Extending the cryoDRGN toolkit with a scalable conformational landscape analysis

Ellen D. Zhong[1,2,*,†], **Ashwin Narayan**[3], **Xue Fei**[4], **Robert T. Sauer**[4], **Joseph H. Davis**[1,4,*], **and Bonnie Berger**[2,3,*]

[1]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA
[4]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

CryoDRGN and other recent deep learning-based cryo-EM reconstruction algorithms are capable of reconstructing complex distributions of heterogeneous structures. However, the downstream analysis of the resulting density maps typically relies on manual visualization and inspection, which is impractical for modern deep generative modeling approaches that can produce large ensembles (> 100) of structures. We present an efficient and automated volume analysis framework for quantitative analysis of a trained cryodrgn model, including assigning discrete conformational states and visualizing continuous conformational landscapes. Our framework uses a combination of sketching, clustering, and dimensionality reduction techniques on the set of reconstructed volumes, which provides a more grounded physical interpretation over cryoDRGN's latent variable representation. On a previously published dataset of the ClpXP protease, we newly identified a substrate-engaged state that was missed in traditional 3D classification. We provide both an automated tool and interactive notebooks that implement this analysis in the cryoDRGN software.

1 Introduction

Single particle cryo-electron microscopy (cryo-EM) is uniquely poised to study complex structural ensembles of large, dynamic biomolecular complexes[1], and several advanced tools for heterogeneous reconstruction have recently been proposed towards this promise[2–5]. In the cryoDRGN method, heterogeneous reconstruction is framed as unsupervised learning of a deep generative model of 3D density maps parameterized by a neural field[6] representation of structure[2,7]. Central to the cryoDRGN approach is learning a generic latent variable model for structural heterogeneity, which has been empirically shown to model both discrete and continuous forms of structural variability, for example compositional changes from co-factor binding during ribosome assembly[2] and large-scale continuous motions of dynein motor protein complexes[8]. In cryoDRGN's framework of generative modeling, once a model is trained, an arbitrary number of volumes may be reconstructed at sampled values of the latent variable, thus tools are needed to comprehensively explore the reconstructed distribution.

To accommodate the diverse sources of heterogeneity present in cryo-EM data, cryoDRGN possesses a number of interactive and automated processing approaches for analyzing cryoDRGN results. Existing approaches have focused on visualization of the low-dimensional latent embeddings coupled with user-guided exploration of the volume ensemble[2]. However, while the distribution of latent space embeddings may possess interpretable features reflective of the underlying structural ensemble, the objective function of training a cryoDRGN model aims to reconstruct the distribution of the imaged particles without any guarantees that their latent space representation is (visually) interpretable. Here, we instead focus our analysis of the learned distribution on the high-dimensional output space of volumes.

Briefly, we first summarize the volume distribution with a k-means-based sketching of the latent space to enable computationally tractable analysis. Two types of analyses are then performed on the sketch of volumes: 1) an aggolmerative clustering algorithm which produces a small number of summary volumes that can be interpreted as *discrete* conformation states and 2) principal component analysis (PCA), where the estimated eigenvectors (or, "eigenvolumes") define *continuous* reaction coordinates that may be used to interpret the full ensemble of particles. The rationale behind these choices is described in the next section.

Applied on a previously published dataset of the ClpXP protease[9], we automatically identified a new substrate-engaged state comprising 1,255 particles (0.3% of the dataset) that was both missed in traditional 3D classification and was not immediately apparent from visualizing the cryoDRGN latent space. By applying PCA on the volume ensemble, we produced reaction

coordinates that provide a more interpretable visualization of the ensemble than cryoDRGN's latent variable representation. A software tool that implements this "landscape analysis" is openly available in cryoDRGN software version 1.0.

2 Methods

In this section, we describe the computational pipeline and design choices for analyzing and interpreting a cryoDRGN model (Figure 1, top), using the ClpXP protease dataset from Fei *et al.*[9] as an instructive example (Figure 1, bottom).

2.1 Overview of cryoDRGN outputs

Given a dataset of single particle cryo-EM images, $\{X_1, ..., X_N\}$, cryoDRGN performs heterogeneous reconstruction by jointly training an inference model over images, $q_\xi(z|X)$, (the encoder) and a generative model over volumes, $p_\theta(V|z)$, (the decoder). Once trained, the model may be used to predict a latent variable representation for each image in the dataset, which we refer to as the "latent embedding":

$$z_i \sim q_\xi(z|X_i) \tag{1}$$

and generate an associated volume representation:

$$V_i \sim p_\theta(V|z_i) \tag{2}$$

In practice, we define the latent embedding as the *maximum a posteriori* estimate of the (Gaussian) posterior $q_\xi(z|X_i)$, which provides a low-dimensional representation of each image, i.e. $z_i \in \mathbb{R}^N$, where $N = 8$ is typical; the volume is rendered on a 3D lattice for downstream visualization tools, i.e. $V_i \in \mathbb{R}^{D\times D\times D}$, where $D = 128$ or $D = 256$ is typical.

2.2 Motivation of volume space analysis

The set of latent embeddings of the dataset $\{z_i\}$ gives a low-dimensional vector representation of the dataset that can be visualized in 2-D with dimensionality reduction algorithms such as PCA, t-SNE, or Uniform Manifold Approximate and Projection (UMAP)[10] (Figure 1A). As shown in Zhong et al.[2], the resulting features of the distribution of latent embeddings can be reflective of structural heterogeneity, such as clusters that correspond to different compositional states. While this may suggest an interpretation of the latent embeddings as an energy landscape, (e.g. where regions of higher/lower particle density correspond to low/high energy states), this interpretation is flawed. Namely the layout of the latent space is arbitrary (hindering interpretation), distances in latent space are not meaningful (z are passed through a nonlinear decoder), and empty regions of latent space do not in general correspond to high energy configurations. Thus, the interpretation of the latent embeddings typically requires annotations from user-guided exploration of the volume ensemble. For example, after training a 8-D latent variable model on the ClpXP dataset (Section A), we visualized the final latent embeddings with UMAP (Figure 1A). Although there are "features" in the UMAP visualization (e.g. regions of higher and lower particle density), their interpretation requires manual inspection of volumes to annotate the various regions. Furthermore, the interpretability of UMAP distances is not reliable[11].

