Pacific Northwest
NATIONAL LABORATORY

PNNL-33162

Structures via Reasoning - Applying AI to Cryo Electron Microscopy to Reveal Structural Variability

January 2022

Doo Nam Kim
Andrew August
Henry Kvinge
James Evans

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-33162

Structures via Reasoning - Applying AI to Cryo Electron Microscopy to Reveal Structural Variability

January 2022

Doo Nam Kim
Andrew August
Henry Kvinge
James Evans

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

There have been breakthroughs of latest cryo electron microscopy (cryo-EM) data analysis algorithms to classify cryo-EM image data. However, most of these cryo-EM reconstruction methods have focused on classifying distinctly different biomolecule structures. Here, we present our approaches of deep learning to differentiate homologous structures that are distinguishable only with inner morphological differences. We succeeded supervised classification of these subtly different homologues. However, we could not differentiate them with unsupervised methods. Here we discuss what further approaches are likely needed for successful unsupervised classification.

Summary

Ever since structural biology methods have played a major role, all these methods require time consuming sample preparation process. This limitation has been especially problematic for homologous structure determination. Here, we show that deep learning approach overcomes this hurdle by deciphering inner pixel densities. Determination of biological homologous structure will enable deeper mechanistic understanding and delicate control of pathway fate.

Acknowledgments

This project is funded by MARS (The Mathematics for Artificial Reasoning in Science) initiative in PNNL. We appreciate Ellen Zhong, an author of *cryoDRGN*, who advised us some features of *cryoDRGN*. We re-factored many codes from the *cryoDRGN*.

This research was supported by the Initiative/Investment at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

CNN: Convolutional Neural Network

cryoDRGN: Cryo-EM Deep Reconstructing Generative Networks

cryoSPARC: Cryo-EM Single Particle Ab-Initio Reconstruction and Classification

Cryo-EM: Cryo Electron Microscopy

Grad-CAM: Gradient-weighted Class Activation Mapping

MLP: Multiple Layered Perceptron

MNIST: Modified National Institute of Standards and Technology database

PDX: Pyridoxal 5'-phosphate synthase

UMAP: Uniform Manifold Approximation and Projection

U-NET: U-shaped Network

VAE: Variational AutoEncoder

Contents

Abstract ii
Summary iii
Acknowledgments iv
Acronyms and Abbreviations v
1.0 Introduction A.1
2.0 Pixel Subtraction A.3
2.1 Background A.3
2.2 Data Set Preparation A.3
2.3 Method A.3
2.4 Result A.3
3.0 CNN without Residual Connection A.5
3.1 Background A.5
3.2 Method A.5
3.3 Result A.5
4.0 CNN with Residual Connection A.6
4.1 Method A.6
4.2 Result A.6
4.3 Interpretability A.6
4.4 Discussion A.7
5.0 VAE with *cryoDRGN* A.8
5.1 Background A.8
5.2 Result A.8
6.0 Beta-VAE A.10
7.0 VAE with Cartesian Coordinates A.11
7.1 Background A.11
7.2 Method A.11
7.3 Result A.11
7.4 Discussion A.11
8.0 VAE after Manual Masking A.12
8.1 Background A.12
8.2 Method A.13
8.3 Result A.13
9.0 VAE Based on *cryoDRGN* version 2 ..... A.14
9.1 Background A.14
9.2 Current Development Status A.14
10.0 VAE Based on U-NET A.15
10.1 Background A.15

10.2 Current Development Status A.15
11.0 Conclusion and Future Direction A.17
12.0 References A.18
Appendix A – Data availability A.19

Figures

Figure 1. **Heteromeric assembly mechanism of PDX.** A.1
Figure 2. **Representative native mass spectrometry spectrum for PDX co-expression complex 9:1, zoomed into the 12mer region.** A.2
Figure 3. **Superimposed and distinct regions between PDX1.2 and PDX1.3 homologs.** A.2
Figure 4. **Visualization of region of interest.** A.7
Figure 5. ***CryoDRGN* based classification cryo-EM data with MLP and beta=1.** A.8
Figure 6. **VAE classification cryo-EM data after modification of *cryoDRGN*.** A.10
Figure 7. **An example that our current GPU hardware cannot minimize loss quickly.** A.12
Figure 8. **Manual masking to leave region of difference only.** A.13
Figure 9. **Example of U-NET Based Segmentation.** A.15

Tables

Table 1. **Classification prediction with 50k simulated cryo-EM images of 4 PDX pseudo-enzyme/enzyme assembly states using mean squared error of averaged pixels per each homolog.** A.4
Table 2. **Resnet Based Classification** A.6
Table 3. **Comparison between Methods We Tried** A.17

1.0 Introduction

Single particle analysis of cryo electron microscopy (cryo-EM) has been advanced by many technical improvements in both equipment and computational modeling[1]. Especially, 2 dimensional (2D) and 3D image classification algorithms have empowered to extract meaningful heterogenous compositional and conformational analysis[2 3 4]. These breakthroughs have enabled to extract more realistic understanding of biological samples that are captured in vitreous ice. Especially, one of the recent deep learning based methods, *cryoDRGN* (Cryo-EM Deep Reconstructing Generative Networks), had pioneered heterogeneous cryo-EM reconstruction that models a continuous distribution over 3D structures by using a representation for the volume[5 6]. This algorithm is unique since it shows to model continuous and discrete heterogeneity that has not been easily achieved even with the state-of-the-art 3D reconstruction method. However, including *cryoDRGN,* most of these reconstruction methods have focused on to classify distinctly different biomolecule structures.

