estimation with electron cryo-microscopy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):706–718, 2017. 4

[31] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker. CryoSPARC: Algorithms for rapid unsupervised cryo-em structure determination. *Nature Methods*, 14:290–296, 2017. 3

[32] A. Punjani, H. Zhang, and D. J. Fleet. Non-uniform refinement: Adaptive regularization improves single particle cryo-EM reconstruction. *bioRxiv*, 2019. 10, 15

[33] A. S. Ramírez, J. Kowal, and K. P. Locher. Cryo–electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B. *Science*, 366(6471):1372 LP – 1375, dec 2019. 21

[34] K. B. Rogala, X. Gu, J. F. Kedir, M. Abu-Remaileh, L. F. Bianchi1, A. M. Bottino1, R. Dueholm1, A. Niehaus1, D. Overwijn1, A. C. Priso Fils1, S. X. Zhou1, D. Leary, N. N. Laqtom1, E. J. Brignole, and D. M. Sabatini. Structural basis for the docking of mTORC1 on the lysosomal surface. *Science*, 366(6464):468–475, oct 2019. 21

[35] S. Roweis. EM algorithms for PCA and SPCA. *Advances in Neural Information Processing Systems*, pages 626–632, 1998. 6, 7, 9

[36] S. H. Scheres. Processing of structurally heterogeneous cryo-em data in RELION. *Methods Enzymol.*, 579:125–157, 2016. 2, 3, 4

[37] S. H. W. Scheres. A Bayesian view on cryo-em structure determination. *Journal of Molecular Biology*, 415:406–418, 2012. 4

[38] S. H. W. Scheres. RELION: Implementation of a Bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519 – 530, 2012. 2, 3, 4

[39] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods*, 4(1):27–29, 2007. 2, 3, 4

[40] S. Schilbach, M. Hantsche, D. Tegunov, C. Dienemann, C. Wigge, H. Urlaub, and P. Cramer. Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature*, 551(7679):204–209, 2017. 4

[41] T. B. Stanishneva-Konovalova, P. I. Semenyuk, L. P. Kurochkina, E. B. Pichkur, A. L. Vasilyev, M. V. Kovalchuk, M. P. Kirpichnikov, and O. S. Sokolova. Cryo-EM reveals an asymmetry in a novel single-ring viral chaperonin. *Journal of Structural Biology*, 209(2):107439, 2020. 21

[42] H. D. Tagare, A. Kucukelbir, F. J. Sigworth, H. Wang, and M. Rao. Directly reconstructing principal components of heterogeneous particles from cryo-em images. *J Structural Biology*, 191(2):245–262, 2015. 2, 3, 7

[43] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J Royal Statistical Society, series B*, 61:611–622, 1999. 6, 9

[44] M. van Heel, R. V. Portugal, and M. Schatz. Multivariate Statistical Analysis of Large Datasets: Single Particle Electron Microscopy. *Open Journal of Statistics*, 06(04):701–739, jul 2016. 2, 6

[45] W. Wong, X.-c. Bai, A. Brown, I. S. Fernandez, E. Hanssen, M. Condron, Y. H. Tan, J. Baum, and S. H. W. Scheres. Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, jun 2014. 12

[46] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, and J. S. McLellan. Cryo-em structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483):1260–1263, 2020. 3, 21

[47] H. Xu, T. Li, A. Rohou, C. P. Arthur, F. Tzakoniati, E. Wong, A. Estevez, C. Kugel, Y. Franke, J. Chen, C. Ciferri, D. H. Hackos, C. M. Koth, and J. Payandeh. Structural basis of Nav1.7 inhibition by a gating-modifier spider toxin. *Cell*, 2019. 15

[48] E. D. Zhong, T. Bepler, B. Berger, and J. H. Davis. CryoDRGN: Reconstruction of heterogeneous structures from cryo-electron micrographs using neural networks. *bioRxiv*, 2020. 3, 12, 17, 20

[49] E. D. Zhong, T. Bepler, J. H. Davis, and B. Berger. Reconstructing continuously heterogeneous structures from single particle cryo-em with deep generative models. *International Conference on Learning Representations*, 2020. 3, 12, 17