Our motivation here is to provide an automated and comprehensive analysis approach for the entire ensemble of volumes $\{V_i\}$ to facilitate interpretation of the trained model. However, it is computationally intractable to generate the entire ensemble of volumes $\{V_i\}$ associated with each particle in the dataset due to the computational cost of rendering a volume (seconds per volume) and the storage requirement for the voxel arrays (for $10^5 - 10^6$ volumes). Existing cryoDRGN analysis approaches typically generate tens of volumes from different regions of the latent space followed by manual inspection. However, this approach can be time-intensive for the practitioner and prone to missing (rare) states that are not sampled, especially because it requires the practitioner to decide *a priori* the regions worth deeper study.

2.3 Sketching the volume ensemble

We first generate a *sketch*, or a representative subsample, of volumes from the trained cryoDRGN model that will be used for the downstream structural landscape analysis. The general objective of sketching a dataset is to generate a subsample such that some important properties of the original dataset are preserved. For instance, one can consider naive uniform sampling as a method of sketching, where the property preserved is the probability density of the data. However, a major drawback of uniform sampling is that rare classes will often not end up in a sample unless a very large sketch size is used. For example, in order to capture at least one example of rare substrate-engaged state in ClpXP (see Figure 4), which occurs in 0.3% of the samples with a probability of 99%, more than 1,500 samples are needed. See[12] for a discussion of various sketching algorithms for the computational analysis of single-cell RNA-sequencing datasets.

Here, we use a k-means clustering algorithm for sketching: Given a desired sketch size k, we perform k-means clustering on the set of latent embeddings $\{z_i\}$ (Figure 1B). The latent embedding that is closest to each k-means cluster center is then