Therefore, it is imperative to develop a new computation method that can classify even closely resembling homologues. Here, we present our approaches to differentiate molecules that are distinguishable only with subtle inner morphological differences. These approaches include 'classification using pixel based difference', convolutional neural network (CNN) without and with residual connection, and variational autoencoder (VAE). Since our goal is to differentiate subtle structural differences with pixelated differences in 2D space, we first tested its viability by subtracting pixel differences. Specifically, we aim to classify different homologues that are assembled with different ratios (stoichiometries) of PDX1.2 (Pyridoxal 5'-phosphate synthase 1.2) and PDX1.3 monomers[7] (Fig. 1, 2, 3).

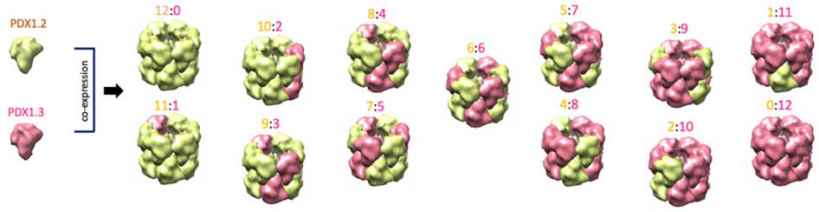

Figure 1. Heteromeric assembly mechanism of PDX.
(Left) monomer of each PDX homologues. (Right) assembled dodecamer after co-expression. Adapted from Novikova et al.[7].

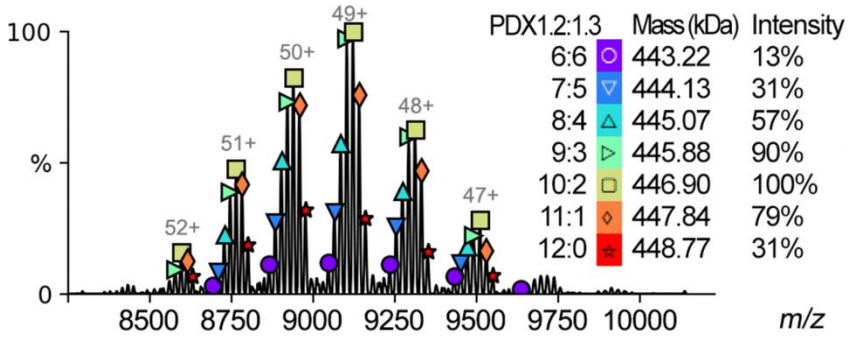

Figure 2. Representative native mass spectrometry spectrum for PDX co-expression complex 9:1, zoomed into the 12mer region.

Each symbol above the peak indicates one 12mer species, with their assignment, mass, and relative intensity shown on the right. Each peak with the same symbol is the same species carrying a different number of charges. Annotation was performed using *UniDec*. Adapted from Novikova et al.[7]

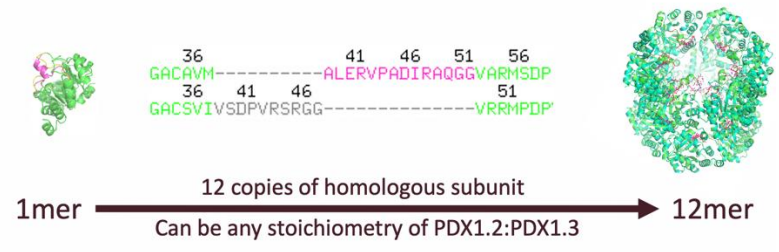

Figure 3. Superimposed and distinct regions between PDX1.2 and PDX1.3 homologs.

Green: indistinguishable superimposition of PDX1.2 and PDX1.3 homologs. Magenta: PDX1.3 specific local structure and sequence. Green & gray: PDX1.2 specific local structure and sequence

2.0 Pixel Subtraction

2.1 Background

To assess whether intact pixel value differences are sufficient to classify homologues, we tried to classify 4 sub-classes with subtracted pixel values only. Since the classification was performed per same Euler angle set (e.g. rot, tilt, and psi), rotation-based obfuscation was not tested. Since the goal of this method is to classify image classes, it may look like contrastive learning. However, contrastive learning tries to learn an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart[8].

2.2 Data Set Preparation

We hypothesized that a model trained with simulated cryo-EM maps can be applied to predict raw experimental cryo-EM map as others have shown[9] [10]. One common challenge with this approach is that raw experimental cryo-EM map has inherent noise that is not easily captured in these synthetic simulated maps. However, Yao et al. have shown that adding 8 different levels of additive white gaussian noise to each simulated map is effective to apply to experimental map later[9]. Similarly, we added different signal to noise ratios (SNR, e.g. 0.5, 0.25, 0.1, 0.05) to our simulated maps using *cryoSPARC*[2] that confers noise as gaussian with zero mean. Since averaged SNR of a cryoEM micrograph is estimated to be around 0.1[11] [12], we believe that our 0.05 SNR set would be enough to capture experimental SNR. Additionally, we used defocus values from 10,000 to 20,000 angstrom to mimic various defocuses which are often enforced to try to enhance contrast during cryo-EM experiments. Specifically, we generated synthetic cryo-EM 2D images of PDX (pyridoxal 5'-phosphate synthase) molecule. Since there are many projected 2D views in experimental cryo-EM data, we needed to simulate ample amount of cryo-EM map data sets for training and testing. Therefore, we prepared pdb structures of 4 sub-classes using *UCSF Chimera.* These sub-classes are PDX1.2 (all monomers are PDX1.2), PDX1.3 (all monomers are PDX1.3), hexagonal (either PDX1.2 or PDX1.3 hexagonal units are stacked with top and bottom assemblies), and alternate (PDX1.2 and 1.3 monomers are assembled in an alternative fashion). Then, we transformed these pdb structure files into mrc starting maps using *SPIDER*[13], and projected into 50,000 2D images in random orientations with various signal to noise ratios (SNR) (Fig. S1) using *cryoSPARC*[2].

