Open access • Posted Content • DOI:10.1101/2020.04.08.032466

3D Variability Analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM — Source link

Ali Punjani, David J. Fleet

Institutions: University of Toronto

Published on: 18 Nov 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Ribosome assembly

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

Ali Punjani[1,2,3] David J. Fleet[1,2]
alipunjani@cs.toronto.edu fleet@cs.toronto.edu

[1]University of Toronto : Department of Computer Sciences, University of Toronto, Canada M5S 3G4
[2]Vector Institute : 710-661 University Ave. Toronto, Canada M5G 1M1
[3]Structura Biotechnology Inc. : 129-100 College Ave. Toronto Canada M5G 1L5

Abstract

Single particle cryo-EM excels in determining static structures of protein molecules, but existing 3D reconstruction methods have been ineffective in modelling flexible proteins. We introduce 3D variability analysis (3DVA), an algorithm that fits a linear subspace model of conformational change to cryo-EM data at high resolution. 3DVA enables the resolution and visualization of detailed molecular motions of both large and small proteins, revealing new biological insight from single particle cryo-EM data. Experimental results 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, large motions in a spliceosome complex, and discrete conformational states of a ribosome assembly. 3DVA is implemented in the *cryoSPARC* software package.

Keywords: Single particle analysis; Electron Microscopy; Principle Components; Continuous Heterogeneity; Image Processing; Expectation Maximization

1 Introduction

Protein dynamics hold the key to understanding function in many molecular machines. 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 thus holds great promise for studying protein dynamics [22]. However, there are significant computational challenges in resolving dynamics from static image data. These difficulties stem from the need to simultaneously estimate the canonical structure of the molecule, the multiple ways in which the structure can change (due to motion, dissociation, etc), and the state of the molecule in each particle image. These problems are exacerbated by 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 particle in each image and the 3D model.

A heterogeneous target molecule will exist in different conformational states within a sample. These 3D structures lie on a manifold in the abstract space of all possible 3D structures. Each observed particle image X provides a noisy, CTF corrupted, 2D projection of the 3D structure $\mathcal{V}$ at a point on the manifold. The manifold may have a non-linear geometry in general, but for many molecules of interest a linear subspace model provides an effective way to represent the 3D structures in a cryo-EM sample. A K-dimensional linear subspace represents 3D structures present in the sample in terms of a base structure V_0 and a weighted sum of K components, V_k:

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

The weight vector, $\mathbf{z} = (z_1, ..., z_K)$, or latent coordinates, represent the conformational state of the molecule in a given particle image. Subspace models, such as Principal Component Analysis (PCA), have been discussed extensively in the cryo-EM literature [1, 30, 45, 47]. They are attractive for their simplicity, analytical properties, and the existence of stable numerical methods for eigenvector computation. Nevertheless they are not widely used in practice (e.g., compared to K-way discrete 3D classification [12, 39, 41, 42]), as existing methods cannot estimate the components in (1) at high resolutions, nor estimate several components simultaneously. As a consequence they fail to resolve the detailed conformational changes that are most important at high resolutions.

We introduce *3D Variability Analysis* (3DVA) for fitting 3D linear subspace models to single particle cryo-EM data. Following Tagare et al [45], the algorithm is formulated as a variant of the Expectation-Maximization algorithm for Probabilistic PCA. By decoupling specific tasks in the numerical optimization, we obtain a method that enables efficient computation of accurate, high-resolution representations. 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. It resolves high-resolution flexibility, with motions of individual α-helices as small as a few angstroms, for both large and small proteins. To our knowledge, this is the first method to resolve flexible dynamics of small proteins, such as the distinct types of flexible motion we find in the sub-50 kDa transmembrane region of a GPCR complex, at resolutions below 4Å.

By resolving continuous conformational changes, 3DVA reveals biological insights from single particle cryo-EM data beyond the reach of existing methods, simplifying the analysis of conformational heterogeneity. We show that it can resolve ratcheting motions of a ribosome, large flexible motion of a pre-catalytic spliceosome complex, and several discrete conformational states of a bacterial large ribosomal subunit assembly. For such large, multi-MDa complexes, recently proposed deep generative models [52, 51] have been experimentally demonstrated to resolve similar continuous and discrete heterogeneity at coarse resolution. However, 3DVA is 20-50x faster to optimize, requiring no manual parameter tuning or hyperparameter optimization on individual datasets.

3DVA was released in the *cryoSPARC* [33] software package v2.9+ and has been used in several structural studies, including the SARS-CoV-2 virus spike protein [49], mTORC1 docked on the lysosome [37], human oligosaccharyltransferase complexes OST-A and OST-B [35], bacterial unfoldase-protease complex ClpXP [36], a viral chaperonin [44], the *Acinetobacter baumannii* 70S Ribosome [25], and Adrenomedullin Receptors AM1 and AM2 [21].

2 Linear Subspaces, PCA, and 3D Variability Analysis

Under the standard cryo-EM image formation model [8, 13, 32, 40], a 2D particle image, X_i, is a corrupted projection of the target 3D density $\mathcal{V}$ from an unobserved pose ϕ_i, plus additive noise, η:

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

where $P(\phi_i)$ is the projection operator and C_i is the contrast transfer function (CTF) for image i. The standard model assumes all images are generated from a single 3D density $\mathcal{V}$. It assumes a homogeneous population of molecules, all in the same conformation, differing only by a rigid coordinate transformation. K-way 3D classification methods [12, 39, 41, 42] extend this mode, assuming each particle image is generated from one of K different, independent, 3D densities. This assumes the sample comprises a small number of distinct structures.

Many protein molecules have a mechanistic function, exhibiting flexible motion across a continuous landscape of energetically favourable conformations. This flexibility can result in a continuum of conformational states present in a frozen sample. In such cases, algorithms based on assumptions of the homogeneous or K-way classification models yield low quality reconstructions that do not shed light on the biological function of molecular motions.

3D variability analysis accounts for continuous conformational flexibility in terms of a K-dimensional linear subspace model. Under a variant of the standard probabilistic PCA model, each experimental particle image is generated from a **mean density**, V_0, plus weighted contributions of several **variability components**, $V_{1:K}$. The model also includes a per-particle scale factor, α_i, to account for varying ice-thickness and contrast level, and additive Gaussian noise, η. In particular, η is assumed to have zero mean and a diagonal covariance matrix in the Fourier domain, with constant variance in annular rings. White noise is a special case. Formally,

$$X_i \;=\; \alpha_i\, C_i\, P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) + \eta \;. \tag{3}$$

Each V_k is a 3D density, capturing a particular type of change to V_0. The vector of per-particle weights, $\mathbf{z}_i = (z_{i1}, ..., z_{iK})$ are also known as the **latent coordinates**.

The predominant method for learning linear subspace models in many domains is Principal Component Analysis (PCA) [2], under the assumption that the basis functions are orthogonal; i.e., $V_i^{\mathrm{T}} V_j \;=\; 1$ if $i \;=\; j$, and 0 otherwise. Given a random data sample (e.g., 3D densities), assumed to be generated independently from the same underlying distribution, for PCA one would computes the sample mean (V_0 in (3)), and the leading eigenvectors of the covariance matrix of the data distribution (i.e., those with the largest eigenvalues). These *principal directions* account for the most significant factors of variation in the data (like $V_{1:K}$ in (3).

Direct application of PCA to single particle cryo-EM data is challenging however. First, each 2D image is a partial observation of the 3D density from one direction, so construction of the full covariance of 3D densities from images is non-trivial. Second, the high dimensionality of the covariance matrix means that it may not be possible to store in memory, or to compute its eigenvectors except at low resolutions. Third, each image is corrupted by a different CTF, violating the PCA assumption that data are generated from the same distribution.

To mitigate these challenges one could restrict analysis to 2D by computing the 2D covariance of particles within a single 2D class or viewing direction. This removes the partial observation problem, but fails to capture 3D variability or account for the CTF [47]. In 3D one can use bootstrapping [30], wherein multiple random subsets of images are used to reconstruct multiple 3D structures under the homogeneous model (Eq. 2). With such ”bootstrapped” 3D densities one avoids the missing information problem due to projection, but it can only operate at low resolutions to avoid high-dimensional covariance matrices. It further suffers from statistical inefficiency due to the random noise injected into each bootstrap sample. Together, these drawbacks limit such methods to resolving coarse resolution eigenvectors that capture gross structural changes of large protein complexes [30]. One can also employ low-dimensional approximations to the covariance, which mitigate storage and computational cost, but such methods have been demonstrated only at coarse resolutions [1].

As explained in the Methods section, 3D Variability Analysis overcomes the challenges with standard PCA techniques, enabling computation of variability components at high resolution and for smaller proteins. 3DVA is a form of Probabilistic PCA [38, 46], which assumes data are drawn from a high dimensional Gaussian distribution. The model assumes Gaussian observation noise in (3), and a Gaussian prior over latent coordinates. The Expectation-Maximization algorithm [7] is used to obtain a Maximum Likelihood (ML) estimate of the variability components $V_{1:K}$, along with ML estimates for the noise covariance and the latent coordinates z. Crucially, this PPCA algorithm accommodates partial observations and data-specific corruption (the CTF in cryo-EM data), and it works well with high-dimensional data without the need to explicitly store or approximate the 3D covariance matrix [38]. It also supports direct estimation of only the top K components, as desired in 3DVA. For a set of M images, the log likelihood objective is, up to an additive constant, a weighted sum of squared residual errors between images, X_i, and model predictions:

$$E(V_{1:K}, \alpha_{1:M}, \mathbf{z}_{1:M}) \;=\; \sum_{i}^{M} \frac{1}{2} \left\| X_i - \alpha_i C_i P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) \right\|_{\Lambda}^{2}, \tag{4}$$

where Λ is the inverse covariance (a.k.a. precision matrix) of the observation noise η, and $\|v\|_{\Lambda}^{2} \equiv v^T \Lambda v$.

