THE UNIVERSITY of TENNESSEE KNOXVILLE

Variational Autoencoders

Sergei V. Kalinin

The VAE Story

- What are (Variational) autoencoders?
  - Encoding and decoding
  - Latent distribution
  - Latent representations
  - Disentanglement of the representations

- Why invariances: rotational, translational, and shear

- Other colors of VAEs:
  - Semi-supervised
  - Conditional
  - Joint

- VAEs for real-world examples
- From VAEs to encoder-decoders (VED)
- BO in the latent space

- Active learning: DKL

Autoencoders: Encoding

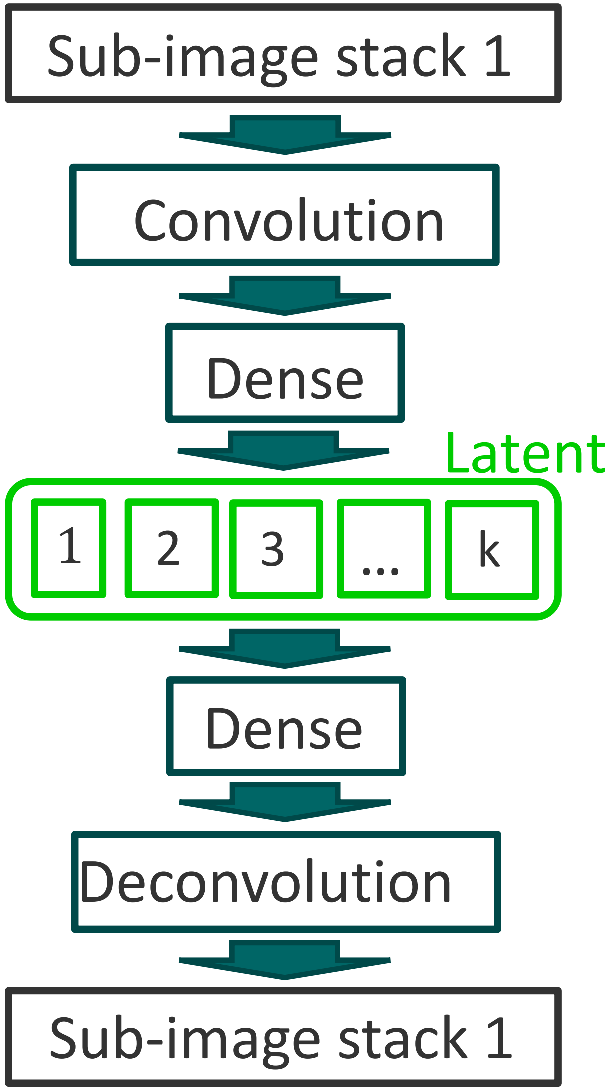
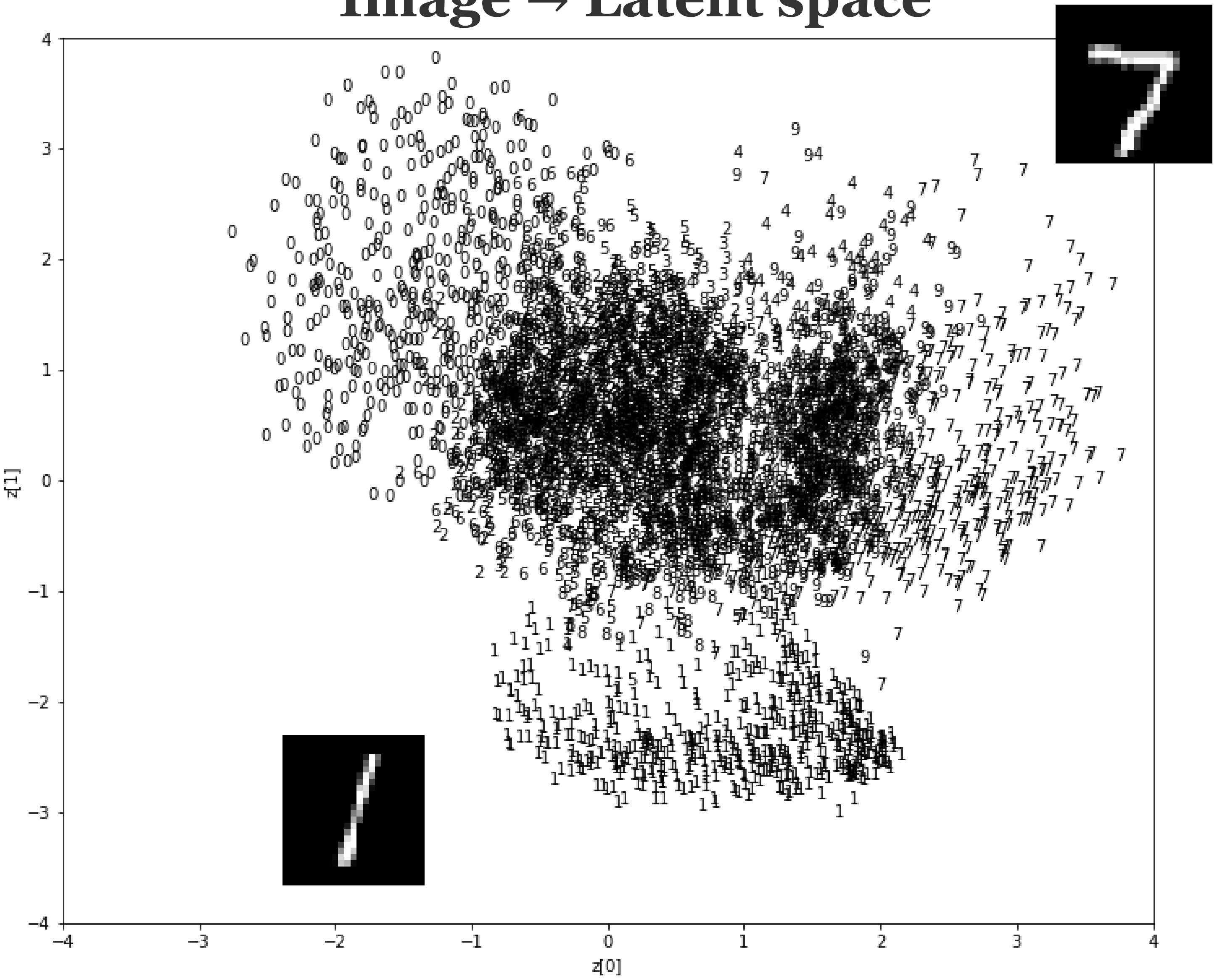
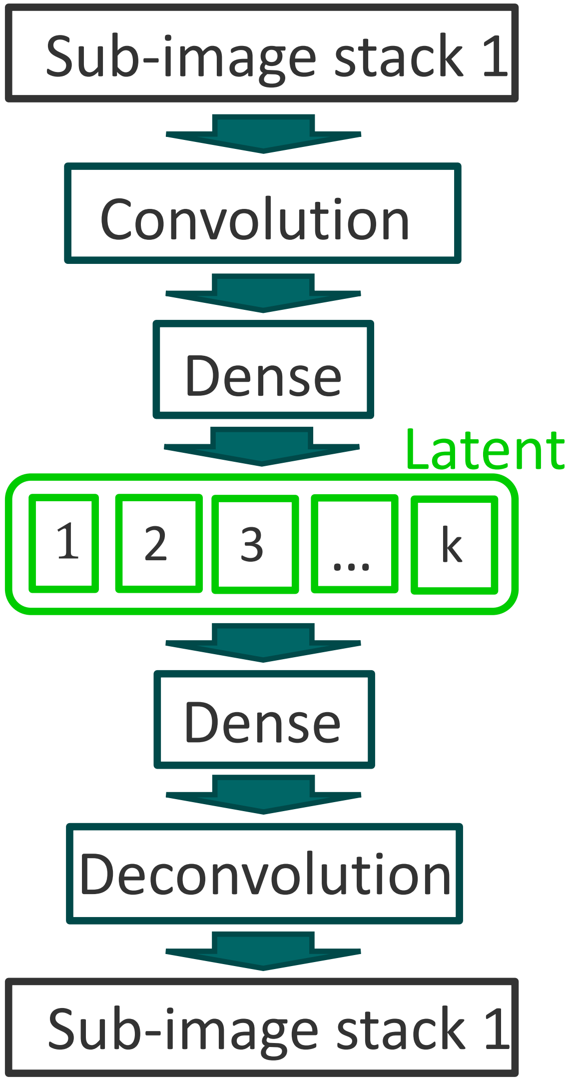


Image → Latent space

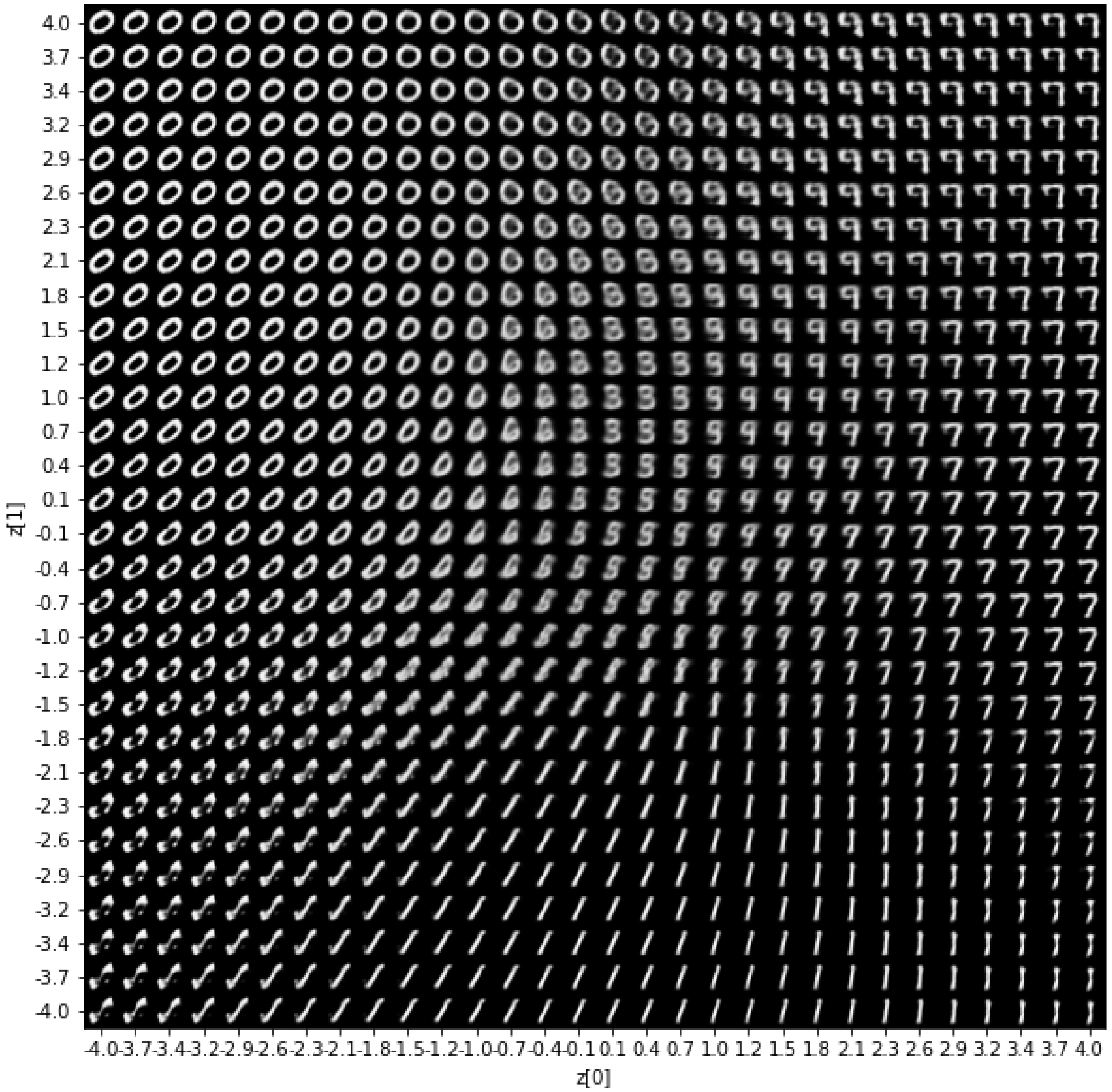


Latent distribution: Encoding the data via low dimensional vector

Autoencoders: Decoding



Latent space → Image



Latent representation: Decoding images from uniform grid in latent space

AE and Variational AE (VAE)

Geoffrey Hinton

FOLLOW

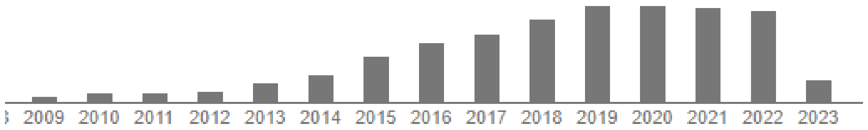
Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google
Verified email at cs.toronto.edu - Homepage

machine learning psychology artificial intelligence cognitive science computer science

TITLE	CITED BY	YEAR
Imagenet classification with deep convolutional neural networks A Krizhevsky, I Sutskever, GE Hinton Communications of the ACM 60 (6), 84-90	130318	2017
Deep learning Y LeCun, Y Bengio, G Hinton Nature 521 (7553), 436-44	62790	2015
Dropout: a simple way to prevent neural networks from overfitting N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov The journal of machine learning research 15 (1), 1929-1958	42078	2014
Visualizing data using t-SNE L van der Maaten, G Hinton Journal of Machine Learning Research 9 (Nov), 2579-2605	35035	2008
Learning representations by back-propagating errors DE Rumelhart, GE Hinton, RJ Williams Nature 323 (6088), 533-536	32239	1986
Learning internal representations by error-propagation DE Rumelhart, GE Hinton, RJ Williams Parallel Distributed Processing: Explorations in the Microstructure of ...	30711	1986
Schemata and sequential thought processes in PDP models. D Rumelhart, P Smolensky, J McClelland, G Hinton Parallel distributed processing: Explorations in the microstructure of ...	28073 *	1986
Learning multiple layers of features from tiny images A Krizhevsky, G Hinton	21876	2009
Rectified linear units improve restricted boltzmann machines V Nair, GE Hinton Proceedings of the 27th international conference on machine learning (ICML ...	21050	2010
Reducing the dimensionality of data with neural networks GE Hinton, RR Salakhutdinov Science 313 (5786), 504-507	19930	2006

Reducing the dimensionality of data with neural networks

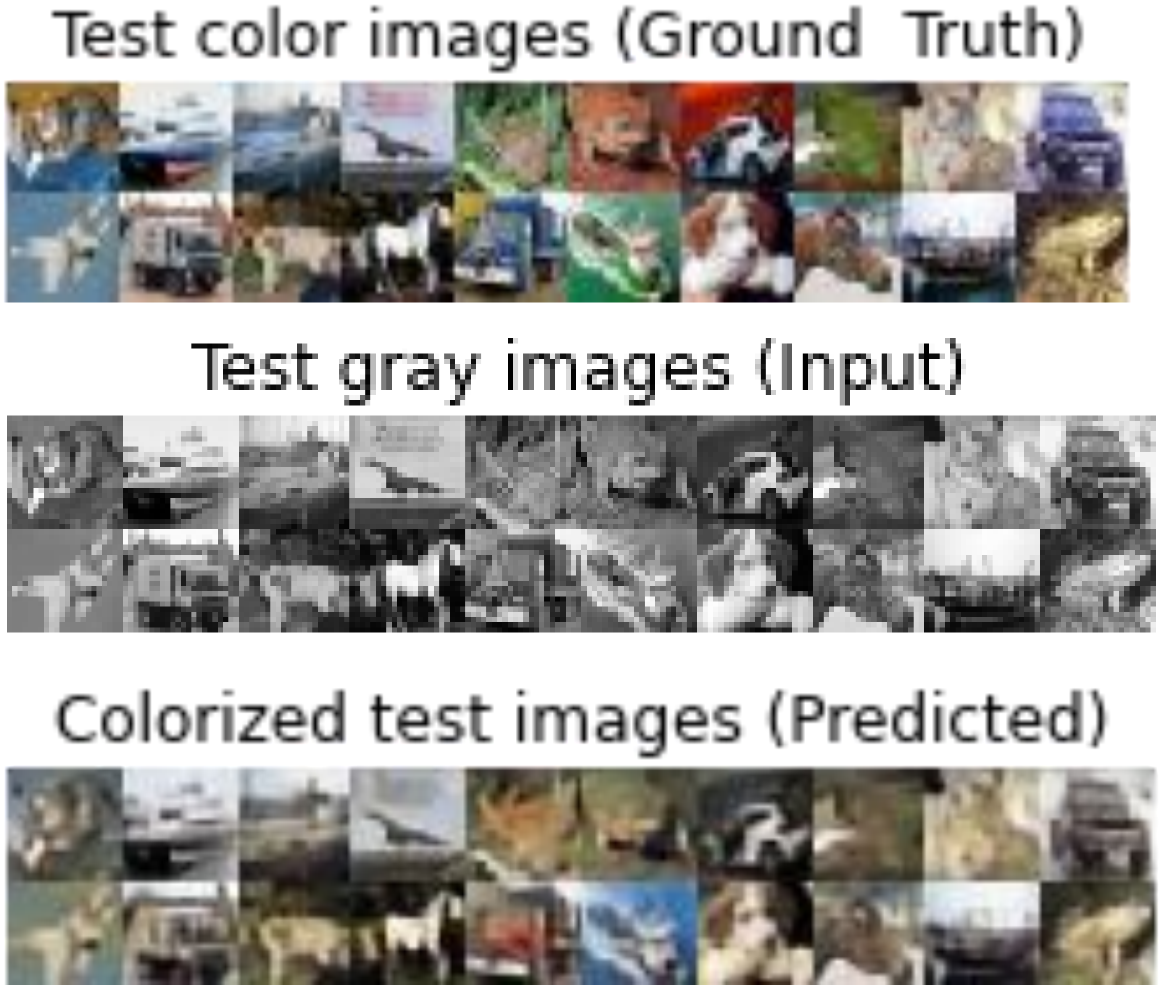
Authors	Geoffrey E Hinton, Ruslan R Salakhutdinov
Publication date	2006/7/28
Journal	Science
Volume	313
Issue	5786
Pages	504-507
Publisher	American Association for the Advancement of Science
Description	High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such “autoencoder” networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.
Total citations	Cited by 19930



Denoising

- **Training:** pairs of the high-noise and low-noise images
- **Application:** new high noise images (from the same distribution)
- **Concern:** has to be from the same distribution

Restoration



- **Training:** pairs of the grayscale and color images
- **Application:** new grayscale images (from the same distribution)
- **Concern:** has to be from the same distribution

AE and Variational AE (VAE)

Diederik P. Kingma

Other names

Research Scientist, Google Brain
Verified email at google.com - Homepage

Machine Learning Deep Learning Neural Networks Generative Models Variational Inference

FOLLOW

TITLE	CITED BY	YEAR
Adam: A Method for Stochastic Optimization DP Kingma, J Ba Proceedings of the 3rd International Conference on Learning Representations ...	141306	2014
Auto-Encoding Variational Bayes DP Kingma, M Welling arXiv preprint arXiv:1312.6114	26540	2013
Semi-Supervised Learning with Deep Generative Models DP Kingma, S Mohamed, DJ Rezende, M Welling Advances in Neural Information Processing Systems, 3581-3589	2946	2014

- Variational Autoencoder (VAE): uses "reparameterization trick" to sample from the latent space
- Can be used for same tasks as AE
- Have a much better-behaved latent space: disentanglement of the representations