2.3 Method

With this synthetic set, we made classification models using mean squared error of averaged pixels per sub-class.

2.4 Result

With higher signal to noise ratio (SNR, e.g. 0.25-0.5) simulated data set, we were able to achieve 93-99% accuracy (Table 1). This high accuracy proves that it is viable to classify target

molecules even with few pixel differences only if we provide Euler angle information. This is encouraging since human eye cannot differentiate these subtle differences. However, for low SNR data set (e.g. SNR=0.05-0.1), this pixel subtraction method achieved merely 28-38% accuracy (note: randomly predicted accuracy is 25% since there are 4 prediction classes). Therefore, we decided to use convolutional neural network (CNN) to exploit advantage of deep neural net.

Table 1. Classification Classification prediction with simulated cryo-EM images using mean squared error of averaged pixels per each homolog.

prediction with 50k simulated cryo-EM images of 4 PDX pseudo-enzyme/enzyme assembly states using mean squared error of averaged pixels per each homolog.

(Upper table) Summary. (Lower table) We summed each case to summarize prediction accuracy for SNR=0.5 case.

Signal-to-noise ratio	Validation set accuracy
0.5	99%
0.25	93%
0.1	38%
0.05	28%

	PDX1.2	PDX1.3	Hexagonal	Alternate
PDX1.2	10031	11	103	1
PDX1.3	120	10110	120	2
Hexagonal	98	11	10037	3
Alternate	7	7	5	10120

3.0 CNN without Residual Connection

3.1 Background

Our possible challenge can be to train models overcoming rotation variance issue. Si et al. have shown that 2D rotation variance obfuscation can be overcame and in fact adding four sets of rotated views (e.g. 0^o, 90^o, 180^o and 270^o) helps to expand training set reliably[10]. However, 3D rotation invariance (inherent features regardless of rot, tilt, psi Euler angles) will confer additional complexity when modeling. Other deep learning researchers argued that CNN can deal as if the data is rotation invariant. However, features freshly extracted from CNN are not scale or rotation invariant. Only max pooling layer introduces such invariants[14].

3.2 Method

We implemented conventional CNN architecture (2 layers each with Convolution, BatchNorm, Relu and MaxPooling operation) with Pytorch.

3.3 Result

We achieved much higher accuracy than pixel difference method. One notable lesson is that our CNN differentiates different stoichiometries catching rotation invariant features, because it performs well not only with focused data set with similar Euler angles, but also with comprehensive data set with diverse Euler angles even without explicit Euler angle information. This rotation invariant features are possibly caught due to MaxPooling operation by reducing input dimension.

4.0 CNN with Residual Connection

4.1 Method

With the same synthetic data set, we applied Residual Network (ResNet) architecture-based CNN since the ResNet (skip network) tends to better take advantage of deeper layers without vanishing gradient problem. Specifically, we employed the latest PyTorch version (1.6.0) in the GPU (graphic processing unit) cluster (*Marianas, Deception*) in Pacific Northwest National Laboratory. Starting code of ResNet comes from PyTorch tutorial[15].

4.2 Result

ResNet achieved much higher accuracy even without any Euler angle information than pixel-based difference classification and slightly higher accuracy than CNN without residual connection (Table 2).

Table 2. ResNet Based Classification

Training set accuracy for each SNR data set is either equal or slightly higher than test set accuracy.

Signal-to-noise ratio	Base model	Run time	Test set accuracy
0.5	ResNet50	23 hrs (35 epochs)	99%
0.25	ResNet50	18 hrs (28 epochs)	99%
0.1	ResNet18	14 hrs (53 epochs)	90%
0.05	Best accuracy model from SNR-0.1 ResNet training	7 hrs (27 epochs)	50%

4.3 Interpretability

Recent CNN model interpretability has been driven by gradient-based visual attention methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) and saliency map. To explain which local area was used to make classification, we visualized region of attention by applying Grad-CAM to our ResNet based classification model (Fig. 4).