The 3DVA algorithm assumes the mean density, V_0, along with the per-particle CTFs, C_i, and poses, ϕ_i, are known. These quantities could be estimated by processing the image data using standard homogeneous refinement. Given these quantities, and random initial values for the variability components, 3DVA uses iterative optimization, each iteration of which has an E-step and an M-step. As explained in Methods, based on a free-energy derivation of the algorithm, the E-step updates the mean of the posterior distribution over the latent coordinates, $\mathbf{z}_i$, independently for each image X_i. The M-step then uses the expected log liklihood to update the per-particle scale parameter α_i, and the variability components, i.e., $V_{1:K}$.

Tagare et al. [45] proposed a similar subspace model, using Maximum Likelihood estimation with a form of Expectation-Maximization. However, their M-step is solved via an approximate iterative conjugate gradient, and each V_k is solved sequentially, requiring K complete rounds of optimization, each passing through the data

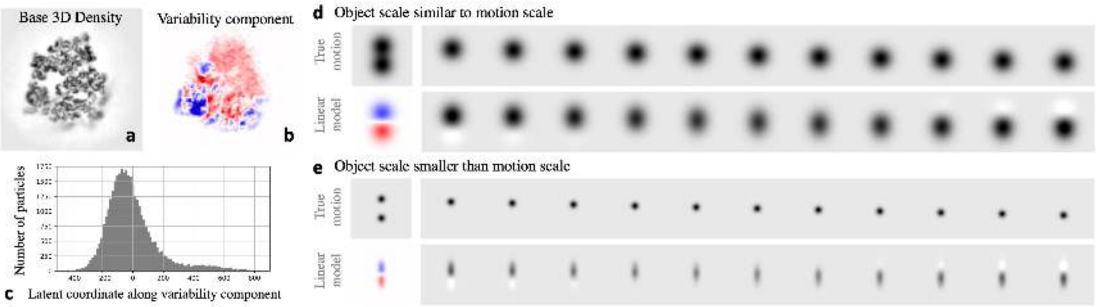

Figure 1: Examples of 3DVA. **a:** A central slice through a 3D density map reconstructed from 80S ribosome particle images. **b:** A central slice of the first 3D variability component for the 80S ribosome reconstruction in **a**. Positive (red) and negative (blue) values correspond to density to be added and subtracted from the mean density in **a** to explain heterogeneity in the image data. This variability component captures motion of the small subunit (bottom left, darker blue) relative to the large subunit. **c:** Histogram showing the distribution of latent coordinates for the component in **b**. **d, e:** Synthetic examples of how a linear subspace model represents object motion. In both cases, the top row shows observed images of an object in motion, while the bottom shows the first variability component (in red / blue) and images generated by the model to explain the motion. When the displacement is small relative to object size (**d**) the model approximates the motion well. When the object is smaller than the motion (**e**), the linear model is less accurate.

repeatedly until convergence for each V_k. Furthermore, the CTF is treated approximately using Wiener filtering, image noise is assumed to be white, and particles are binned into a finite number of 3D pose bins. Thus, they were only able to compute linear subspace models for coarse low-resolution motion and variations; results on experimental data are demonstrated only at 15Å resolution, and only for $K = 2$ variability components.

3DVA computes the M-step exactly via matrix operations, without approximate methods, and all K variability components are solved simultaneously at full resolution. In addition, per-particle scale factors α_i and a colored noise model are optimized in parallel. During optimization, the variability components $V_{1:K}$ are regularized by a low-pass filter and high-pass filter. The low-pass filter attenuates high-frequency noise to prevent over-fitting. An optional high-pass filter removes sources of low-resolution variability caused by contaminants (e.g. denatured protein mass at air-water interfaces).

A GPU implementation of 3DVA in *cryoSPARC* allows one to process large experimental datasets with 10^6 particle images. It resolves multiple variability components simultaneously at resolutions as high as 3Å-4Å, limited only by GPU memory and signal present in the data.

2.1 Interpretation and properties of 3DVA

When 3DVA is applied to single particle EM data, it outputs variability components, $V_{1:K}$, and per-particle latent coordinates, $\mathbf{z}$. Variability components are those directions in the space of 3D density that, from the mean structure, V_0, define a linear subspace that best fits the principal directions of variation in the data. Each component is a 3D voxel grid of density values. These value indicate where density should be added or removed to explain variability amongst the particles. Orthogonality of the K components ensures that each explains a different mode of variation. Fig. 1 shows an example of a variability component on a ribosome dataset. Variability components found by 3DVA often capture flexible motion of the underlying molecule in a sample. Fig. 1d,e show qualitatively how linear subspace models capture object motion. Displacements up to the scale of the object are well approximated. Small objects moving large distances are less well modelled. That is, 3DVA can represent large motion in low resolution regions of a 3D density map, as well as small motions in high resolution regions.

3DVA also yields latent coordinates, $\mathbf{z}$, for each particle image. They indicate the level to which each variability component is present in a given image. If the ith particle image has a large positive value for z_{ik}, this means that it is best described by adding a large amount of V_k to V_0. The variance of each latent coordinate across the images provides a measure of importance, since components with large variance explain the most variability in the data. When a population of particle images contains well-separated clusters of different conformations, 3DVA components will identify the differences between clusters, and the latent coordinates of particles will be similarly clustered, providing insight into discrete heterogeneity.

3DVA is not sensitive to initialization bias. Generally, iterative algorithms that solve non-convex optimization problems (like ML estimation of the linear subspace model) can give incorrect results if initialized poorly. It has been proven, however, that for PPCA models (like 3DVA) any stable local optima is also a global optimum of the objective function [46]. 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 and other non-linear methods, where initialization is critical and multiple runs can often yield significantly different results.

It also notable that 3DVA, using Expectation-Maximization, is a Krylov subspace method [38]. Krylov subspace methods are a general class of iterative optimization algorithms for computing the eigenvectors of matrices, used in popular linear algebra techniques like SVD [10]. Krylov subspace methods like 3DVA have the beneficial property of solving eigenvectors in order of importance. Running 3DVA with $K = 3$ will yield the top three variability components that explain the most variability in the data, and running with $K = 6$ will yield the same three plus the next three important components.

With these properties and the stability of Expectation-Maximization, 3DVA does not require tuning or parameter changes between datasets, other than resolution limits and the number of components, K. Since 3DVA depends on the quality of the mean density, V_0 and the particle alignments, methods that produce high quality consensus structures and alignments in the presence of flexibility and heterogeneity, e.g., non-uniform refinement [34], can improve the results of 3DVA.

3 Results

In what follows, we consider one synthetic dataset and six real-world cryo-EM datasets. Results on the synthetic dataset help to validate the approach and convergence of 3DVA to the true underlying principle linear subspace of the data. Results on the real-world datasets demonstrate the ability of the method to resolve high resolution continuous and discrete conformational changes from single particle cryo-EM.

For the real-world experimental data we first compute a consensus refinement using all particle images. In the case of membrane proteins, non-uniform refinement [34] is used to improve image alignments and overall resolution. The resulting particle images and pose alignments are then used to compute variability components and latent

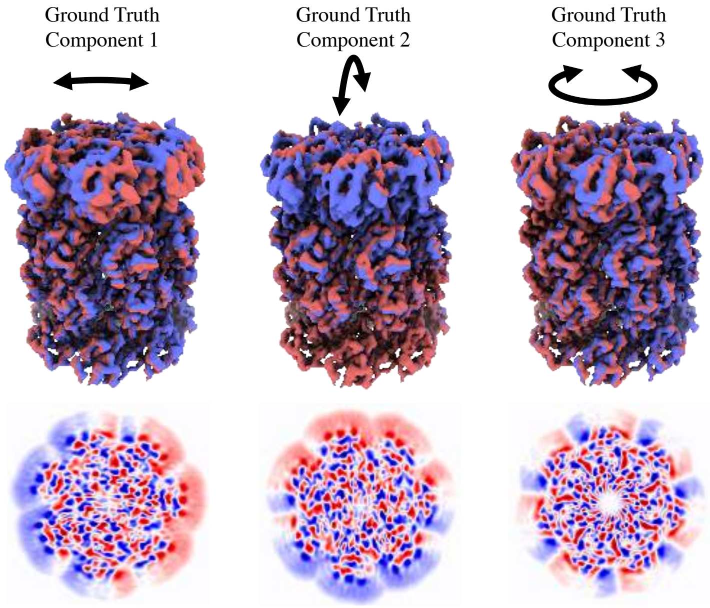

Figure 2: Ground-truth linear subspace directions of a synthetic T20S proteasome molecule used to generate synthetic particle image data. **Top row:** 3D renderings of the consensus density plus (red) and minus (blue) each subspace direction at one standard deviation. The three components represent bending of the molecule around the x-axis, bending around the y-axis, and twisting about the z-axis. **Bottom row:** Slices in the $x - y$ plane of the three subspace directions depicted in the top row. The slices show positive and negative values where density is added and subtracted according to the subspace direction.

coordinates via 3DVA. No prior information is provided about the type or form of heterogeneity in each dataset. 3DVA is run with a mask that excludes solvent, and for membrane proteins the mask also excludes detergent micelle or lipid nanodisc. With all experimental datasets here, 3DVA executes 20 iterations of Expectation-Maximization, starting from random initialization. All experiments are run on a single workstation with one NVIDIA V100 GPU.