VAE Training

Latent manifold -> Image space

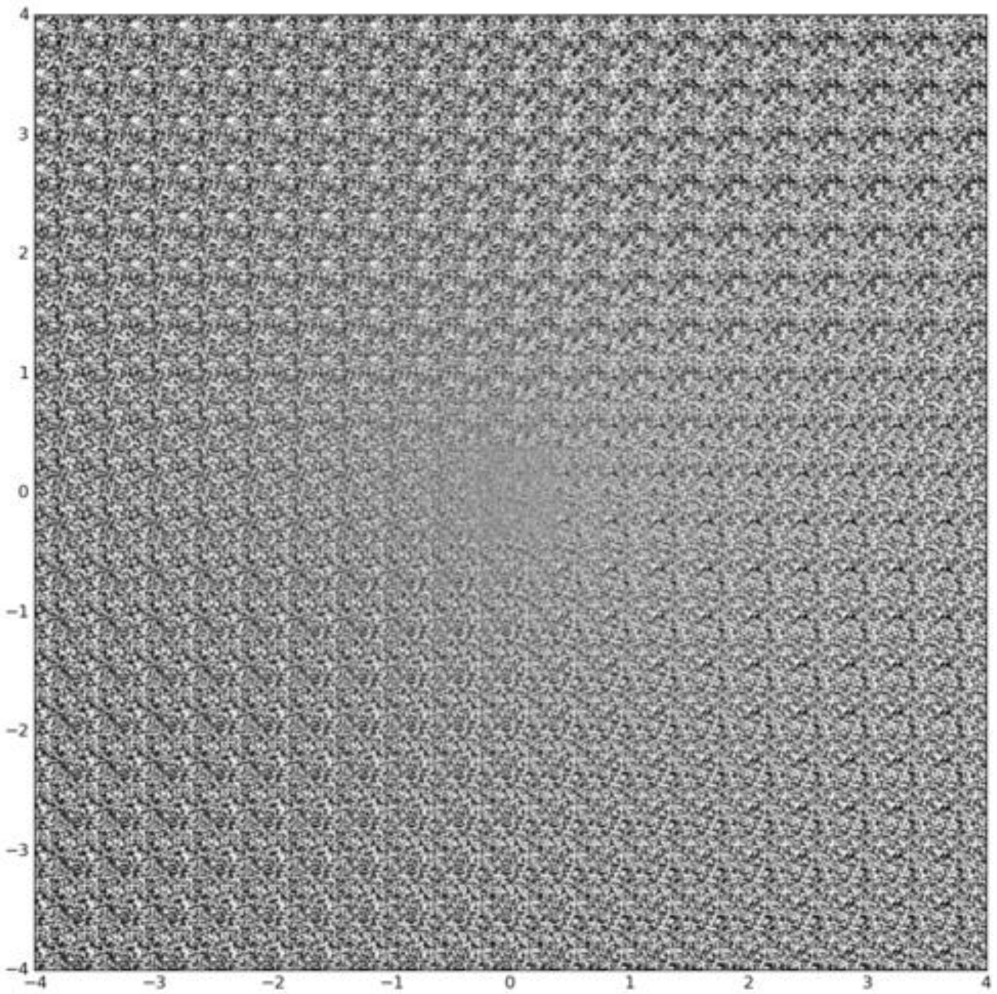
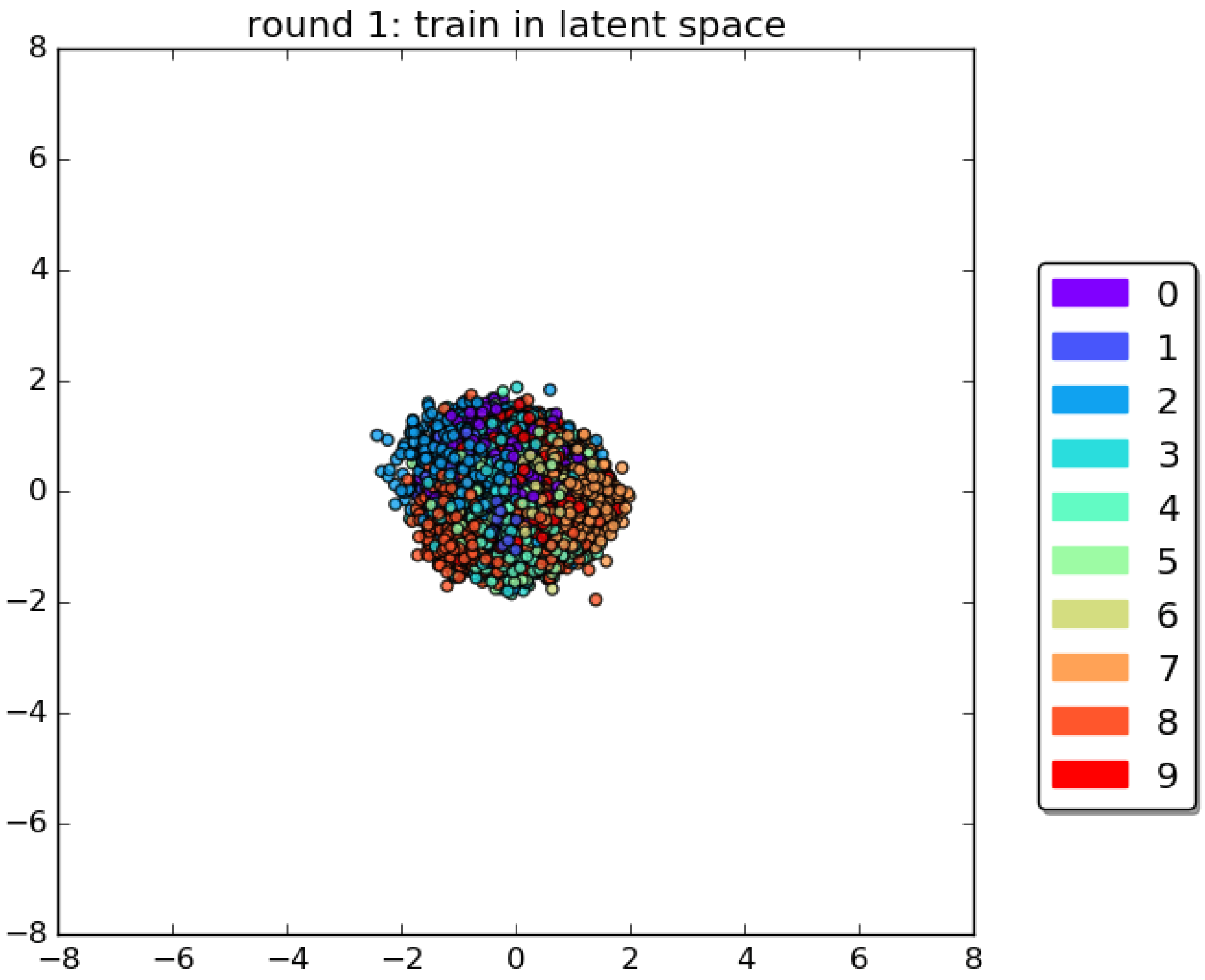
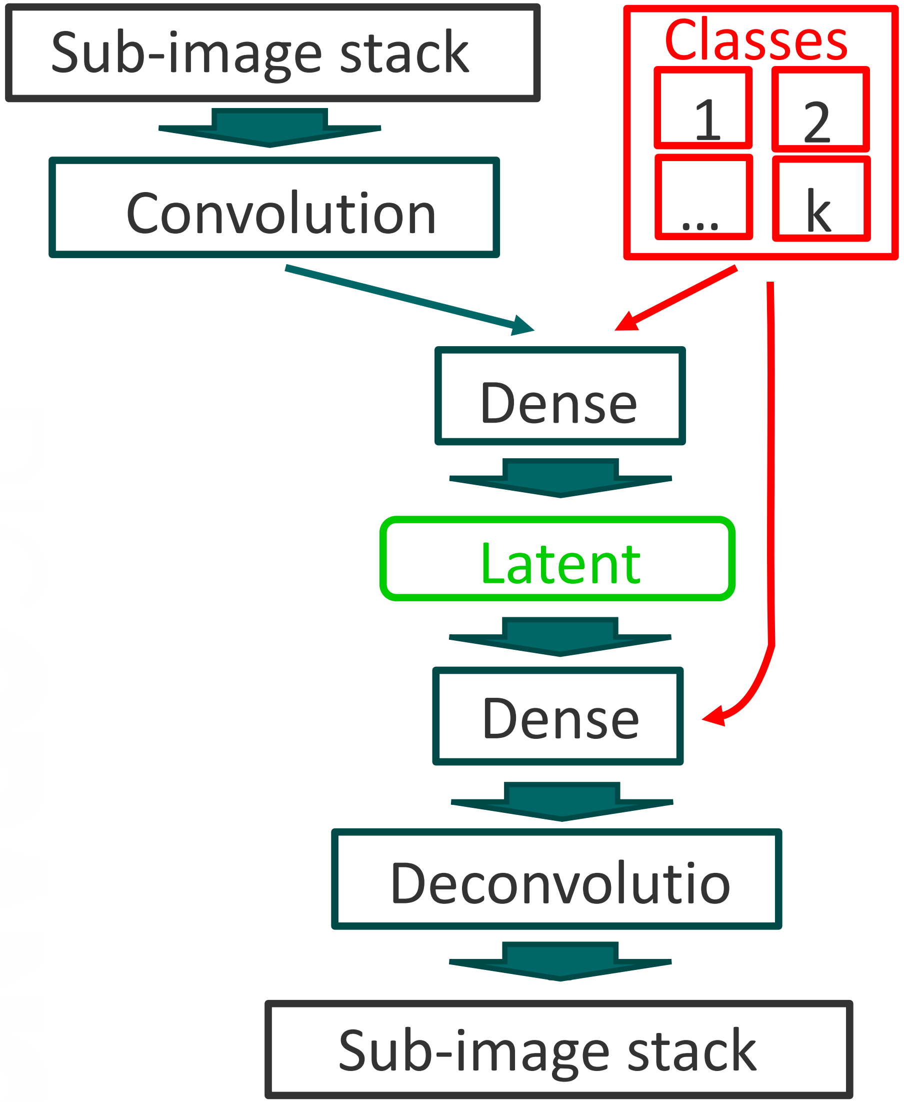
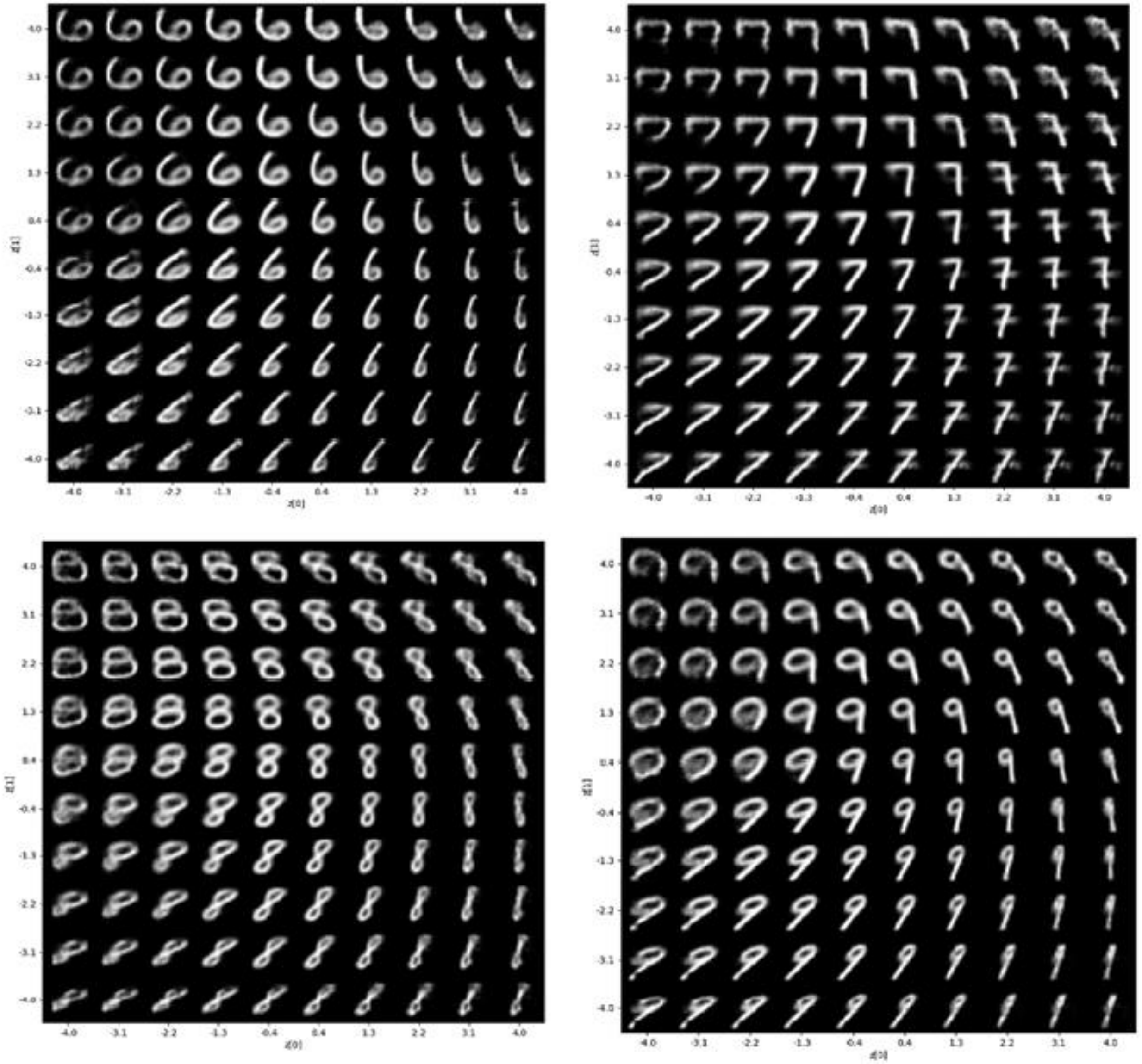


Image space -> Latent space



Conditional VAE

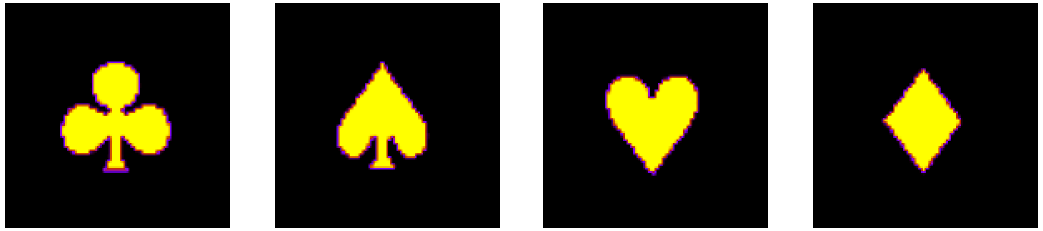




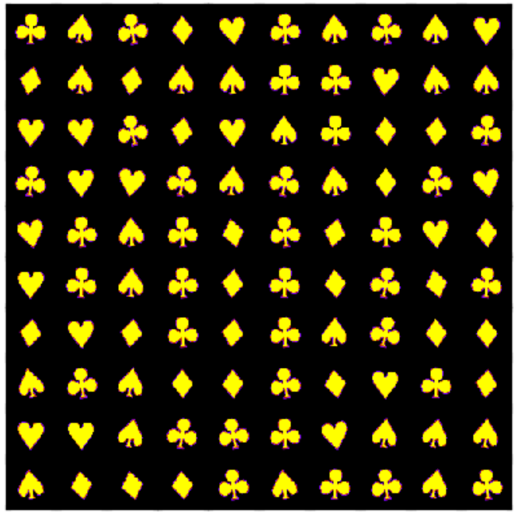
Note the trends in the latent representation for each digit: disentanglement of the representations

(R)VAE on Cards

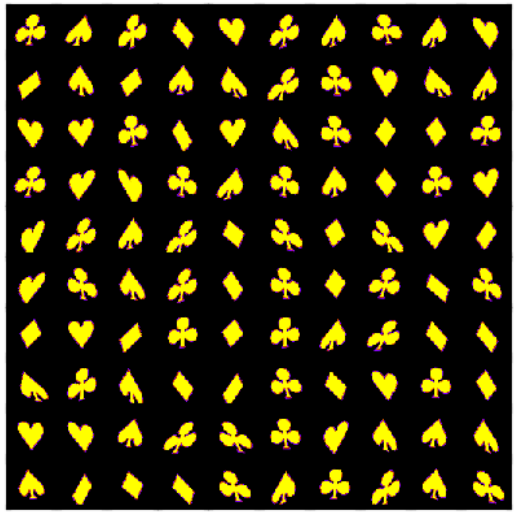
Introduce the **cards** data set:
- Classical 4 hands (diamonds, clubs, pikes, hearts)
- Interesting similarities (pikes and hearts)
- And invariances on affine transforms (e.g. diamonds)



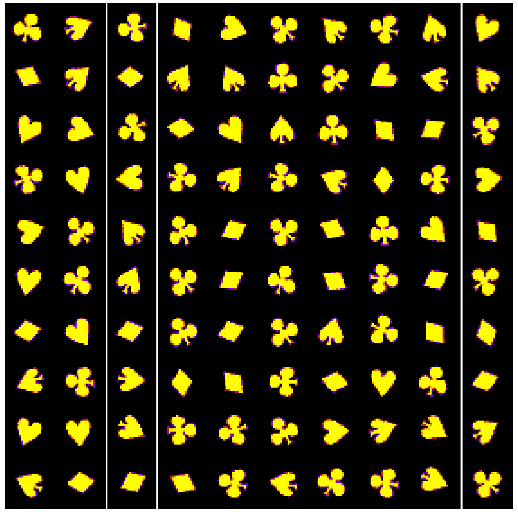
Cards 1: Low R (12 deg) and low S (1 deg)



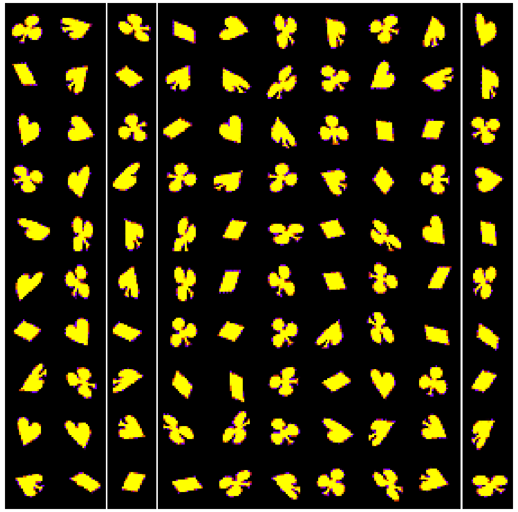
Cards 2: Low R (12 deg) and high S (20 deg)



Cards 3: High R (120 deg) and Low S (1 deg)



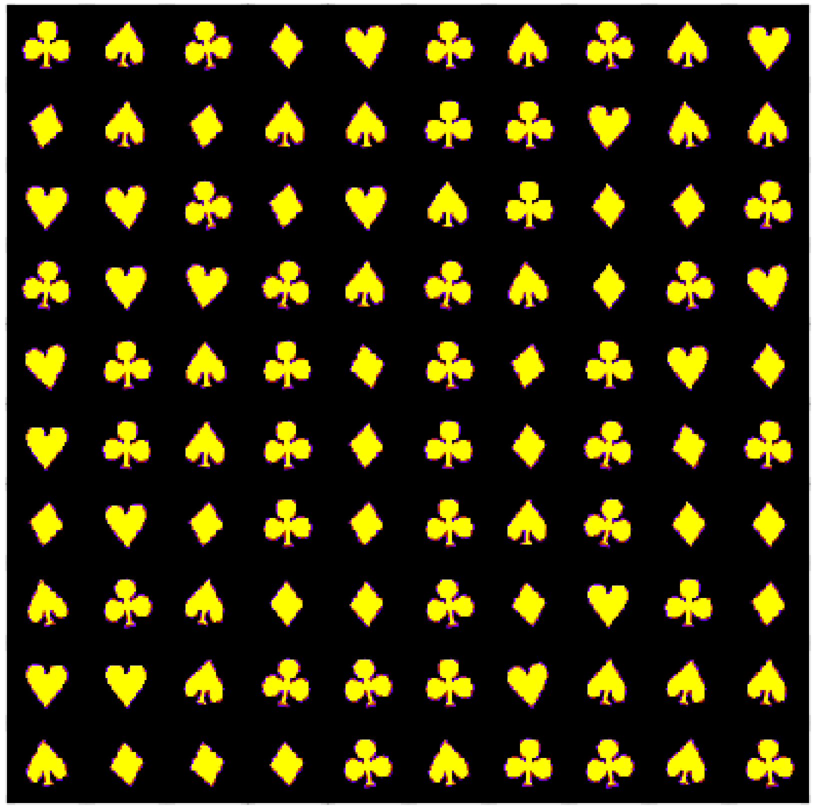
Cards 4: High R (120 deg) and high S (20 deg)



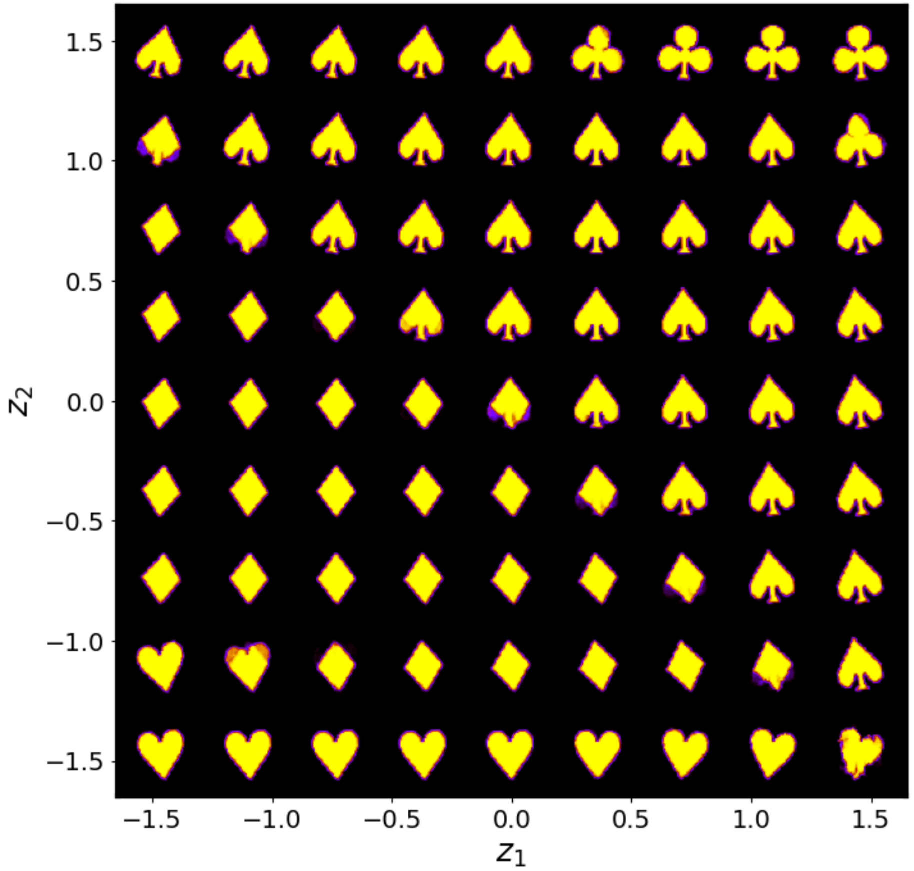
- Shear, rotations, and translations are **known** factors of variability (or traits) in data
- Can VAE disentangle representations and discover these factors of variability

VAE on Cards

Example of data

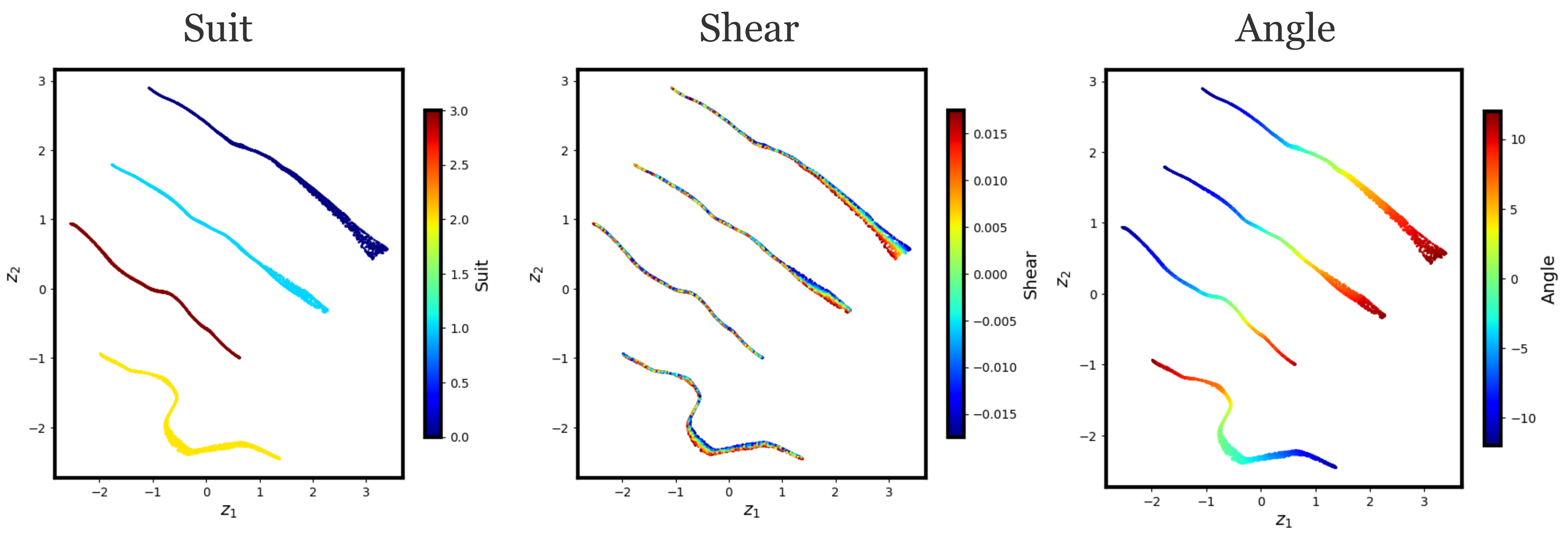


Latent representation



Cards 1: Low rotation (12 deg) and low shear (1 deg)

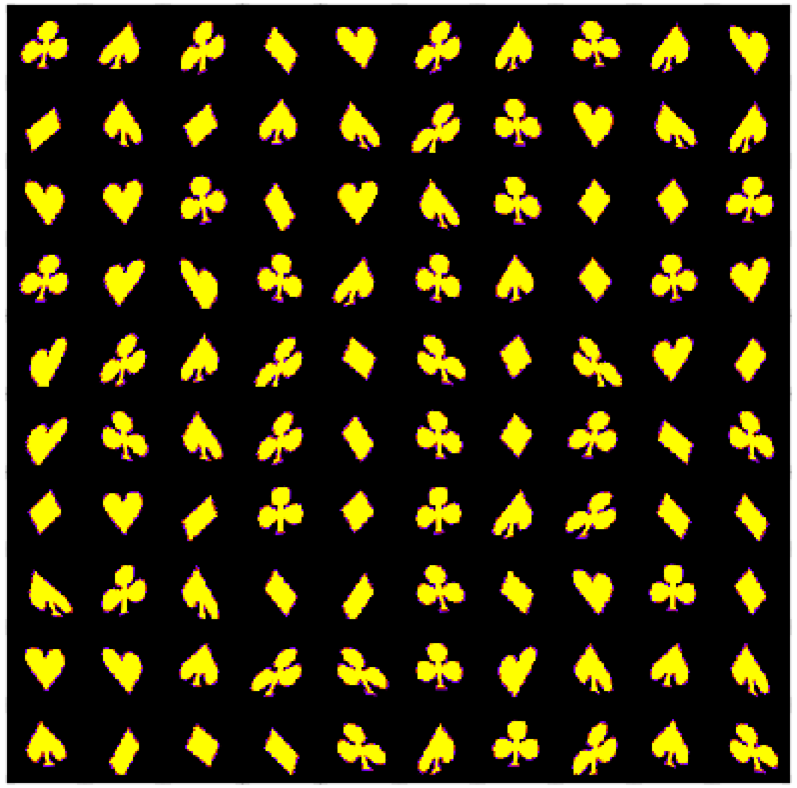
VAE on Cards



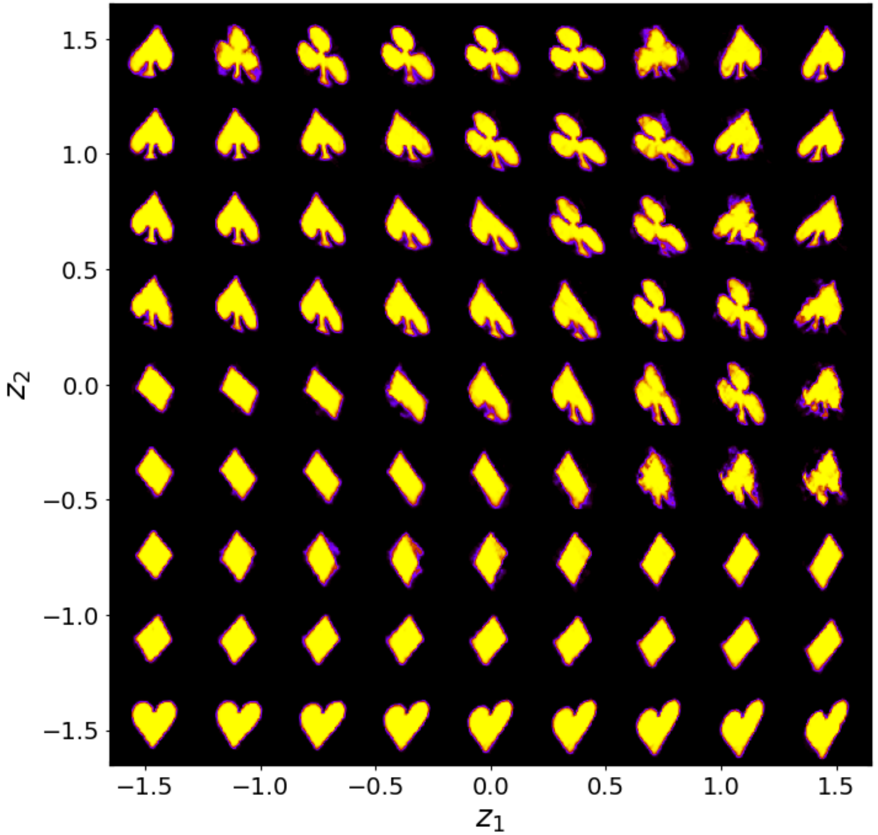
Cards 1: Low rotation (12 deg) and low shear (1 deg)

VAE on Cards

Example of data

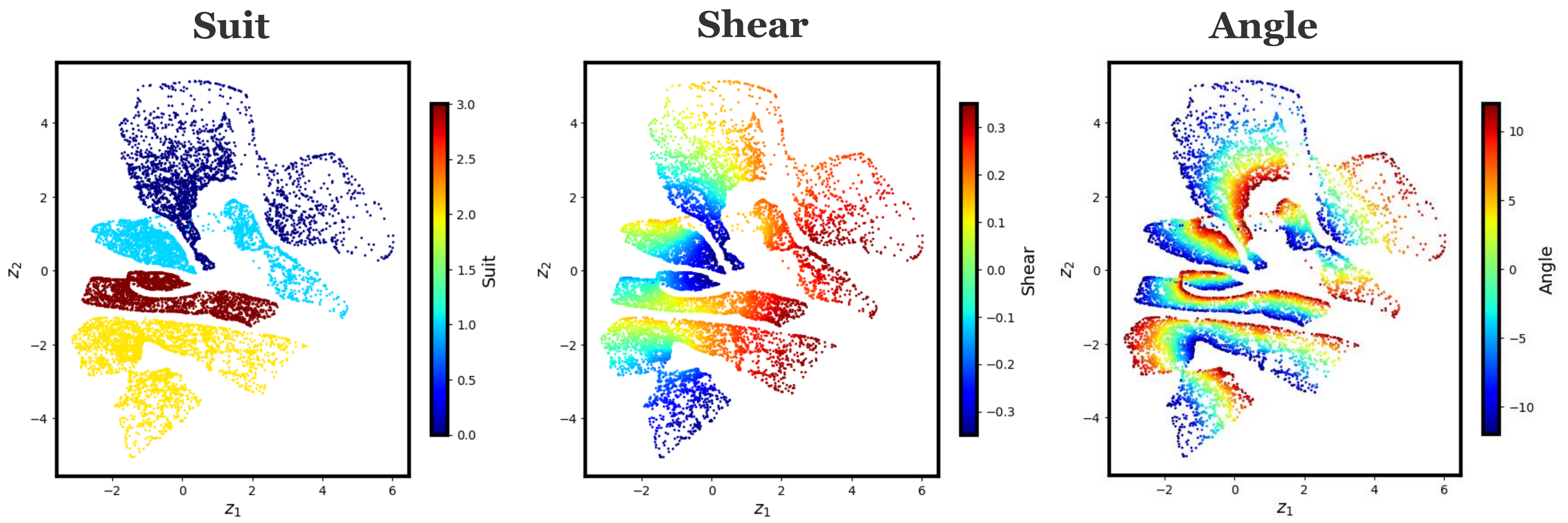


Latent representation



Cards 2: Low rotation (12 deg) and high shear (20 deg)