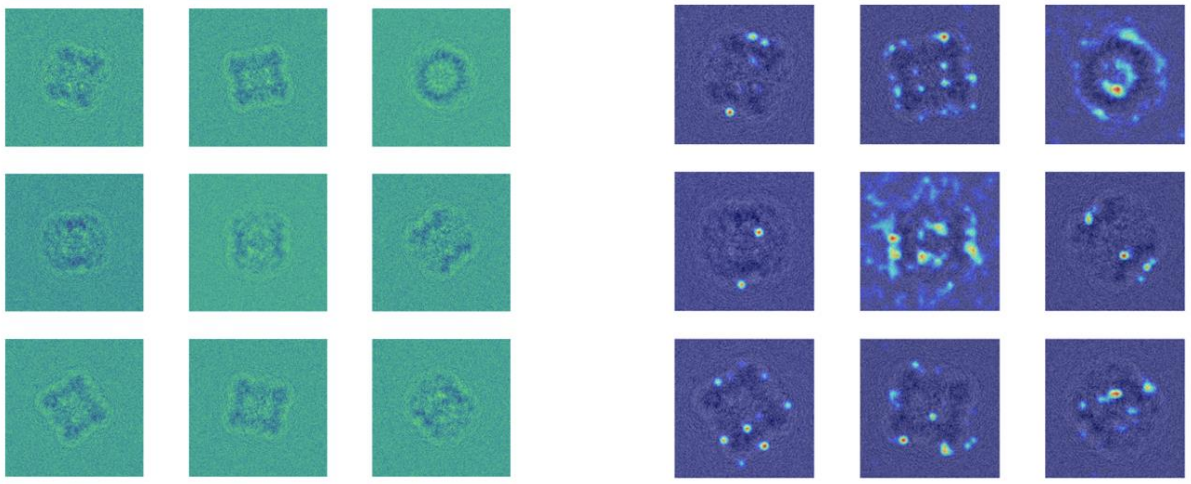

Figure 4. Visualization of region of interest.

(Left) original image of PDX molecules. (Right) Result of Grad-CAM (white dots are estimated region of interest).

4.4 Discussion

When we performed transfer learning from trained ResNet series with *Imagenet* data, it just slightly improved the accuracy over non-transfer learning. This result makes sense since our synthetic dataset is mostly similar to each other with respect to morphology (projected cylinder) and color (grayscale only) while *Imagenet* dataset is consisted with diverse shapes and colors.

Since random prediction accuracy would be 25% with our 4 target classes, it is evident that CNN can predict with decent prediction accuracy. However, we expect that SNR of experimental cryo-EM data is around 0.05~0.1. If we interpolate our naïve application of CNN model that is trained with simulated data, prediction of experimental cryo-EM data would be around 50~90% only (of course since experimental data has more limited Euler angle ranges, the accuracy might be little higher than this). Therefore, either we need to further optimize hyperparameters of CNN to further improve accuracy or we need to use a different deep learning method. More importantly, our ultimate goal is to develop unsupervised method. However, classification with pixel based difference and CNN without and with residual connection are all supervised method.

5.0 VAE with *cryoDRGN*

5.1 Background

We believed that regions of interest in our target (PDX multimers) will be better captured in 3D space. Therefore, we wanted to map between cryo-EM 2D input image and 3D output volume. Spatial-VAE is an ideal foundation architecture for this goal[16] because it is rotation, translation equivariant. In other words, it can model rotation, translation invariant features (e.g. inherent features regardless of random rotation and translation). Starting from the spatial-VAE, *cryoDRGN* reconstructs 3D volume even with 2D projected images as input[5]. This is possible by matching with known Euler angles from another input file (e.g. star) which is generated by homogenous refinement.

Strictly speaking, *cryoDRGN* is not transformer since it lacks any attention model. However, it uses positional encoding to all input coordinate pixels. This positional encoding information is concatenated into latent space when it is fed into decoder. Since this positional encoder is added to all input tokens, decoder knows input token order (as transformer model does). Therefore, this coordinate MLP (or Neural Radiance Field[17]) can reconstruct 3D volumes. Since latent space maps between matching 3D volumes with 2D inputs, latent space clustering of *cryoDRGN* can differentiate obviously different biomolecule structures.

5.2 Result

When we tried the leading VAE method of cryo-EM reconstruction, *cryoDRGN*[5], to classify published obviously different molecules, it classified them well (Fig. 5 Left). However, when we ran it again to our target experimental images (e.g. 284,133 particles of PDXcoexpression) that has various stoichiometries of PDX homologues (Fig. 2), we were unable to classify even after excessive number of epochs (e.g. > 2,400) (Fig. 5 right. A typical number of epoch trials by *cryoDRGN* author is ~25). One notable fact is that we tried fairly exhaustive combinations of UMAP (Uniform Manifold Approximation and Projection) hyperparameters (e.g. metric, n_neighbors, min_dist, and spread) as well. Although we did not vary different a, b hyperparameters due to computational hardware limitation, these fairly exhaustive combinations of UMAP hyperparameters ensure that this lack of classification power does not stem from less ideal set of UMAP hyperparameters. As a reference, *cryoDRGN* application to classify distinctly different protein structures, default usage of UMAP hyperparameters has been sufficient[5].

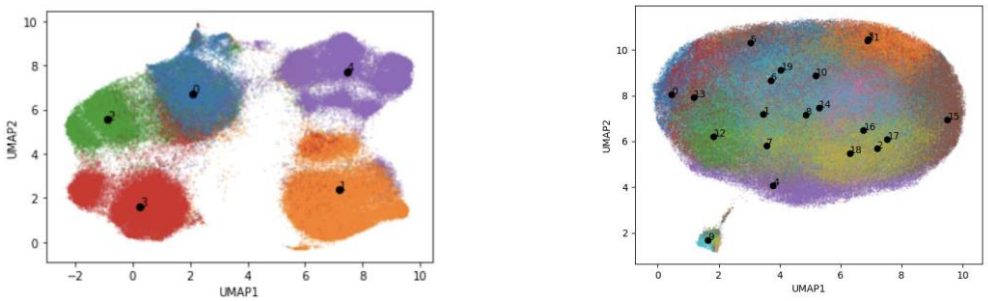

Figure 5. *CryoDRGN* based classification cryo-EM data with MLP and beta=1.