For data with continuous heterogeneity, we render the 3D density generated by the subspace model at positive (blue) and negative (red) values of the latent coordinate for individual variability components, along with a superimposed rendering of both densities to clearly indicate the difference between them. Guide lines (solid and dashed) and feature markers (+) provide visual reference points. For each dataset, videos displaying flexible conformational changes are available as supplementary movies.

3.1 Synthetic T20S

We generate synthetic data by deforming the 3D density map of a T20S proteasome [3] at $3\mathring{A}$. The *ground truth* deformations lie in a three dimensional linear subspace, the basis functions for which correspond to bending along each of the x, y, and z axes. Figure 2 shows 3D renderings of the consensus refinement plus/minus (red/blue) one standard deviation along each of the ground-truth subspace directions (top row), along with slices in the $x - y$ plane through each of the ground-truth subspace directions (bottom row).

Each synthetic dataset comprises $100,000$ particle images. For each particle image, we randomly sample *ground-truth* latent coordinates from a mean-zero Gaussian distribution with a diagonal covariance matrix, the diagonal variances of which are $50^2, 40^2$ and 30^2, corresponding to bending along the x, y, and z axes. The corresponding deformation is then applied to the T20S density map. A viewing direction is then randomly drawn from a uniform distribution over the sphere. For CTF parameters, average defocus is drawn at random uniformly from the interval $[0.5\mu m, 2.0\mu m]$, astigmatism from the interval $[-0.1\mu m, 0.1\mu m]$, and astigmatism angle uniformly around the circle. Accelerating voltage is set to 300kV, spherical aberration to 2.7mm, and per-particle scales to 1.0. Particle images are generated using Eq. 3.

Particle images are collectively normalized to have unit signal variance on average. White noise is them generated and added to the particle images. To this end we use four noise levels, namely, $\sigma = 0$, 4, 10, and 20, corresponding to noiseless data, and SNRs of $1/16$, $1/100$, and $1/400$ respectively. This yields four datasets, each with 100,000 particle images, that are used to test 3DVA (see Figure 3a).

At each noise level, we run 3DVA using the ground-truth poses and CTF parameters for each particle, estimating 3 variability components. A consensus reconstruction of input particles is used as the mean density V_0. Low-pass filtering is not used, and initialization is random. In all cases, 3DVA correctly recovers the true subspace, as well as the latent coordinates of particles. Figure 3b shows slices through the $x - y$ plane for estimated variability component 1 at each noise level. These figures can be compared against Figure 2 (bottom left). In all cases, 3DVA resolves the ground-truth subspace direction. Noise is present in the estimated variability component due to the inherent noise in the input images. Figure 3c shows scatter plots of the true latent coordinates for component 1 against the estimated coordinates for component 1. When the noise level is low or zero (left), latent coordinates are recovered with nearly perfect correlation to the true coordinates. As noise increase the estimates become more varied, but correlation with the true coordinates remains high. Figure 3d shows values of the correlation coefficient between ground-truth subspace directions (x-axis) and estimated variability components (y-axis) at each noise level. These results show that in all cases, 3DVA recovers the three subspace directions in the correct order, and without significant confusion between the directions. In the noiseless case, correlations are nearly perfect. In the three noisy cases, correlation decreases as the variability components are affected by noise.

Despite the noise in input images, 3DVA estimated variability components do recover the true signal in the subspace directions at high resolutions. To see that this is the case, we first compute FSC curves between the consensus reconstruction using particles at each noise level and the ground-truth concensus reconstruction (Figure 4a). This baseline indicates the degree to which noise limits the algorithm's ability to recover signal from the finite number of images. As expected, the noiseless reconstruction has nearly perfect correlation at all frequencies against the ground-truth. As the noise level is increased, resolvable signal decreases and resolution is limited.

Next, we compute FSC curves between the 3DVA estimated variability components and the ground truth subspace directions. Figure 4b-d show the FSC curves for components 1, 2, and 3 respectively. Again, we see that in the noiseless case, 3DVA recovers the true subspace directions with nearly perfect correlation at all frequencies. As noise increases, components are resolved with decreasing resolution. Notably, the decrease in resolution of variability components mirrors the decrease in consensus refinement resolution, indicating that the inherent noise in particle images is the main limiting factor for 3DVA estimates, rather than the 3DVA algorithm. In particular, in the noisiest case of $\sigma = 20$, the consensus reconstruction is limited to an FSC=0.5 value of $4.0\mathring{A}$ while all three subspace components reach nearly the same resolution, $4.1\mathring{A}$.

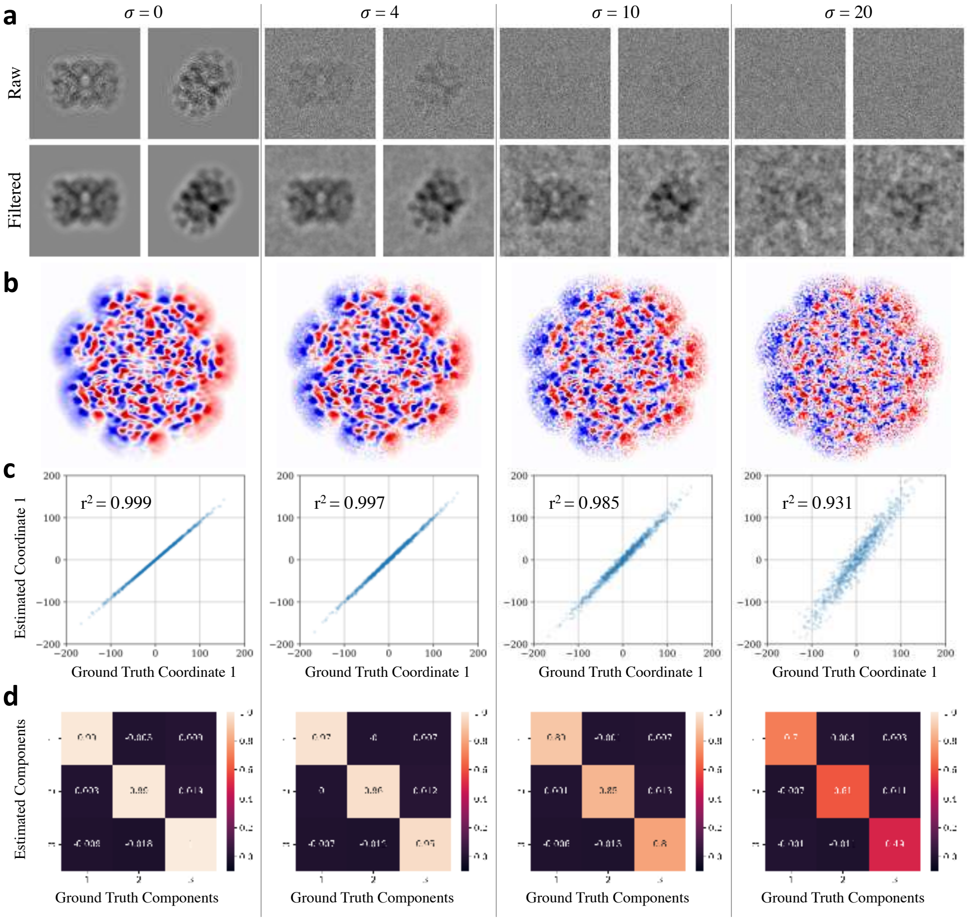

Figure 3: Synthetic input data and validation results of 3DVA. Columns from left to right depict data and results with noiseless data, noisy data with $\sigma = 4$, $\sigma = 10$, and $\sigma = 20$. 3DVA is run with particles at each noise level, solving three variability components. **a:** Example synthetic images in the dataset. Raw particle images are shown above, and filtered images (with low-pass filter of $30\AA$) below. In the noiseless case, ringing from the CTF is visible in the raw particle images. As the noise level is increased, the particle becomes less distinguishable even after filtering. **b:** Results of 3DVA used to recover variability components. Slices in the $x - y$ plane through the estimated first variability component are shown, with red and blue corresponding to positive and negative values respectively. These results can be compared with Figure 2 (bottom left), showing that 3DVA resolves the true underlying linear subspace direction. **c:** Scatter plots showing ground-truth latent coordinates for component 1 (x-axis) vs 3DVA estimated latent coordinates for component 1 (y-axis) across the dataset. Every 100th point is shown. High correlation (r^2 values) are obtained even in noisy cases. **d:** "Confusion matrices" showing the level on alignment between ground-truth and 3DVA estimated variability components. Each value is the correlation coefficient between true (x-axis) and estimated (y-axis) components. Values of 1.0 indicate perfect correlation and 0.0 no correlation. In all cases, 3DVA resolves the true subspace directions in the correct order from most signficant to least significant, with nearly zero confusion between the subspace directions. Correlations decrease as noise is added.

