CryoEM 2D classification with translational and rotational invariant features

Yong Ming Da [1] **Joel Yeo** [2] **Yixiao Yang** [3] **N. Duane Loh** [1 2]

[1]*Department of Physics, National University of Singapore, 117551 Singapore, Singapore,* [2]*Department of Biological Sciences, National University of Singapore, 117558 Singapore, Singapore,* [3]*School of Information and Electronics, Beijing Institute of technology, Beijing, China, 100081.* Correspondence to: N. Duane Loh duaneloh@nus.edu.sg.

1. Introduction

Cryogenic electron microscopy (CryoEM) is a nobel-prize winning method for reconstructing 3D models of biological particles. Reconstructing a high-fidelity 3D model requires selecting the images that contains the particle of interest that is of quality. Due to the large number and noisy nature of CryoEM images, a 2D classification is used to select the images for 3D reconstruction. However, little is known about the discarded images beyond their inability to produce coherent averages. This challenge is further compounded by variability among particle images that share the same projection view but differ in translations, in-plane rotations, and defocus parameters. To address these issues, we propose an efficient image featurization that is invariant to translations and rotations. This featurization enables unsupervised learning of structural motifs directly of CryoEM images. The resulting feature manifold can be used to vector quantize the images and form a map of CryoEM data, via unsupervised clustering methods such as Gaussian Mixture Models (GMMs). The manifold allows us to visually separate junk from good particles, as well as an alternative way of selecting images for 3D reconstruction. Furthermore, the manifold is generalizable for different CryoEM data, and is able to separate images containing different particles without prior knowledge of the 3D structures.

2. Manifold of CryoEM data with translational and rotational invariance

The utility of translational and rotational invariant features is that the data manifold formed from images containing the same particle, featurized this way, would cluster together even with different projected views of the particle. Because features are invariant to in-plane transformations, images representing the same viewing angle collapse to the same point. As particles rotate out of plane, these points trace a continuous closed orbit on the manifold – a topological representation of the particle's projection space. Distinct particles or "junk" images, which possess different structural features, will likely occupy separate manifolds (or distinct regions of a manifold). This topological separation enables the effective isolation of high-quality particles from "junk" images. Furthermore, because these manifolds are already invariant to translations and rotations, unsupervised methods like Gaussian Mixture Models (GMMs) can cluster the images to form class averages for vector quantization.

To obtain the features given an image $g(x, y)$, we first take the amplitude spectrum of the Fourier transform of the image $|G(u, v)| = |F^{-1}\{g(x, y)\}|$. Then we take the inverse Fourier transform of $|G(u, v)|$. This will give features that are invariant to translations and commutes with rotations:

$$T(x, y) = F^{-1}\{|G(u, v)|\}. \tag{1}$$

Next, we take the Zernike transform of $T(x, y)$ that is rotationally invariant: [1][2]

$$T(x, y) \equiv \sum_{m,n} a_{m,n} R_n^m(\rho) e^{im\theta}. \tag{2}$$

The absolute value of the Zernike coefficients, $|a_{m,n}|$ are rotationally and translationally invariant. $R_n^m(\rho)$ is the radial part of the Zernike polynomial and $e^{im\theta}$ is the angular part of the Zernike polynomial.

2.1 Experimental & simulated results

We first tested the method on the simulated T20S[3] and 5MAC[4] CryoEM images with random orientations. We simulated the images 256×256 with a pixel size of 1Å per pixel, a dose of 20 $e^-/\text{Å}^2$, energy 300keV and a defocus range of 10000Å to 25000Å. We first phase flipped the images to remove aberrations, then downsampled the images to 128×128 pixels then extracted the translational and rotational invariant features.

We also tested the method using 3 different EMPIAR datasets, EMPIAR-10025[3], EMPIAR-10028[5], EMPIAR-11377[6]. EMPIAR is a database for CryoEM data.[7] We first cropped the images from the micrographs using CryoSPARC (an existing software for CryoEM 3D reconstruction) [8], then processed the images similarly with a downsampling of 64×64 pixels.