(Left) It clearly clusters ribosome 50S data. (Right) We could not classify PDX coexpression data set whose homologues are similar to each other except few internal differences among them.

6.0 Beta-VAE

Since *cryoDRGN* uses beta coefficient for KL (Kullback–Leibler) divergence = 1 by default, we tried various beta values that are greater than 1 to realize betaVAE[18]. The purpose of betaVAE is to better disentangle latent features even if it sacrifices some reconstruction quality as shown with MNIST (Modified National Institute of Standards and Technology database) and human face examples. This newer VAE seemed worth a try since our *cryoDRGN* trials reconstructed expected dodecamer structures, we may be able to forfeit some reconstruction loss minimization. However, even with the betaVAE (e.g. beta=2, 4), we still could not classify closely resembling homologues (Fig. 6 left).

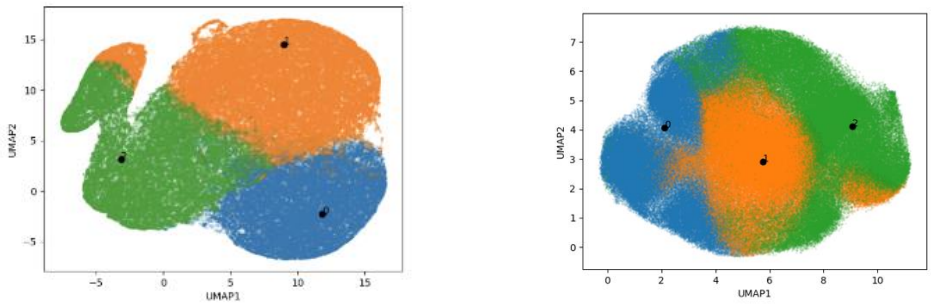

Figure 6. VAE classification cryo-EM data after modification of *cryoDRGN*.

(Left) beta-VAE with *cryoDRGN* (vae128_z8_e1600_beta_4_amp/analyze.1599/kmeans3).

(Right) *cryoDRGN* with CNN encoding (vae128_z10_e100_conv/analyze.75/kmeans3).

7.0 VAE with Cartesian Coordinates

7.1 Background

For faster computation, most cryo-EM programs (including *cryoDRGN*) use coordinates in fourier space during the most intensive parts (e.g. pose search, refinement). Of course, the very first input and very last outputs are in real space so that human can understand more intuitively. We hypothesized that if we feed real space input into encoder, then the VAE may better differentiate incoming image data preserving same differences among different stoichiometries that may be lost/obscured in fourier space. Eventually, processing data in real space will be better interpreted by human anyway.

7.2 Method

We modified *cryoDRGN* to use real space encoder with two options. One is to use original MLP (MultiLayer Perceptron) encoder. The other is to use CNN encoder. For CNN encoder, since we observed high prediction accuracy with ResNet based CNN, we coded residual connection based VAE. For decoding, we used the same fourier transform MLP version (e.g. FTPositionalDecoder). With this new CNN encoding model, trainable parameters in model are increased to 4 million (4,790,422) from 3 million (3,790,354) in MLP. Typically, the number of parameters of MLP should be much higher than the one for CNN since MLP uses fully connected layer while CNN abstracts parameters with Pooling (often Maxpool) layers. The possible reason behind this increased number of parameters with CNN could be that we used a larger architecture of CNN to use residual attention layer that will be visualized by Grad-CAM to highlight crucial local regions that are more responsible for classification. Otherwise, MLP in the original *cryoDRGN* may not end up using fully connected layer after all. Possibly due to increased number of trainable parameters, our CNN encoding model ran 10 times slower than MLP encoding model.

7.3 Result

Usage of real space encoder still could not classify closely resembling homologues both with MLP and CNN encoders (Fig. 6 right). It is notable that even larger number of parameters and various trials of z dimensions (for latent space) and epoch numbers for CNN encoder did not help to differentiate fine details among homologues.

7.4 Discussion

High numbers of trainable parameters of this new CNN encoding VAE may seem large. However, *ResNet18* and *ResNet50* that we used have even more number of parameters (e.g. 11 and 23 million respectively). Furthermore, the number of trainable parameters of our 2D only input and output InfoVAE models are 134 million. Therefore, it is evident that simple increase of number of parameters does not guarantee to classify subtle structural differences. What is more often critical is whether the model captures essential properties. Indeed, *Deepmind* reported that 25 times fewer parameters of *Retrieval-Enhanced Transformer (RETRO)* achieved similar performance of GPT-3 (that used 178 billion parameters)[19].

8.0 VAE after Manual Masking

8.1 Background

Other than simple subtraction of pixel values among sub-classes, all our approaches are deep learning based (e.g. deep neural network). We suspected that our unsupervised deep learning approach lacks resolving power since supervised deep learning approaches successfully classified subtle differences among homologues. Additional resolving burden with unsupervised learning may have complicated pixel resolution in latent space. However, our computing power is limited to try very deep layers (Fig. 7). Therefore, we wanted to check whether manually masked region of interest (differing parts between different stoichiometries) alone can be captured by our current VAE architecture.