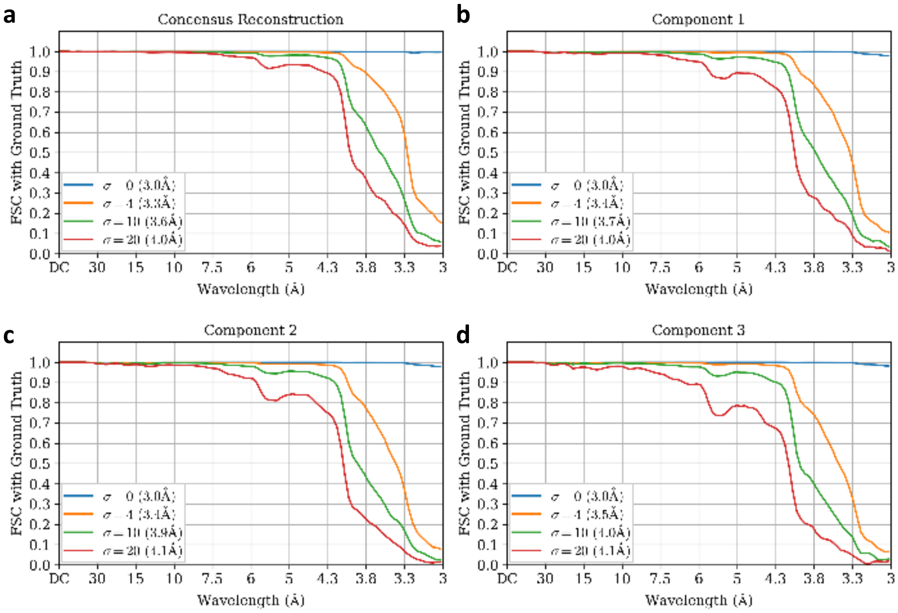

Figure 4: FSC curves showing the ability of 3DVA to resolve variability components to high resolution even in the presence of substantial noise, on the synthetic T20S datasets. **a:** FSC curves between the ground-truth consensus density map and consensus reconstruction from particle images, at each noise level. These curves serve as a baseline, indicating the level to which reconstruction from the finite, noisy particle set can yield resolvable signal. As expected, noiseless data reconstructs nearly perfectly at all resolutions. As noise is added, resolvable resolution decreases. **b,c,d:** FSC curves between the ground-truth linear subspace direction and the 3DVA estimated variability component, for components 1, 2, and 3 respectively. In the noiseless case, 3DVA recovers the true subspace directions nearly perfectly. As noise as added, resolution of variability components decreases, mirroring the consensus resolution. Notably, variability components can be resolved by 3DVA at nearly the same resolutions as the consensus reconstruction, indicating that noise is the primary limiting factor rather than the 3DVA algorithm.

3.2 Cannabinoid GPCR: High resolution flexible motion of small proteins

GPCRs are small membrane proteins responsible for cell signalling and transmission [15]. 3DVA is applied to a dataset of Cannabinoid Receptor 1-G GPCR complex particles [19], containing the CB1 GPCR, G protein, and scFv. Raw microscope images (EMPIAR-10288) are processed in *cryoSPARC* v2 to obtain a consensus refinement of the entire structure (Fig. 5a) from 250,649 particle images. When 3DVA is first run on the entire 3D map, the variability components are dominated by changes in the shape and position of the micelle (Fig. 5b and 5c). To inspect the heterogeneity of the protein itself, a mask is used to exclude solvent and detergent micelle (Fig. 5d). 3DVA is then run using three variability components and a low-pass filter resolution of 3Å.

For the Cannabinoid receptor (Fig. 5e-f), 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 small subregion of the complex. The entire CB1 protein is only 53 kDa [17] and the CB1 transmembrane domain is even smaller.

3DVA is, to our knowledge, the first method capable of resolving high resolution continuous flexibility for a protein as small as a GPCR complex. An early version of 3DVA in *cryoSPARC* was also used successfully to understand the differences in conformational dynamics between adrenomedullin 1 (AM1) and adrenomedullin 2 (AM2) receptors [21].

3.3 80S Ribosome: Solving multiple types of heterogeneity

Applied to 105,247 *Pf*80S ribosome particles (EMPIAR-10028 [48]), 3DVA with four variability components resolves four types of heterogeneity (Fig. 6). The first component identifies the presence or absence of a portion of the “head” region of the 40S small subunit. The second resolves rotational motion of the entire 40S subunit along an axis that connects the 40S and 60S subunits. The third component resolves lateral motion of the 40S subunit. The fourth 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 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 particles images, with a 360 pixel box size, taking 71 minutes on a single NVIDIA V100 GPU. As a comparison, on this dataset the recently proposed cryoDRGN deep generative model [52] was reported to resolve conformational changes similar to variability components 1, 2, and 4 [51], but required model training for 150 epochs taking over 60 hours (i.e., $50\times$ more than 3DVA) on the same hardware.

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

The $\mathrm{Na_V}$1.7 channel [50] 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,

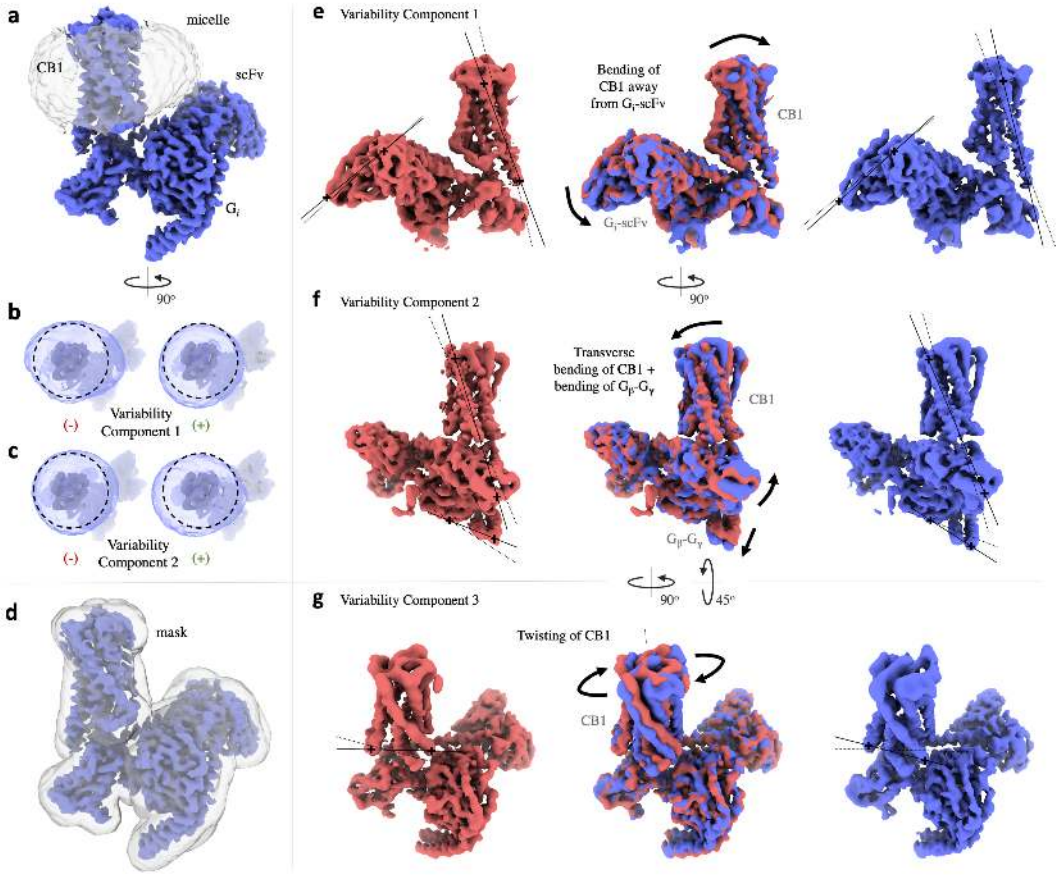

Figure 5: Results of 3DVA with three variability components on 250,649 particle images of a Cannabinoid Receptor 1-G GPCR complex [19]. This result demonstrates the capacity for 3DVA to resolve detailed high resolution motion of small proteins and small subregions of proteins. A video version of these results are available in supplementary materials. **a**: Consensus 3D refinement density of all particle images. 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 latent 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. **e**: Component 1 resolves bending of the CB1 transmembrane region away/towards the G-protein. Bending also affects the pose of the scFv subunit. **f**: 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. **g**: Component 3 resolves twisting of the CB1 transmembrane domain around an axis perpendicular to the membrane.

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 with 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^o around the symmetric axis (i.e. symmetry expansion). The resulting 601,518 particles are then input to 3DVA, with six components and a low-pass filter resolution of 3Å.