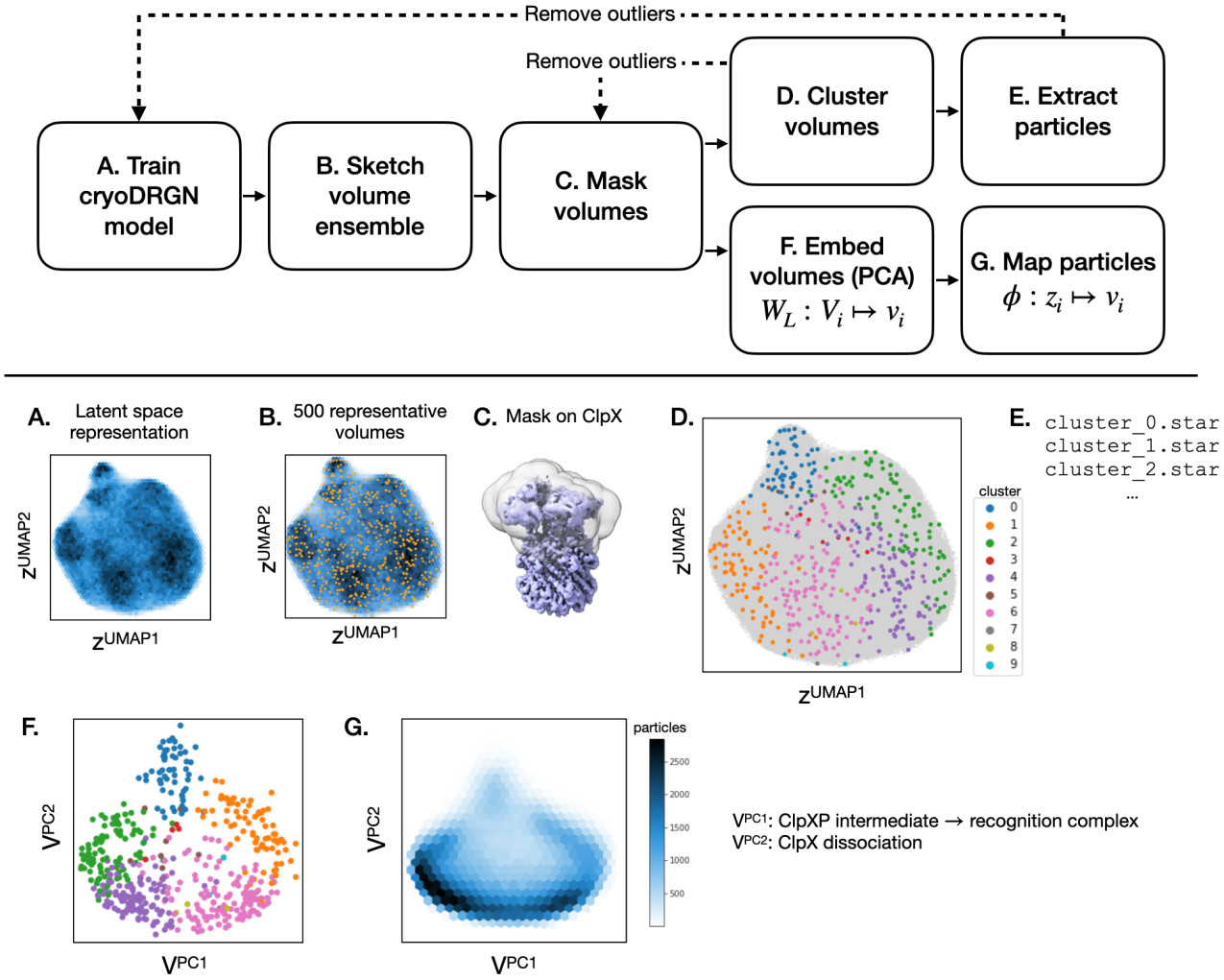

Figure 1. Overview of the landscape analysis pipeline: We show the general schematic of landscape analysis (top) and its application to a single particle dataset of ClpXP protease[9] (bottom). **A.** First, a cryoDRGN model is trained, so that a latent variable representation z_i can be generated for each image i in the original dataset. While the latent space representation describes the heterogeneity in the dataset, []. **B.** Once trained, an arbitrary number of volumes may be generated from the resulting model. For downstream analysis of the volume distribution, we *sketch* the set of latent embeddings to find k representative volumes (shown as orange points). **C.** A mask is applied on the sketched volumes to reduce noise from the background. A user-specified mask can be provided to focus on a subset of the volume; here, a mask covers ClpX, the mobile region of the complex. **D.** The sketched volumes are then clustered to characterize *discrete* conformational states. **E.** Particles associated to each cluster can be exported for refinement. **F.** The set of sketched volumes can be visualized with principal component analysis (PCA), which produces a linear map W_L for estimating low-dimensional volume embeddings v_i; the principal components (PC) indicate high variance modes of continuous motion in the structure and can be used to interpret $\{v_i\}$. Cluster assignments from **(D)** are also plotted (colors). **G.** A conformational landscape for the full dataset can be visualized by mapping all particles from their latent representation z_i to their PC-embedded volume representation v_i. We train a multilayer perceptron (MLP) ϕ to learn this mapping because generating volumes for the entire dataset is intractable. **Arrows:** Clusters can be inspected for artifacts (e.g. from junk particles); the underlying volumes or image data can be excluded when re-analyzing the volumes or retraining a cryoDRGN model, respectively.

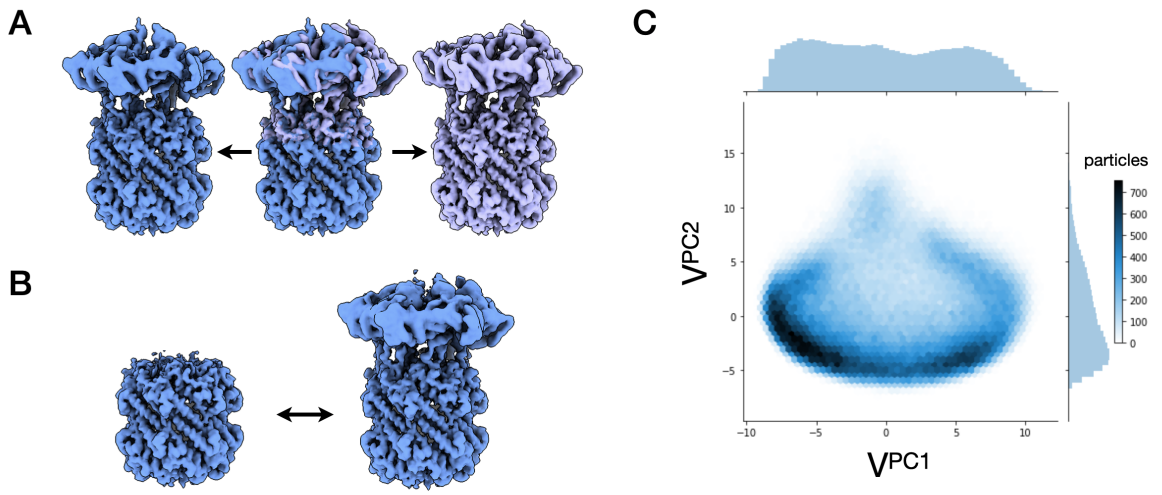

Figure 2. Conformational landscape of ClpXP inferred from PCA of the cryoDRGN volume distribution: A. Structures traversing principal component 1 shows the transition from the recognition to the intermediate complex. **B.** Structures traversing principal component 2 shows the dissociation of ClpX. **C.** A conformational landscape visualization of all particles mapped to the volume PC space. The methods and motivations for these analyses are described in Section 2.5

used to generate a volume for the volume sketch. Because k-means attempts to minimize the total variance of all clusters, with sufficiently large k, most points in the dataset should be relatively close to a point in the sketch. We validate that each each sketched latent embedding yields a representative *volume* for each cluster for a reasonable choice of k, e.g. $k = 500, 1000$. Since each cluster of latent embeddings also represents a set of volumes, we measure the volume-space homogeneity by computing the pairwise L2 distance of volumes for a randomly sampled cluster (Figure S2). We use a sketch size of $k = 500$ for all downstream analysis of the ClpXP dataset.

2.4 Masking for feature selection

Once the volume sketch is generated, we mask out the voxels of the sketched volumes that correspond to the either background or a user-defined region prior to performing the downstream clustering and dimensionality reduction analysis (Figure 1C). A benefit of working in volume space is that the "features" of the volume vector are voxels in the volume representation. This allows us to remove voxels that are known to be irrelevant (by default, the background), which reduces the variation introduced by (random or structured) noise in the masked out region. The remaining variance is thus more likely to be meaningful, which is especially important since variance is fundamental to both our downstream clustering (Ward linkage minimizes cluster variance, see Section 2.6) and dimensionality-reduction (PCA finds the directions of maximum variance) analyses.

For each of the 500 sketched volumes, we define the region to exclude as the voxels whose density is less than half of the maximum density of the volume; the final binary mask applied to all volumes is the union of the masked out region for each volume. The user can also define the mask on their own, which is especially useful if there is prior information on part of the complex that should be focused on. In the ClpXP example, the ClpX subunit is dynamic, and so that region is manually chosen to analyze in our pipeline (Figure 1C). Masking also has the computational benefit of reducing the feature space from 2,097,152 voxels in the 128^3 volume to the 162,210 voxels that are in the mask.

2.5 Visualizing a conformational landscape

Taking inspiration from previous work[13], we apply principal component analysis (PCA) on the set of masked volumes and use the resulting eigenvector decomposition to visualize the entire ensemble of reconstructed density maps (Figure 2). The PCA analysis provides two benefits: 1) the resulting eigenvectors (i.e. "eigenvolumes") produce trajectories along the axes of maximum variation which can be used to summarize the major modes of motion (Figure 2A,B); and 2) the top N principal components (PCs) produce a low-dimensional *volume-space* embedding that can be used to visualize the entire cryoDRGN ensemble (Figure 2C).

For the ClpXP protease, traversing the first PC in volume space corresponds to the transition between the ClpXP intermediate and recognition complexes (Figure 2A); the second corresponds to dissociation of ClpX (Figure 2B); and the third corresponds to appearance of the GFP substrate. Because the PCs form a linear, orthogonal basis, the location of each sketched volume in

the PC embedding space can be more easily interpreted (i.e. as the linear combination of the basis vectors). This is in contrast to the latent variable representation (shown in Figure S5), where traversals along the PC axes in *latent* space or UMAP coordinates are not guaranteed to provide comprehensive summaries of the *volume* ensemble, due to the nonlinear nature of the decoder.

On the ClpXP protease, the top three principal components capture 65% of the variance in the data (Figure S1), making even a three-dimensional representation reasonable. The set of volumes in the sketch may be visualized as a scatterplot in the volume PC space. For example, Figure 3A shows a scatter plot of PC1 and PC2 for the sketched volumes, and Figure 4A shows PC2 and PC3.

As the PC decomposition is estimated on the 500 volumes in the sketch, we next apply this decomposition to the entire dataset of 344,069 volumes in order to visualize the entire conformational landscape of all the imaged particles (shown in Figure 2C). Instead of generating all 344,069 volumes, which is computationally intractable, we learn a function ϕ that maps latent space coordinates for each particle i *directly* to volume embedding space. Specifically, ϕ takes on the form of a simple multilayer perceptron (MLP) network. Having computed the transformation to principal components on the initial 500 sketched volumes, we then compute the volumes of an additional sample of 25,000 latent representations z_i, $1 \leq i \leq 25,000$ and map those into PCA space v_i; these pairs (z_i, v_i) are used to train the MLP ϕ (additional methods in Section A). Once trained, the volume embedding representation of any point in the original dataset z can be computed as $\phi(z)$.

Finally, linear methods do have limitations for visualization, especially in cases where most of the variance is *not* contained in the top few PCs, or where the linear approximation to the underlying nonlinear motions is inaccurate. In those cases, nonlinear methods for visualization can be considered. We show two popular methods for visualization on the ClpXP volume sketch in Figure S3, multidimensional scaling (MDS) and UMAP.

2.6 Identifying conformational states with agglomerative clustering

We also cluster the volume sketch to summarize the major conformational states of the reconstructed ensemble. We use an agglomerative clustering algorithm and allow the user to vary the number of clusters M, the linkage criterion, and the distance metric. We note that different choices of the clustering hyperparameters emphasize different priors and definitions of what a "cluster" should be. We use agglomerative clustering with the goal that this bottom-up clustering algorithm may be effective at identifying outlier states, including rare states (or junk particles), which would "look different" (under the e.g. L2 distance metric) than the rest of the ensemble, and thus be agglomerated last. Unlike top-down clustering algorithms such as k-means, which first define the differences between clusters, agglomerative clustering does not impose any geometric priors on the shape or size of the clusters. On the ClpXP dataset, agglomerative clustering with $M = 10$ target clusters, a Euclidean distance metric, and an "average" linkage criterion, which minimizes the average distance between the two sets when merging clusters, yields five well-populated and five sparse clusters (Figure 3). An example of clustering results with $M = 20$ target clusters is shown in Figure S3 and with Ward linkage (which minimizes the variance within each cluster) is shown in Figure S6.

We compute the centroid of each cluster (i.e. an average of the volumes in the cluster) as a representative structure; these summary states can be quickly and efficiently inspected to evaluate the diversity of the dataset (Figure S4). In the case of ClpXP, we find that cluster **0** represents the complex with the ClpX hexamer absent. Since we are interested in ClpX variability, the volumes from cluster **0** can be excluded to avoid the consideration of this state when re-running landscape analysis (not shown) or the underlying particles may be removed from any further cryoDRGN training.

After inspecting the sparse clusters, we found that cluster **8** reflects the substrate-engaged state of ClpXP (see Figure 4). This state was both missed in the original 3D classification of this dataset[9] and by expert-guided inspection of the cryoDRGN ensemble, yet is biochemically known to be in the sample. Since this cluster represents only 0.3% of the dataset, our focus on ensuring the diversity of possible conformations was covered in our sketching step was crucial for its discovery. We combined the 1,255 particles of the volumes associated with cluster **8** and performed a homogeneous refinement in RELION to validate the presence of this structure (Figure 3C). An atomic structure of the GFP substrate was able to be docked into the resulting density map (Figure 3D).

3 Results

4 Discussion

The ability of cryoDRGN to model complex structural distributions has raised new questions on how its underlying deep generative model should be interpreted to yield testable structural hypotheses. In particular, the ability of cryoDRGN to reconstruct an arbitrary number of structures, rather than a single or discrete set of structures (tens of structures), presents a novel challenge since examining each structure V_i of the dataset individually is both computationally and manually intractable. Here, we have introduced a "landscape analysis" pipeline that aims to summarize the full diversity of structures in a trained cryoDRGN model for the practitioner. The method is implemented as a tool in the cryoDRGN software for automated analysis,