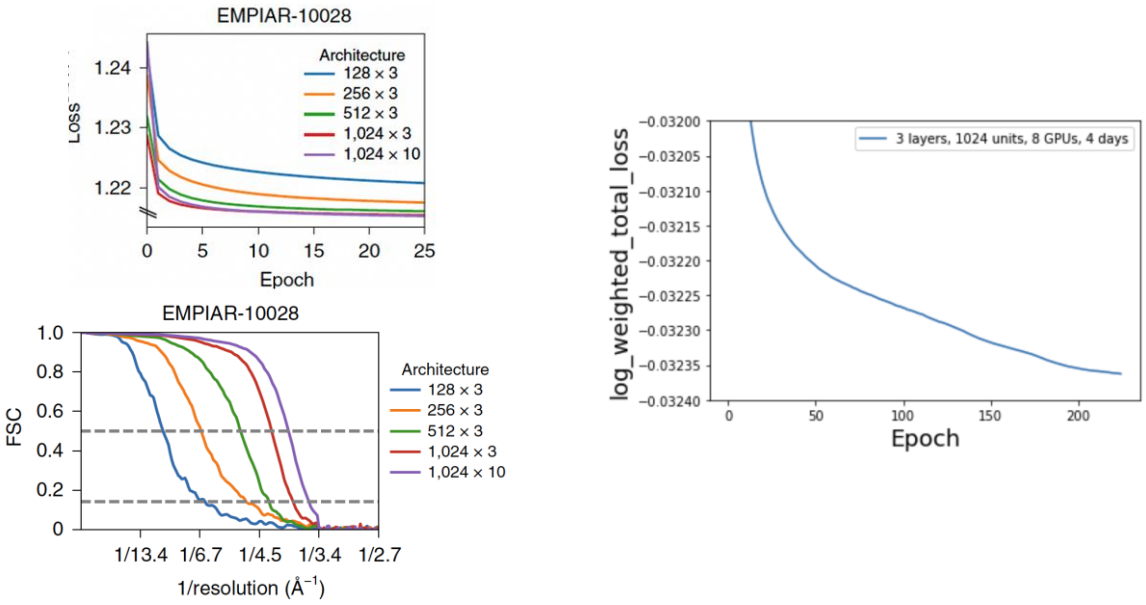

Figure 7. An example that our current GPU hardware cannot minimize loss quickly.

(Left) Better loss minimization of *cryoDRGN* enables higher/better resolution reconstruction (adapted from Zhong et al.[5]).

(Right) However, even 8 GPUs are not reaching optimization plateau within 4 days with merely 64x64 images for 3 layers and 1,024 hidden units.

8.2 Method

After aligning corresponding pdb files, we converted them into cryo-EM maps with *molmap* command (*pdb2mrc* by *eman2*) in *UCSF Chimera*. Then, we subtracted these maps by *vop subtract* command in *UCSF Chimera* (Fig. 8). These subtracted 3D maps are projected into 2D images by *relion_project*.

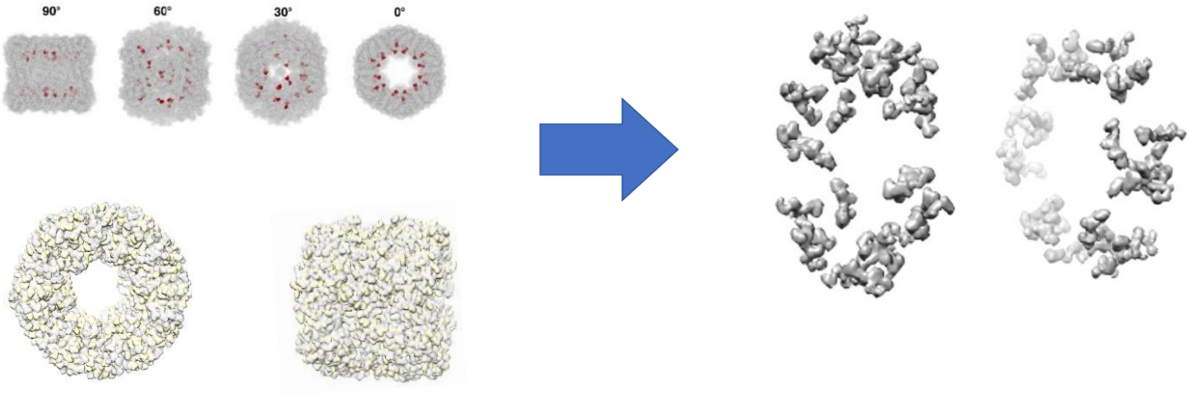

Figure 8. Manual masking to leave region of difference only.

(Upper Left) Red dots are 'region of interest' (e.g. difference between two different homologous structures.

(Lower Left) gray → PDX1.2 homologue, yellow → PDX1.3 homologue

(Right) Masked region of interest after hiding dust.

8.3 Result

Even with these masked input images (e.g. superimposed regions between two target homologues are omitted/discarded), unsupervised VAE could not classify expected stoichiometries still (e.g. it classified targets randomly). This result was unexpected since resolving power by unsupervised method should be solely devoted to region of interest only this time (and these same regions of interests were well classified by supervised CNN). This result suggests us that latent space information that is supposed to extract essential information (e.g. without noise/tangential structural information) is generated with middle/low resolution information only.