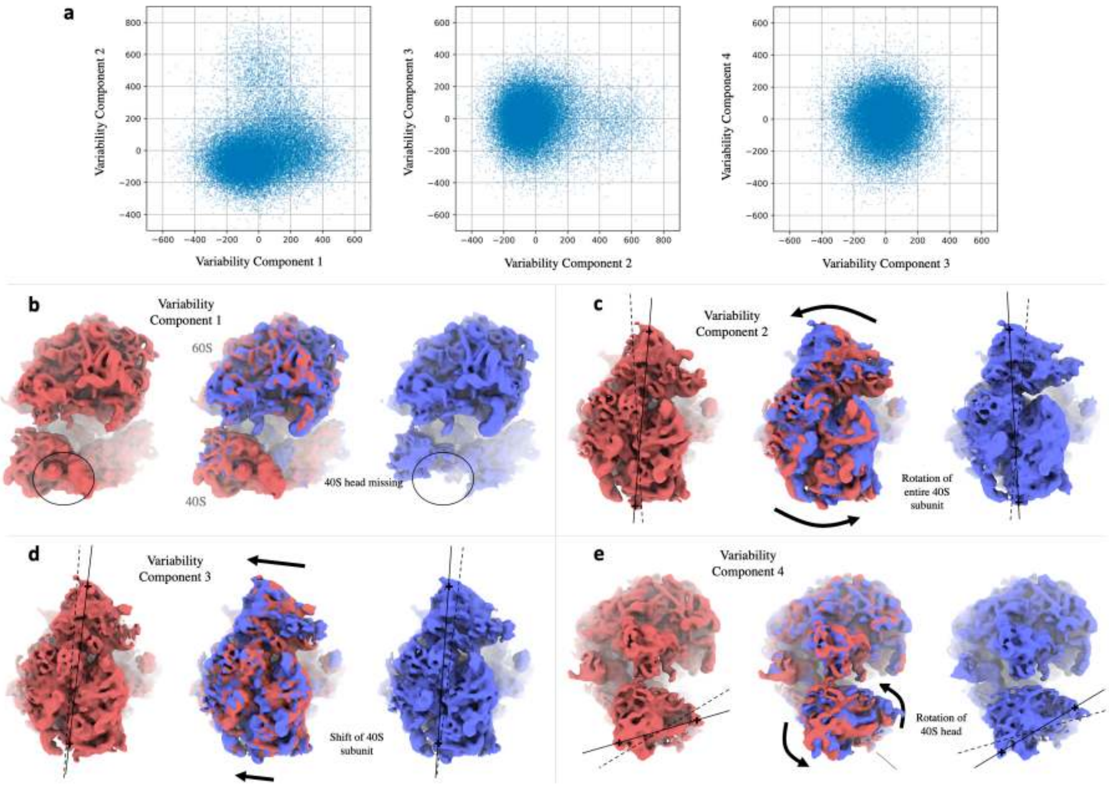

Figure 6: Results of 3DVA with four variability components applied to 105,247 particle images of *Pf* 80S Ribosome [48]. A video of these results is available in supplementary materials. **a**: 2-D scatter plots of 4-D latent coordinates of individual particles are solved by 3DVA. The scatter plots indicate 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.

Of the six components, Figure 7a-c displays three. The first (component 1), resolves bending of two of the transmembrane subunits of the tetrameric protein, along with motion of the bound Fabs. 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 data, refinement resolution of the peripheral transmembrane helices is limited [34]; 3DVA provides insight into the flexibility that causes this limitation.

3.5 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. sym-

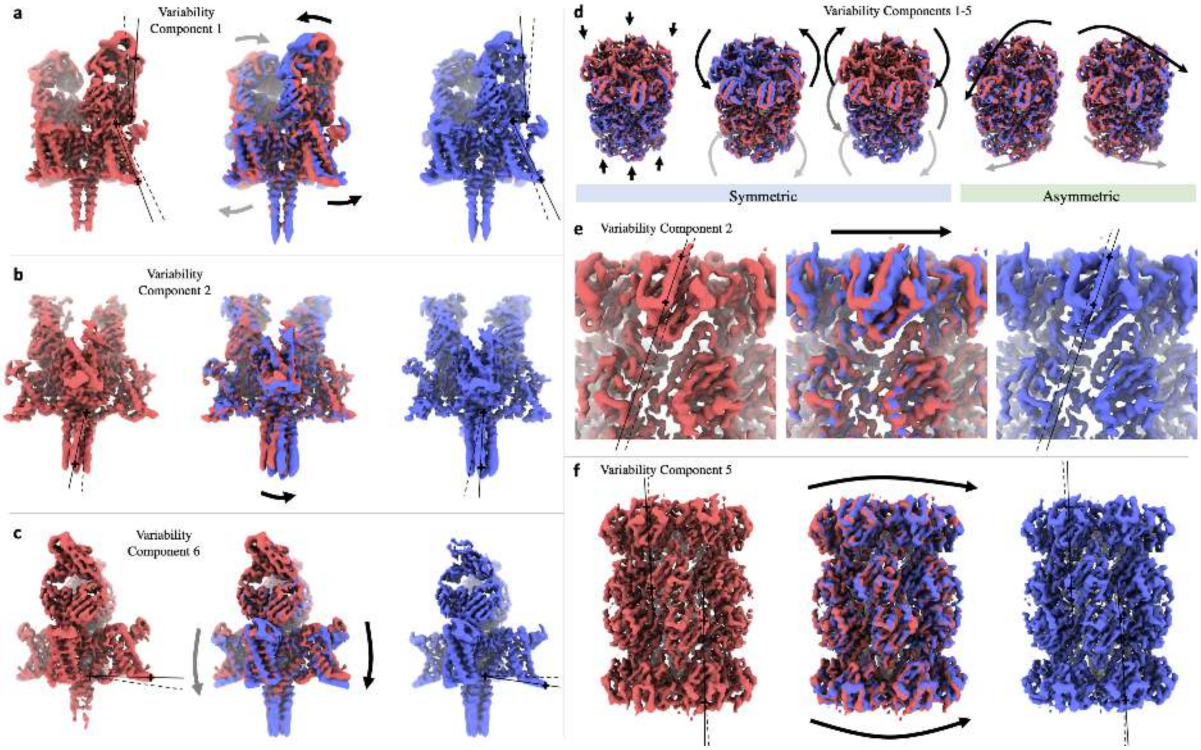

Figure 7: **a-c**: Results of 3DVA with six variability components (three shown) on 300,759 particle images of the Na$_V$1.7 channel membrane protein [50]. Particles are subject to symmetry expansion around a C2 symmetry. A video version of these results are available in supplementary materials. **a**: 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. **d-f**:Results of 3DVA with five variability components on 84,604 particle images of the T20S Proteasome [3]. Particles are subject to symmetry expansion around a D7 symmetry. A video version of these results are available in supplementary materials. **d**: Overlayed renderings of 3D density maps generated by 3DVA at negative (red) and positive (blue) latent 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. **e**: Detail view of variability component 2, showing rotational motion of the top region of the barrel. **f**: Detail view of variability component 5, showing bending of the entire molecule, breaking D7 symmetry.

metry 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 detects five types of continuous bending flexibility in the molecule (Fig. 7d-f). This is an interesting case due to the high symmetry; three of the variability components (Fig. 7d, left) are symmetric, corresponding to symmetric motion of all 14 subunits. The first component resolves a stretching-compression motion of the top and bottom regions of the barrel. The second component (Fig. 7e) 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 last two components are asymmetric (Fig. 7d, right), resolving bending of the entire barrel in two different directions (detail of component 4 in Fig. 7f). Importantly, this shows that 3DVA can resolve several orthogonal modes of detailed high resolution flexible motion of larger molecules. It can also help detect pseudo-symmetries created by asymmetric flexibility of symmetric molecules.

3.6 Spliceosome: Large flexible motions of large complexes

On 327,490 particle images of a pre-catalytic spliceosome complex (EMPIAR-10180 [31]), 3DVA was run with two components and a low-pass filter of 8Å, resolving two large motions of multiple parts of the complex (Fig. 8). The first component (Fig. 8b) 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 (Fig. 8c) resolves side-to-side rotation of SF3b, diagonal rotation of the helicase, and slight side-to-side bending of the foot region.

The linear subspace model captures both types of large motion, and the latent coordinates of individual particles provide an estimate for the position of the helicase and SF3b regions in each image (Fig. 8a, top). Nevertheless, the linear subspace model underlying 3DVA is limited in its ability to faithfully represent large motions (Fig. 1), so a simple local weighting scheme is used to create intermediate 3D reconstructions along each variability component. Particles are weighted based on their position along each latent coordinate (Fig. 8a, bottom) and 3D reconstructions are created using those weighted particles. This type of local neighborhood weighting is commonly used in manifold embedding applications, and has been used previously in methods for manifold embedding of cryo-EM data [5]. The resulting 3D density maps depict the molecule at different positions along its continuous flexibility (Fig. 8d,e).

The pre-catalytic spliceosome data was processed in 3DVA in 176 minutes. By comparison, on this data, the recently proposed cryoDRGN deep generative model [52] was reported to resolve conformational changes similar to the two variability components [51], but required model training for 30 hours on less than half the number of particles (i.e., 20× slower than 3DVA) on the same hardware.

3.7 Bacterial Large Ribosomal Subunit: directly resolving discrete heterogeneity

The bacterial large ribosomal subunit (LSU) is a large macromolecular complex; the cryo-EM dataset contains a mixture of assembly intermediates (EMPIAR-10076 [6]). 3DVA is applied to 131,899 particles, with four variability components and a low-pass filter of 5Å. The 3DVA components lie in a subspace that spans several discrete 3D states. Within that subspace (Fig. 9a) the latent coordinates of particle images are well clustered in multiple dimensions. We find that each cluster corresponds to a different partial assembly state of the complete LSU complex.