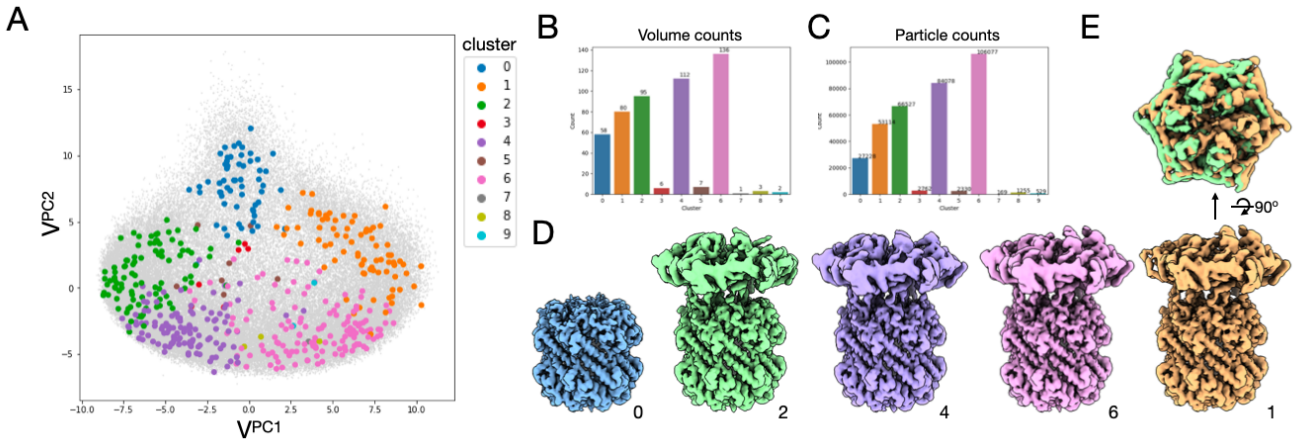

Figure 3. Conformational states of ClpXP inferred from clustering the CryoDRGN volume distribution: Agglomerative clustering ($M = 10$ clusters) produces five well-populated and five sparse clusters. **A.** The sketched datapoints are colored by their assigned cluster and plotted in volume PC space (from Figure 2). **B.** and **C.** The number of volumes and the number of particles for each cluster. Note that some clusters have very few counts, indicating they are outlier groups that might be artifacts or interesting rare conformations. **D.** Representative structures (the centroid of the cluster) for the five most populated clusters. Additional structures are shown in Figure S4. **E.** The top-down view of the cluster 1 and cluster 2 volumes from **D** superimposed, highlighting the conformational change between the ClpXP recognition and intermediate complex.

and we have found this approach to be useful for quickly analyzing models, especially in cases where the latent embeddings are visually uninformative.

This landscape analysis pipeline performs two separate but complementary approaches for summarizing the learned distribution: 1) as a small number of *discrete* conformational states (and their constituent particles for further refinement), including rare states of interest or 2) with *continuous* reaction coordinates inferred from PCA that provide an interpretable conformational landscape visualization of the full dataset. The interpretation as discrete or continuous variability, while seemingly at odds, work well together. For one, the choice of method can be tailored to specific structural hypotheses surrounding the dataset of interest. But more generally, the two interpretations may be complementary when both compositional and conformational heterogeneity are present in the dataset, such as in the ClpXP protease, e.g. dissociation of ClpX (compositional) and conformational transitions between the intermediate and recognition complexes (conformational). Even when there is a conformational continuum, it may be useful to discretize the continuum for summary structures. We emphasize that these analyses place different structural assumptions on the ensemble of volumes, and ultimately the choice of interpretation is made by the practitioner.

Unlike PCA-based approaches for reconstruction, such as in Tagare et al.[14] and 3D Variability Analysis[3], PCA is used here to summarize features of a full-rank set of volumes reconstructed by cryoDRGN. While cryoDRGN and other nonlinear methods for heterogeneity analysis can produce complex distributions of density maps, the latent representation is not directly interpretable due to the nonlinear nature of the mapping from latent space to volumes. Here, by separating reconstruction from the downstream volumetric analysis, we can take advantage of both cryoDRGN's powerful nonlinear representation of 3D density maps and established dimensionality reduction techniques to obtain interpretable features.

The analysis of large sets of vectorized volumes (i.e. high-dimensional vector arrays) is a general problem in large-scale, high-dimensional data analysis, and many other algorithms are transferable to this space. For example, this landscape analysis framework can be easily modified to use a different sketching algorithm, clustering algorithm, or volume embedding algorithm. This approach may also be tailored to analyze the results from other heterogeneous reconstruction methods that generate large ensembles of volumes, and may be especially relevant for the growing number of reconstruction methods based on deep learning[4,15]. Finally, this approach is a purely data-driven approach for analyzing the ensemble of volumes (aside from any user-provided masks), and thus will be less biased, but perhaps less informative than other methods that guide the analysis of the ensemble based on an atomic model[16,17].