9.0 VAE Based on *cryoDRGN* version 2

9.1 Background

Based on classification failure even with manually masked input, we came to believe that our VAE models that are based on *cryoDRGN ver. 1* could not deal high resolution structural information (without very time-consuming deep learning architectures). Indeed, a recent report of *cryoDRGN ver. 2*[20] shows that *cryoDRGN ver. 1* (Branch and Bound pose search, e.g. BNB) based VAE could not reconstruct into high resolution map. Additionally, *cryoDRGN ver. 1* based volume model (MLP / NeRF: Neural Radiance Fields) is much slower to render than voxel-based models.

However, *cryoDRGN ver. 2* could search pose 172 times more accurately 2~4 times faster. This boost of performance of *ver. 2* becomes possible by refactoring the pose search (e.g. enable tractable joint inference of pose and volumes in the context of an MLP representation of volume). Specifically, there are two improvements. First, it alternates epochs between pose search and reusing the latest computed pose (instead of pose search epochs only). Second, it resets the coordinate MLP model and optimizer state intermittently. This resetting coordinate may seem like a conventional residual connection as seen in skip connection of ResNet. However, it is not residual connection. Other than these improvements, *cryoDRGN ver. 2* is fundamentally same as *cryoDRGN ver. 1* (e.g. same positional encoding for 3D volume reconstruction).

9.2 Current Development Status

As of 12/27/2021, *cryoDRGN* ver. 2 author announced that she will release the working code after 1/1/2022. As the code is released, we plan to use it with 4 combinations (*cryoDRGN* ver. 2 original, *cryoDRGN* ver. 2 + InfoVAE, *cryoDRGN* ver. 2 + BetaVAE and *cryoDRGN* ver. 2 + InfoVAE + BetaVAE).

10.0 VAE Based on U-NET

10.1 Background

All deep learning based image classifications by big techs (including Nvidia, Facebook) have been focused on large differences between targets (such as Cifar100, Imagenet, Google's own proprietary image set). However, we need a different deep learning approach to deal with high resolution information (Fig. 3). We believe that U-NET (U-shaped network) architecture is ideal for this purpose (Fig. 9). Specifically, by replacing pooling operators with upsampling ones, the U-NET does not lose spatial information (not compressing image size). Additional benefit of U-NET is that it requires few images/classes to train (few-shot learning). This may be particularly useful since cryo-EM experimental 2D images may not be enough.

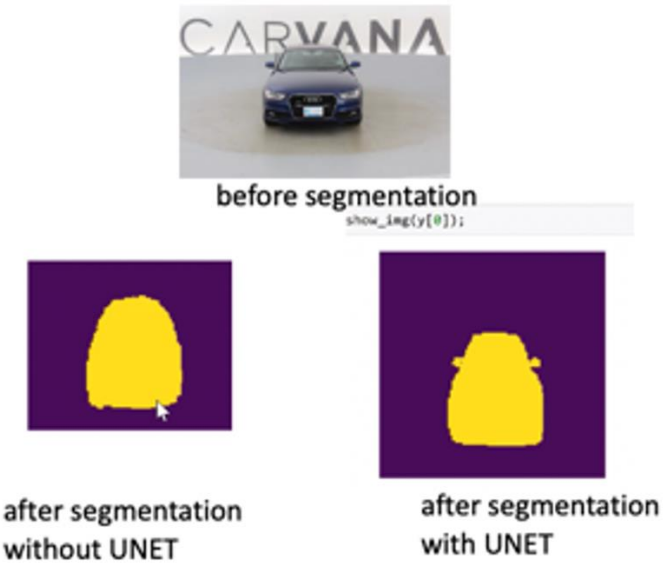

Figure 9. Example of U-NET Based Segmentation.

U-NET can segment up to high resolution information. Adapted from U-NET tutorial.

10.2 Current Development Status

As there are many successful U-NET applications to MLP and CNN only deep learning architecture, there are U-NET based VAE model as well both in 2D[21] and 3D[22] representation. However, unlike concatenating VAE output into regular U-NET bottleneck[23] (like some VAE-GAN models), we are modifying our current 2D input, 3D output VAE[5] to have U-NET based skip connection instead of vanilla MLP. Specifically, we identified that resizing of concatenated z latent space into oriented image coordinates has been arbitrarily done. These image coordinates are encapsulated in sine and cosine of input 2D images in Fourier space which is already multiplied

by "geom_lowf" frequency. For example, torch.size[3, 192] (coordinates dimension and dimension of input mrcs) are resized by ‘sum of all encoded mask (e.g. circularly masked input that are eventually fed into VAE) – z space dimension’. Most U-NET codes have shown that in practice encoder layer concatenation into decoder is done after arbitrary cropping of encoder just to fit dimensions of corresponding decoder layer. Therefore, we are adding coordinates of encoder layer into decoder (without consideration of perfect symmetry between encoder and decoder) which will be halved later so that existing *cryoDRGN* based 3D volume reconstruction will work without further code update. Since cryo-EM data does not fit into single GPU memory, current *cryoDRGN* code uses generator. Per each batch of this generator based input feeding, we are deciding which encoder layer to concatenate into decoder in this generator scheme.