To analyze and understand the clustering of particles in the 3DVA latent coordinate space, we first fit an 8-component Gaussian Mixture Model (GMM) to the latent coordinates (Fig. 9b). This clusters particles into 8 major groupings. The particles from each cluster are then used to perform a 3D reconstruction; Fig. 9e shows the resulting 3D densities. Cluster 3 contains the least density, and each other cluster adds on a part to the assembly. Cluster

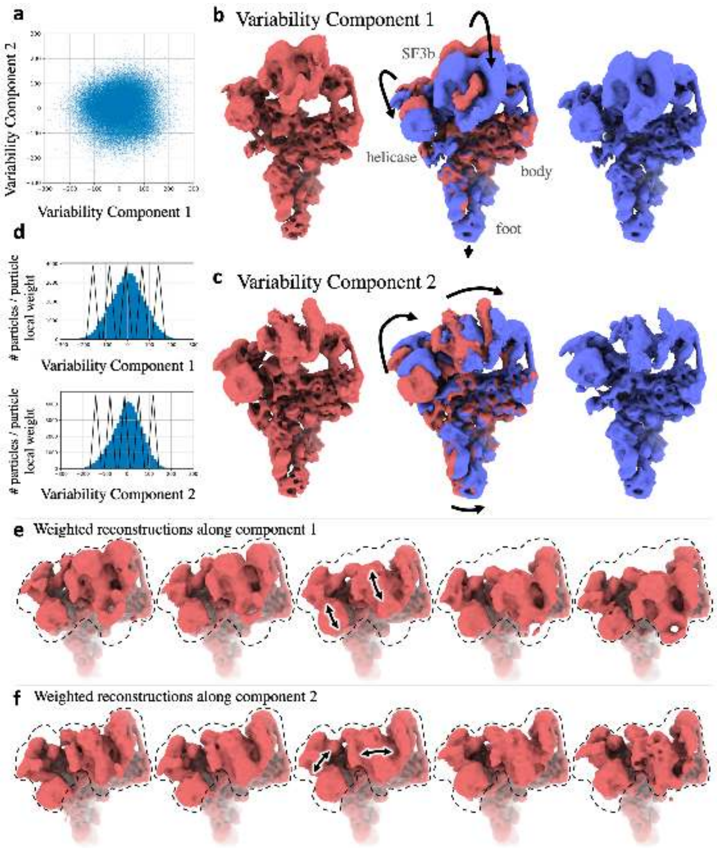
a
Variability Component 2
Variability Component 1
b Variability Component 1
SF3b
helicase
body
foot
c Variability Component 2
d
# particles / particle local weight
Variability Component 1
# particles / particle local weight
Variability Component 2
e Weighted reconstructions along component 1
f Weighted reconstructions along component 2

Figure 8: Results of 3DVA with two variabiltiy components on 327,490 particle images of a pre-catalytic spliceosome complex [31]. A video version of these results are available in supplementary materials. **a**: Particle images are spread over two latent coordinates. **b-c**: Renderings of 3D density maps generated by 3DVA at negative (red) and positive (blue) latent 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.

5 displays the entire LSU complex. Clusters 1, 2, 3 and 6 correspond to the four large 3D class subpopulations reported in the original study (C, D, B, E respectively) [6]. Cluster 8 appears to contain mainly outlier particles (contaminants and junk particles). The presence of outliers was also noted in the original study [6]. The identity of this cluster as outliers can be seen in the histogram of estimated per-particle contrast scales α in Fig. 9e, 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-D per-particle latent coordinates. First, a Variational Autoencoder (VAE) [18] is trained to reduce the 4-D latent coordinates to a 1-D embedding where individual clusters can be directly visually identified. The VAE is constructed using *PyTorch* [29], with a single hidden layer of size

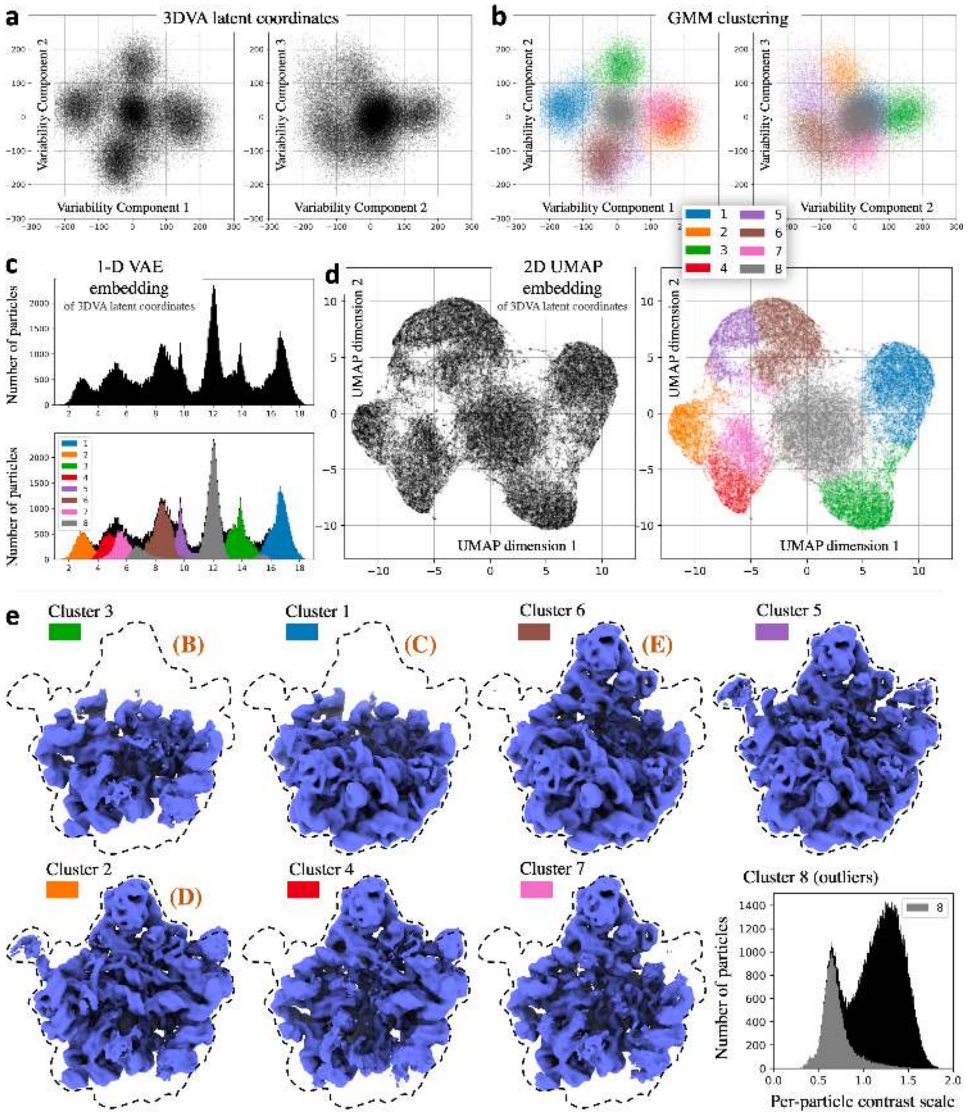

Figure 9: 3DVA with four components on 131,899 particle images of a bacterial ribosomal large subunit [6]. 3DVA identifies clusters corresponding to discrete conformational states. **a**: 3DVA particle latent coordinates, as 2D scatter plots of sequential pairs of dimensions. **b**: Latent coordinates colored by 8-way Gaussian Mixture Model (GMM) clustering. **c**: 1D histogram of Variational Autoencoder (VAE) embedding of latent 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 [24] embedding of latent 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 complex, and other clusters have more subunits assembled. The 8th cluster contains outliers for which 3DVA estimates lower per-particle scale factor α (histogram).

1024 in both encoder and decoder architectures. Training took 2 minutes for 40 epochs. The resulting 1-D latent encoding is shown in Fig. 9c as a histogram of particle positions. The peaks correspond to the multiple discrete states and the outlier cluster detected by GMM fitting (with the same colors as Fig. 9b). This result can be directly compared with the 1-D Spatial-VAE embedding carried out in processing of this same dataset using the cryoDRGN generative model [51], where a similar histogram is generated that resolves four major sub-states, plus an outlier cluster.

A second non-linear embedding technique, UMAP [24], is applied to the 4-D latent coordinates (Fig. 9d), displaying the geometry of clusters in the 4-D space, e.g., that clusters 2, 4, and 7 are sub-clusters of a larger cluster. This can also be compared with results in [51] to illustrate the ability of 3DVA to resolve nuanced discrete conformational heterogeneity, despite the apparent simplicity of the linear subspace model.

4 Discussion

The 3DVA algorithm enables one to fit high-resolution linear subspace models to single particle cryo-EM data. The output variability components and associated latent 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. It can can also be used to identify discrete conformational states of proteins from heterogeneous samples.

Related methods for manifold fitting

Linear subspace models, such as PCA, Multivariate Statistical Analysis (MSA), and eigenanalysis have been explored for single particle cryo-EM. Early methods used bootstrapping 3D reconstructions from random subsets of images to coarsely simulate sampling from the conformational distribution of a target molecule [30]. Others attempt 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 [45]. Such methods are most closely related to 3DVA, as discussed in Results.

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 [5, 9]. These methods have been demonstrated on experimental data of large proteins, and recent work aims to improve the difficult process of combining 2D manifolds to form a 3D model [23]. Non-linear methods employing the graph Laplacian of similar 2D particle images have also been developed [26] and applied to synthetic data. Most recently, non-linear, deep generative models, have been proposed and shown to resolve coarse-scale conformational changes from experimental cryo-EM data of large complexes [51, 52]. Methods that directly model protein motion have also been proposed, but only demonstrated on simplified synthetic data [11, 20].

Relationship to other SPA heterogeneity techniques

The predominant method currently used to resolve heterogeneity in cryo-EM data is K-way 3D classification [12, 39, 41, 42]. By comparison, a linear subspace model does not constrain particles to exist in a (small) finite number of conformational states. Rather, the subspace components span a continuous space of possible states. It is well-known in the cryo-EM field that clustering methods are not effective for resolving continuous flexibility.

Local refinement [12, 41] and multi-body refinement [27] methods allow high resolution refinement of some regions of flexible molecules, by assuming the molecule is composed of a small number of rigid parts. For successful 2D-3D alignment, such methods require that each region has sufficient mass for accurate local alignment, independent of the rest of the structure. 3DVA 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 [43]. In contrast, 3DVA does not presuppose knowledge of the energy landscape, bending, or flexing of the molecule. Rather, it learns this from the image data.

Finally, linear subspace models have also been useful in cryo-electron tomography (cryo-ET), specifically in sub-tomogram averaging where multiple 3D reconstructions of individual molecules, are obtained. By averaging such reconstructions one can improve resolution. In this case, PCA can be directly applied [14, 16] since tomography provides complete 3D observations for each individual particle.