To visualize features we used PCA[9] first, keeping 95% of the variance, then use UMAP[10] to form a lower dimensional embedding of the features. We then used GMMs[9] to cluster the images in the embedded space and then calculated their respective class averages.

In both Fig. 1 and 2, we can indeed see that the images containing the different particles are separated without prior knowledge of the 3D structures. The inset images on the scatter points are the class averages of the corresponding patch. We see that vector quantizing the images in the embedded space is able to form stable class averages and it is able to detect junk classes. In Fig. 2, the junk classes are separated from the main patch of good images.

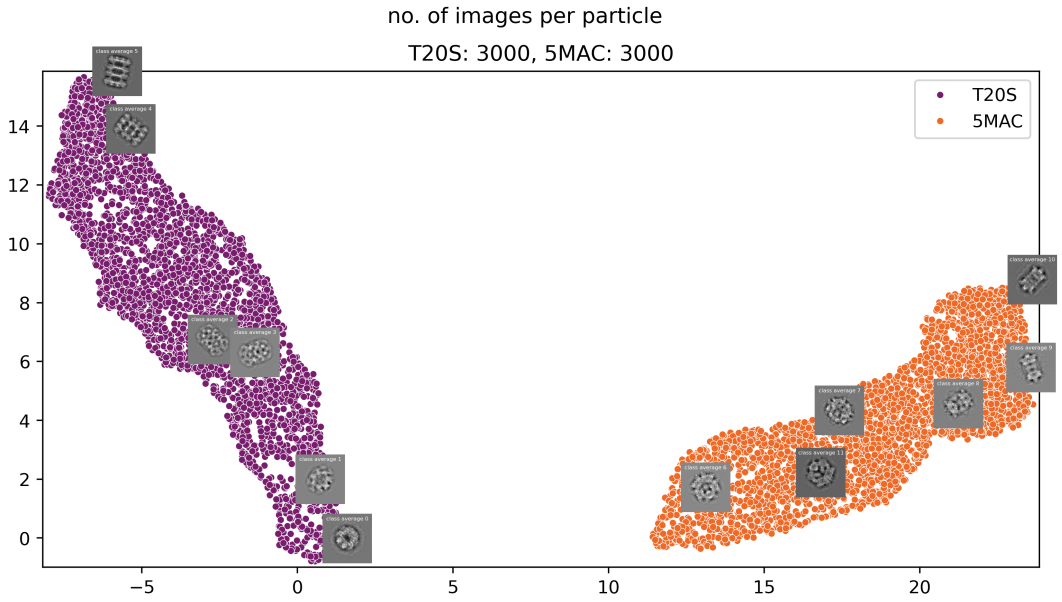

Fig. 1: Shown are the embeddings of simulated T20S and 5MAC images. Each scatter point represents the embedded translationally and rotationally invariant features extracted from an individual image. The inset images in the scatter points are the GMMs class averages of the corresponding patch. Under identical experimental conditions, images containing different particles form distinct clusters in the embedding space. Furthermore, images with similar particle projections are identified via vector quantization in the embedding space, yielding stable class averages.