11.0 Conclusion and Future Direction

We were able to affirm that supervised deep learning approach classifies even inner structural differences between homologous structures much better than simple pixel values-based classification does. However, we have not been able to classify same targets with unsupervised deep learning approaches even after different architectures of VAE and exhaustive searches of VAE and UMAP hyperparameters (Table 3). It turns out dealing higher resolution structural information is essential for our goal based on VAE result with manually masked image. To model higher resolution structural information, *Siamese* neural net that is specialized to identify tiny difference among input images seems worth a try[24]. Instead of popular cross-entropy, the Siamese neural net uses contrastive loss (e.g. contrastive learning)[8 25]. Its pretraining property enables one shot learning and fast application to external dataset as well. Of course, U-NET based VAE, and *cryoDRGN* 2 based InfoVAE are expected to model higher resolution information as well. For faster development of these updates, *PytorchLightning*[26] and *SimpleTransformer*[27] that wrap long lines of PyTorch and Transformer architecture will save time.

Table 3. Comparison between Methods We Tried

Method	Pros	Cons	Lesson
Pixel difference	Training logic is better understood by human	Less accurate than deep learning approach	Identifying inner structural differences is feasible by pixel density difference identification with high SNR data
CNN	Once trained, deployment to test set is fast	Pre-labelled data is required (supervised)	Identifying inner structural differences is feasible by deep learning architecture even without Euler angle information
VAE	Pre-labelled data is not required (unsupervised)	Development of successful classification tends to be more challenging than supervised approaches	Conventional VAE architecture cannot differentiate tiny pixel differences among sub-classes

12.0 References

1. Kim, D. N., Gront, D. & Sanbonmatsu, K. Practical considerations for atomistic structure modeling with Cryo-EM maps. *J. Chem. Inf. Model.* **60**, 2436–2442 (2020).

2. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

3. Fernandez-leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. Sect. D Struct. Biol.* **73**, 1–14 (2017).

4. Ludtke, S. J. Single-Particle Refinement and Variability Analysis in EMAN2.1. *Methods Enzymol.* **579**, 159–189 (2016).

5. 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).

6. Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, B. B. Reconstructing continuous distributions of 3D protein structure from cryo-EM images. *arxiv.org* (2019).

7. Novikova, I. V *et al.* Tunable Heteroassembly of a Plant Pseudoenzyme–Enzyme Complex. *ACS Chem. Biol.* **16**, 2315–2325 (2021).

8. Koch, G. R. Siamese Neural Networks for One-Shot Image Recognition. in (Proceedings of the 32 nd International Conference on Machine Learning, 2015).

9. Yao, R., Qian, J. & Huang, Q. Deep-learning with synthetic data enables automated picking of cryo-EM particle images of biological macromolecules. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz728.

10. Si, D. *et al.* Deep Learning to Predict Protein Backbone Structure from High-Resolution Cryo-EM Density Maps. *Sci. Rep.* **10**, 4282 (2020).

11. Baxter, W. T., Grassucci, R. A., Gao, H. & Frank, J. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. *J. Struct. Biol.* **166**, 126–132 (2009).

12. Bepler, T., Kelley, K., Noble, A. J. & Berger, B. Topaz-Denoise: general deep denoising models for cryoEM and cryoET. *Nat. Commun.* **11**, 5208 (2020).

13. Shaikh, T. R. *et al.* SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nat. Protoc.* **3**, 1941–1974 (2008).

14. About CNN, kernels and scale/rotation invariance. https://stats.stackexchange.com/questions/239076/about-cnn-kernels-and-scale-rotation-invariance (2016).

15. TRANSFER LEARNING FOR COMPUTER VISION TUTORIAL. https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html.

16. Bepler, T., Zhong, E., Kelley, K., Brignole, E. & Berger, B. Explicitly disentangling image content from translation and rotation with spatial-VAE. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).

17. Xie, Y. *et al.* Neural Fields in Visual Computing and Beyond. (2021).

18. Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, A. L. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR 2017 Conf.* (2016).

19. Borgeaud, S. *et al.* Improving language models by retrieving from trillions of tokens. (2021).

20. Zhong, E. D., Lerer, A., Davis, J. H. & Berger, B. CryoDRGN2: Ab Initio Neural Reconstruction of 3D Protein Structures From Real Cryo-EM Images. in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 4066–4075 (2021).

21. Kohl, S. A. A. *et al.* A Probabilistic U-Net for Segmentation of Ambiguous Images. *CoRR* **abs/1806.0**, (2018).

22. Nikolas Adaloglou. 3D Medical image segmentation with transformers tutorial. https://theaisummer.com/medical-segmentation-transformers/ (2021).

23. Esser, P., Sutter, E. & Ommer, B. A Variational U-Net for Conditional Appearance and Shape Generation. (2018).

24. Dhillon, I. Siamese Neural Nets and Contrastive Losses. https://www.youtube.com/watch?v=2t8dOv0RzpA&ab_channel=InderbirDhillon (2021).

25. Khac, P. H. L., Healy, G. & Smeaton, A. F. Contrastive Representation Learning: A Framework and Review. *IEEE Access* **8**, 193907–193934 (2020).

26. pytorchlightning. https://www.pytorchlightning.ai/.

27. simpletransformers. https://simpletransformers.ai/.

Appendix A – Data availability

All codes used for this project are available at https://gitlab.pnnl.gov/kimd999/mars_cryo. Contact authors for datasets.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov