AFFINITY-VAE FOR CLUSTERING AND CLASSIFICATION OF OBJECTS IN MULTIDIMENSIONAL IMAGE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work we present affinity-VAE: a framework for automatic clustering and classification of objects in multidimensional image data based on their similarity. The method expands on the concept of β-VAEs with an informed similarity-based loss component driven by an affinity matrix. The affinity-VAE is able to create rotationally-invariant, morphologically homogeneous clusters in the latent representation, with improved cluster separation compared with a standard β-VAE. We explore the extent of latent disentanglement and continuity of the latent spaces on both 2D and 3D image data, including simulated biological electron cryo-tomography (cryo-ET) volumes as an example of a scientific application.

1 INTRODUCTION

Lying at the core of machine learning research, representation learning is one of the most important problems in our data-driven world. In recent decades, the performance of visual tasks has been coupled with appropriately chosen representation of an input, whether handcrafted or learned (Kingma & Welling, 2014). However, the interpretability of such factorised representations plays an important role, especially with scientific data. Furthermore, unlike the standard benchmark data in machine learning, real-life or scientific scenarios are often open problems with no existing annotations or ground truth. The development of methods for interpretable, factorised representations of data without supervision is therefore crucial to future scientific discoveries.

In the recent years, the β-VAE has become a prominent method for factorised representations in visual tasks (Higgins et al., 2017). However, despite significant developments (Bepler et al., 2019; Ziatdinov et al., 2021), learning continuous latent representations that capture and disentangle translation and rotation along with object semantics remains an open problem to this date, especially in 3D. Moreover, as such representations were designed to be completely data-agnostic, the approach makes no assumptions about the data resulting in little control over the learned semantics. In certain domains, however, such as visual tasks, common visual affinities (e.g. shape) could aid the process of representation learning (Huth et al., 2012; Op de Beeck et al., 2008).

We illustrate the problem on an open scientific challenge: the identification of molecules in volumetric cryogenic Electron Tomography (cryo-ET) image data. Cryo-ET is an emerging high resolution imaging technique that has the potential to revolutionize our understanding of molecular and cellular biology. Although powerful in their own right, structures of isolated and purified proteins convey little to no information on spatial distribution and interactions between the cellular systems in their native environments. Cryo-ET is uniquely capable of 3D *in situ* imaging, spanning molecular to cellular scales. The main promise of cryo-ET is to deliver such spatial mapping of a cellular landscape (Oikonomou et al., 2016), otherwise known as visual proteomics (Bäuerlein & Baumeister, 2021; Sali et al., 2003).

Cryo-ET tomograms are generated by collecting a tilt series of a frozen specimen in a transmission electron microscope (TEM). The individual 2D projection images are aligned and back-projected to generate the 3D tomogram (Turk & Baumeister, 2020). It enables resolution of the entire proteome of molecules inside whole cells in 3D (Murata et al., 2010; Jin et al., 2008), with further promise of increasing the precision of this method to 20 Å (Kühlbrandt, 2014; Wan & Briggs, 2016b). However, recent advances in instrumentation have not been matched by equivalent methodological develop-

ments for extraction of contextual information from reconstructed volumes (Bäuerlein & Baumeister, 2021). Such analysis routinely includes recognition and classification of particles of the same class, followed by subtomogram averaging within a class to obtain structures with higher local resolution and signal-to-noise ratios (Wan & Briggs, 2016b; Castaño-Díez & Zanetti, 2019; Wan & Briggs, 2016a). However, particle localisation, recognition and classification are inherently challenging for several reasons including low signal-to-noise ratios (SNR), molecular crowding, compositional and conformational heterogeneity, the random orientation of molecules and the abundance of different protein types

Many existing computational strategies have been developed to enable subtomogram target identification, most notably template matching which is a computationally expensive algorithm relying on the availability of high-resolution template libraries. CNNs have also been used in this context for multi-class classification, including work in domain adaptation and semi-supervised learning, however their reliance on manually-labelled data urges further development (Moebel et al., 2021; Che et al., 2018). Min et al. (2015) explored template- and label-free methods using pattern recognition, however they rely on high abundances of proteins for successful classification. More recently, some representation learning frameworks emerged providing coarse characterisation of cellular features (Zeng et al., 2018b). Related are also works of Rice et al. (2022); Zeng et al. (2018a) which focus on organising feature embeddings from CNNs and classical autoencoders. Recently, there has also been an increased interest in rotationally equivariant VAEs, which we are also exploring in this work (Bepler et al., 2019; Ziatdinov et al., 2021; Ziatdinov & Kalinin, 2021). Finally, our work also relates to contrastive learning frameworks (Chen et al., 2020) and deep subspace clustering Ji et al. (2017), though in contrast to our work, those methods include pairwise affinities as learnable parameters in a non-variational framework.

In this work, we introduce Affinity Variational Autoencoder (affinity-VAE), a deep neural network for automatic clustering and classification of multidimensional objects based on their similarity – in our case, their morphological similarity. We focus on affinity-based latent space regularisation in addition to a standard β-VAE loss function. We introduce an affinity-based γ-parameterised loss component along with the addition of a pose layer and an automatically generated affinity matrix to learn the pose of the sample in an unsupervised manner during training. The performance of this method is first investigated on a 2D alphanumeric dataset and then the application of the method is demonstrated on a simulated 3D cryo-ET example. The preliminary success of this novel approach in comparison with the existing β-VAE framework demonstrates its potential for application with experimental cryo-ET data.