5 Code Availability

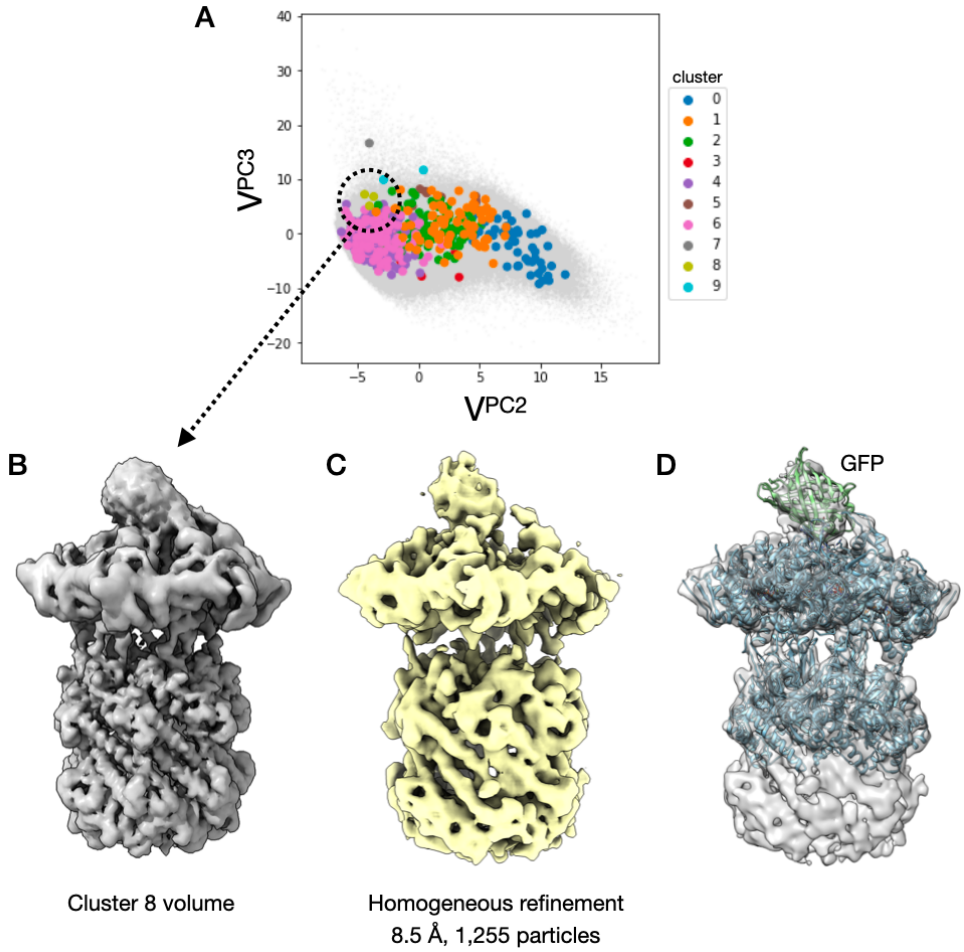

Figure 4. Identification of the ClpXP substrate-engaged state: Inspecting the cluster **8** structure from Figure 3 revealed the ClpXP substrate-engaged state. **A.** Cluster **8** can be identified on the volume PCA plot when comparing PC2 and PC3, where this cluster is more separated. **B.** The representative volume (cluster centroid) for cluster **8**. **C.** Homogeneous refinement in RELION of the 1,255 particles within this cluster. **D.** The density map from (C) with the atomic model docked. Although the GFP substrate is low-resolution, the density of GFP is well aligned with the atomic model.

6 Acknowledgements

We thank the MIT-IBM Satori team for GPU computing resources and support. This work was funded by the NSF GRFP Fellowship (award 1122374) to E.D.Z. , NIH grant R01-GM081871 to B.B., NSFCAREER-2046778 and NIH grant R01-GM144542 to J.H.D., and a grant from the MIT J-Clinic for Machine Learning and Health to J.H.D. and B.B.

7 Author contributions

8 Competing Interests

References

1. Cheng, Y. Single-particle cryo-EM—how did it get here and where will it go. *Science* **361** (2018).
2. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. methods* **18**, 176–185 (2021).
3. Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *J. Struct. Biol.* **213**, 107702 (2021).
4. Chen, M., Ludtke, S. & Marrs, V. Deep learning based mixed-dimensional GMM for characterizing variability in CryoEM. *arXiv* (2021).
5. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife* **7**, e36861 (2018).
6. Xie, Y. *et al.* Neural fields in visual computing and beyond. (2021). 2111.11426.
7. Zhong, E. D., Bepler, T., Davis, J. H. & Berger, B. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *ICLR* (2020).
8. Gui, M. *et al.* Structures of radial spokes and associated complexes important for ciliary motility. *Nat. structural & molecular biology* (2020).
9. Fei, X., Bell, T. A., Barkow, S. R., Baker, T. A. & Sauer, R. T. Structural basis of ClpXP recognition and unfolding of ssra-tagged substrates. *Elife* **9** (2020).
10. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. (2018). 1802.03426.
11. Narayan, A., Berger, B. & Cho, H. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat. Biotechnol.* **39**, 765–774 (2021).
12. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric sketching compactly summarizes the Single-Cell transcriptomic landscape. *Cell Syst* **8**, 483–493.e7 (2019).
13. Haselbach, D. *et al.* Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell* **172**, 454–464.e11 (2018).
14. Tagare, H. D., Kucukelbir, A., Sigworth, F. J., Wang, H. & Rao, M. Directly reconstructing principal components of heterogeneous particles from cryo-EM images. *J. Struct. Biol.* **191**, 245–262 (2015).
15. Punjani, A. & Fleet, D. J. 3d flexible refinement: Structure and motion of flexible proteins from cryo-em. *bioRxiv* (2021).
16. Giraldo-Barreto, J. *et al.* A bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci. Rep.* **11**, 13657 (2021).
17. Davis, J. H. *et al.* Modular Assembly of the Bacterial Large Ribosomal Subunit. *Cell* **167**, 1610–1622.e15 (2016).

A Supplemental Methods

A.1 cryoDRGN training

CryoDRGN version 0.3.2 models were trained on 344,069 single-particle images of ClpXP from Fei et al.[9] downsampled to an image size of 128×128 (2.71875 Angstroms per pixel), with their corresponding poses assigned from a consensus reconstruction in RELION. All reconstructions used an MLP architecture with 3 hidden layers of width for the encoder and decoder networks. The latent variable dimension was 8. Training was performed in minibatches of 8 images using the Adam optimizer and a learning rate of 0.0001. Training was performed on a single V100 GPU and lasted 9 hours and 22 minutes.

A.2 Volume mapping

Given a PCA transformation W_L which keeps the top L components, we train a simple MLP network to learn the mapping from latent embeddings z_i to volume embeddings $v_i = V_i W_L$ to avoid generating V_i for all images in the dataset. A training set of (z_i, v_i) pairs is first generated: 25,000 latent embeddings are sampled from the dataset and used to generate their associated volumes through the decoder. Each volume is generated on the fly and embedded to avoid storing 25,000 voxel arrays. The MLP is trained using a 3:1 training set to validation set split, where the loss on the held out validation set is monitored to prevent overfitting. The MLP is trained for 50 epochs in minibatches of size 64 with the Adam optimizer and a learning rate of 0.001. The generation of the training set lasted 7 hours and 48 min. Training ϕ for 50 epochs lasted 4 minutes on a single Nvidia V100 GPU.

B Supplemental Figures

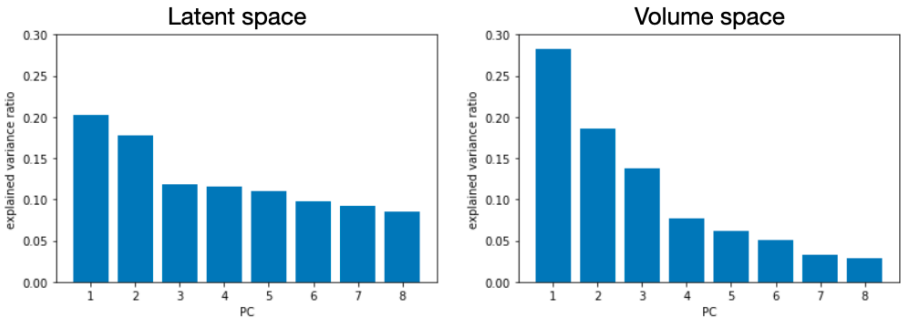

Figure S1. Explained variance of the top 8 principal components of the set of latent space embeddings and the volume sketch.