Extensions and next steps

The 3DVA algorithm uses a linear subspace model of 3D structures that comprise the conformational landscape of a protein molecule. The true manifold can have non-linear and complex geometry, for which non-linear manifold models are likely 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 aggregate structural information across continuous conformational states along a manifold, to solve higher resolution 3D density maps of flexible molecules.

5 Methods

Expectation-Maximization and Free Energy

Following Neal and Hinton [28], the Expectation-Maximization algorithm can be formulated in terms of the minimization of a free energy, which serves as an upper bound the negative log likelihood of the data under the model. Given data x, model parameters θ, and latent (or missing) paramters z, from Jensen's inequality, that the free energy $\mathcal{F}$ bounds the negative log likelihood:

$$\mathcal{L} \equiv -\log p(x|\theta) \leq \int q(z) \log \frac{p(x,z|\theta)}{q(z)} \equiv \mathcal{F} \, . \tag{5}$$

The bound is tight when q is the posterior over the latent variables, $q(z) = p(z|x,\theta)$. In that case, $\mathcal{F} = Q + H$, where H is the differential entropy of the posterior, and Q is the expected complete log likelihood under the posterior:

$$Q(\theta) = \int p(z|x,\theta) \, \log p(x,z|\theta) \, . \tag{6}$$

For exponential families the posterior is specified in terms of sufficient statistics ψ (eg its first and second moments, or mean and covariance). Given the true posterior parameters, the bound is tight and the free energy equals the negative log likelihood; Ie, $\min_\psi \mathcal{F}(\theta, \psi) = Q + H$. It follows that minima of $\mathcal{F}$ with respect to ψ and θ are also minima of $\mathcal{L}$ with respect to θ. The EM algorithm can thus be expressed as coordinate descent on $\mathcal{F}$. At iteration t the E step optimizes the tightness of the bound by computing $\psi^t = \arg\min_\psi \mathcal{F}(\theta^{t-1}, \psi)$. The M step then optimizes the bound, i.e., $\theta^t = \arg\min_\theta \mathcal{F}(\theta, \psi^t)$.

PPCA and PCA

The usual generative model for probabilistic PCA is

$$x \;=\; \mu + Wz + \eta$$

where $z \sim \mathcal{N}(0, I)$, and $\eta \sim \mathcal{N}(0, \epsilon I)$ is isotropic white noise. The model parameters θ include the matrix W and the noise variance ϵ. Under this model, the complete likelihood, over x and z, is

$$p(z, x|\theta) \;=\; \mathcal{N}((0, \mu), \begin{bmatrix} I & W^T \\ W & WW^T + \epsilon I \end{bmatrix}) \tag{7}$$

From this one can show that posterior $p(z|x)$, a.k.a., the inference distribution, has the form:

$$p(z|x, \theta) \;=\; N(m_z, C_z) \tag{8}$$

$$m_z \;=\; W^T(WW^T + \epsilon I)^{-1}(x - \mu) \tag{9}$$

$$C_z \;=\; I - W^T(WW^T + \epsilon I)^{-1}W \tag{10}$$

For this PPCA modeel, it is straightforward to derive the E-step and M-steps for estimating θ [38, 46].

It has been shown that PCA and a corresponding EM algorithm are the limiting case of PPCA as $e \to 0$ [38]. In this case one can show that the posterior covariance reduces to $(W^TW)W^T$, and the posterior distribution, in the limit, is a Dirac delta function:

$$p(z|x, \theta) \;=\; \delta(z - (W^TW)^{-1}W^T(x - \mu)) \,. \tag{11}$$

As a consequence, given W^{t-1} from the previous EM iteration, the posterior mean in the E step is compute per data point as $z_i^t = \arg\min_z \, ||x_i - (W^{t-1}z + \mu)||^2$. The M step then updates $W^t = \arg\min_W \sum_i ||x_i - (Wz_i^t + \mu)||^2$.

Notably, it has been shown that both PCA and PPCA (with finite ϵ) produce the same subspace [38]. The main difference between PCA and PPCA is the estimation of the noise variance, which also entails shrinkage of the latent coordinates.

Finally, it is also important to note that for M observations of D-dimensional data and a K-dimensional subspace, EM entails $O(MDK)$ operations per iteration, instead of the $O(MD^2)$ operations needed for the formation of the full covariance matrix [38]. In terms of a box with linear dimension (box-size) N, the difference is $O(MN^6)$ versus $O(MN^3K)$. Since $N \gg K$ this difference is significant. Finally, we note that, although many subspace models assume W is orthogonal, this assumption is not strictly necessary; the formulation of the posterior mean above does not assume orthogonality.

3D Variability Analysis

The generative model for 3DVA is given in (3). The noise and latent coordinates are assumed to be independent and Gaussian, while α and V are unknown parameters we wish to estimate. In particular, we assume a mean-zero, isotropic Gaussian prior for the latent coordinates $z \sim \mathcal{N}(0, I_K)$, and we assume isotropic mean-zero noise $\eta \sim \mathcal{N}(0, \epsilon I_{N^2})$. In most cryo-EM datsets of interest a colored noise model is more appropriate than an isotropic noise model. This is important as it changes the relative influence of model fitting as a function of spectral wavenumber. In practice in 3DVA we use a colored noise model. However, for convenience in what follows we assume the noise has been whitened. For example, if the noise is mean-zero Gaussian with a diagonal precision matrix Λ, then we simply multiply the image x and the CTF by $\Lambda^{1/2}$. Formally, the whitened model is

$$X'_i \;=\; \alpha_i\, C'_i\, P(\phi_i) \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) + \eta' \,. \tag{12}$$

where $X'_i \equiv \Lambda^{1/2}\, X_i$, and $C'_i \equiv \Lambda^{1/2}\, C_i$, and the noise is now isotropic $\eta' \sim \mathcal{N}(0, \epsilon I)$.

Finally, for 3DVA we derive the algorithm under the assumption that the noise variance ϵ tends to zero. While it may seem counter-intuitive to assume that noise tends to zero for cryo-EM data, this allows us to use the limited case of the PPCA model above, which reduces to PCA. Crucially, as noted, PPCA with finite ϵ and PPCA with $\epsilon \to 0$ will both yield the same linear subspace [38]. This means that the variability components recovered by 3DVA will be unchanged by this assumption. Under the zero noise assumption, the E step only computes the posterior mean over the latent parameters, and the M step simplifies because the free energy reduces to the log marginal likelihood. Both the E and M steps used to fit the subspace model in 3DVA entail the minimization of a single energy function equal to the sum of squared residual errors between the observed images, X_i, and model predictions, all represented in the Fourier domain. In what follows for notational convenience we drop the prime in X and C, taking for granted that in the case of colored noise the particles images and CTFs have been modulated by the square root of the precision matrix of the noise $\Lambda^{1/2}$. We assume colored Gaussian observation noise, with a diagonal covariance in the Fourier domain. We further assume that the noise variances for different wavenumbers (i.e., along the diagonal), are constant in spherical shells.

Accordingly, the E-step and M-step can be viewed as a form of iterative coordinate descent. The E-step minimizes the free energy to obtain the posterior mean for each particle image. The M-step then decreases the energy using coordinate descent with respect to the variability component $V_{1:K}$ and the per particle scale factors. Given M particle images, the energy function can be rewritten as follows:

$$E(V_{1:K}, \alpha_{1:M}, \mathbf{z}_{1:M}) \;=\; \frac{1}{2} \sum_{i=1}^{M} \left\| \, X'_i - \alpha_i A_i \left(V_0 + \sum_{k=1}^{K} z_{ik} V_k \right) \right\|^2 , \tag{13}$$

where $A_i \equiv C'_i\, P(\phi_i)$. The main challenges with the optimization stems from the large number of unknowns.

In the E-step, we estimate the posterior means, with one optimization per particle image, given the variability components $V_{1:K}$ and the scale factors α_i. To this end, let $W_{ik} = A_i V_k$ be the scaled projection of kth variability component according to the CTF and pose of the ith particle image. Then, the energy for the ith image is

$$E_i(V_{1:K}, \alpha_i, \mathbf{z}_i) \;=\; \frac{1}{2} \left\| \, X'_i - \alpha_i W_{i0} - \alpha_i \mathbf{W}_i \mathbf{z}_i \, \right\|^2 \tag{14}$$

where $\mathbf{W}_i \equiv [W_{i1}, ..., W_{iK}]$ and $\mathbf{z}_i \equiv (z_{i1}, ..., z_{iK})^T$. The columns of $\mathbf{W}_i$ are the projections of the K variability components according to the CTF and pose of the ith particle. The dimension of each column is N^2, ii.e., the size of the image X_i. So $\mathbf{W}$ is a $N^2 \times K$ matrix, and $\mathbf{W}_i^H \mathbf{W}_i$ is a $K \times K$ symmetric positive-definite matrix, where $\mathbf{W}^H$ denotes the conjugate transpose of $\mathbf{W}$. The posterior mean is given in closed form by

$$\mathbf{z}_i \;=\; \frac{1}{\alpha_i} (\mathbf{W}_i^H \mathbf{W}_i)^{-1} \mathbf{W}_i^H \, (X_i' - \alpha_i W_{i0}) \, . \tag{15}$$

The first phase of the M-step minimizes the energy associated with each particle image in (14) to solve for the updated scale factors (given the most recent latent coordinates). This is given by

$$\alpha_i \;=\; \frac{(W_{i0} + \mathbf{W}_i \mathbf{z}_i)^H X_i'}{(W_{i0} + \mathbf{W}_i \mathbf{z}_i)^H (W_{i0} + \mathbf{W}_i \mathbf{z}_i)} \, . \tag{16}$$