2 METHODS

β-VAE is an iteration of the VAE framework introducing a hyperparameter β which modulates the learning constraints applied to the model (Higgins et al., 2017). Values of $\beta > 1$ put effective constraints on the capacity of the latent z bottleneck encouraging factorised representations (Burgess et al., 2018). The β-VAE approach is therefore most commonly used for unsupervised factorised representation learning.

The goal of the training is to minimise the objective function L,

$$\underset{x}{\operatorname{argmin}}\, L = ||x - y||^2 + \beta \times D_{KL}\left[\mathcal{N}(\mu_z, \sigma_z), \mathcal{N}(0, 1)\right] \tag{1}$$

Where x and $y = d(e(x))$ denote the input data and the output reconstruction, $z = e(x)$ is the encoded latent representation and e and d are the encoder and decoder parts of the neural network, respectively. D_{KL} refers to the Kullback-Leibler divergence between the prior $\mathcal{N}(0, 1)$ and posterior $\mathcal{N}(\mu, \sigma)$ distributions. The first term of the equation is the reconstruction loss minimising the difference between the inputs and the decoded outputs. The second term, parameterised by β, is the variational term that regularises the latent space.

Affinity-based loss component: In addition to the reconstruction and KL divergence terms of the standard VAE loss function, we introduced a new shape regularisation term $S(z)$. The hyperparameter γ provides fine control of the influence of this regularisation term (in a similar manner to β):

$$L = ||x - y||^2 + \beta \times D_{KL}\left[N(\mu_z, \sigma_z), N(0,1)\right] + \gamma \times S(z) \tag{2}$$

where $S(z)$ is the L1 norm of the difference between a pre-calculated affinity matrix $\mathbf{A}$ and the cosine similarity of the latent representations:

$$S(z) = \frac{\sum_{i,j}^{N} \left|\left| \mathbf{A}_{ij} - \frac{z_i \cdot z_j}{||z_i|| \cdot ||z_j||} \right|\right|}{N} \tag{3}$$

with z denoting the latent variables, (i, j) the indices corresponding to the pairwise combinations of the n input classes (in the batch) and N the batch size.

The cosine similarity measures a distance between two latent points, whereas the pre-computed affinity matrix ($\mathbf{A}$) provides feedback on their actual pairwise similarity. This effectively organises the latent space so that similar objects (regardless of their pose), as described by the similarity descriptor in the affinity matrix, are placed close together in the latent space. The pre-calculated affinity matrix is generated automatically by computing pairwise similarity scores for all classes in the training set with a target function. In our case this is SOAP for the 2D alphanumeric data and FSC for 3D protein data (more on data generation in section 2, data simulation), but different metrics could be chosen to organise the latent space by different factors or to facilitate other types of data (see section 2, affinity metrics). Furthermore, the affinity matrix is only used in the training stage, therefore a pre-trained network can easily be applied to the discovery of new classes or species.

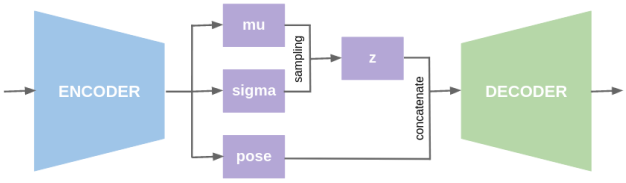

Figure 1: Architecture of the affinity-VAE.

Pose feature space: The affinity-VAE architecture is shown in Figure 1. In addition to a standard β-VAE, we introduce a third fully connected layer (parallel to μ and σ) to represent the pose of the object. By providing the same affinity values for all instances of the same class regardless of their orientation (or other class-specific variation), we discourage any pose-related variance from the latent representation z. However, in order to maximise the reconstruction certainty, such variation must be represented elsewhere in the model, in this case through the non-affinity-regularised pose feature space.

Latent map: To ensure that our method is generalising to new data, so that the network does not need retraining every time it analyses a new data set, we introduce the "latent map" approach. The latent space is pre-seeded with a non-redundant but exhaustive set of existing objects during training to create a "latent map", which serves as a reference space during inference. The "latent map" offers the promise of applying affinity-VAE as an unsupervised template-and-label-free approach to the classification and detection of novel structures in cryo-ET tomograms.

Latent embeddings In this work we use t-SNE (Van der Maaten & Hinton, 2008) to visualise the multi-dimensional latent spaces in 3D data and UMAP (McInnes et al., 2018) in 2D data.

Clustering evaluation: The accuracy of the method was evaluated by computing cluster centres in the latent space and assigning the class of the closest cluster for each unseen data point.

Data simulation: Three datasets were used in the evaluation of the methods: 1) 2D alphanumeric data, 2) 3D tetramino data, and 3) simulated 3D protein data.

Images in the alphanumeric dataset are constructed from a selection of letters and digits ($x \in \{\texttt{a, e, b, d, p, i, j, z, 2, k, x, u}\}$), rotated at various angles (Figure 2 left). Images are rotated to a defined angle θ (where $\{\theta \in \mathbb{Z} | -45 < \theta < 45\}$), and converted to a binary two-dimensional array. We use an $80/20$ split for training and testing. The tetramino data (Figure 2 middle) was designed to emulate an easier, but similar, 3D scenario to real cryo-ET data. Tetraminos were made of combinations of four identical cubes connected wall-to-wall. They were

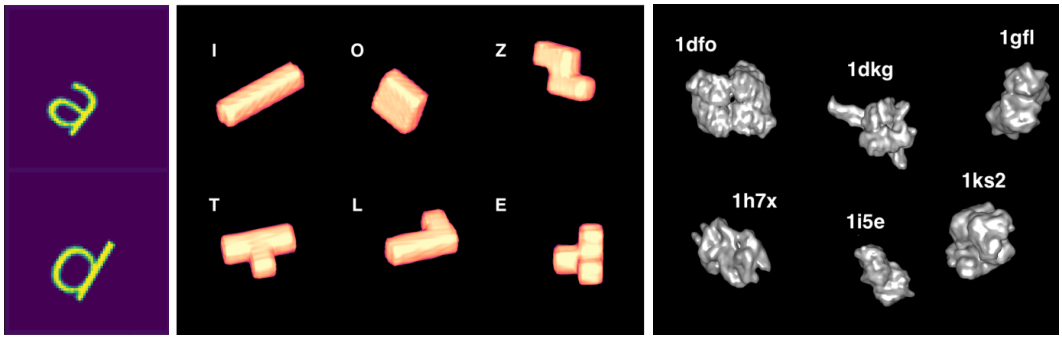

Figure 2: Sample data, left: alphanumeric, middle: tetramino blocks and right: protein volumes from the list in McGuffee & Elcock (2010).

generated on-the-fly with controllable parameters such as rotation and other morphological adjustments (e.g. elongation), allowing us to generate any desired shape combinations to test the semantic disentangling power of the method. The training data was constructed from 6 morphologically different tetraminos which were used to pre-seed the latent map, while for evaluation we constructed a new, previously unseen class that was morphologically similar to two other classes from the training set. The protein data (Figure 2 right) was generated from the list of 50 most abundant *E. coli* proteins (McGuffee & Elcock, 2010). UCSF Chimera (Pettersen et al., 2004) was used to generate a synthetic 3D density map from each protein on the list. The maps were generated at 10 Å resolution without taking atomic B-factors into account. The training data was constructed from n randomly selected classes (protein types), which were used to pre-seed the latent map, and a different, previously unseen class was selected for evaluation. All 3D data was augmented with rotations randomly sampled at $\theta = \{10, 20, ..., 360\}$ in 3 different planes. All images were resized prior to training to 32^3 and 64^3 voxels in tetramino and protein data respectively. Test and validation data constituted 10% and 20% of the whole dataset respectively.

Affinity metrics: Affine regularisation of the latent space ensures that semantically similar objects are encoded in proximity in the latent space. However, the choice of affinity descriptor should be made with respect to the property of the data intended to achieve the desired data separation. In our case we aim to organise the latent space by structural similarity, therefore we use Smooth Overlap of Atomic Positions (SOAP) descriptor (Bartók et al., 2013) in 2D data and average Fourier Shell Correlation (FSC) in 3D data (Harauz & van Heel, 1986).

SOAP (Bartók et al., 2010; 2013) is a shape descriptor that uses a combination of radial and spherical harmonics. In our model, we are treating every pixel that is not background as an "atom". The SOAP descriptor places a (three-dimensional) Gaussian density distribution at the location of each selected pixel. The SOAP kernel is then defined as the overlap of the two local nearest neighbouring densities integrated over all three-dimensional rotations.

FSC (Harauz & van Heel, 1986) curves are the standard metric for resolution estimation of cryo-EM maps. The method calculates the similarity of two images as a function of spatial frequency, by calculating the correlation between the Fourier coefficients of each image in thin spherical shells:

$$FSC(k) = \frac{\sum F_1(k) F_2(k)^*}{\sqrt{\sum |F_1(k)|^2 \sum |F_2(k)|^2}} \tag{4}$$

where $F_1(k)$ and $F_2(k)$ are the (complex) coefficients of the Fourier transforms of the two structures within a spherical shell at radius k. In this work, to obtain a single value for use as an affinity metric, we take an average of the FSC across all spatial frequency shells, weighted by the number of Fourier coefficients in each shell according to the method described by Brown et al. (2015). This gives a measure of similarity between the two 3D objects with a value between $+1$ and -1, with the former indicating a strong agreement.

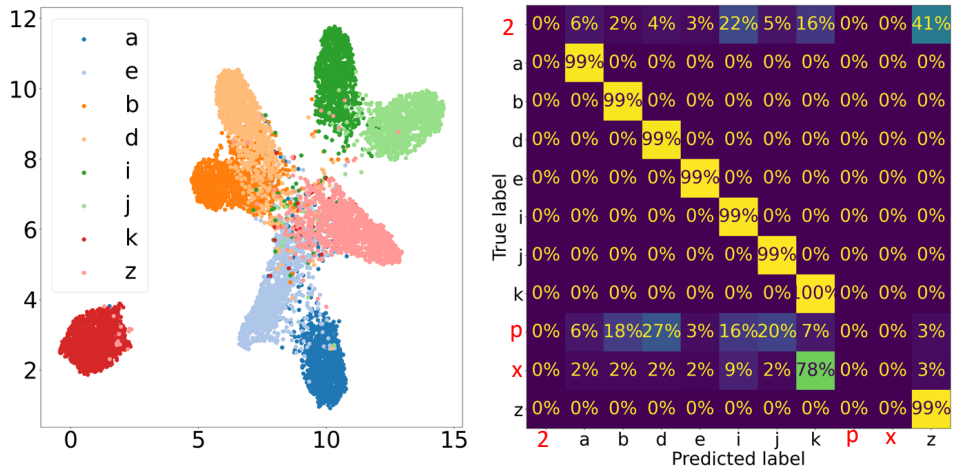

Figure 3: Left: UMAP embedding of latent space trained with 10,000 randomly rotated samples of alphanumeric data. Right: The confusion matrix built from the predictions for 200 samples of seen (a, b, d, e, i, j, z and k) and unseen data (2, x and p). For this calculation the choice of hyperparameters include $\gamma = 8$ and $\beta = 5$.

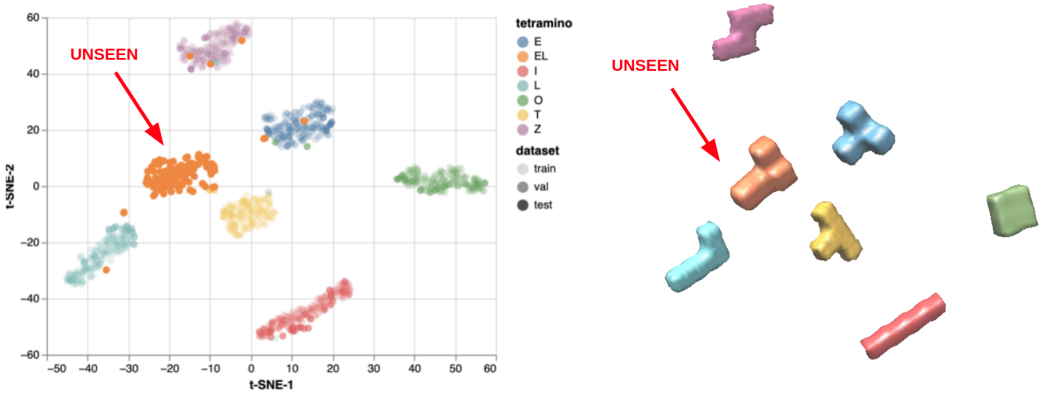

Figure 4: Left: t-SNE embedding of 1,000 (rotationally augmented) tetraminos from a set of morphologies. Different colours correspond to different morphologies and opacity indicates train/validation/test set. Right: Morphology of different tetramino types, placed in positions corresponding to their location in the latent space embedding, and with corresponding colors. One tetramino unseen during training and validation indicated with red arrow.

3 Results

3.1 Latent clustering for 2D alpha-numeric data

The left panel of Figure 3 illustrates the UMAP embedding of 10,000 rotations from a set of samples from the alphanumeric data with $\gamma = 8$ and $\beta = 5$. The right panel shows the confusion matrix constructed from the prediction for 200 samples of the seen data (a, b, d, e, i, j, z and k) and unseen data (2, x and p). The confusion matrix shows that the model predicts a strong affinity between letters with higher shape similarity for the seen and unseen data (for example, the unseen numeral 2 shows the closest match to the letter z from the training set).

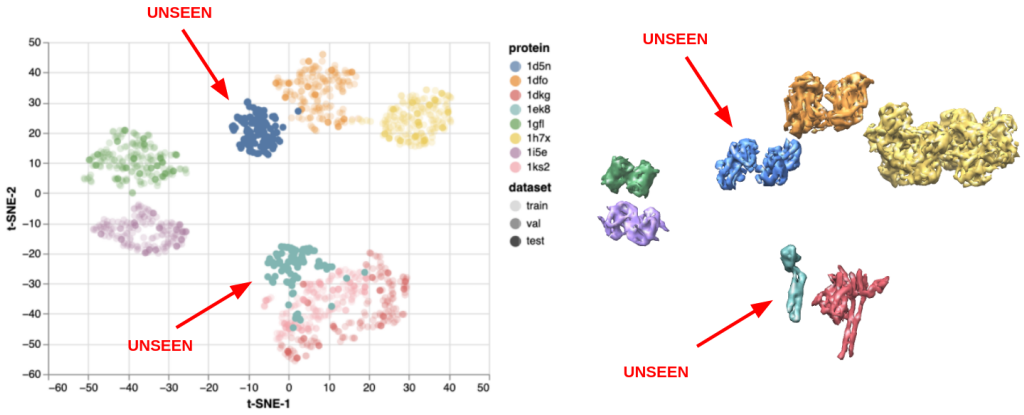

Figure 5: Left: t-SNE embedding of 1,000 (rotationally augmented) proteins from a set of 8 randomly chosen morphologies. Different colours correspond to different morphologies and opacity indicates train/validation/test set. Right: Morphology of different protein types, placed in positions corresponding to their location in the latent space embedding, and with corresponding colors. Proteins unseen during training and validation indicated with red arrows.