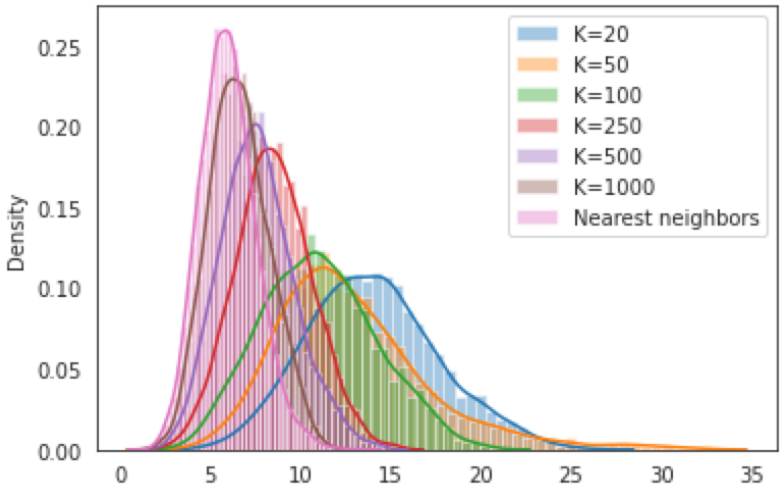

Figure S2. Distribution of pairwise L2 distances for the set of volumes in a sketched cluster for different values of k in k-means sketching.

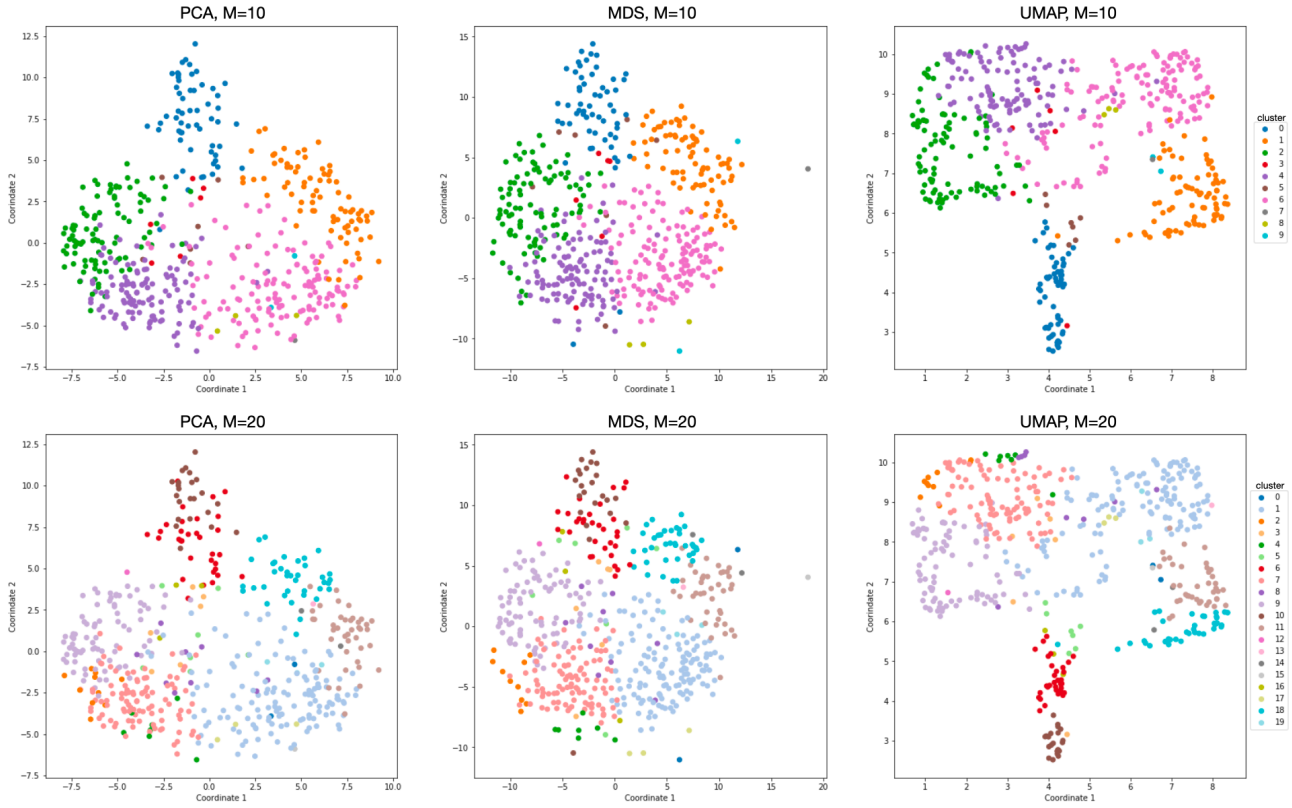

Figure S3. Different volume embedding algorithms applied on the sketch of volumes (PCA, MDS, UMAP from left to right). Different choices in in the number of clusters M (top row $M = 10$, bottom row $M = 20$)

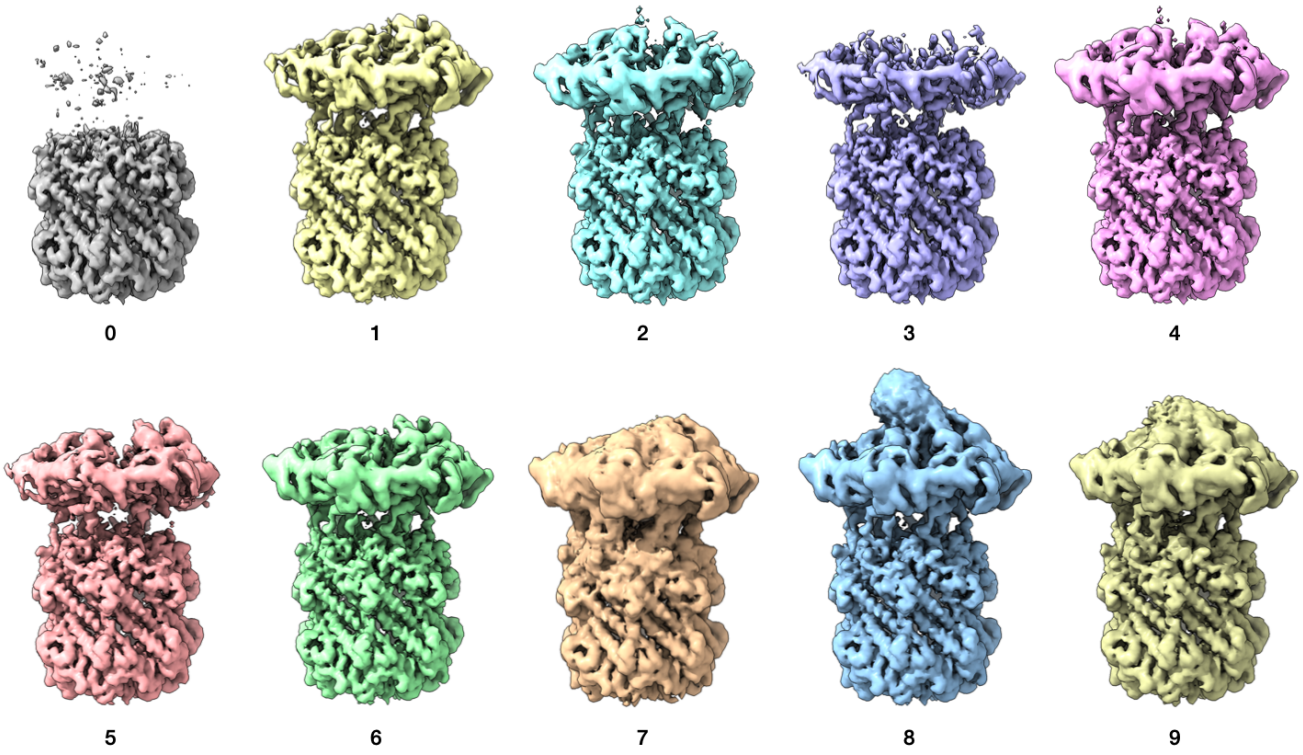

Figure S4. Cluster centroids after agglomerative clustering of the ClpXP volume sketch with $M = 10$, an average linkage criterion, and a Euclidean distance metric.

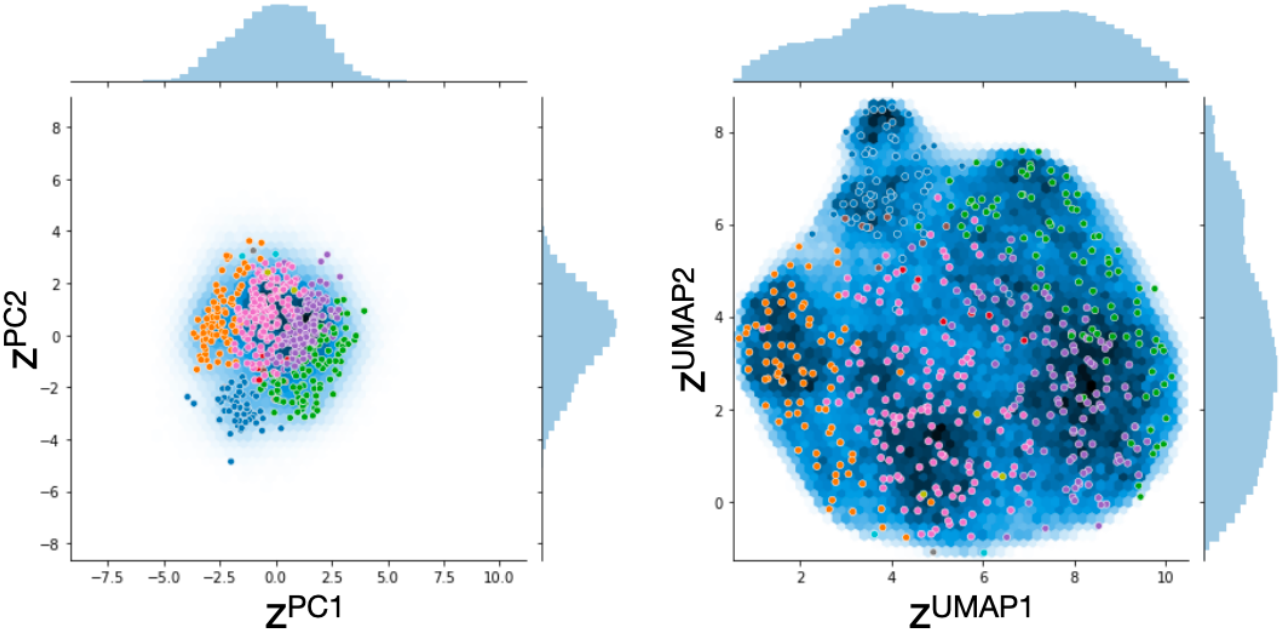

Figure S5. Clusters from Figure 3 visualized in the latent space representation of the dataset (PCA left, UMAP right)

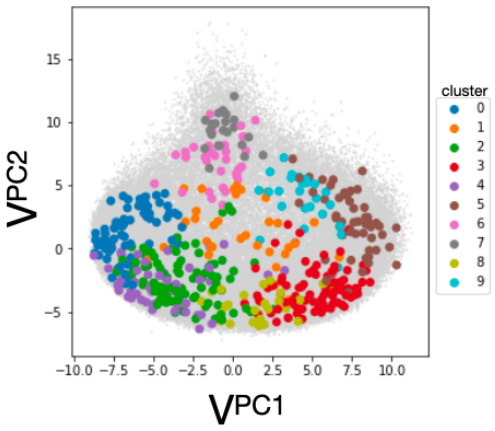

Figure S6. Agglomerative clustering of the volume sketch with $M = 10$, a Ward linkage criterion, and a Euclidean distance metric, visualized in the volume space representation of the dataset.