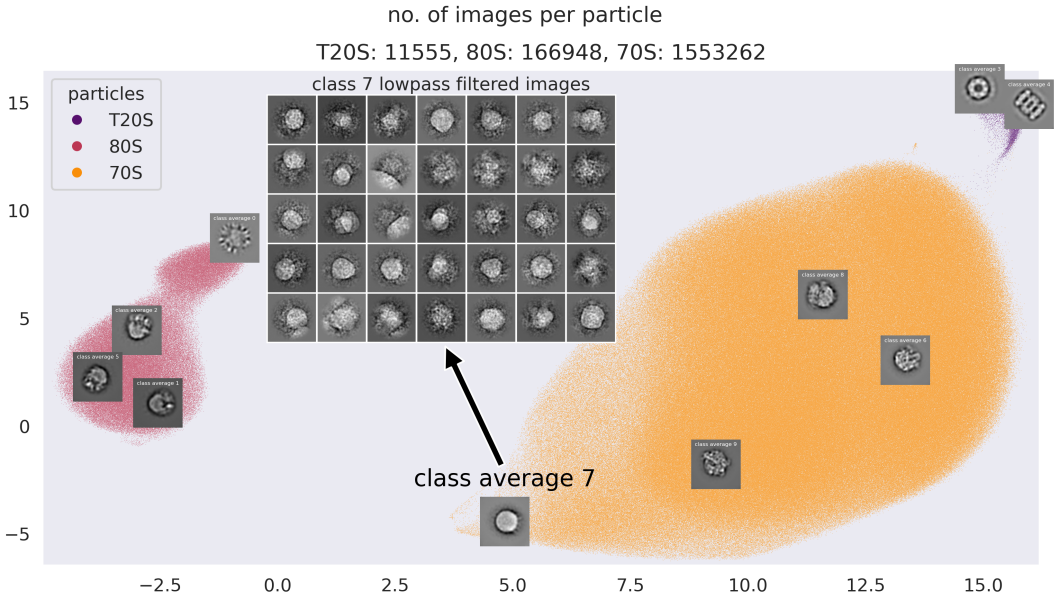

Fig. 2: Shown are the embeddings of real experimental data from EMPIAR-10025[3], EMPIAR-10028[5] and EMPIAR-11377[6]. Here we verify that the method works on real experimental data, yielding stable class averages after vector quantization and clear separation of images containing different particles. Notably, experimental junk images are separated from the main cluster of high-quality particles. Inspecting images from one such cluster (class 7) confirms that they correspond to ice-related artifacts.

3. Conclusion

We have presented a featurization framework that is invariant to translations and rotations. We investigated the structure of the resulting feature manifold for CryoEM images and observed that images containing the same particle cluster together. This property can be exploited to separate junk images from particle-containing images through vector quantization in the embedded space. For future works, we plan to incorporate a broader range of CryoEM datasets to construct a comprehensive data-atlas of currently available CryoEM data. We anticipate that such an atlas could provide insights about experimental conditions that contribute to particle loss.

Acknowledgments

We thank Dr. Yeo Zhen Yuan for helping with the averaging algorithm that is used to calculate the class averages of the images. Y.Y. acknowledges funding support from MOE AcRF Tier 1 (A-8000465-00-00), and J.Y. acknowledges funding support from both A*STAR Graduate Scholarship and the Chan Zuckerberg Imaging Institute. We also thank HPC support from NUS Centre for Bio-imaging Sciences.

References

[1] Jiadong Dan, Xiaoxu Zhao, Shoucong Ning, Jiong Lu, Kian Ping Loh, Qian He, N Duane Loh, and Stephen J Pennycook. Learning motifs and their hierarchies in atomic resolution microscopy. *Sci Adv*, 8(15):eabk1005, April 2022.

[2] Jiadong Dan, Cheng Zhang, Xiaoxu Zhao, and N Duane Loh. Symmetry quantification and segmentation in stem imaging through zernike moments. *Chinese Physics B*, 2024.

[3] Melody G Campbell, David Veesler, Anchi Cheng, Clinton S Potter, and Bridget Carragher. 2.8 Å resolution reconstruction of the thermoplasma acidophilum 20s proteasome using cryo-electron microscopy. *eLife*, 4, March 2015.

[4] Laura H Gunn, Karin Valegård, and Inger Andersson. A unique structural domain in methanococcoides burtonii ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) acts as a small subunit mimic. *Journal of Biological Chemistry*, 292(16):6838–6850, April 2017.

[5] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors H W Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *eLife*, 3, June 2014.

[6] Simon A. Fromm, Kate M. O'Connor, Michael Purdy, Pramod R. Bhatt, Gary Loughran, John F. Atkins, Ahmad Jomaa, and Simone Mattei. The translating bacterial ribosome at 1.55 å resolution by open access cryo-em. *bioRxiv*, 2022.

[7] Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 11 2022.

[8] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods*, 14(3):290–296, February 2017.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.