3.2 Latent clustering for 3D tetramino and protein data

The results on the tetramino data are illustrated in the latent embedding in Figure 4 and on the protein data in the latent embedding in Figure 5. In both datasets, the cluster separation between different morphologies was very good. Rotated objects of the same morphology were placed in homogeneous clusters regardless of their orientation. Additionally, in the tetramino data the clusters were arranged so that morphologically similar objects (e.g. E and T, L and I) were closer together in the latent space than dissimilar objects (e.g. I and E). A similar trend was observed in the protein data, where dimeric (two subunit) proteins were all arranged close together and ordered by the size of the protein, whereas elongated monomeric (single-subunit) proteins were placed separately forming a more homogeneous area in the latent space.

When the network was presented with samples previously unseen during training, they formed separate clusters in the embedding positioned close to similar morphologies, which suggests that the learned latent spaces are continuous (see more in subsection 3.3) and offers potential use of the method for discovery of new morphologies. At the same time, cluster homogeneity was preserved within the unseen clusters, regardless of the object orientation. In the case of the tetramino data, we introduced a new morphology during evaluation that was a fusion of two similar morphologies existing in training (EL), which, as expected, clustered between the two similar classes (E and L). In the case of the protein data, the two introduced morphologies were selected randomly from a list containing an exhaustive set of proteins from the *E. coli* cytoplasm. The dimeric protein (blue) was placed near other dimers and positioned on a line between other smaller and larger dimers, as expected if the network is encoding the size along a continuous axis. The monomeric protein with an elongated domain was placed overlapping with another cluster of proteins with elongated domains.

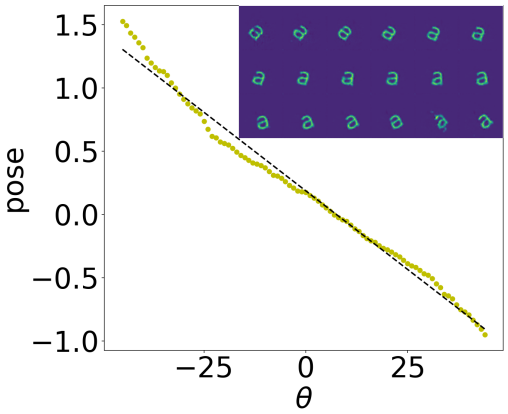

Figure 6: Linear relationship between encoded pose and the associated angle of rotation of the input. The nested panel shows the reconstruction using a given latent encoding (z) and varying the value of the pose corresponding to 5 degree intervals.

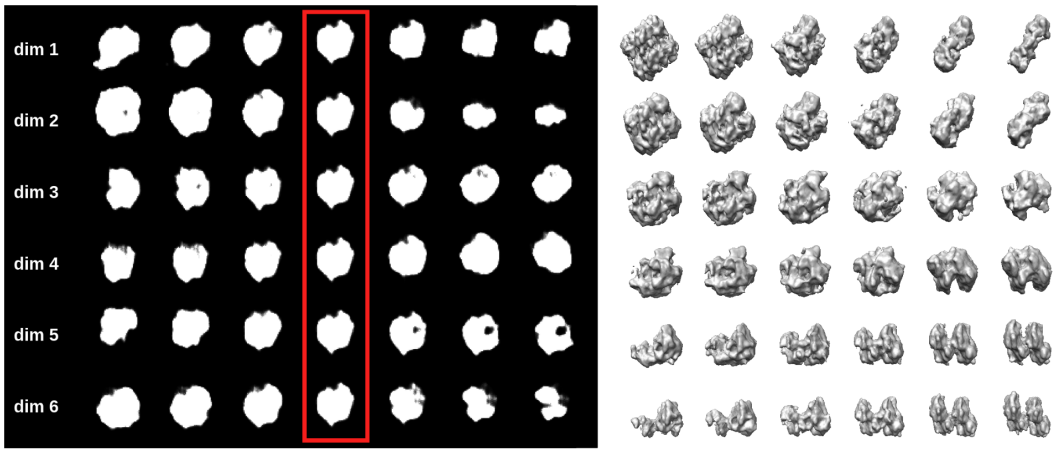

Figure 7: Left: Latent interpolations on protein data in a 6-dimensional latent space. Rows correspond to dimensions and columns correspond to interpolated values, where the central column is the encoded input. Right: Latent interpolations between 4 existing encoded proteins (corners). Decoded reconstructions in between the 4 corners are taken from the latent space at evenly sampled multi-dimensional strides between the encodings of existing proteins.

3.3 Latent and pose interpolations

We performed interpolations across the outputs of the pose component on alphanumeric data. The results of the interpolations are illustrated in Figure 6. We observed a linear correlation between the pose value and the angle of the rotation, which demonstrates that pose component does indeed capture information about the pose of the object.

We also explored the extent of disentanglement present in the generated latent spaces. Upon visual inspection we were able to identify morphological semantics across different dimensions (Figure 7 left). In this example using the protein data, dimension 2 appeared to capture the size, whereas dimension 6 described whether the protein was a dimer (two subunits) or a monomer (single unit). Other dimensions supported other morphological features, including elongation (dim 1), smoothness (dim 3 and 4), and toroidal geometry (dim 5).

Additionally, we performed latent interpolations across all dimensions between four existing (encoded) data points in the latent space (Figure 7 right). Non-existing (not encoded) points from the latent space generated realistic reconstructions and there was a smooth transition between different morphologies, including when transitioning between the number of protein subunits. This shows that the generated latent spaces are continuous and suitable to discovery of new morphologies unseen during training.

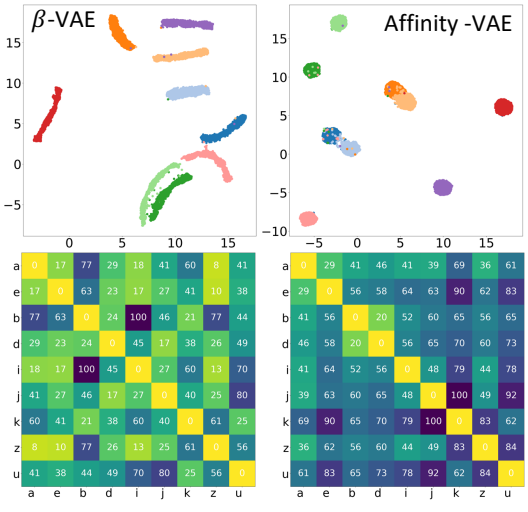

Figure 8: Latent space representation for the alphanumeric dataset using VAE where $\gamma = 0$ and pose component is off (left column) and affinity-VAE where $\gamma = 10$ (right column). Bottom row: latent space proximity matrix displaying the distance between cluster centres. The values for the distances are normalised to be between 0 and 100 in both matrices. The colour map is used as a guide for the eye to emphasise the largest and smallest numbers displayed on the matrix.

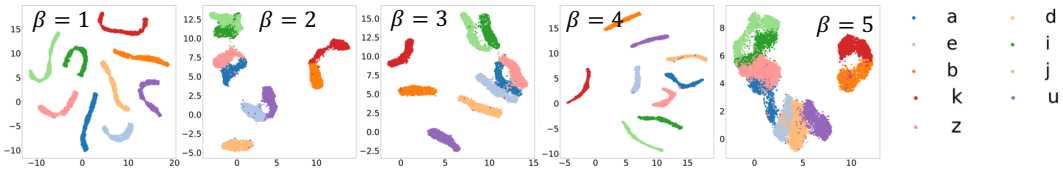

Figure 9: The latent space is presented for $\beta = 1, 2, 3, 4$ and 5 using original VAE framework ($\gamma = 0$). A more effective categorisation of classes is observed for values $\beta > 1$.

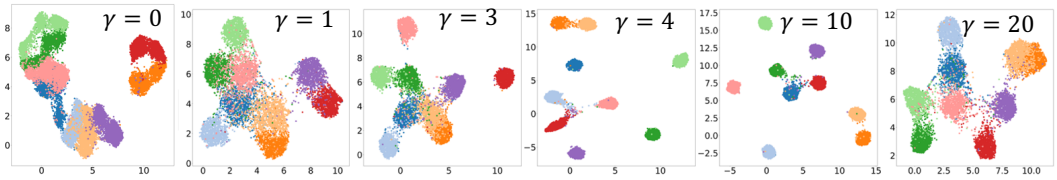

Figure 10: For $\beta = 5$ we have explored the effect of shape-affinity regularisation of latent space for $\gamma = 0, 1, 3, 4, 6, 10$ and 20. The colour code is the same as Figure 9.

3.4 Affinity-based loss component and the influence of γ

Figure 8 illustrates a comparison between affinity-VAE and a standard β-VAE ($\gamma = 0$ and no pose component) on alphanumeric data, including the latent space representations (top row) as well as the proximity matrix where the distances between the centres of the clusters are displayed (bottom row). Inspection of the two latent space representations shows that affinity-VAE ($\gamma = 10$) is more successful at relating the clusters with higher affinity than the β-VAE framework ($\gamma = 0$). This is confirmed in the proximity matrix where the distances between different clusters (i.e. the off-diagonal elements) are generally much higher in affinity-VAE than in the β-VAE, which would be expected to improve classification rates due to less cluster proximity contamination. Secondly, letters with dissimilar morphology (e.g. z and i/j pair) are pushed apart in affinity-VAE. Furthermore, rotationally-equivalent classes were in closer proximity in the latent map in affinity-VAE compared with the β-VAE (e.g. a and e). Interestingly, affinity-VAE also became invariant to other transforms such as symmetry (e.g. b and d).

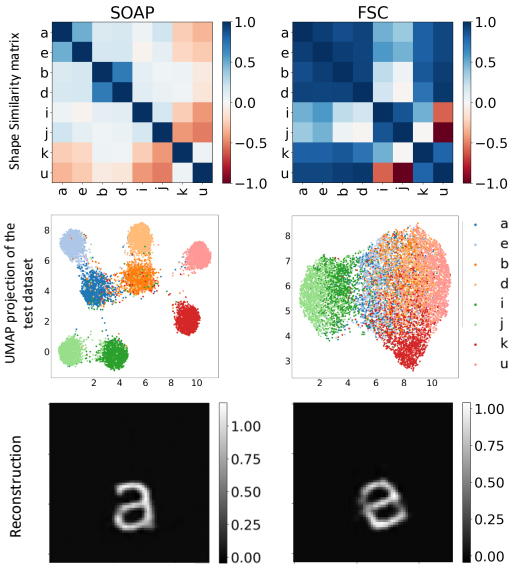

Figure 11: The left and right panel show a comparison of the latent space representation and the reconstruction of the alphanumeric data for the SOAP and FSC shape descriptors respectively.

3.5 The influence of β, γ and z

Figures 9 and 10 explore the effect of regularisation on the latent space. By switching on the shape affinity ($\gamma > 0$) the clusters are grouped together based on their structural affinity (Figure 10) unlike in the β-VAE framework where, while the classes do form homogeneous clusters, their proximity is not related to morphological similarity (Figure 9). As the emphasis on the affinity increases, the latent space becomes increasingly sparse pushing the clusters further apart. Therefore a careful balance should be exercised while selecting the β-γ trade-off.

3.6 SIMILARITY FUNCTION

To illustrate the influence of the choice of similarity metric used to calculate the affinity matrix on the overall cluster separation we compared various metrics across different datasets.

While the SOAP metric was used for 2D alphanumeric data and FSC for 3D data, a comparison between the two descriptors for the alphanumeric data is provided in Figure 11. SOAP provided better cluster separation as well as increased certainty of reconstruction on the alphanumeric data.

We also explored the choice of similarity metric in 3D tetramino data. Figure 12 shows a comparison between no affinity descriptor, mean difference and FSC metrics. Mean difference, unlike average FSC, is a real space descriptor which is not frequency weighted (Equation 5).

$$\frac{1}{N}\sum_{i=0}^{i=N} x_i - y_i \tag{5}$$

where x_i and y_i are voxels in comparison, and N stands for the number of voxels (image size).

While mean difference also improved the cluster separation over a standard β-VAE, some clusters (e.g. L and I) still remained unseparated. On the other hand, after employing the FSC as a similarity metric not only did the cluster separation improve, but also the organisation of clusters was morphologically aligned (e.g. elongated shapes like L, I and T clustered near each other).

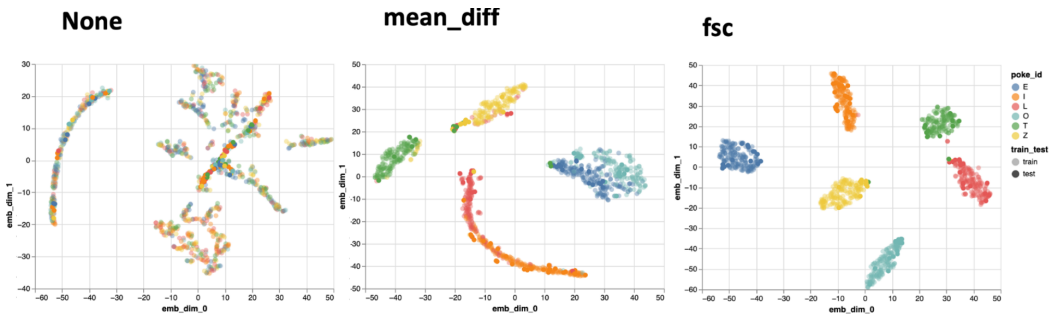

Figure 12: The left, middle and right panel show a comparison of the t-SNE latent space embeddings for the tetramino data for no affinity descriptor, mean difference and FSC respectively.

4 DISCUSSION

In this work we have introduced affinity-VAE, a neural network capable of organising the latent representation based on the similarity of the object. While a β-VAE is capable of cluster separation and latent semantic disentanglement, it offers little control over the learned factorised representations and captured semantics. We have shown that with guidance from an automatically generated affinity matrix we can create more homogeneous, rotationally-invariant clusters that could improve the classification accuracy. Furthermore, the affinity metric can be tailored to the data or domain of interest, improving the generality of the method. Since the affinity metric is only used during the training phase (Equation 2) and we have demonstrated that the learned latent spaces can be continuous, a pre-trained network can easily be applied to unseen data in order to discover new classes or species.

We have demonstrated the potential of the method in a scientific application using example of subtomogram target identification in volumetric cryo-ET data. While the results are promising and show the potential of the method for discovery of new species in experimental data, more experiments are required to test the effectiveness of the method on a full tomogram and non-simulated data.

Additionally, as we start introducing more class-specific sources of variance (e.g. structural heterogeneity or missing wedge artefacts in addition to rotation), the pose component could be phrased in variational terms as an independently parameterised component of the loss function. This would encourage disentangled pose representations that are separated from the content of the classes.

REFERENCES

Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104:136403, Apr 2010. doi: 10.1103/PhysRevLett.104.136403. URL https://link.aps.org/doi/10.1103/PhysRevLett.104.136403.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013. doi: 10.1103/PhysRevB.87.184115. URL https://link.aps.org/doi/10.1103/PhysRevB.87.184115.

Felix JB Bäuerlein and Wolfgang Baumeister. Towards visual proteomics at high resolution. *Journal of Molecular Biology*, 433(20):167187, 2021.

Tristan Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from translation and rotation with spatial-vae. *Advances in Neural Information Processing Systems*, 32, 2019.

Alan Brown, Fei Long, Robert A Nicholls, Jaan Toots, Paul Emsley, and Garib Murshudov. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallographica Section D: Biological Crystallography*, 71(1):136–153, 2015.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *CoRR*, abs/1804.03599, 2018. URL http://arxiv.org/abs/1804.03599.

Daniel Castaño-Díez and Giulia Zanetti. In situ structure determination by subtomogram averaging. *Current Opinion in Structural Biology*, 58:68–75, oct 2019. doi: 10.1016/j.sbi.2019.05.011. URL https://doi.org/10.1016%2Fj.sbi.2019.05.011.

Chengqian Che, Ruogu Lin, Xiangrui Zeng, Karim Elmaaroufi, John Galeotti, and Min Xu. Improved deep learning-based macromolecules structure classification from electron cryotomograms. *Machine vision and applications*, 29(8):1227–1236, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

George Harauz and Marin van Heel. Exact filters for general geometry three dimensional reconstruction. *Optik.*, 73(4):146–156, 1986.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2012.10.014. URL https://www.sciencedirect.com/science/article/pii/S0896627312009348.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. *Advances in neural information processing systems*, 30, 2017.

Liang Jin, Anna-Clare Milazzo, Stuart Kleinfelder, Shengdong Li, Philippe Leblanc, Fred Duttweiler, James C Bouwer, Steven T Peltier, Mark H Ellisman, and Nguyen-Huu Xuong. Applications of direct detection device in transmission electron microscopy. *Journal of structural biology*, 161(3):352–358, 2008.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. International Conference on Learning Representations (ICLR) 2014, 2014. URL https://hdl.handle.net/11245/1.434281.

Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, mar 2014. doi: 10.1126/science.1251652. URL https://doi.org/10.1126%2Fscience.1251652.

Sean R. McGuffee and Adrian H. Elcock. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLOS Computational Biology*, 6(3):1–18, 03 2010. doi: 10.1371/journal.pcbi.1000694. URL https://doi.org/10.1371/journal.pcbi.1000694.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Xu Min, I Tocheva Elitza, Chang Yi-Wei, J Jensen Grant, and Alber Frank. De novo visual proteomics in single cells through pattern mining. *arXiv preprint arXiv:1512.09347*, 2015.

Emmanuel Moebel, Antonio Martinez-Sanchez, Lorenz Lamm, Ricardo D Righetto, Wojciech Wietrzynski, Sahradha Albert, Damien Larivière, Eric Fourmentin, Stefan Pfeffer, Julio Ortiz, et al. Deep learning improves macromolecule identification in 3d cellular cryo-electron tomograms. *Nature methods*, 18(11):1386–1394, 2021.

Kazuyoshi Murata, Xiangan Liu, Radostin Danev, Joanita Jakana, Michael F Schmid, Jonathan King, Kuniaki Nagayama, and Wah Chiu. Zernike phase contrast cryo-electron microscopy and tomography for structure determination at nanometer and subnanometer resolutions. *Structure*, 18(8):903–912, 2010.

Catherine M Oikonomou, Yi-Wei Chang, and Grant J Jensen. A new view into prokaryotic cell biology from electron cryotomography. *Nature Reviews Microbiology*, 14(4):205–220, 2016.

Hans P. Op de Beeck, Katrien Torfs, and Johan Wagemans. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *Journal of Neuroscience*, 28(40):10111–10123, 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2511-08.2008. URL https://www.jneurosci.org/content/28/40/10111.

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

Gavin Rice, Thorsten Wagner, Markus Stabrin, and Stefan Raunser. TomoTwin: Generalized 3d localization of macromolecules in cryo-electron tomograms with structural data mining. jun 2022. doi: 10.1101/2022.06.24.497279.

Andrej Sali, Robert Glaeser, Thomas Earnest, and Wolfgang Baumeister. From words to literature in structural proteomics. *Nature*, 422(6928):216–225, 2003.

Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. *FEBS Letters*, 594(20):3243–3261, 2020. doi: https://doi.org/10.1002/1873-3468.13948. URL https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.13948.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

W. Wan and J.A.G. Briggs. Cryo-electron tomography and subtomogram averaging. In *Methods in Enzymology*, pp. 329–367. Elsevier, 2016a. doi: 10.1016/bs.mie.2016.04.014. URL https://doi.org/10.1016%2Fbs.mie.2016.04.014.

W Wan and John AG Briggs. Cryo-electron tomography and subtomogram averaging. *Methods in enzymology*, 579:329–367, 2016b.

Xiangrui Zeng, Miguel Ricardo Leung, Tzviya Zeev-Ben-Mordehai, and Min Xu. A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. *Journal of structural biology*, 202(2):150–160, 2018a.

Xiangrui Zeng, Miguel Ricardo Leung, Tzviya Zeev-Ben-Mordehai, and Min Xu. A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. *Journal of structural biology*, 202(2):150–160, 2018b.

Maxim Ziatdinov and Sergei Kalinin. Atomai: Open-source software for applications of deep learning to microscopy data. *Microscopy and Microanalysis*, 27(S1):3000–3002, 2021. doi: 10.1017/S1431927621010436.

Maxim Ziatdinov, Ayana Ghosh, Tommy Wong, and Sergei V. Kalinin. Atomai: A deep learning framework for analysis of image and spectroscopy data in (scanning) transmission electron microscopy and beyond. *arXiv preprint arXiv:2105.07485*, 2021.