The second computation in the M-step involves the estimation of the variability component updates. To this end it is useful to rewrite the energy, now summed over all images, in a somewhat simpler form:

$$E(V_{1:K}, \, \alpha_{1:M}, \, \mathbf{z}_{1:M}) \;=\; \frac{1}{2} \sum_{i=1}^{M} \left\| D_i - \sum_k z_{ik} \, A_i \, V_k \right\|^2 \tag{17}$$

where $D_i \equiv X_i' - \alpha_i W_{i0}$, and z_{ik} is the kth latent coordinate for the ith particle image. To solve for variability components, we require the gradient of E with respect to each V_j:

$$\frac{\partial E(V_{1:K}, \, \{\mathbf{w}_i\})}{\partial V_j} \;=\; -\sum_i z_{ij} {A_i}^H \left(D_i - \sum_k z_{ik} A_i V_k \right)$$

Setting the gradient to zero and rearranging terms yields a linear system of equations,

$$\sum_k H_{jk} \, V_k \;=\; h_j \; , \tag{18}$$

where $H_{jk} = \sum_i z_{ij} z_{ik} A_i^H A_i$ and $h_j = \sum_i z_{ij} {A_i}^H \, D_i$. Importantly, H_{jk} can be represented as a diagonal matrix. The CTF operator C_i' is diagonal, and depending on the form of the interpolation used, the product of $P_i^T P_i$ is either diagonal or approximately diagonal. As is standard in single-particle EM reconstruction algorithms, we approximate this product as a diagonal matrix.

Differentiating E with respect to all variability components, we obtain a block matrix system of equations:

$$\begin{bmatrix} H_{11} & H_{12} & \dots & H_{1K} \\ H_{21} & H_{22} & \dots & H_{2K} \\ \vdots & & \ddots & \vdots \\ H_{K1} & H_{K2} & \dots & H_{KK} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_K \end{bmatrix} = \begin{bmatrix} h_1 \\ h_1 \\ \vdots \\ h_K \end{bmatrix} \tag{19}$$

Exploiting the diagonal structure of individual blocks, H_{jk}, we rearrange the rows and column of (19) to form a block diagonal system having N^3 decoupled, $K{\times}K$ linear systems, one for each wavenumber. Each $K{\times}K$ systems is used to estimate the Fourier coefficients at a specific wavenumber for each of the K variability components.

To rearrange the LHS of (19), the first K columns will contain the first column from each of the K blocks in the large block H matrix in (19). The next K columns comprise the second column of each of the K blocks and so on. The rows of the large variability component vector are rearranged correspondingly. This places the Fourier coefficients associated with the first wavenumber together, followed by those associated with the second wavenumber next, and so on. Next we rearrange the rows of the matrix and the corresponding RHS elements The first K rows comprise the first row from each of the K blocks of the H matrix on the LHS of (19). The next K rows comprise the second row of each block, and so on. We also rearrange the corresponding elements on the RHS. This rearrangement yields a block diagonal system, decoupling the estimation of different wavenumbers.

Finally should we want to impose orthogonality constraints, we could use Lagrange multipliers or constrained optimization, but with Expectation-Maximization it is sufficient to orthogonalize the variability components at the end of each iteration. It is done with the Gram-Schmidt algorithm [10], which is efficient as it only requires inner products between variability components, and sclaing each variability component to have unit length, i.e., $V_j^H V_j = 1$. While not technically required by the iterative solution of a linear subspace model, orthogonality improves computational stability by removing degrees of freedom to which the solution is invariant.

Iterations continue until convergence, or for a fixed number of iterations. The experiments here used 20 iterations. Once completed, we form the $K \times K$ covariance matrix of latent coordinates $S = \sum_i \mathbf{z}_i \mathbf{z}_i^T$. The eigenvectors of S specify a rotation of the variability components and the latent coordinates that aligns the variability components with the principal directions of the data and ensures the latent coordinates are independent under the Gaussian model, as with PPCA. The per-particle scale factors α_i are unchanged under this rotation.

Code Availability: The cryoSPARC software package is available for non-profit academic use at `www.cryosparc.com`.

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. Resources used in this research were provided, in part, by the Province of Ontario, the Government of Canada through NSERC and CIFAR, and companies sponsoring the Vector Institute.

Author Contributions: A.P. developed the method, created the implementation and performed experiments. A.P. and D.J.F. created the formulation and wrote the paper.

Competing Interests: A.P. is CEO of Stuctura Biotechnology Inc. which builds the *cryoSPARC* software package, distributed freely for academic non-profit use with software licenses available for commercial use. D.J.F. is an advisor to Stuctura Biotechnology Inc. The novel aspects of the method presented are described in a provisional patent application.

References

[1] J. Andén and A. Singer. Structural Variability from Noisy Tomographic Projections. *SIAM Journal on Imaging Sciences*, 11(2):1441–1492, 2018. 2, 4, 18

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 3

[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. 7, 13, 14

[4] Q. Cui and I. Bahar. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Taylor and Francis, 2005. 19

[5] 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. 15, 18

[6] 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. 15, 16, 17

[7] 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. 4

[8] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, 2007. 3

[9] J. Frank and A. Ourmazd. Continuous changes in structure mapped by manifold embedding of single-particle data. *Methods*, 100:61–67, 2016. 18

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996. 6, 23

[11] Y. Gong and P. C. Doerschuk. 3-d understanding of electron microscopy images of nano bio objects by computing generative mechanical models. In *IEEE Int. Conf. Image Processing*, pages 3161–3165, 2016. 18

[12] T. Grant, A. Rohou, and N. Grigorieff. *cis*TEM, user-friendly software for single-particle image processing. *eLife*, 7:e35383, 2018. 2, 3, 18, 19

[13] N. Grigorieff. FREEALIGN: High resolution refinement of single particle structures. *Journal of Structural Biology*, 157:117–125, 2007. 3

[14] J. M. Heumann, A. Hoenger, and D. N. Mastronarde. Clustering and variance maps for cryo-electron tomography using wedge-masked differences. *Journal of Structural Biology*, 175(3):288–299, Sep 2011. 19

[15] D. Hilger, M. Masureel, and B. K. Kobilka. Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology*, 25(1):4–12, 2018. 11

[16] B. A. Himes and P. Zhang. emClarity: Software for high-resolution cryo-electron tomography and subtomogram averaging. *Nature Methods*, 15:955–961, 2018. 19

[17] 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. 11

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *Proc. Int. Conf. Learn. Rep.*, 2014. 16

[19] 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. 11, 12

[20] R. R. Lederman and A. Singer. Continuously heterogeneous hyper-objects in cryo-em and 3-d movies of many temporal dimensions. *ArXiv 1704.02899*, 2017. 18

[21] 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. 3, 11

[22] D. Lyumkis. Challenges and opportunities in cryo-EM single-particle analysis. *Journal of Biological Chemistry*, 294(13):5181–5197, Mar 2019. 1

[23] 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. 18

[24] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv 1802.03426*, 2018. 17, 18

[25] 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. 3

[26] A. Moscovich, A. Halevi, J. Anden, and A. Singer. Cryo-em reconstruction of continuous heterogeneity by laplacian spectral volumes. *Inverse Problems*, 36 024003, 2020. 18

[27] 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. 19

[28] R. M. Neal and G. E. Hinton. *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*. Learning in Graphical Models, NATO ASI Series (Series D), Vol. 89. Springer, 1998. 19

[29] 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. *NeurIPS*, pages 8024–8035, 2019. 16

[30] P. A. Penczek, M. Kimmel, and C. M. T. Spahn. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images. *Structure*, 19(11):1582–1590, Nov 2011. 2, 4, 18

[31] C. Plaschka, P.-C. Lin, and K. Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, 2017. 15, 16

[32] A. Punjani, M. A. Brubaker, and D. J. Fleet. Building proteins in a day: Efficient 3d molecular structure estimation with electron cryo-microscopy. *IEEE Trans. PAMI*, 39(4):706–718, 2017. 3

[33] 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. 2

[34] A. Punjani, H. Zhang, and D. J. Fleet. Non-uniform refinement: Adaptive regularization improves single particle cryo-EM reconstruction. *bioRxiv*, 2019. 6, 13

[35] 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. 2

[36] Z. A. Ripstein, S. Vahidi, W. A. Houry, J. L. Rubinstein, and L. E. Kay. A processive rotary mechanism couples substrate unfolding and proteolysis in the clpxp degradation machinery. *eLife*, 9:e52158, Jan 2020. 2

[37] 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. 2

[38] S. Roweis. EM algorithms for PCA and SPCA. *Proc. NIPS*, pages 626–632, 1998. 4, 6, 20, 21

[39] S. H. Scheres. Processing of structurally heterogeneous cryo-em data in RELION. *Methods Enzymol.*, 579:125–157, 2016. 2, 3, 18

[40] S. H. W. Scheres. A Bayesian view on cryo-em structure determination. *Journal of Molecular Biology*, 415:406–418, 2012. 3

[41] 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, 18, 19

[42] 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, 18

[43] 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. 19

[44] 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. 3

[45] 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, 4, 18

[46] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J Royal Statistical Society, series B*, 61:611–622, 1999. 4, 6, 20

[47] 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, 4

[48] 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. 11, 13

[49] 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. 2

[50] 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. 11, 14

[51] 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. 2, 11, 15, 18

[52] E. D. Zhong, T. Bepler, J. H. Davis, and B. Berger. Reconstructing continuously heterogeneous structures from single particle cryo-em with deep generative models. *Int. Conf. Learning Rep.*, 2020. 2, 11, 15, 18