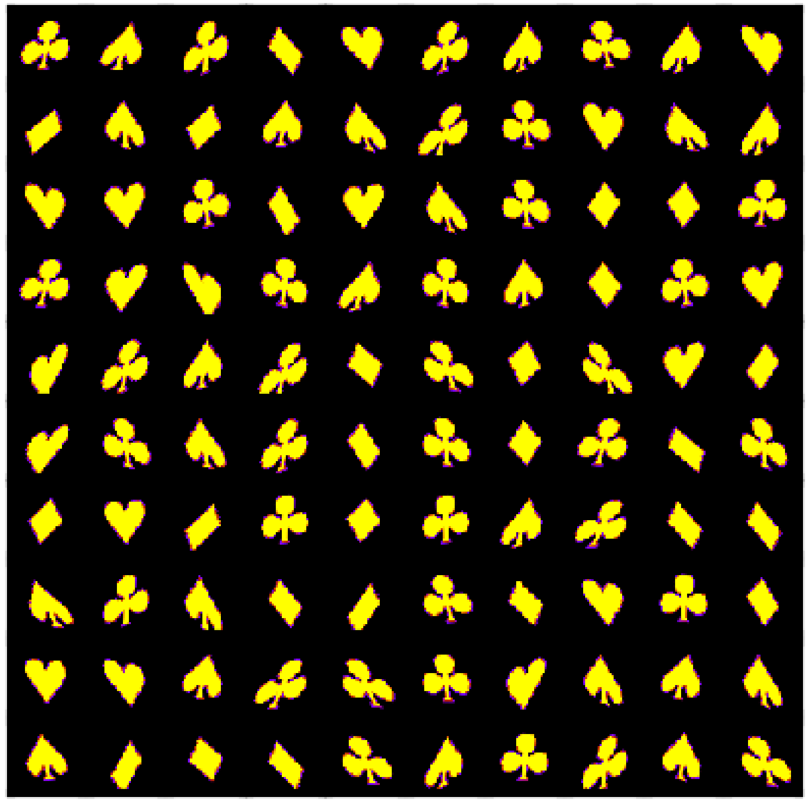
VAE on Cards



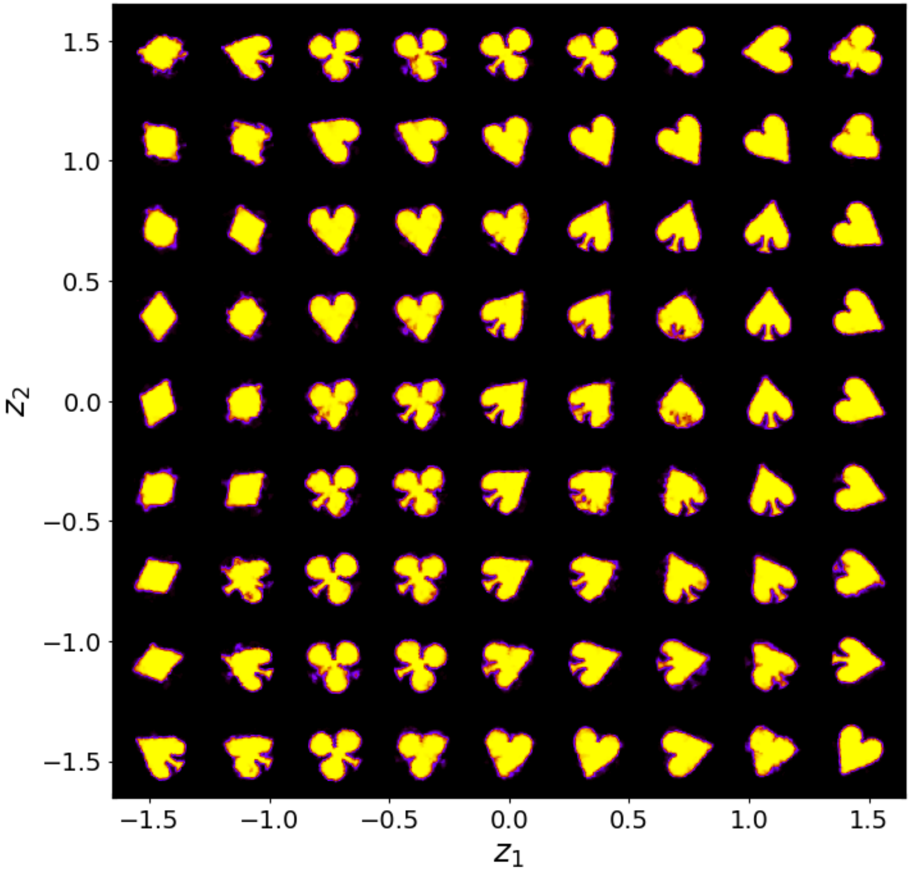
Cards 2: Low rotation (12 deg) and high shear (20 deg)

VAE on Cards

Example of data

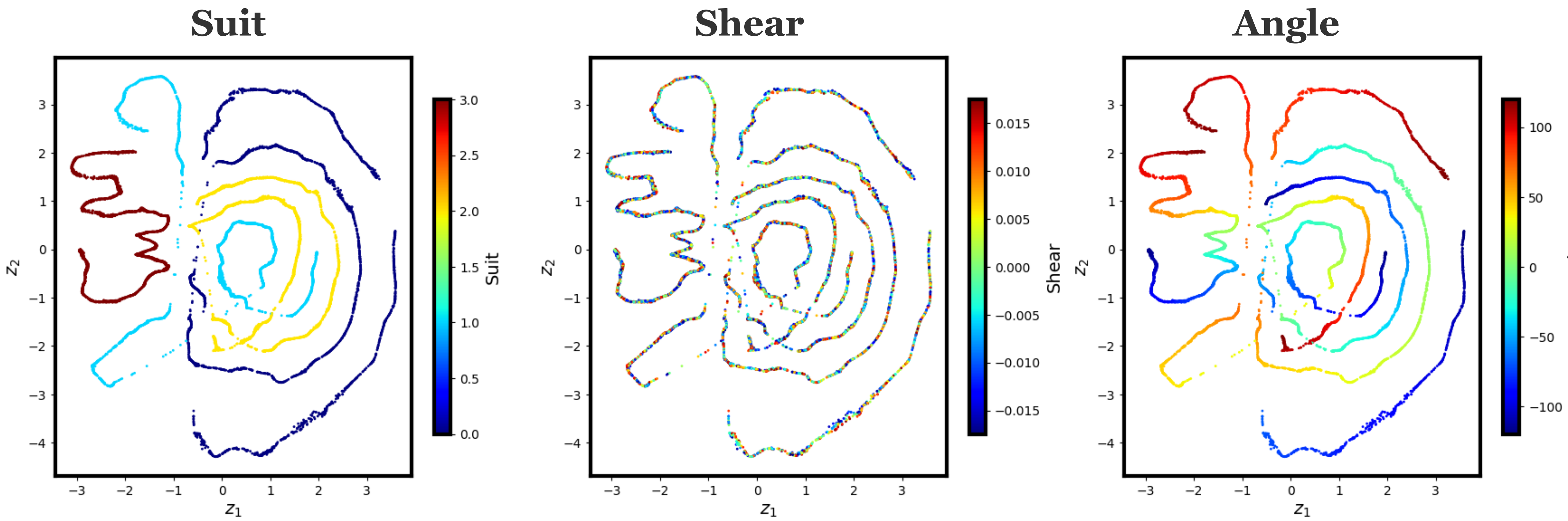


Latent representation



Cards 3: High rotation (120 deg) and low shear (1 deg)

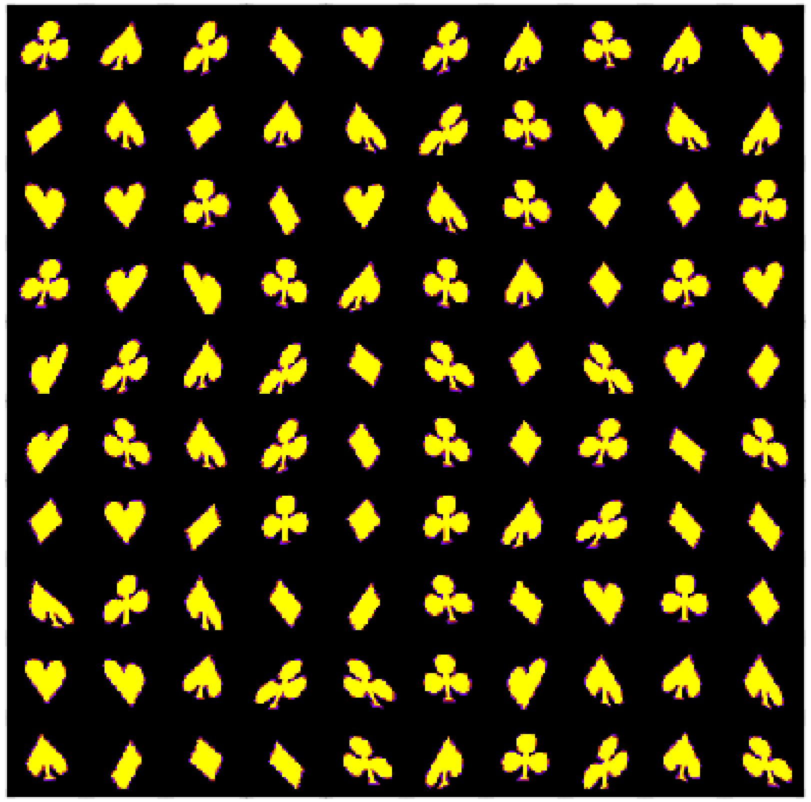
VAE on Cards



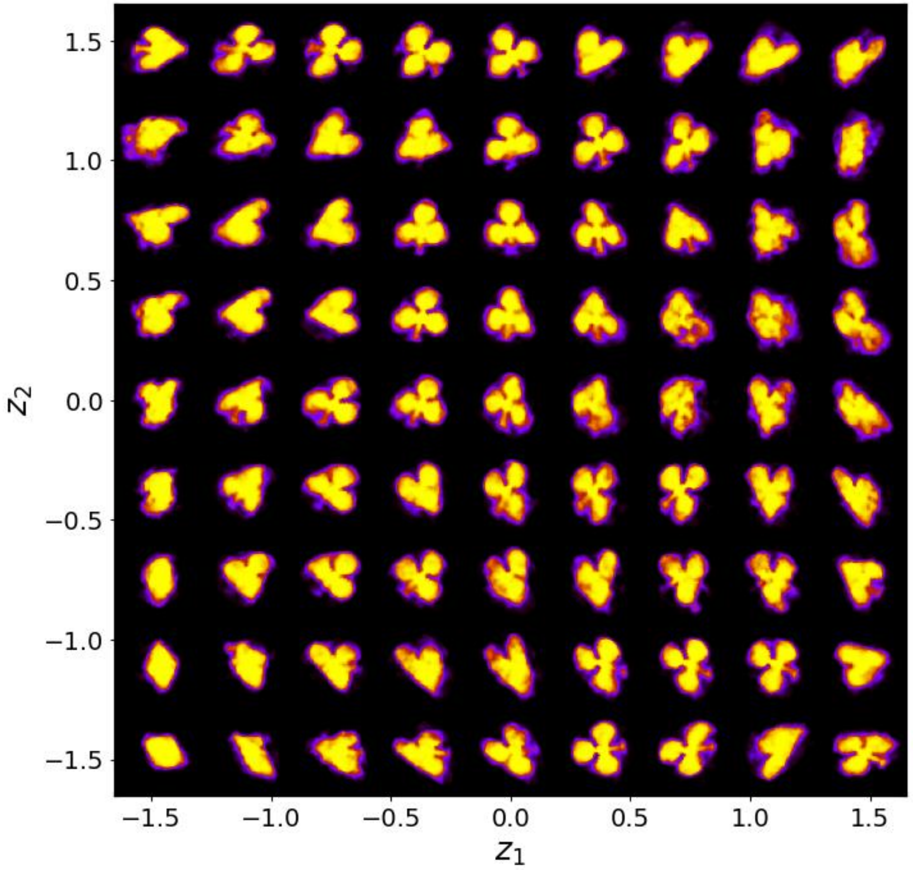
Cards 3: High rotation (120 deg) and low shear (1 deg)

VAE on Cards

Example of data

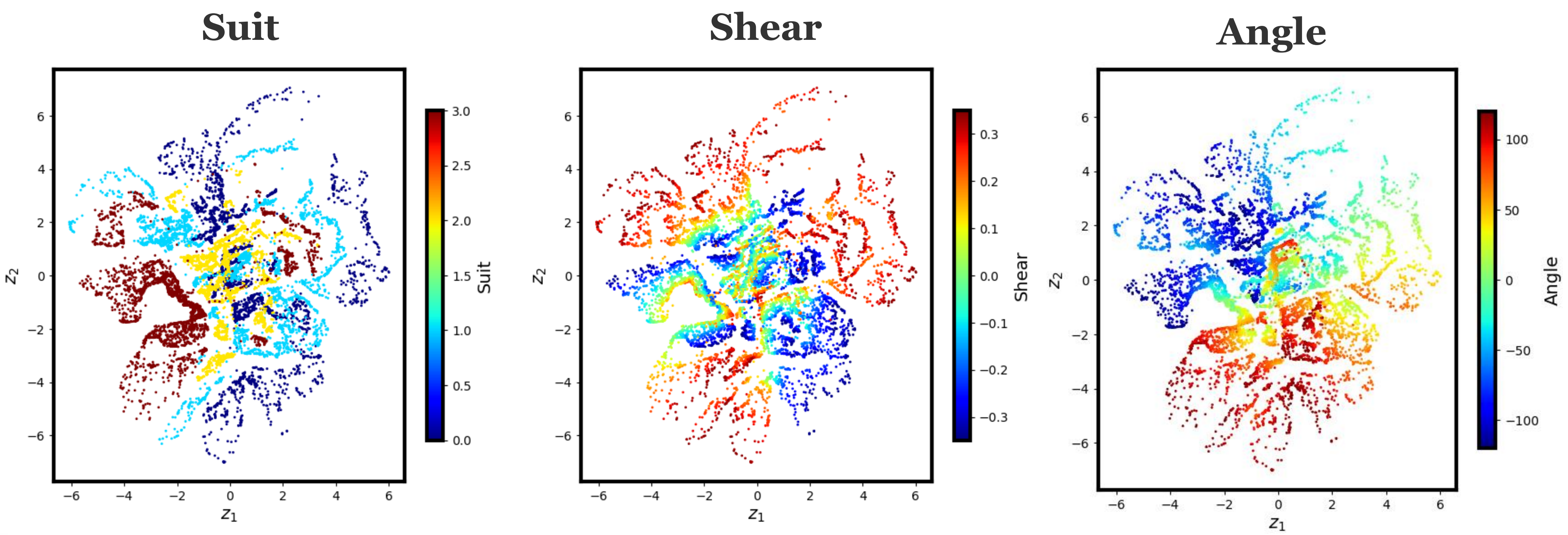


Latent representation



Cards 4: High rotation (120 deg) and high shear (20 deg)

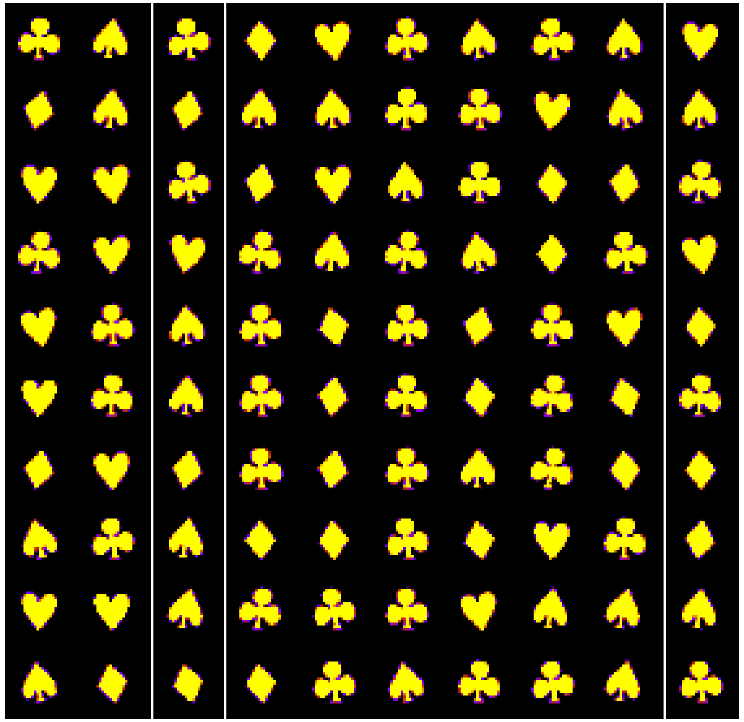
VAE on Cards



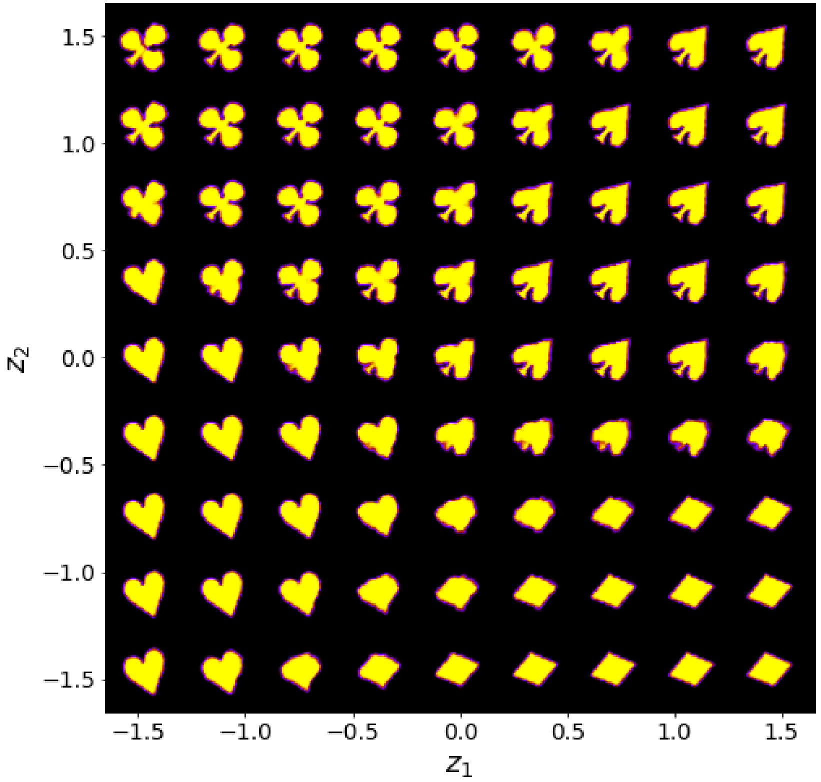
Cards 4: High rotation (120 deg) and high shear (20 deg)

rVAE on Cards

Example of data

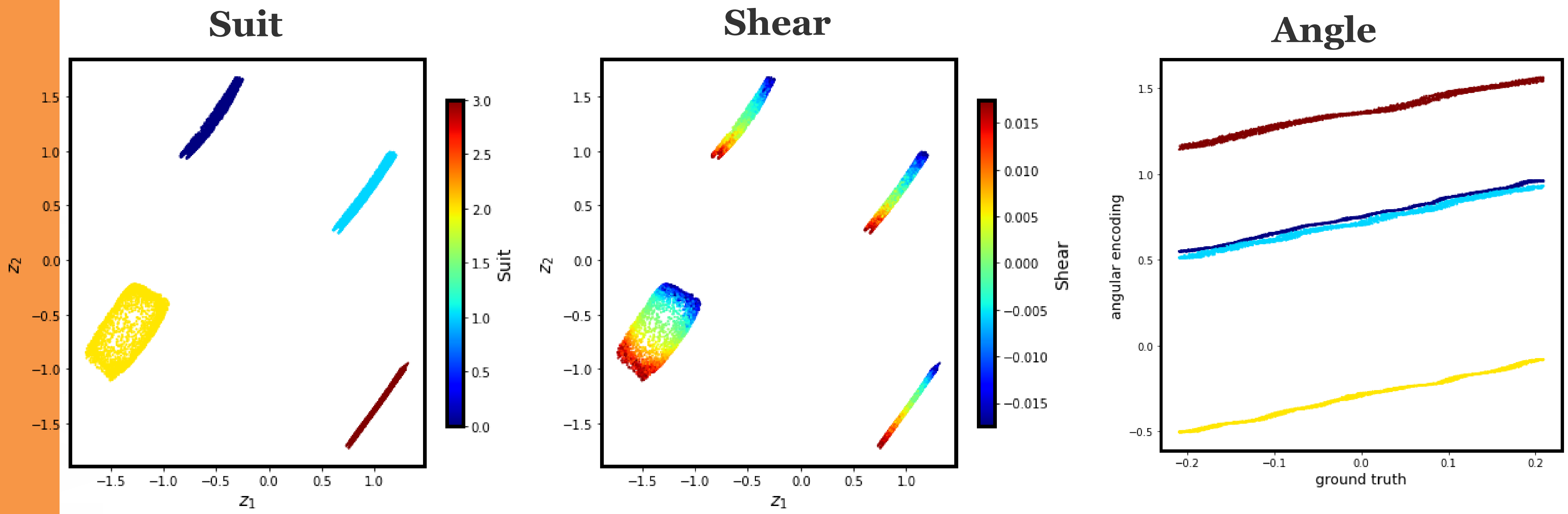


Latent representation



Cards 1: Low rotation (12 deg) and low shear (1 deg)

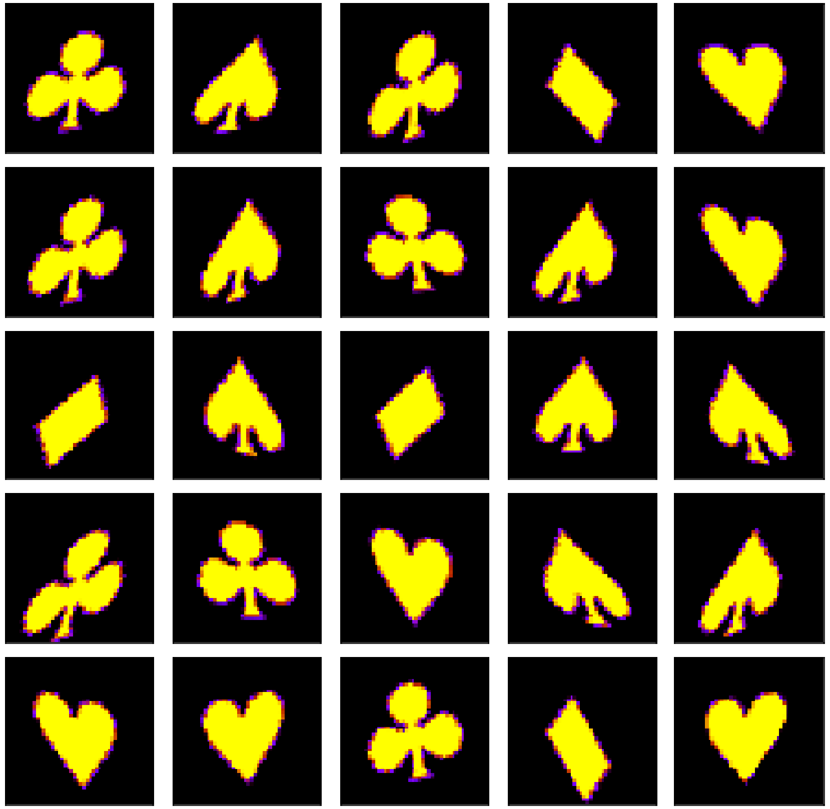
rVAE on Cards



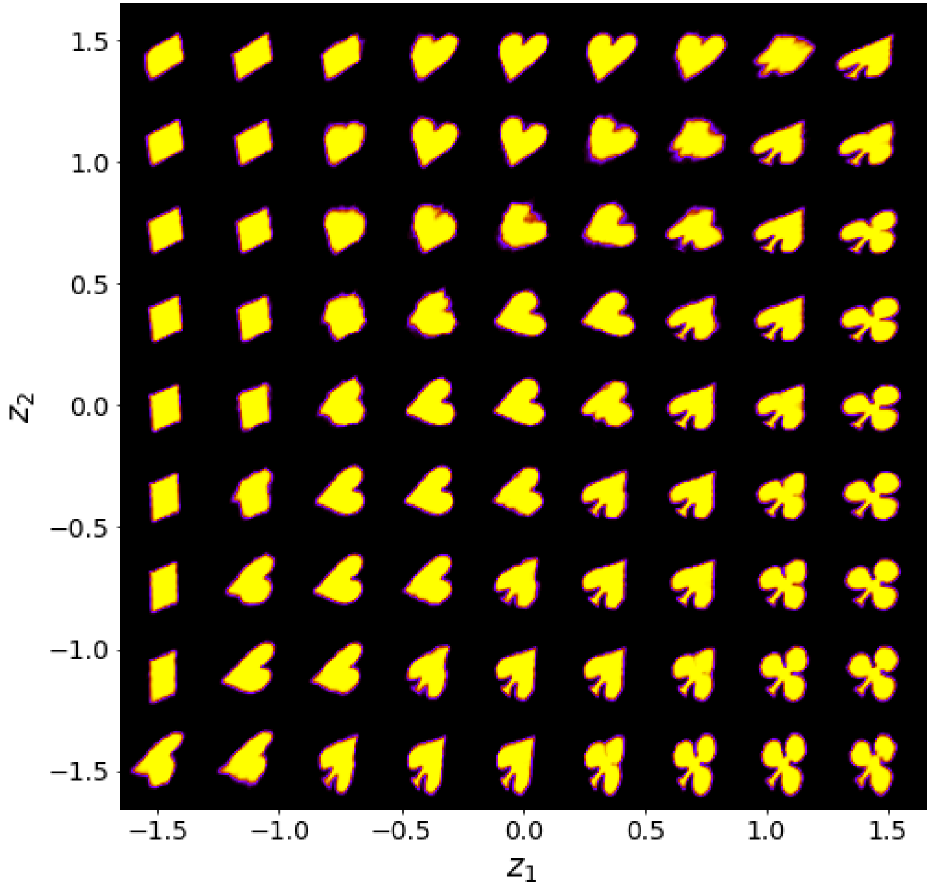
Cards 1: Low rotation (12 deg) and low shear (1 deg)

rVAE on Cards

Example of data

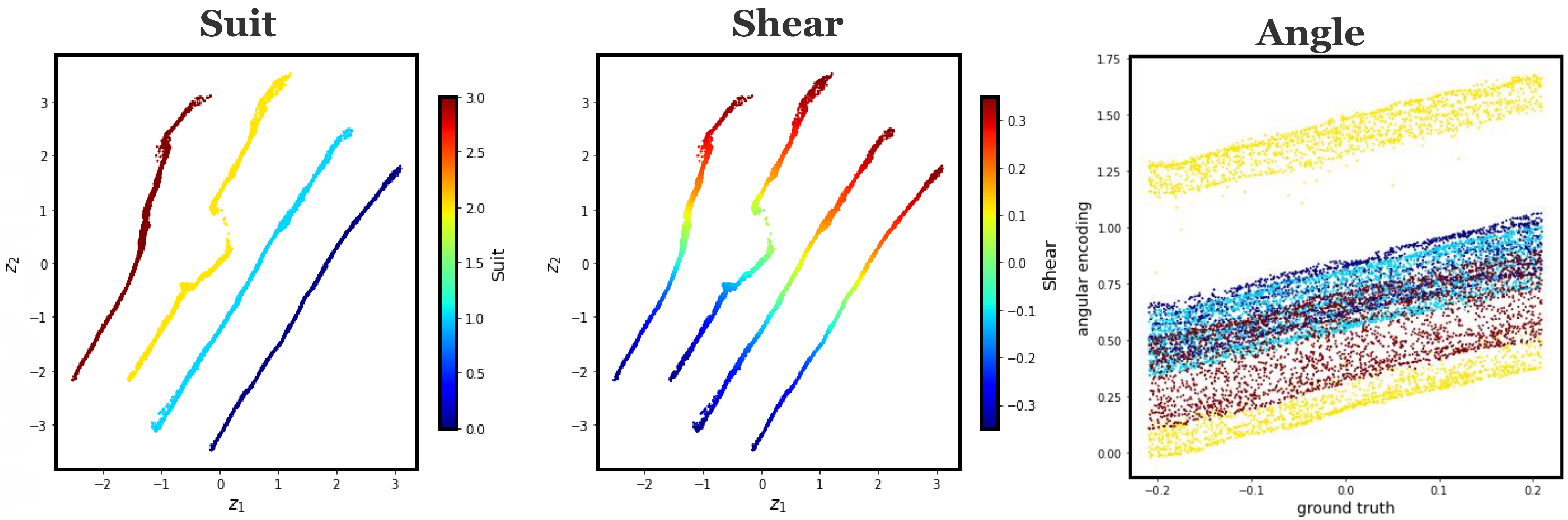


Latent representation



Cards 2: Low rotation (12 deg) and high shear (20 deg)

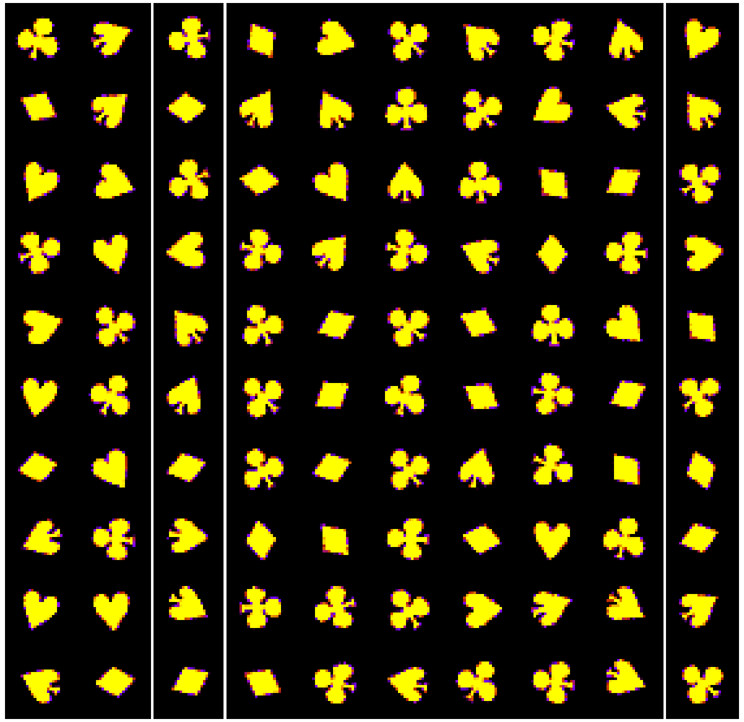
rVAE on Cards



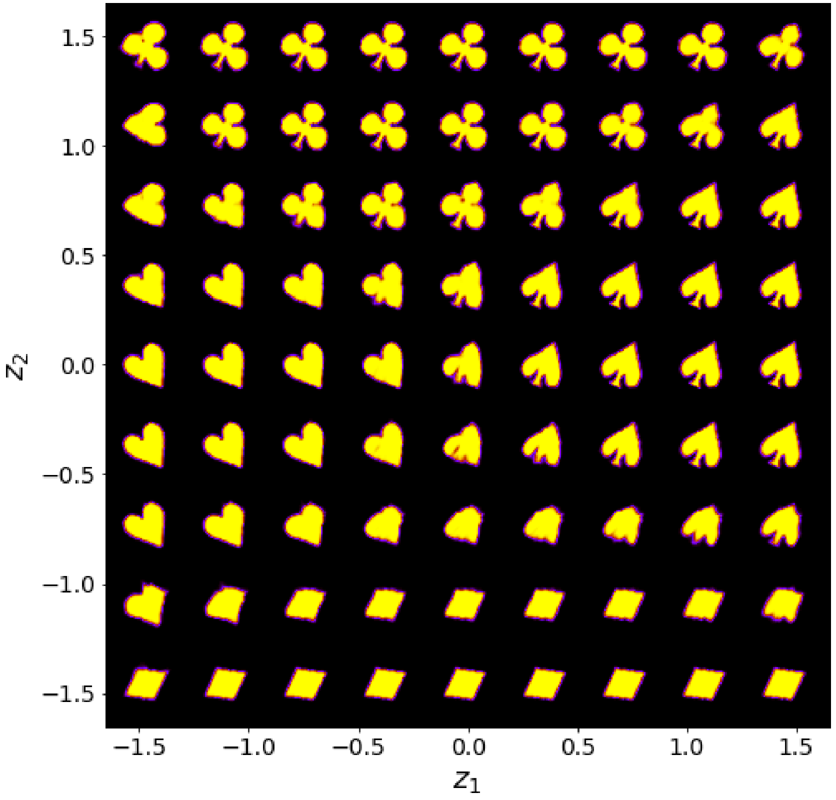
Cards 2: Low rotation (12 deg) and high shear (20 deg)

rVAE on Cards

Example of data

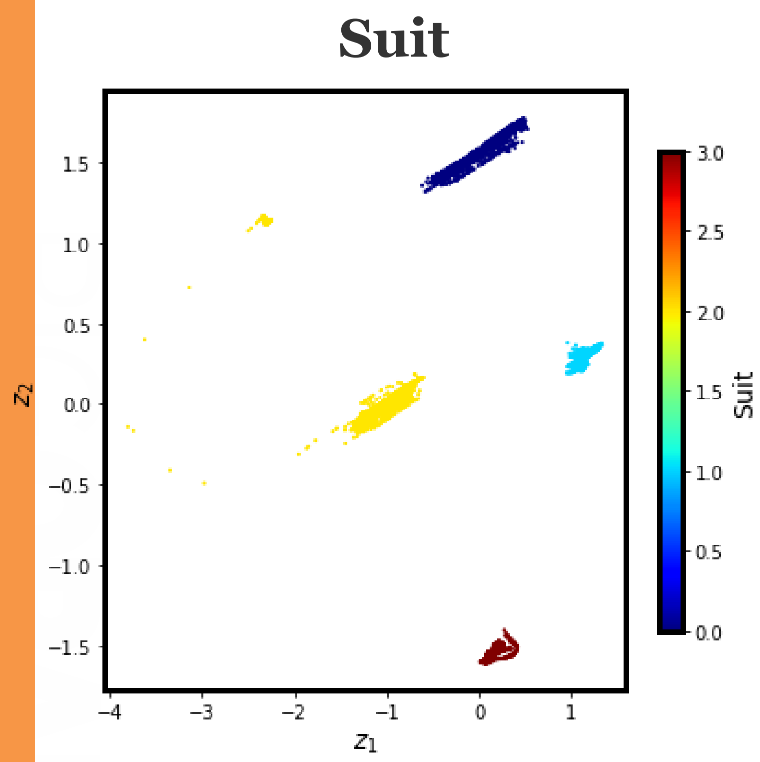


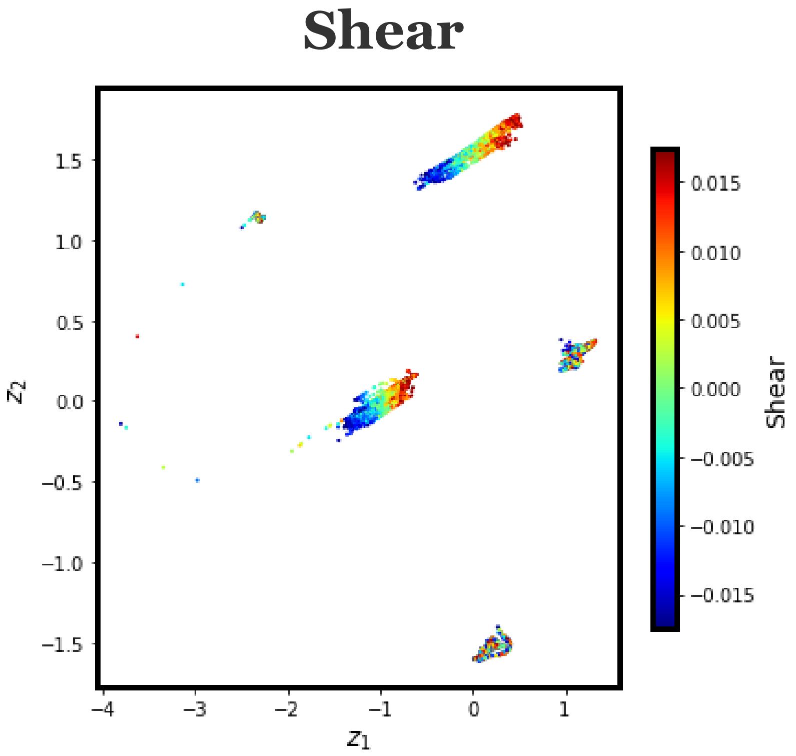
Latent representation

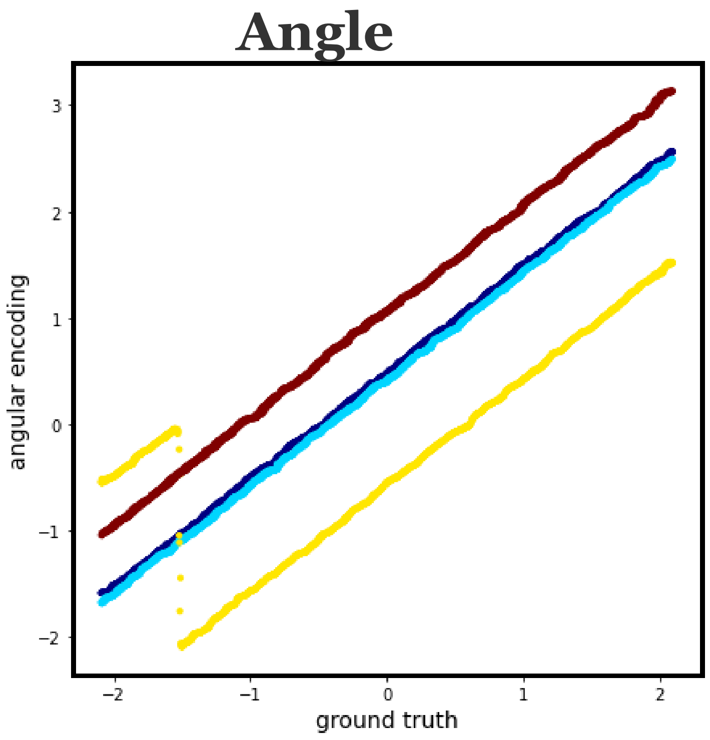


Cards 3: High rotation (120 deg) and low shear (1 deg)

rVAE on Cards



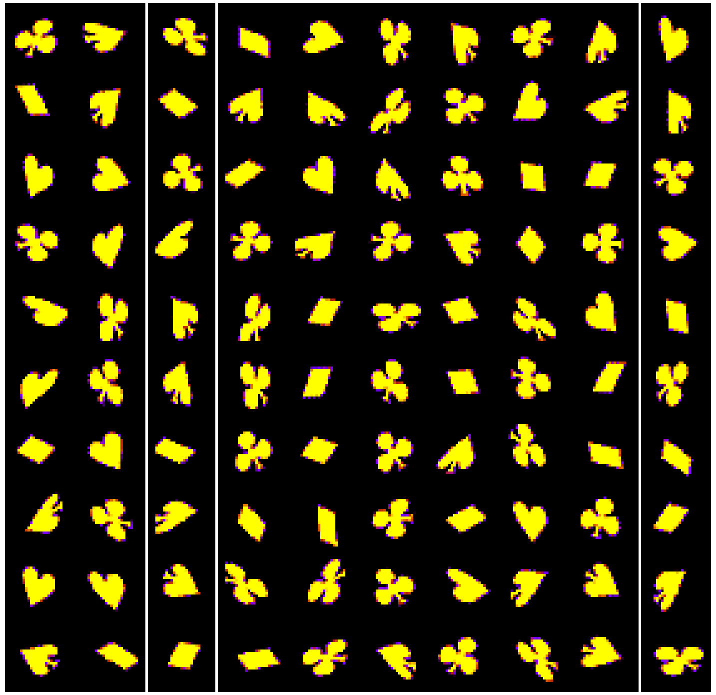




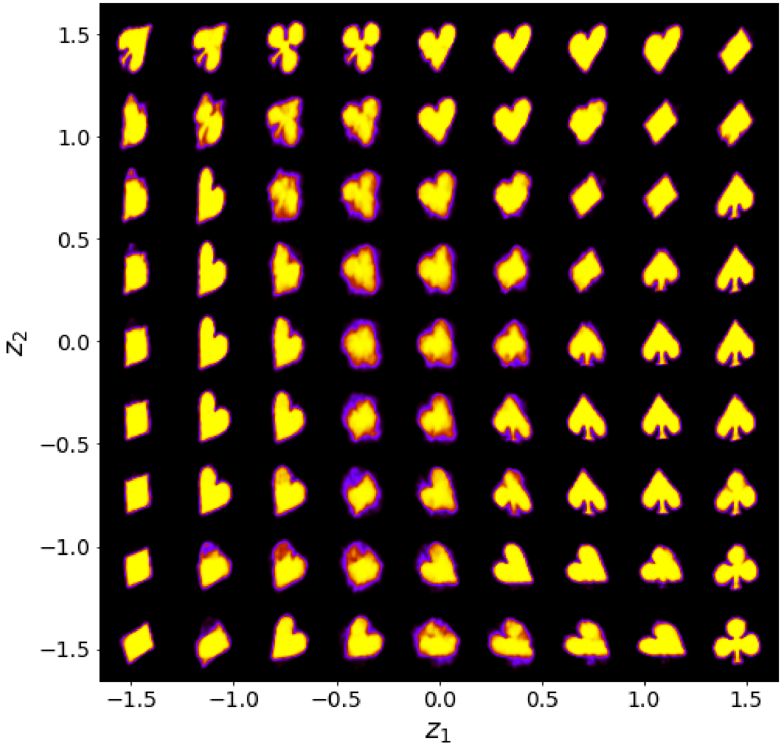
Cards 3: High rotation (120 deg) and low shear (1 deg)

rVAE on Cards

Example of data



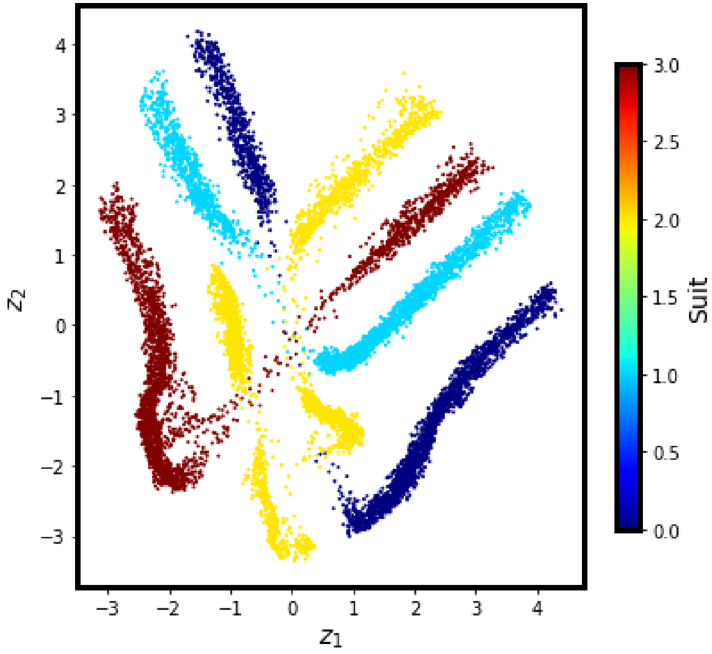
Latent representation



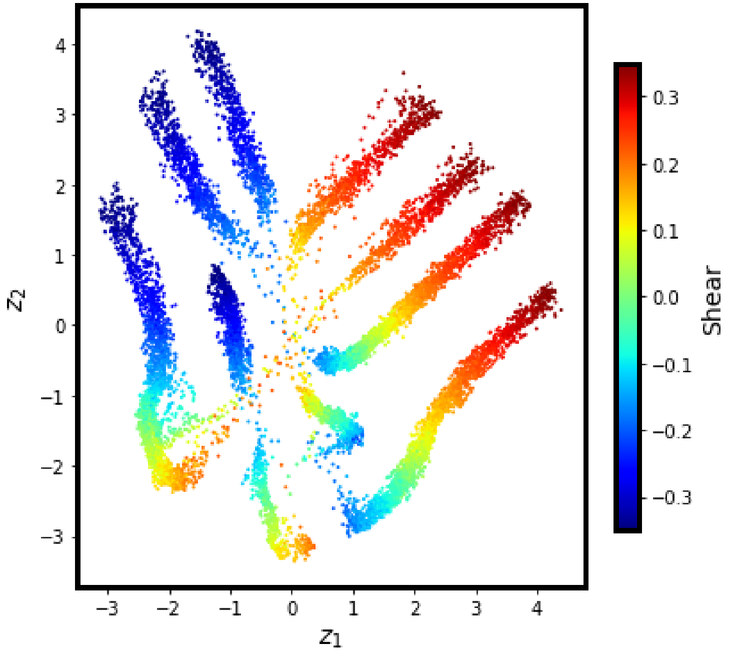
Cards 4: High rotation (120 deg) and high shear (20 deg)

rVAE on Cards

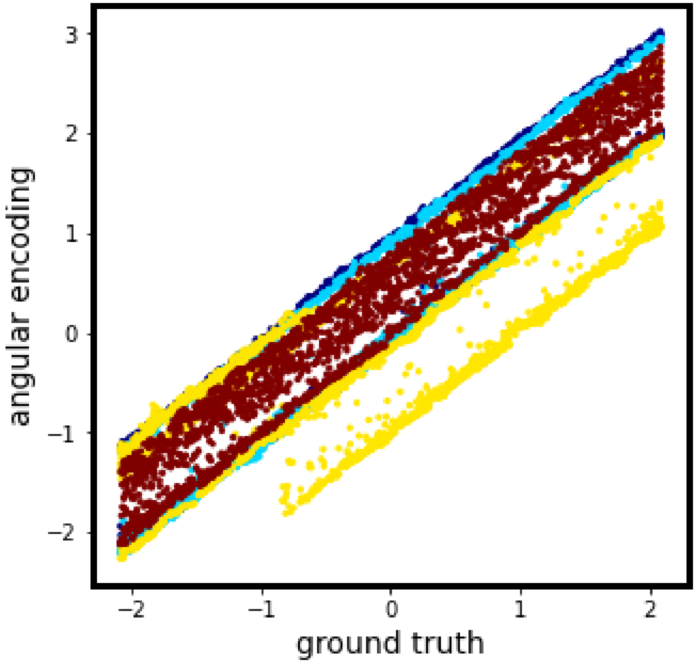
Suit



Shear



Angle



Cards 4: High rotation (120 deg) and high shear (20 deg)

PURPOSE

- (Super-brief) introduction into Neural Networks
- What are (Variational) autoencoders?
- Key notions:
  - Encoding and decoding
  - Latent distribution
  - Latent representations
- Why invariances: rotational, translational, and scale
- Other colors of VAEs:
  - Semi-supervised
  - Conditional
  - Joint
- From VAEs to encoder-decoders (VED)
- Further opportunities:
  - Physics constraints
  - Representation learning
- Active learning: DKL

What if we have multiple classes?

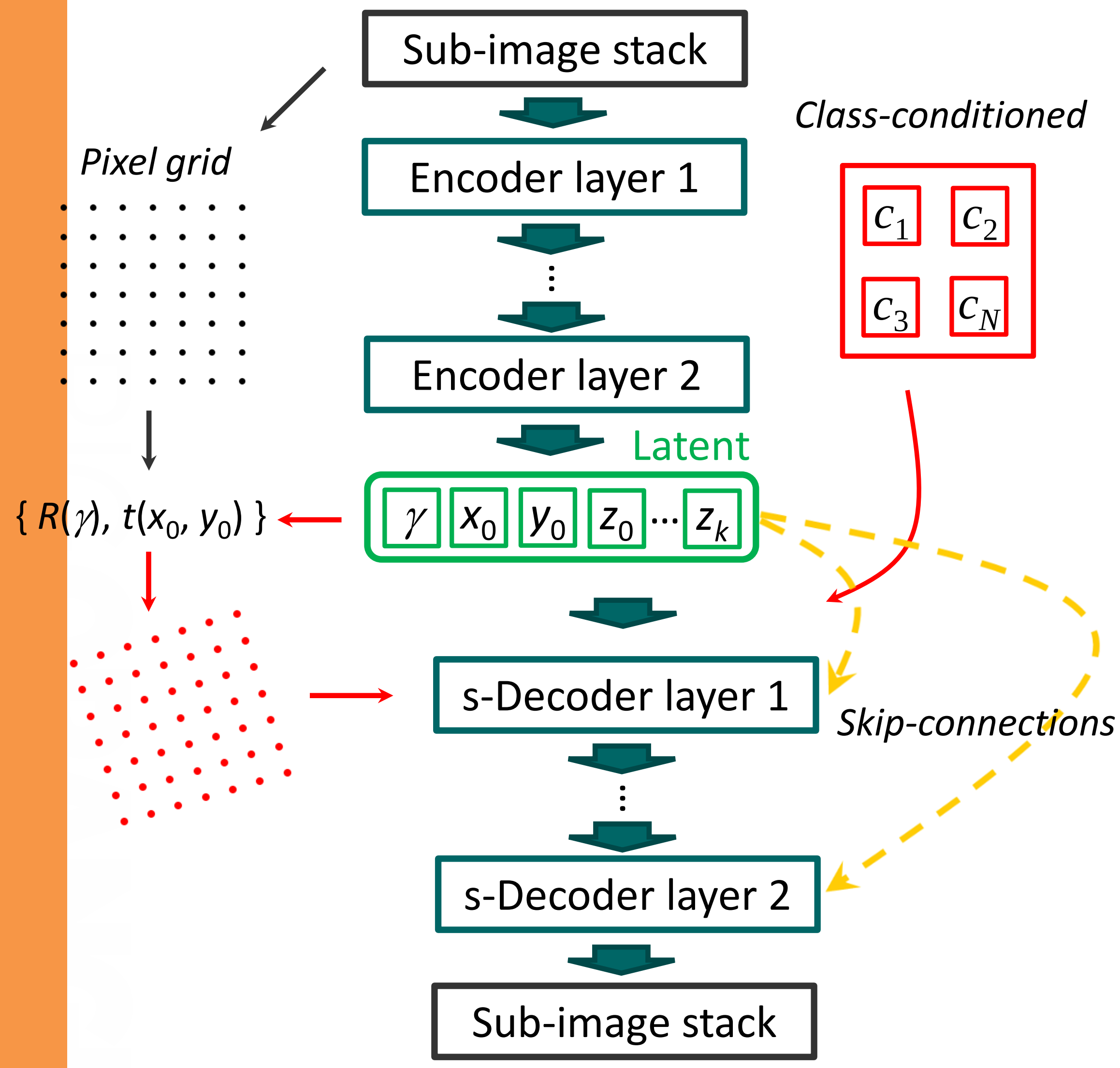
1. Classes are known: conditional (discrete) VAE

2. Factors of variability are known: conditional (continuous) VAE

3. Some classes are known: semi-supervised VAE

4. Number of classes are known: joint VAE

Conditional VAE



- Generative model is a function of spatial coordinate
- 3 additional latent variables to absorb rotations and shifts
- Disentangles rotations and translations from image content
- Ideal for analyzing microscopy sub-images on atomic level

$$\begin{aligned} ELBO &= Reconstruction\ Loss \\ &- D_{KL}\big(q(z|x)\|\mathcal{N}(0,I)\big) \end{aligned} \Big\} \text{ Regular VAE}$$

$$- D_{KL}\left(q(\gamma|x)\|\mathcal{N}\left(0, s_\gamma^2\right)\right) \quad \text{Rotation}$$

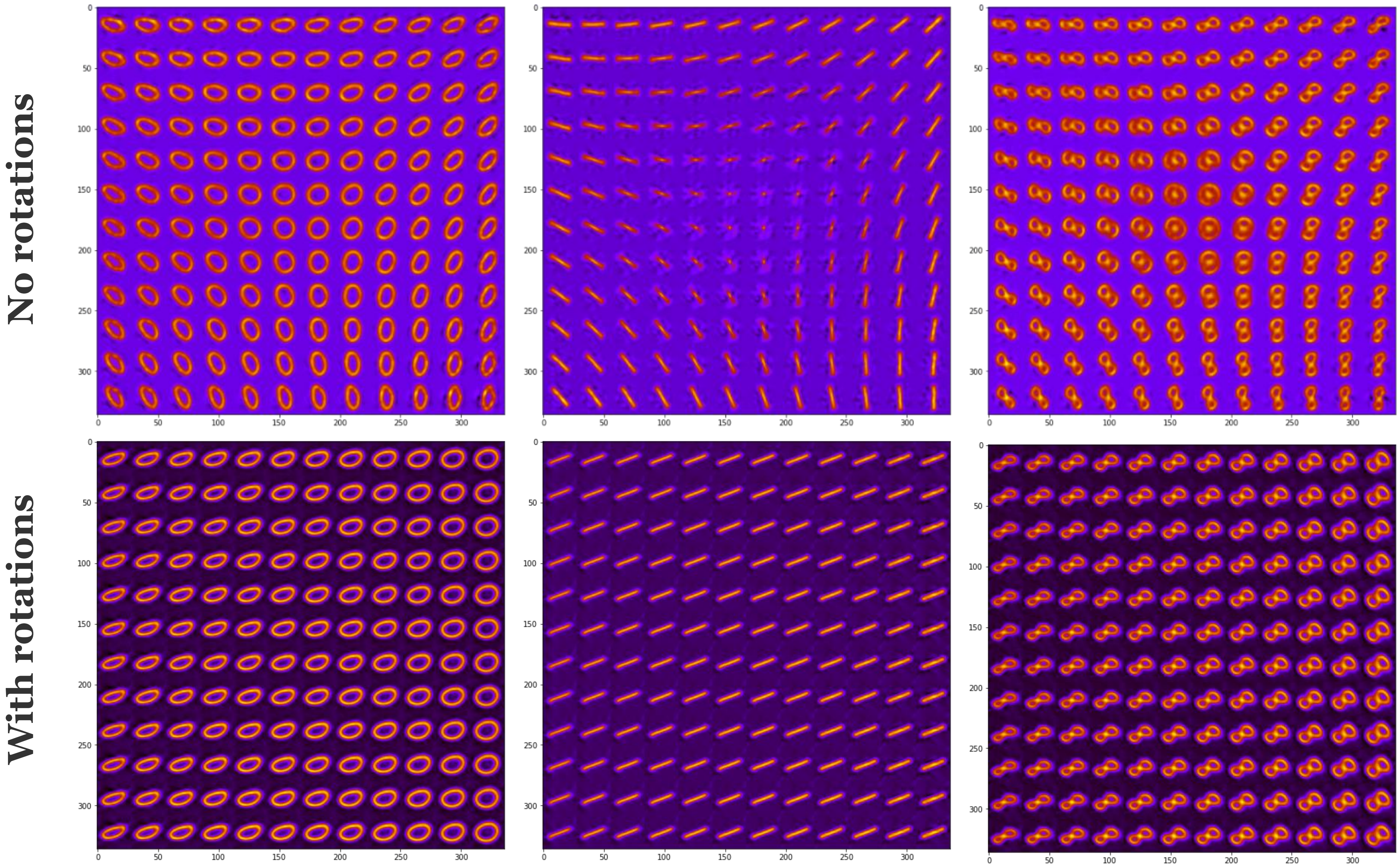
$$- D_{KL}\left(q(\Delta r|x)\|\mathcal{N}\left(0, s_{\Delta r}^2\right)\right) \quad \text{Translation}$$

+ D_{KL}(physics-based “priors”) ?

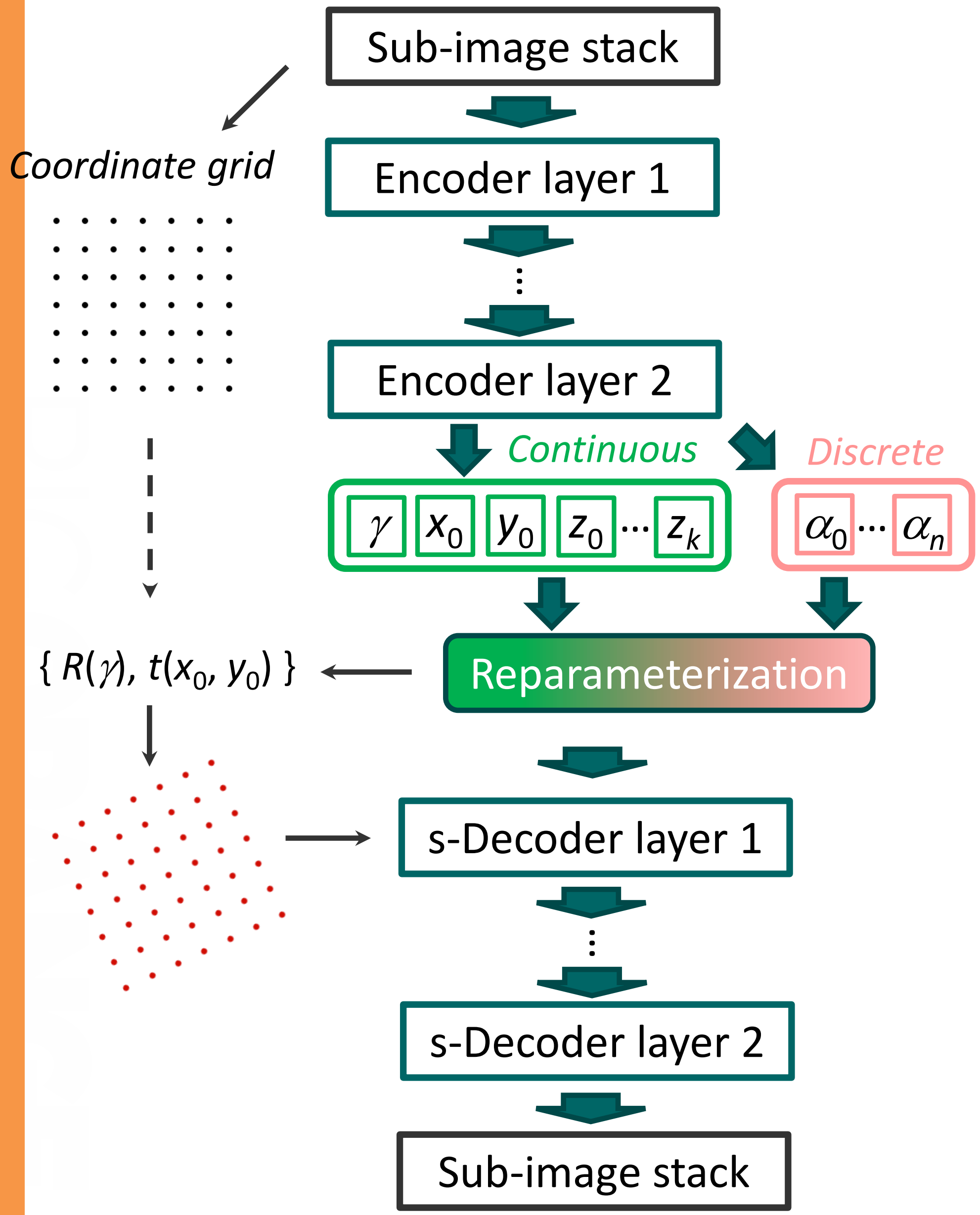
+ D (physics) ?

Bepler *et al.*, NeurIPS 2019

MNIST: cVAE



Joint VAE



- Generative model is a function of spatial coordinate (*e.g.*, via spatial broadcasting)
- 3 additional latent variables to absorb rotations and shifts
- Disentangles rotations and translations from image content
- Learns discrete classes in unsupervised fashion
- Well-suited for analyzing microscopy (sub-)images on atomic and molecular levels

$ELBO =$

$-\,Reconstruction\ Loss$

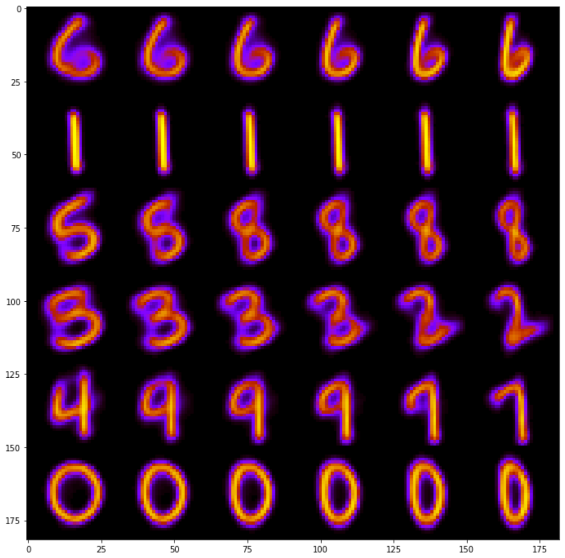
Structure Rotation

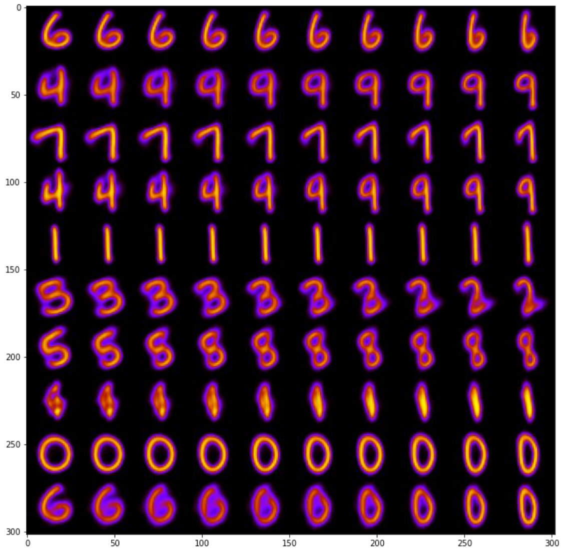
$-\beta_c(t)\left|\left(D_{KL}(q(z|x)\|p(z)) + D_{KL}(q(\gamma|x)\|p(\gamma)\right) - C_z\right|$ Continuous

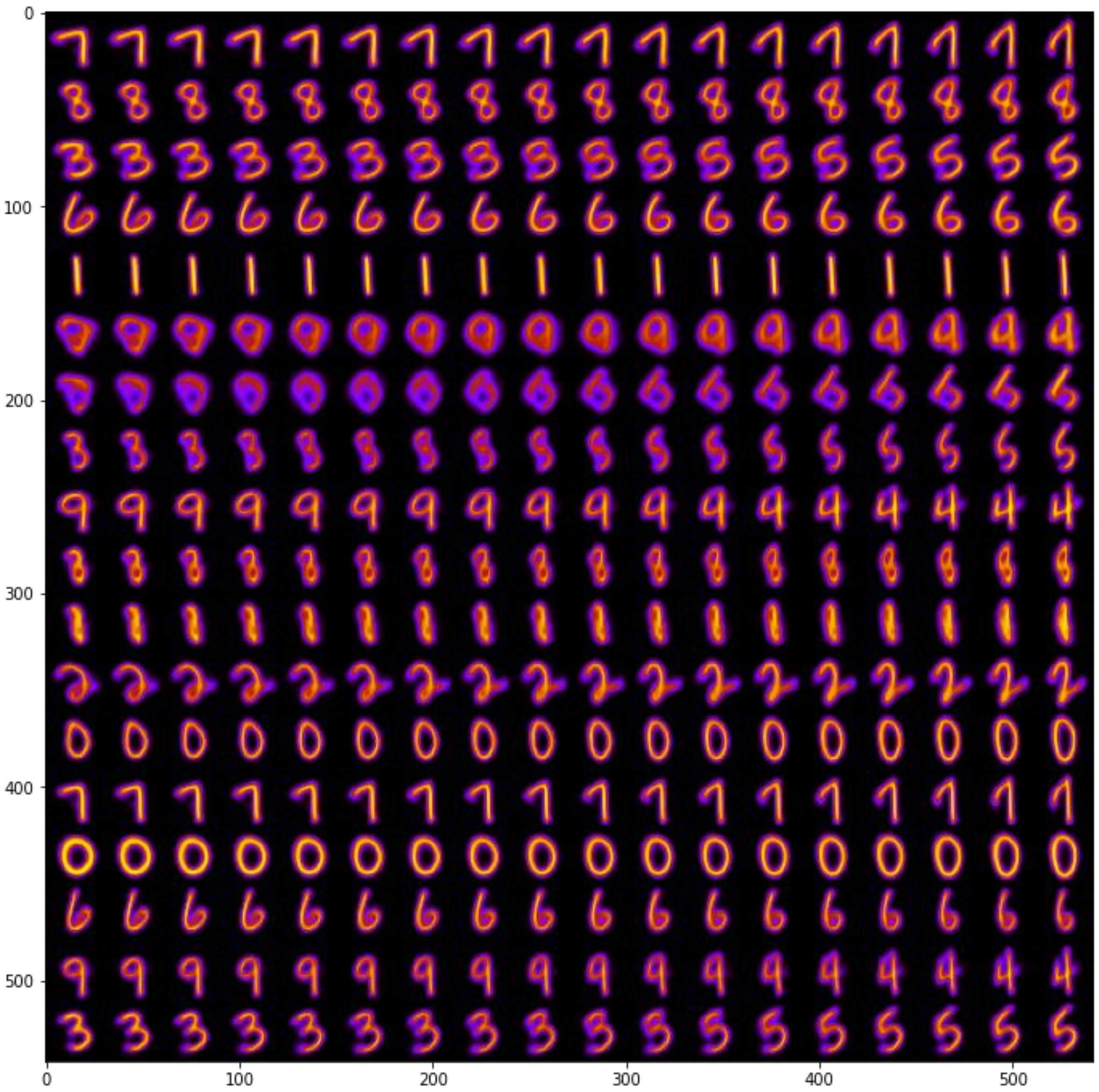
$-\beta_d(t)\left|D_{KL}\left(q(\alpha|x)\|p(\alpha)\right) - C_\alpha\right|$ Discrete

+ physics-based "loss" ?

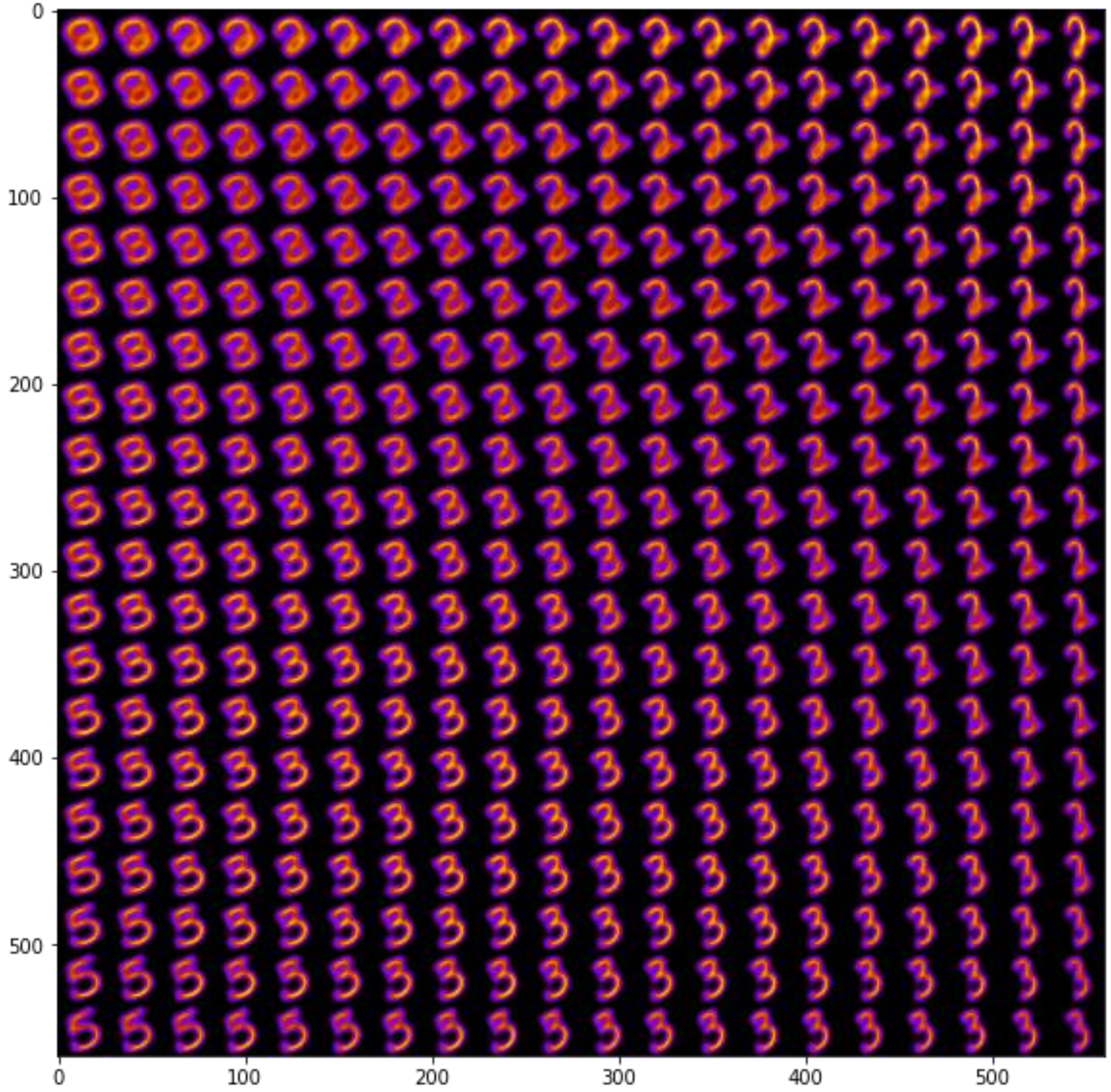
jVAE of MNIST





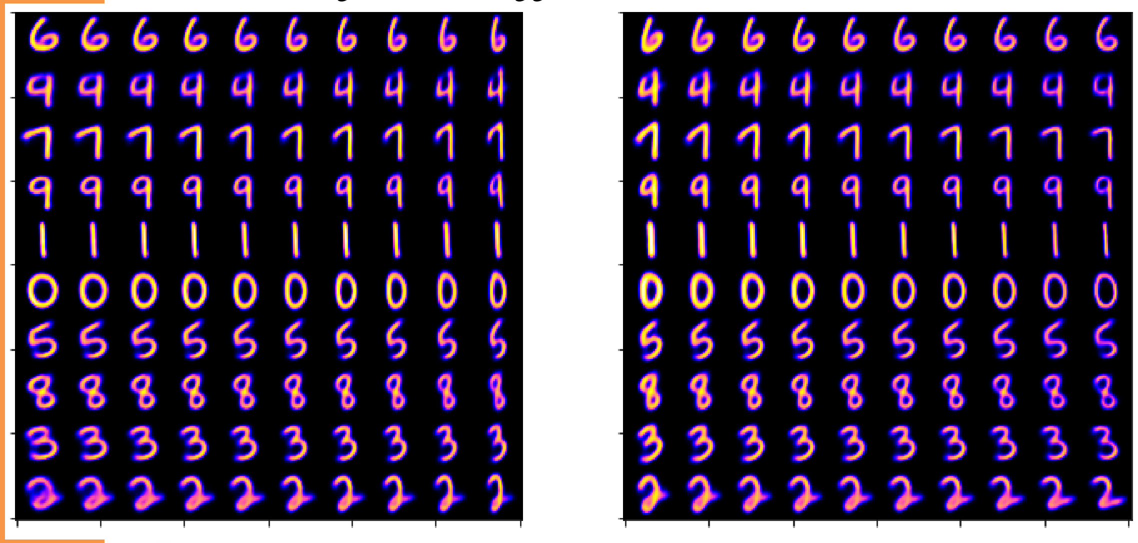


Latent representation

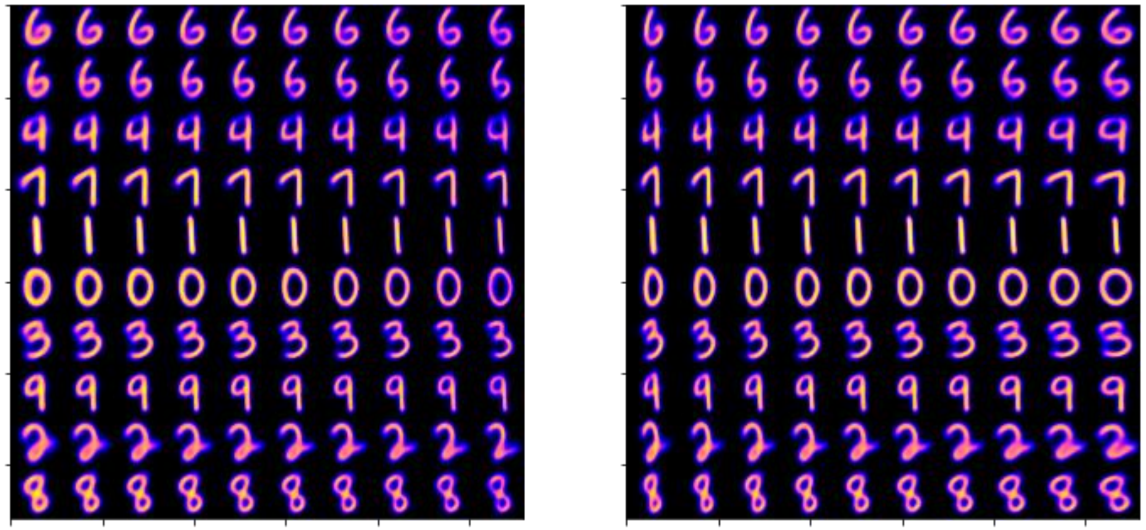


Ensemble jVAE

Predictions from different ensemble models

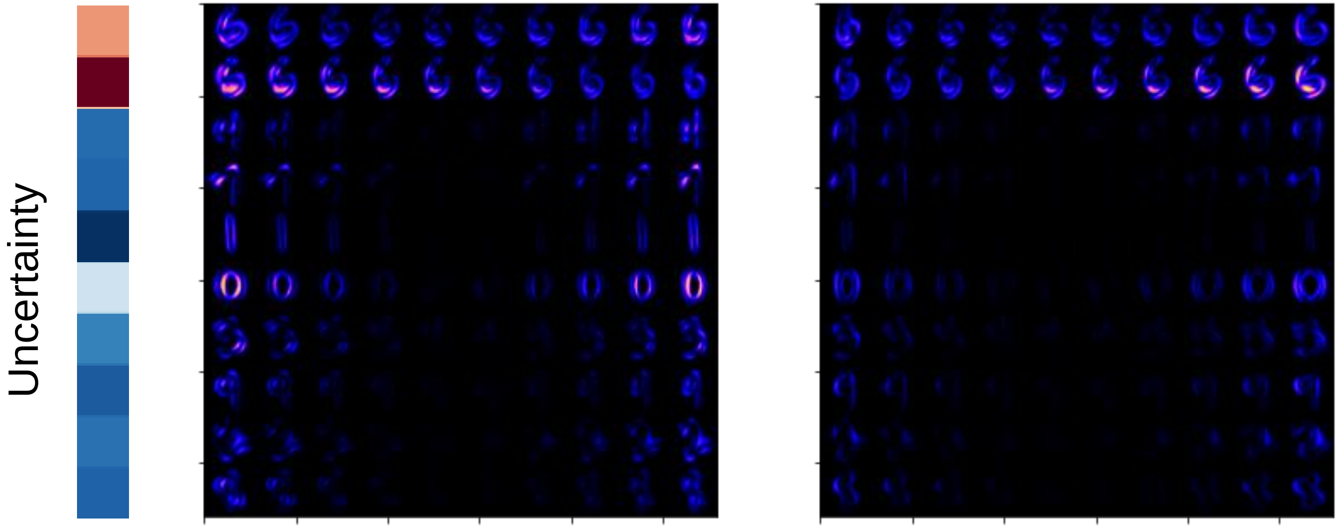


Mean prediction



Baseline: 10 epochs
Ensemble models: 8

Dispersion in predictions ('uncertainty')



- The unstable classes show the largest "uncertainty"
- Indication of the quality of separation and/or a guide for selection of the number of classes

What can (unsupervised) classification give us

- Our research deals with complex data sets containing information on physics of objects we seek to understand
- This can be spectral data sets (EELS in STEM, CITS in STM, complex spectroscopies in PFM) or single, multimodal, or hyperspectral images
- Often, we seek approaches to reduce dimensionality and explore similarities in these data sets.

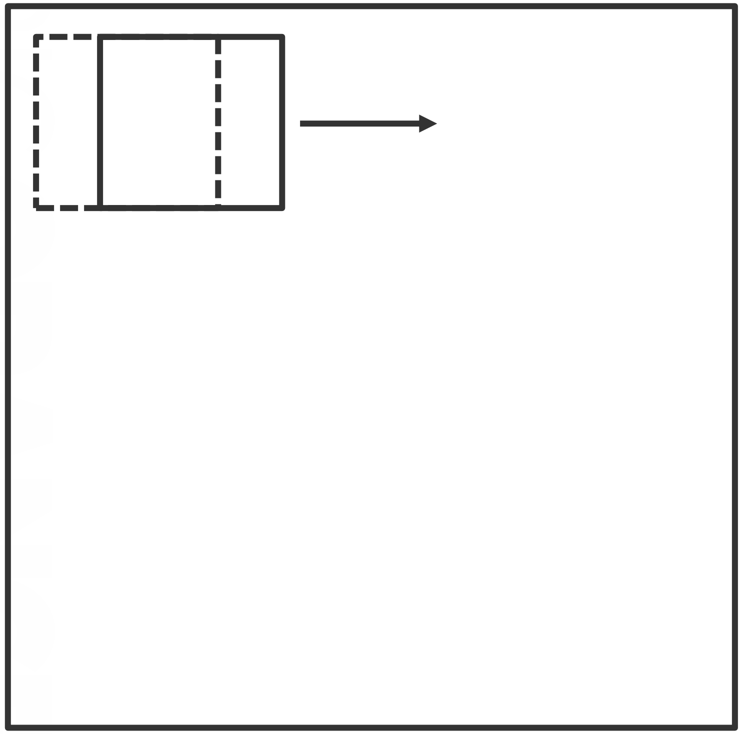
- When working with such data sets, two things matter: descriptors and ML method
- In analysis of EELS or CITS data, very often our descriptor is just the spectrum at each pixel. Typical analysis will be either linear or non-linear dimensionality reduction or clustering:
  - Linear dimensionality reduction: PCA, NMF, BLU
  - Clustering: k-means, GMM
  - Manifold learning: ISO, UMAP, tSNE, DBSCAN
  - Neural nets: SOFM, AEs, VAEs
- Typical result will be the components (representing behavior), and loading maps representing spatial variability of these behaviors. By construct, components will not depend on the relative spatial positions of pixel.

- What about images?

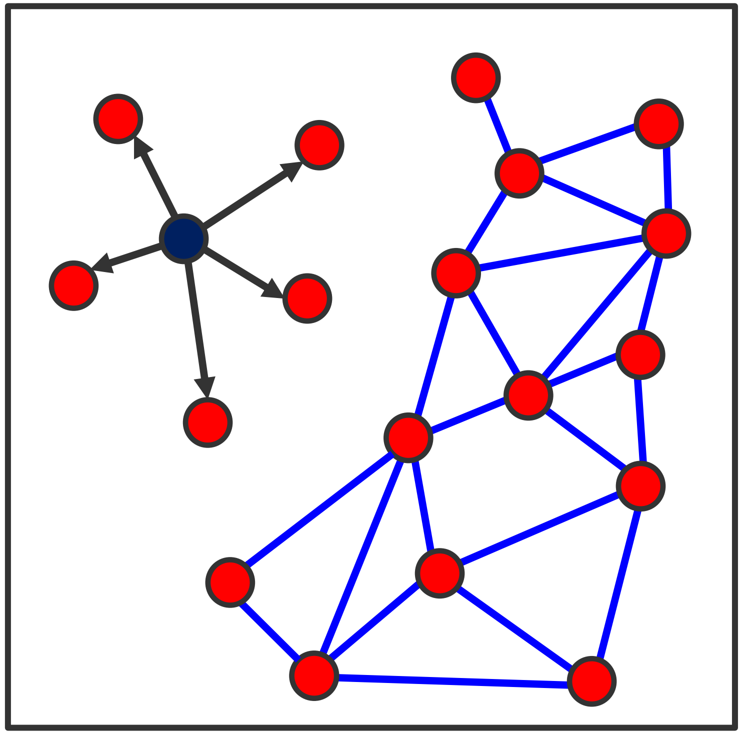
Describing the building blocks

- The classical physical descriptions (symmetry, etc) can be defined locally only in Bayesian sens
- We can argue that local descriptors are simple, if not necessarily known
- And the rules that guide their emergence are also simple, if not known

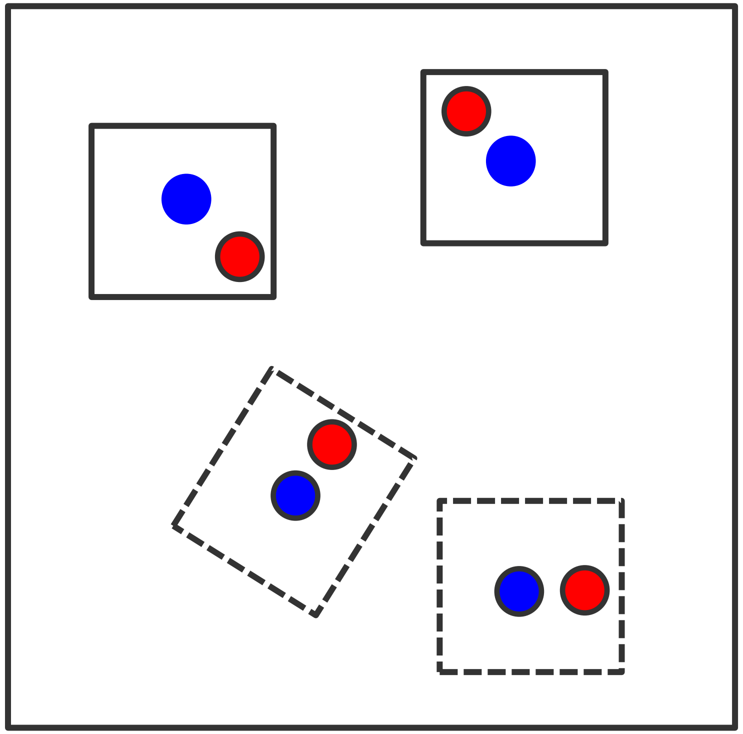
Continuous translational symmetry



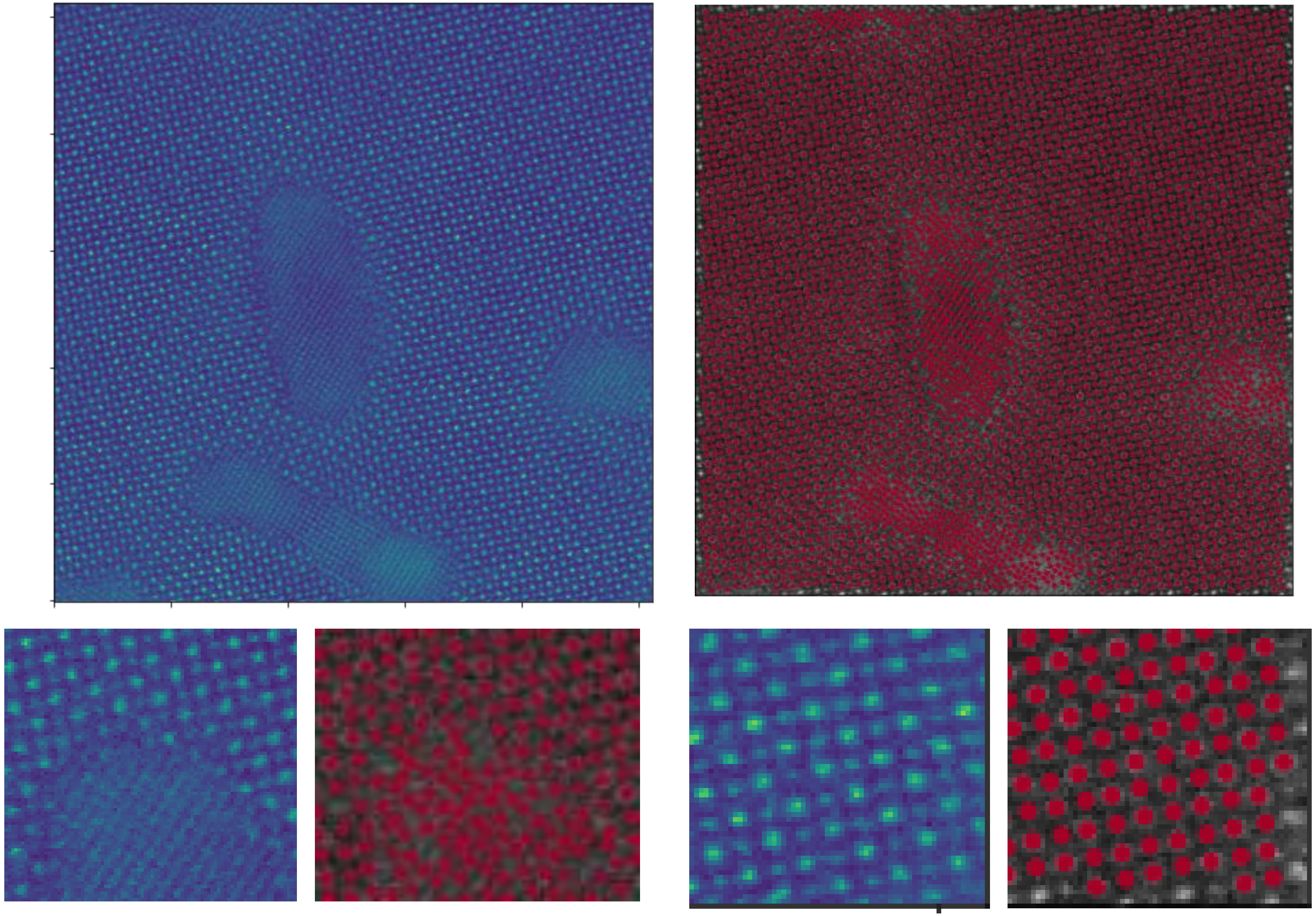
Atom based descriptions



Localized sub-images

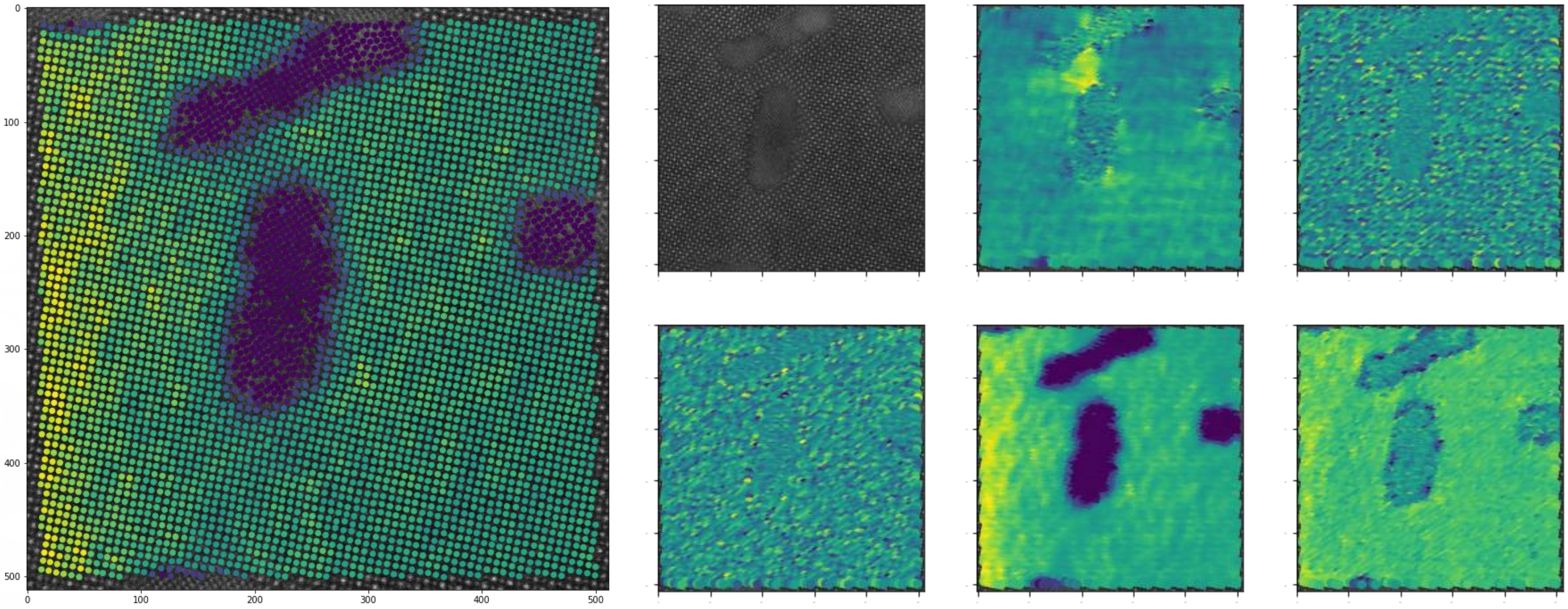


Let’s put it all together!



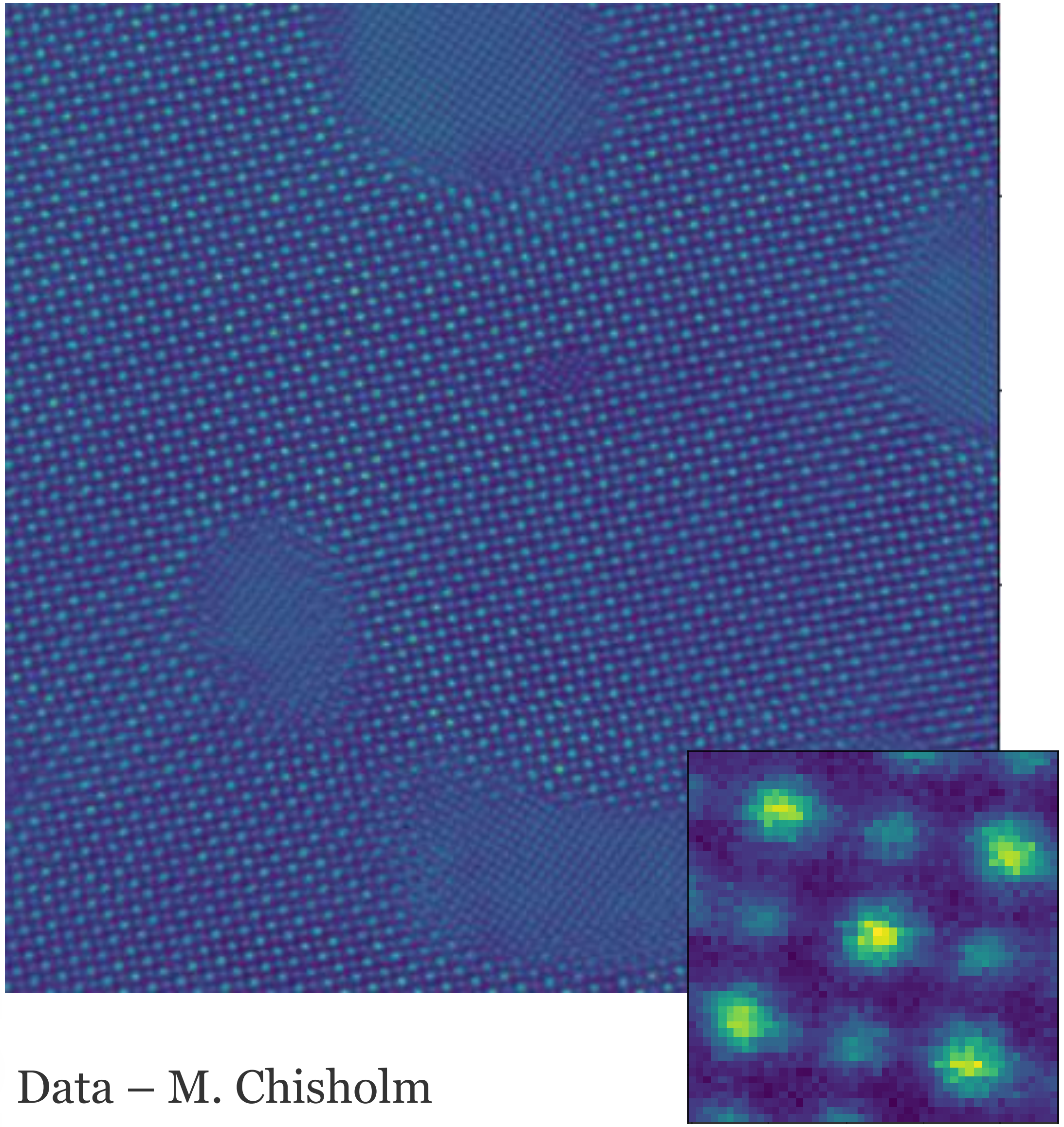
Step 1: Find all atoms (or all that you can) – use maximum finders, blob-log, or DCNNs
Step 2: Create descriptors – patches centered on atoms. Keep track on what part of image (or stack) it came from

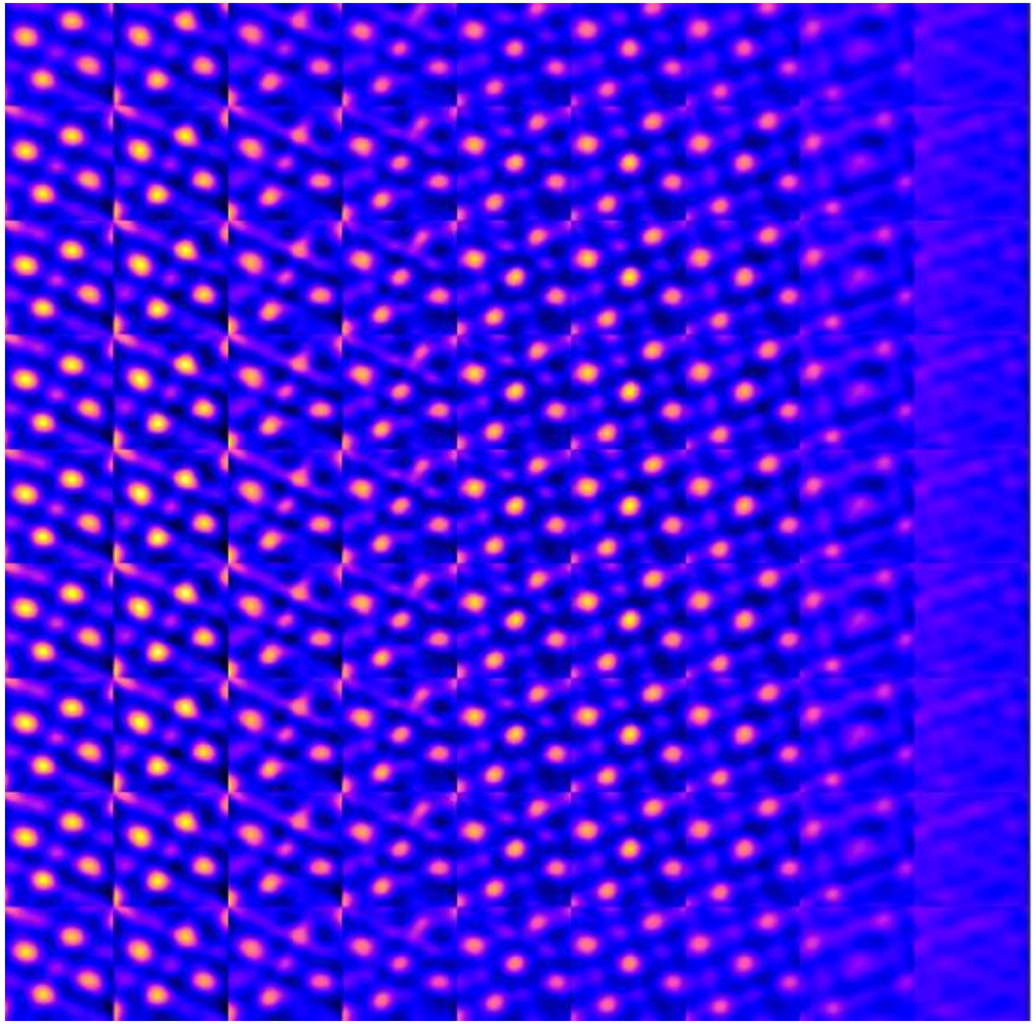
Step 3: rVAE



Output: Latent variable corresponding to local structure of each atomic site. Can be visualized on top of the original atomically resolved image, or as 2D maps (but – not rectangular array!)

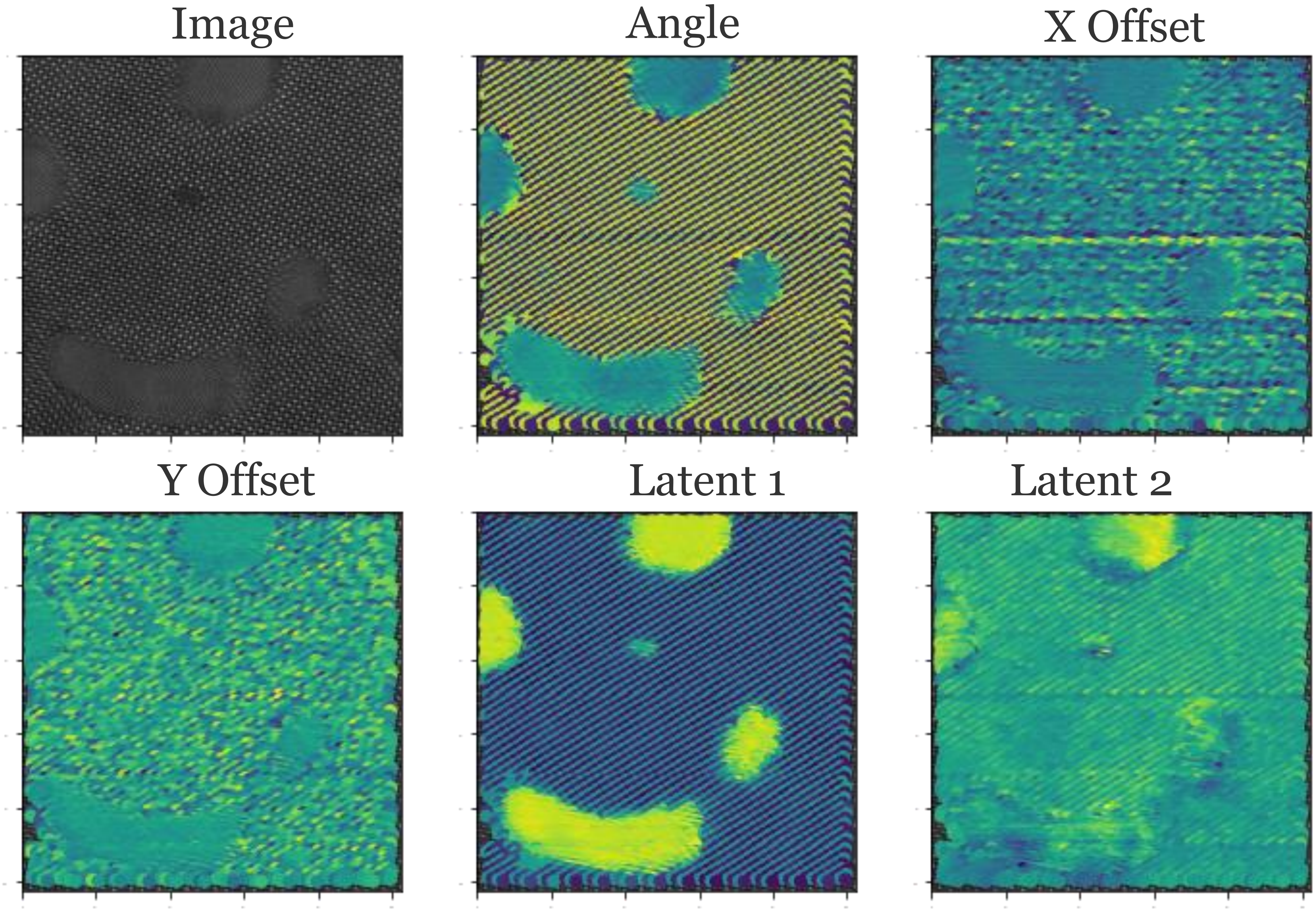
Analysis of the NiO-LSMO



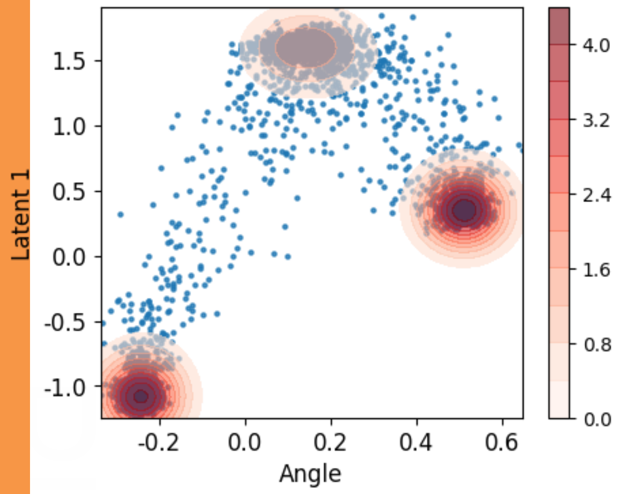


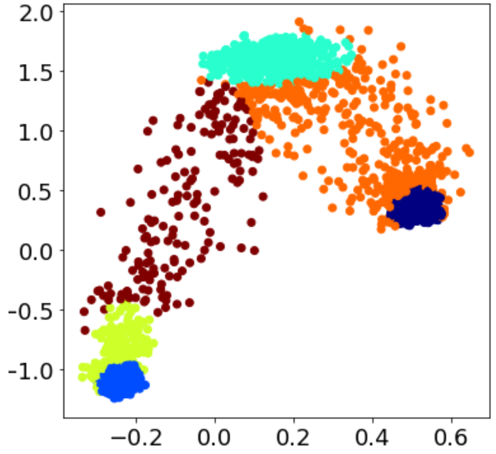
Data – M. Chisholm

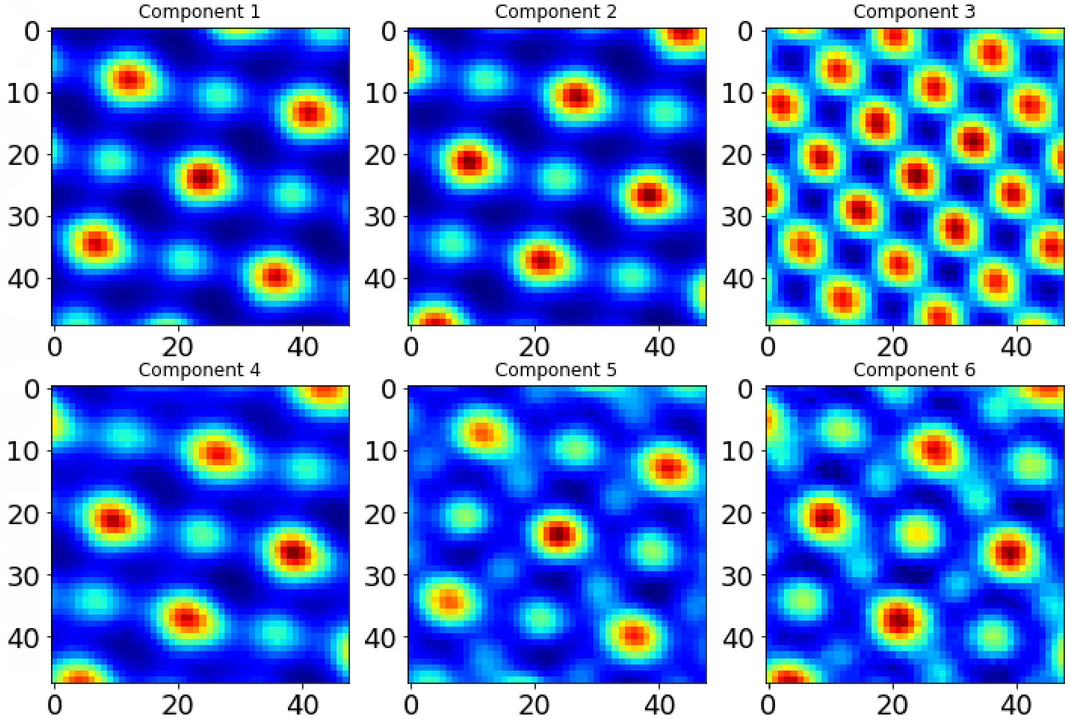
Let’s look at latent space



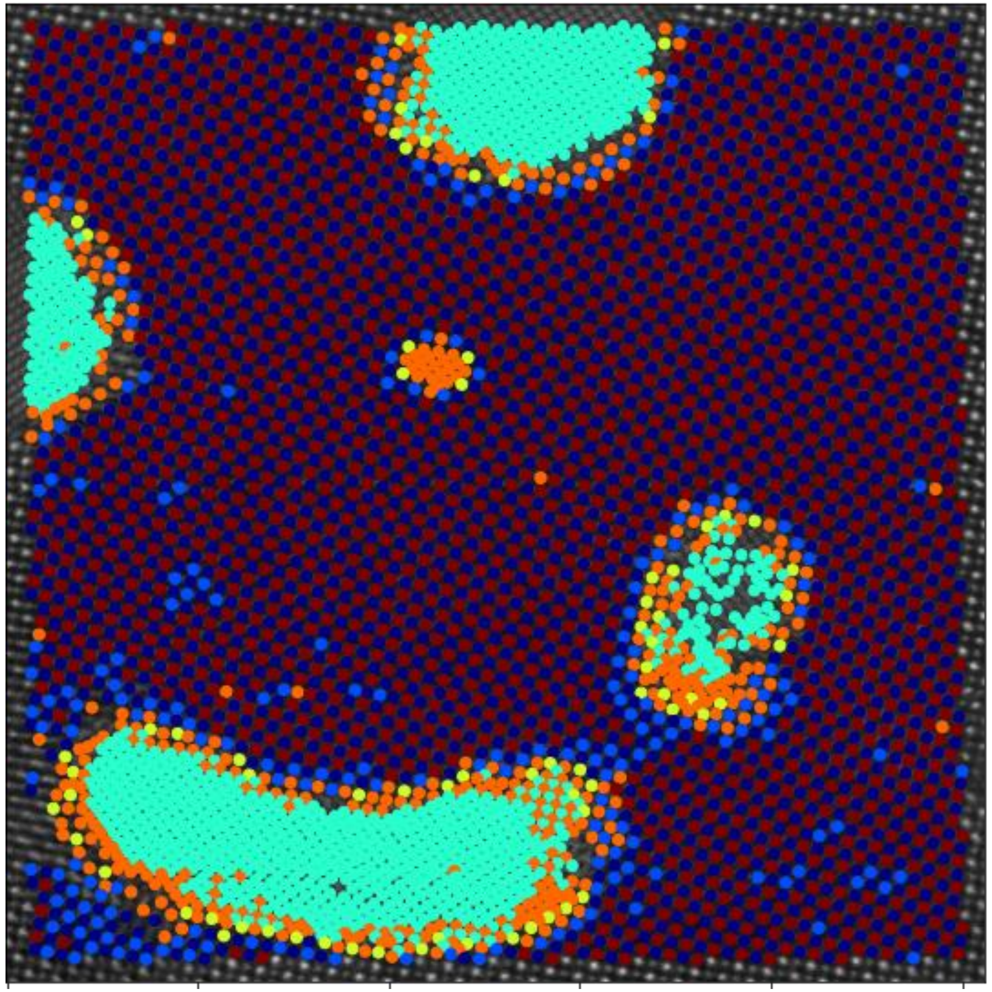
Exploring latent distributions







Labeled image

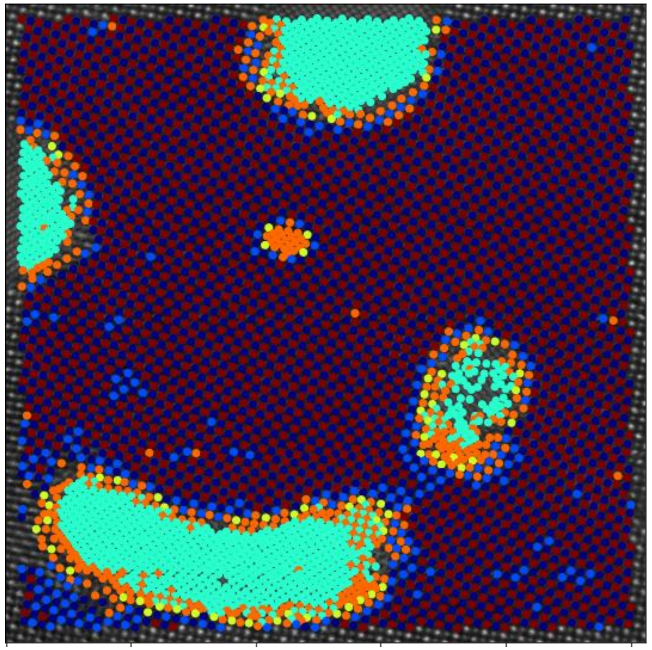


- Classes and variability are mixed in a single latent space
- Disentangling of representation

That’s where the jVAE has come from

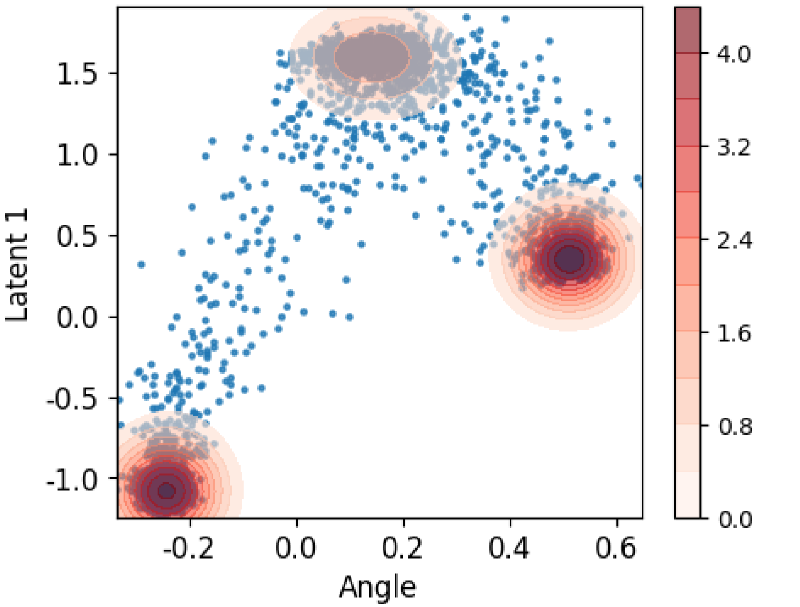
Currently, we have variants of invariant VAE that include:

- Convolutional or dense layers (reconfigurable via **kwargs)
- Rotational invariance
- With and without offsets (as latent variables)
- Multilayer inputs

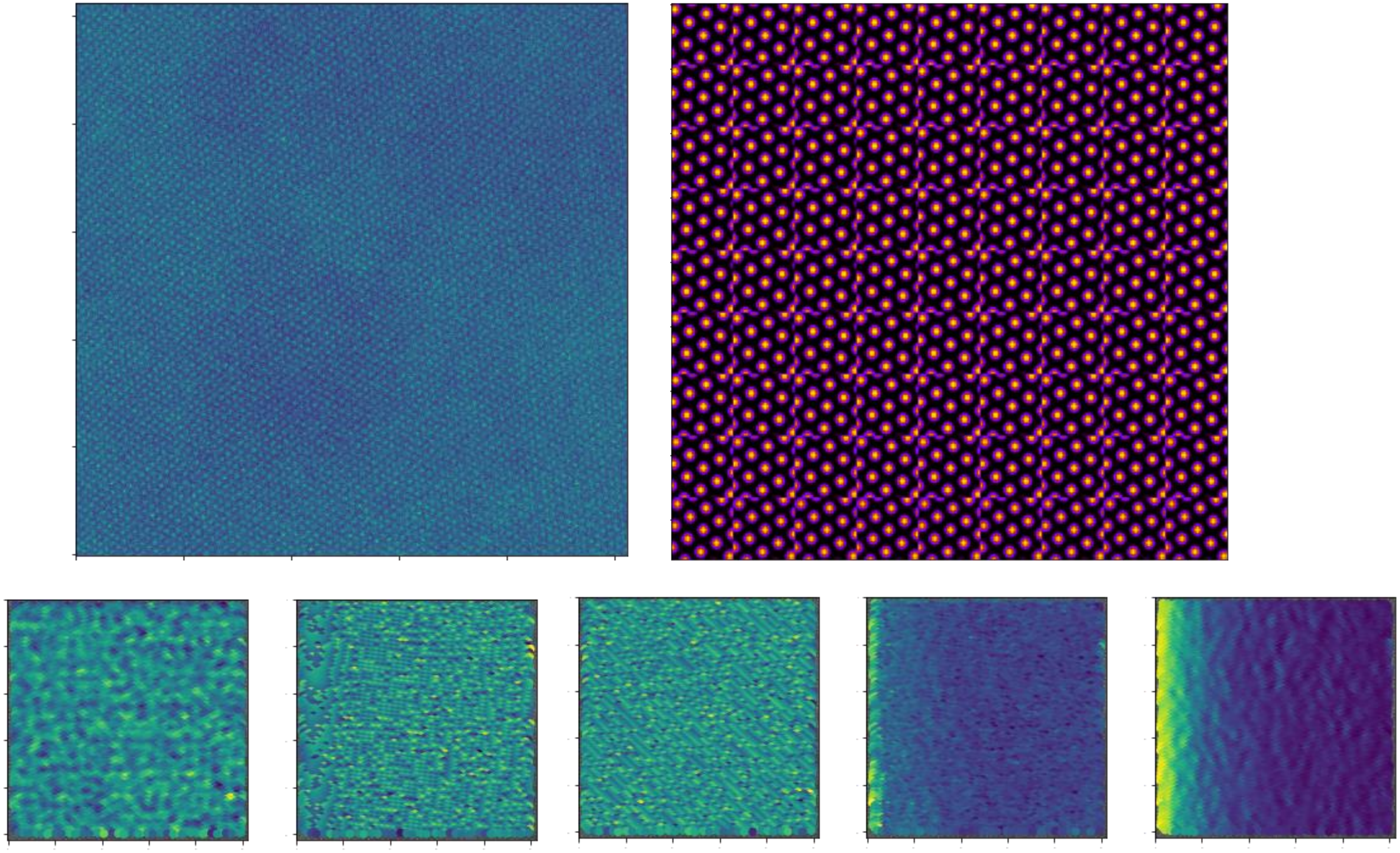


However, our rVAE collects everything in a single latent space. Realistically, very often we deal with system where we expect the presence of finite number of classes that may be known, partially known, or unknown, with certain continuous traits within classes.

- **Graphene and MX2:** structural units (discrete) and strain states
- **Crystalline solids:** phases and ferroic variants, strain states
- **Plasmonic EELS:** particle spectra, off-particle spectra, edge states
- **CITS:** lattice and defects, strain states

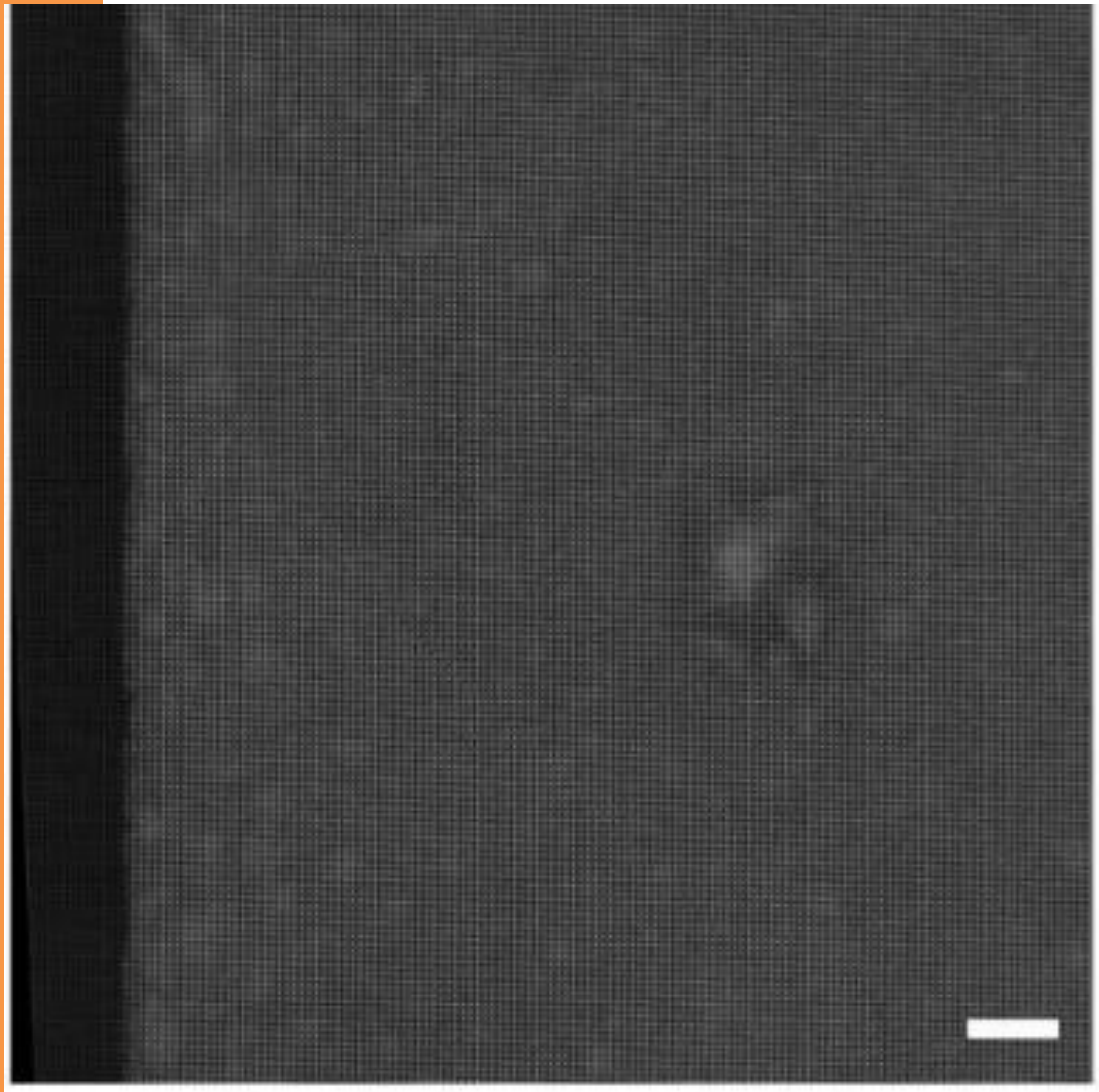


Out of curiosity: single crystal?

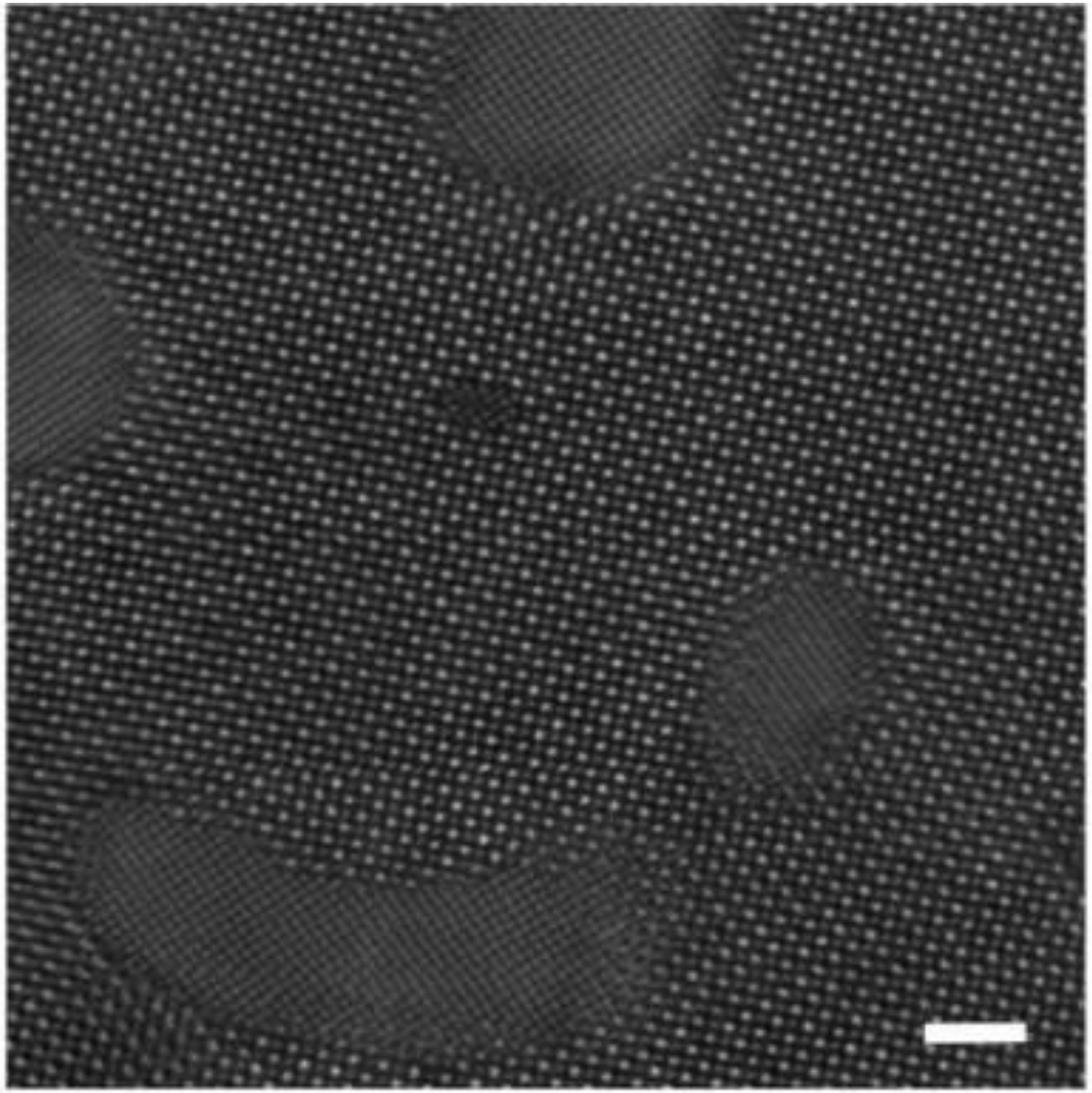


VAE without Atom Finding

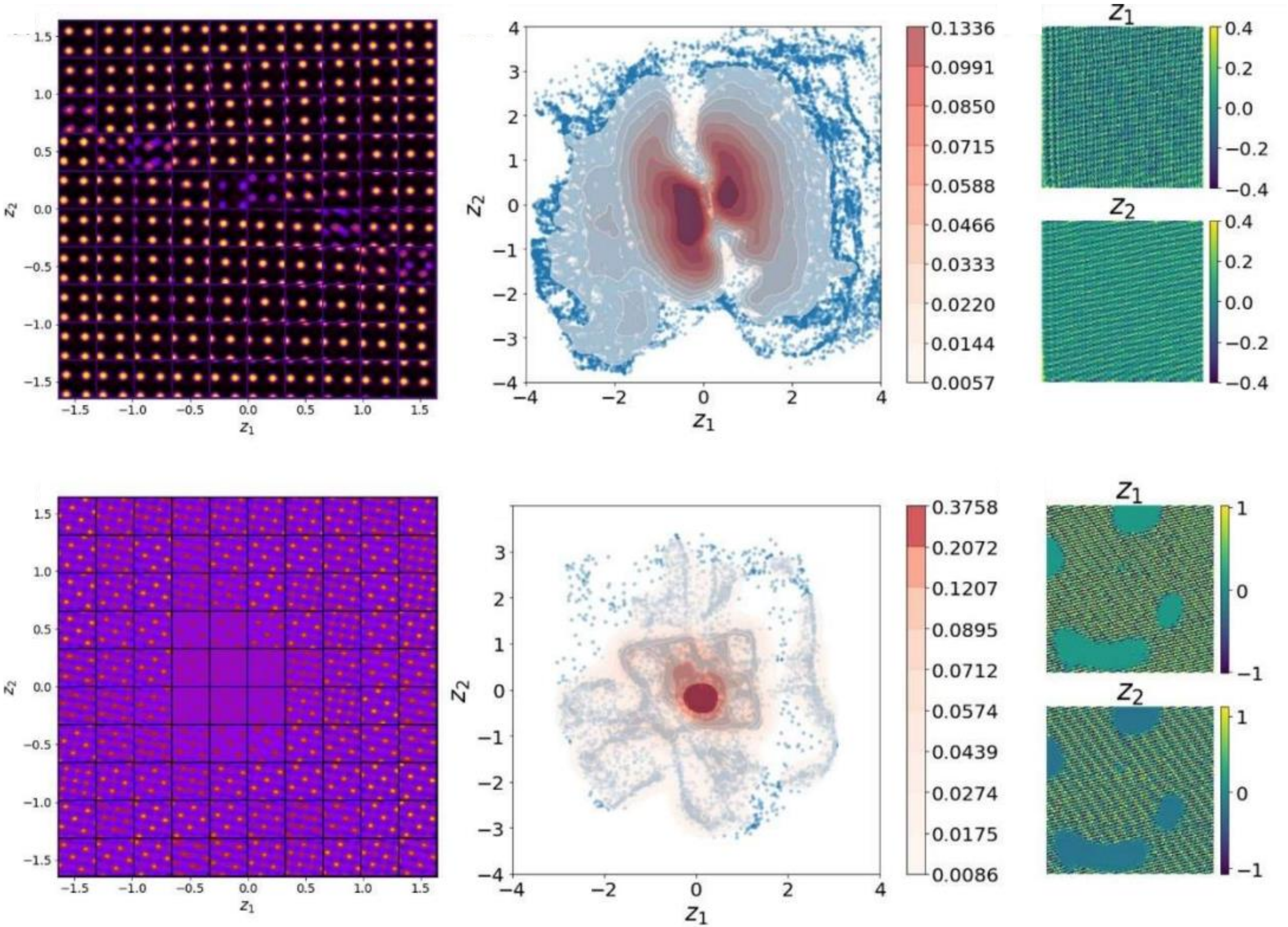
Ferroelectric $BiFeO_3$



$NiO - La_xSr_{1-x}MnO_3$

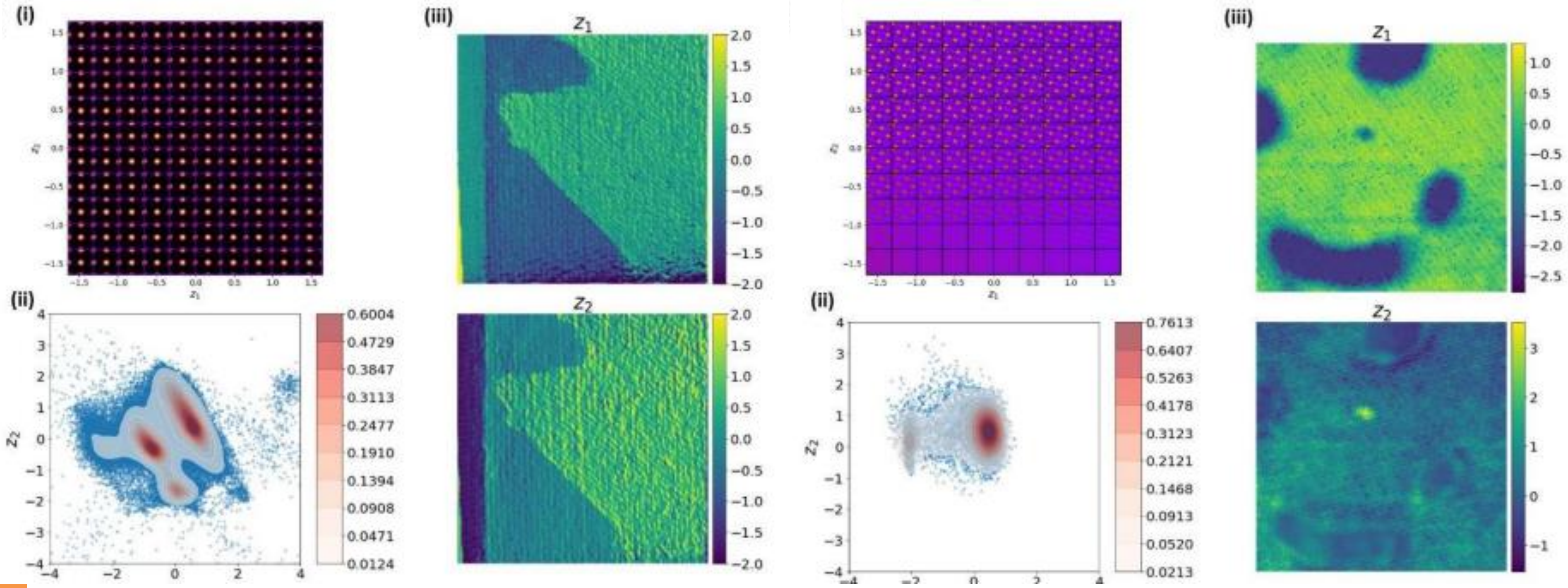


Simple VAE

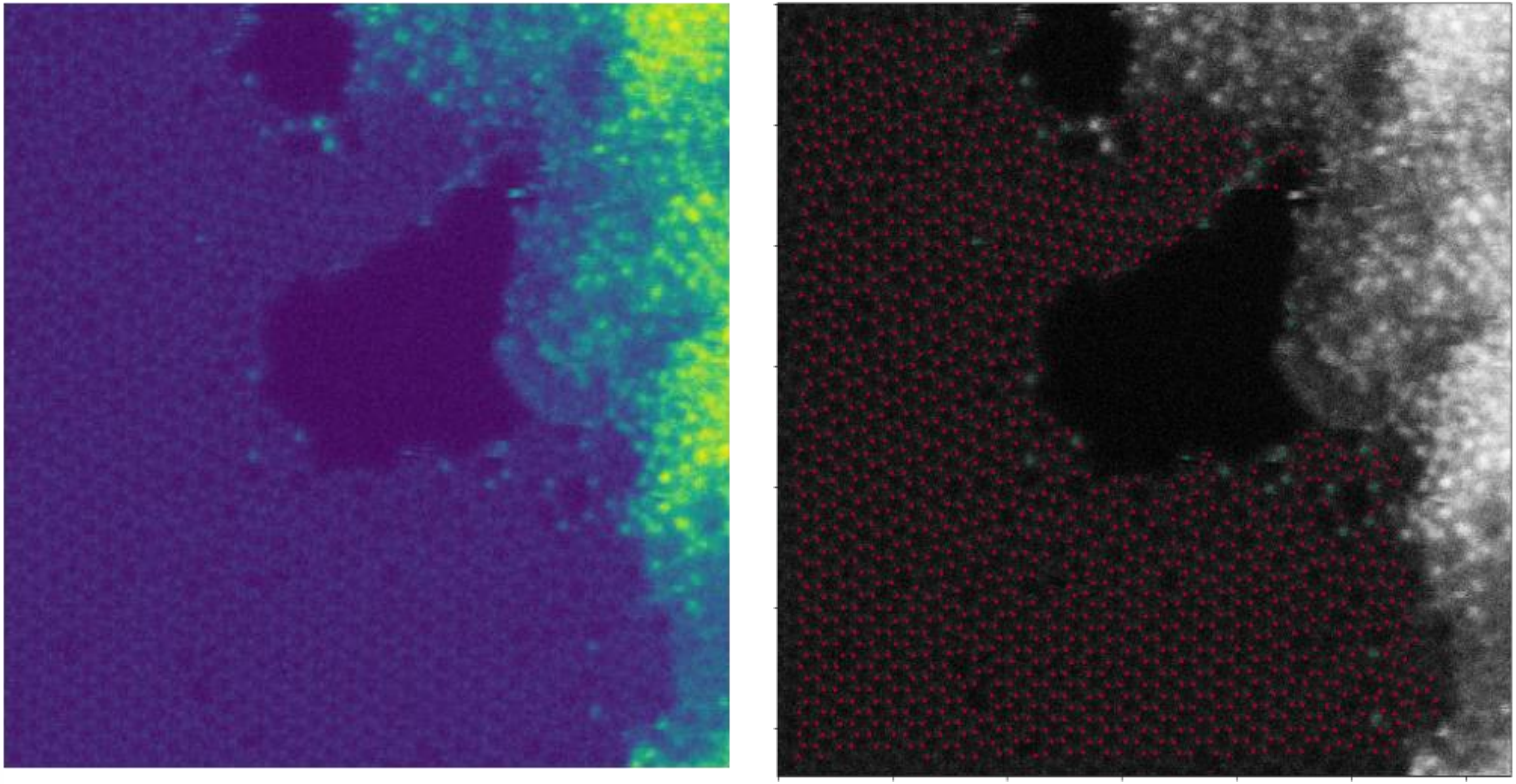


z_1
z_2
0.1336
0.0991
0.0850
0.0715
0.0588
0.0466
0.0333
0.0220
0.0144
0.0057
0.3758
0.2072
0.1207
0.0895
0.0712
0.0574
0.0439
0.0274
0.0175
0.0086

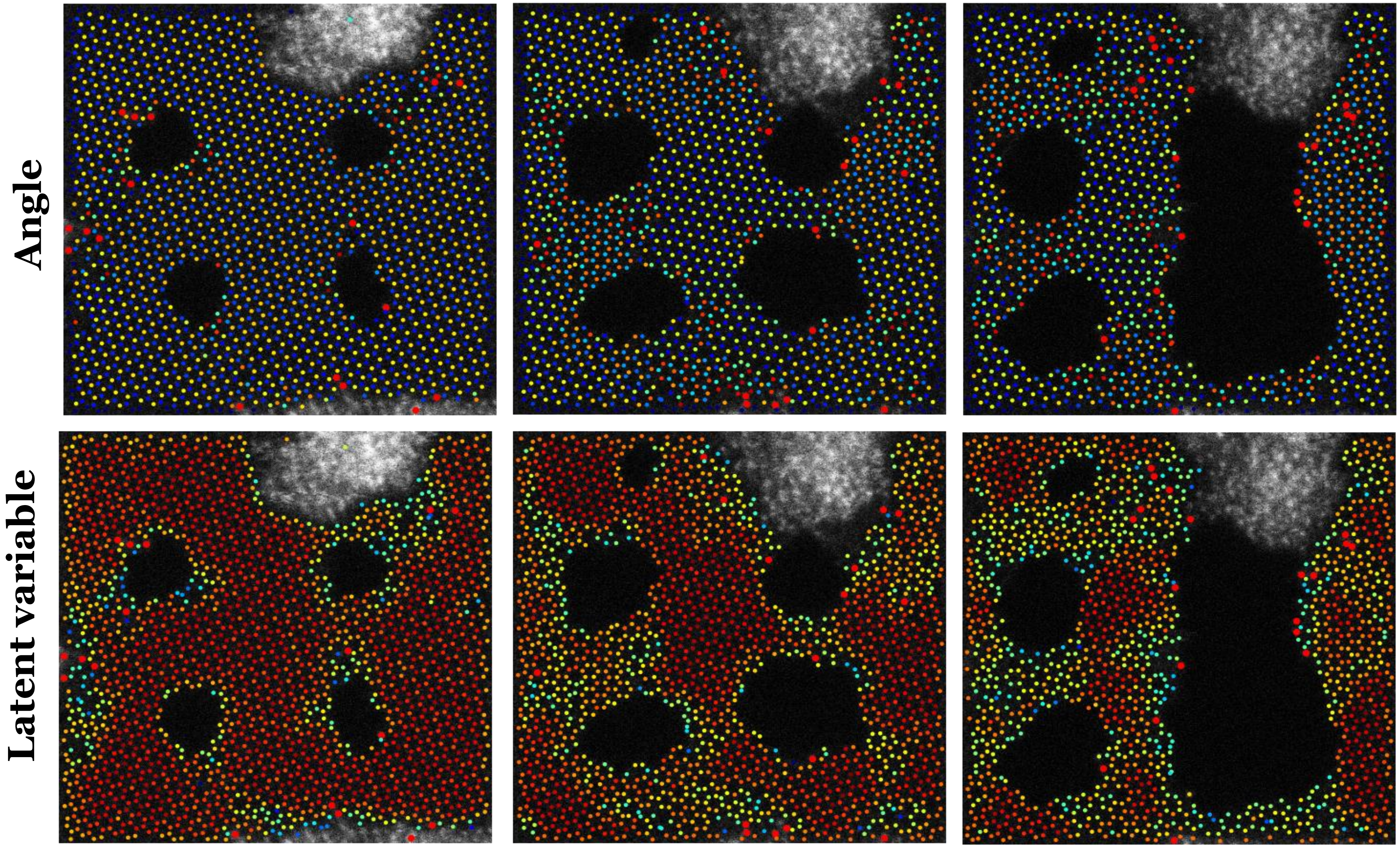
Shift VAE: Translational Invariance



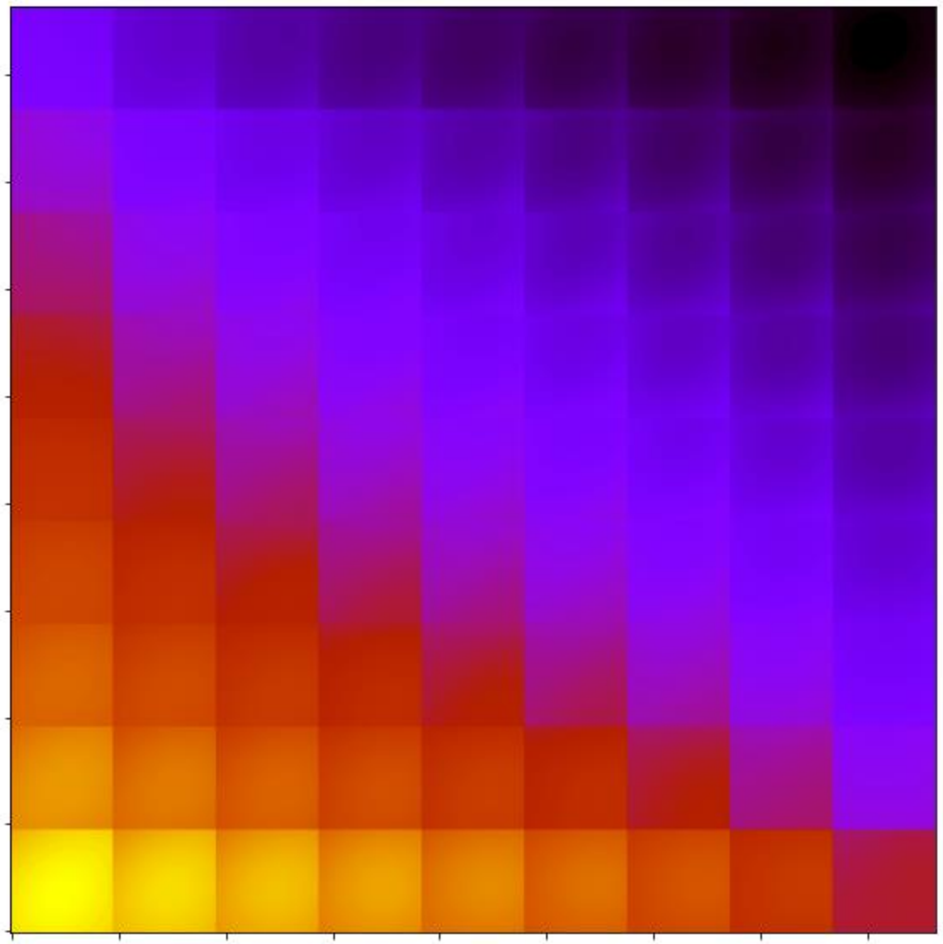
Off to chemically-disordered systems

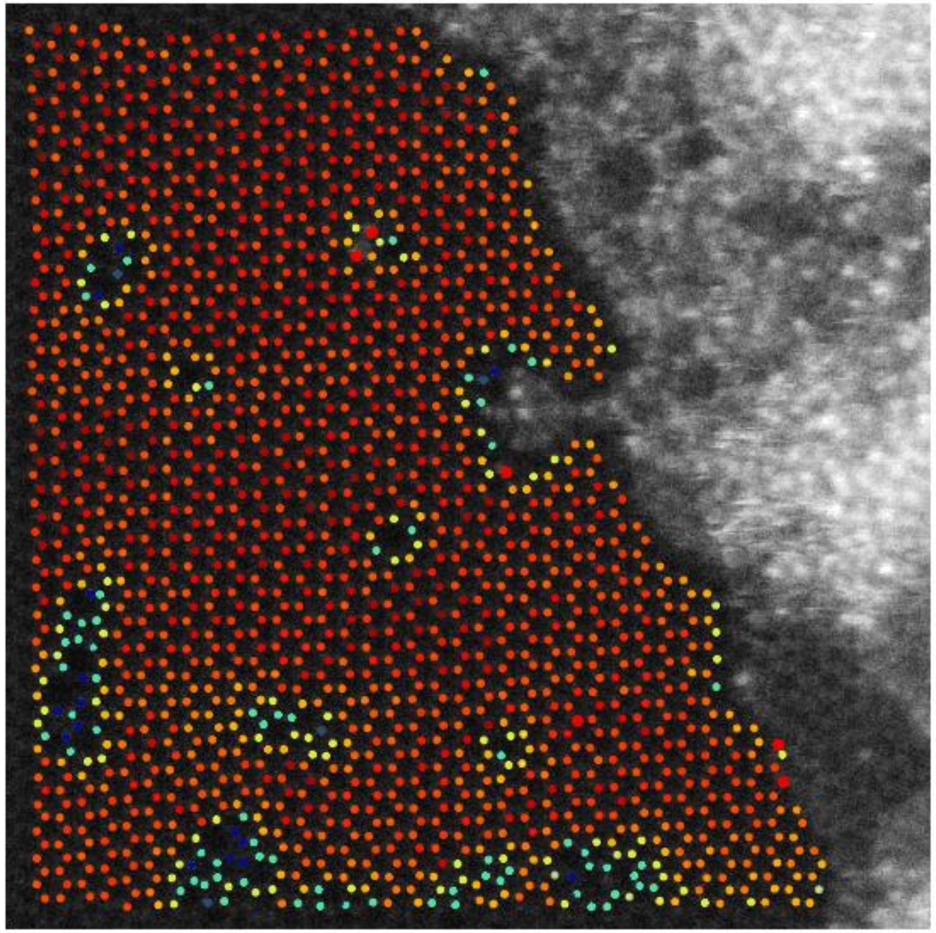


rVAE analysis at different time steps

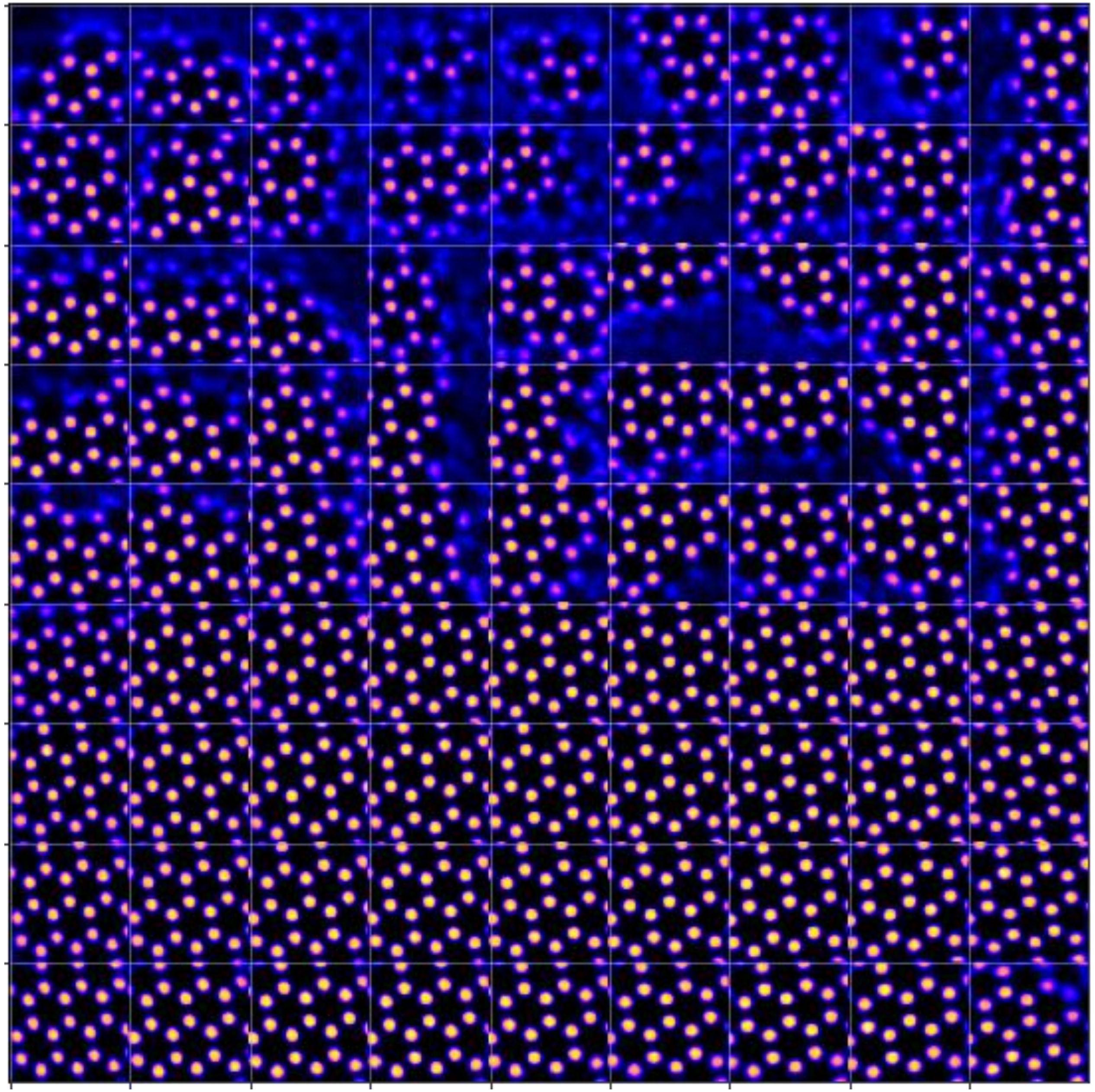


There is nothing as beautiful as training VAE

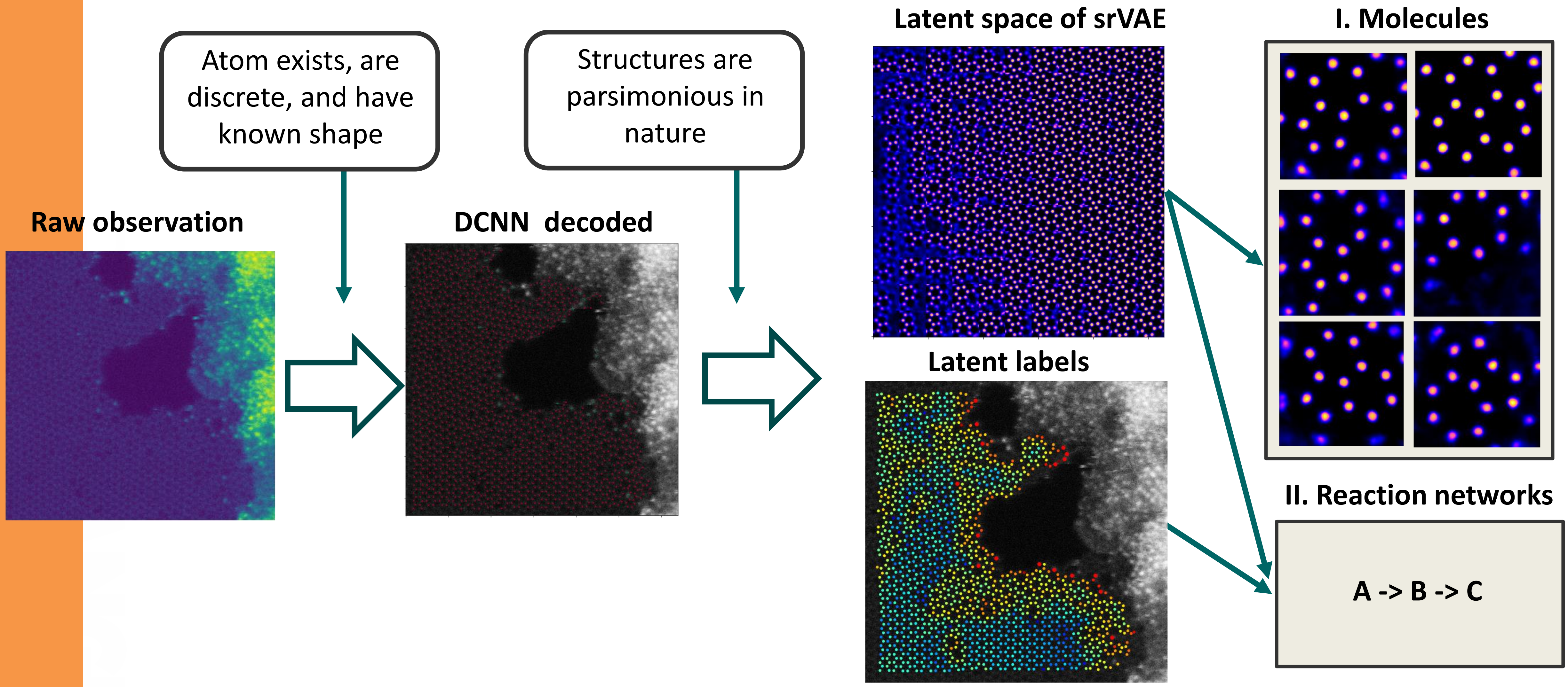




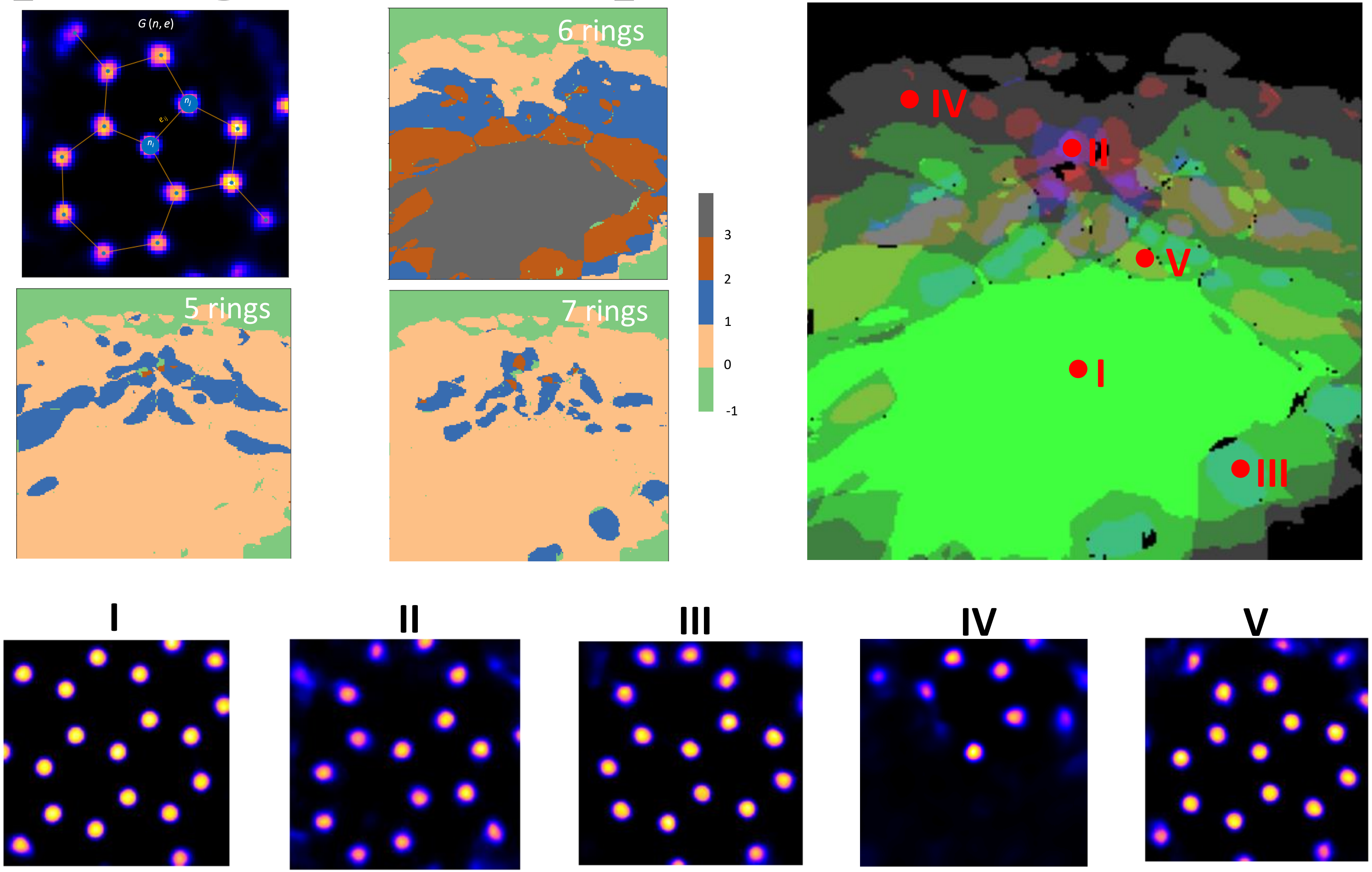
Next step: skip-rVAE



Unsupervised discovery of molecules

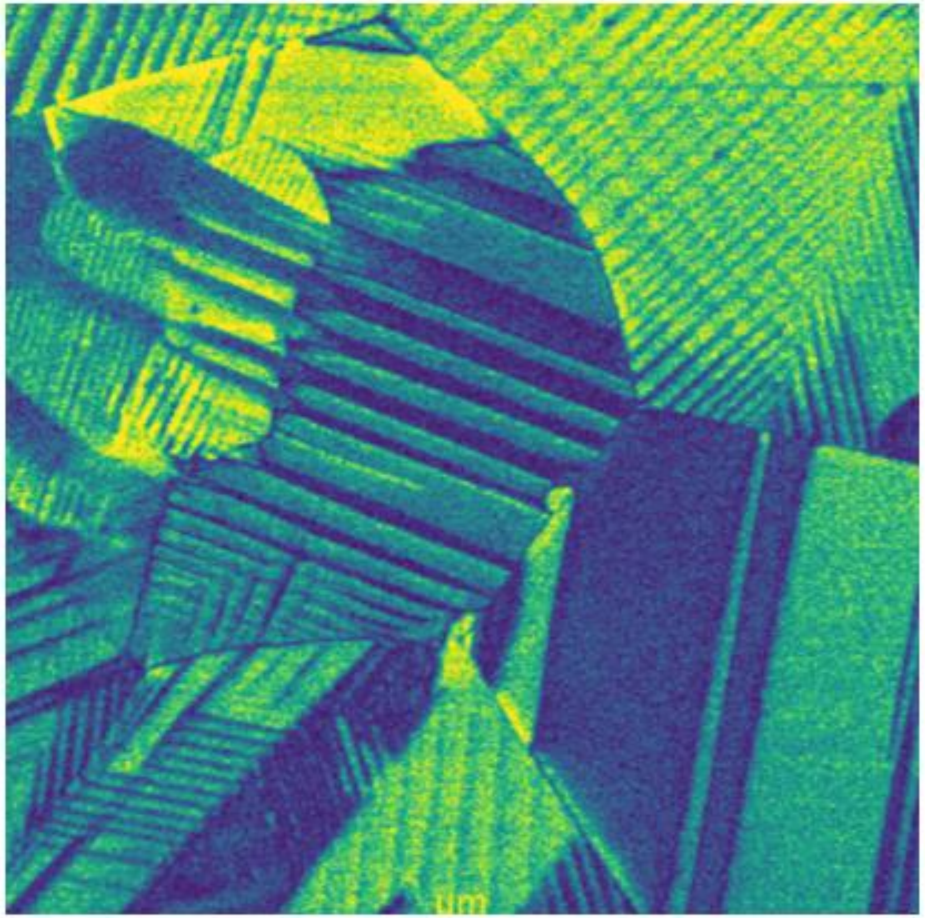


Exploring the latent space structure

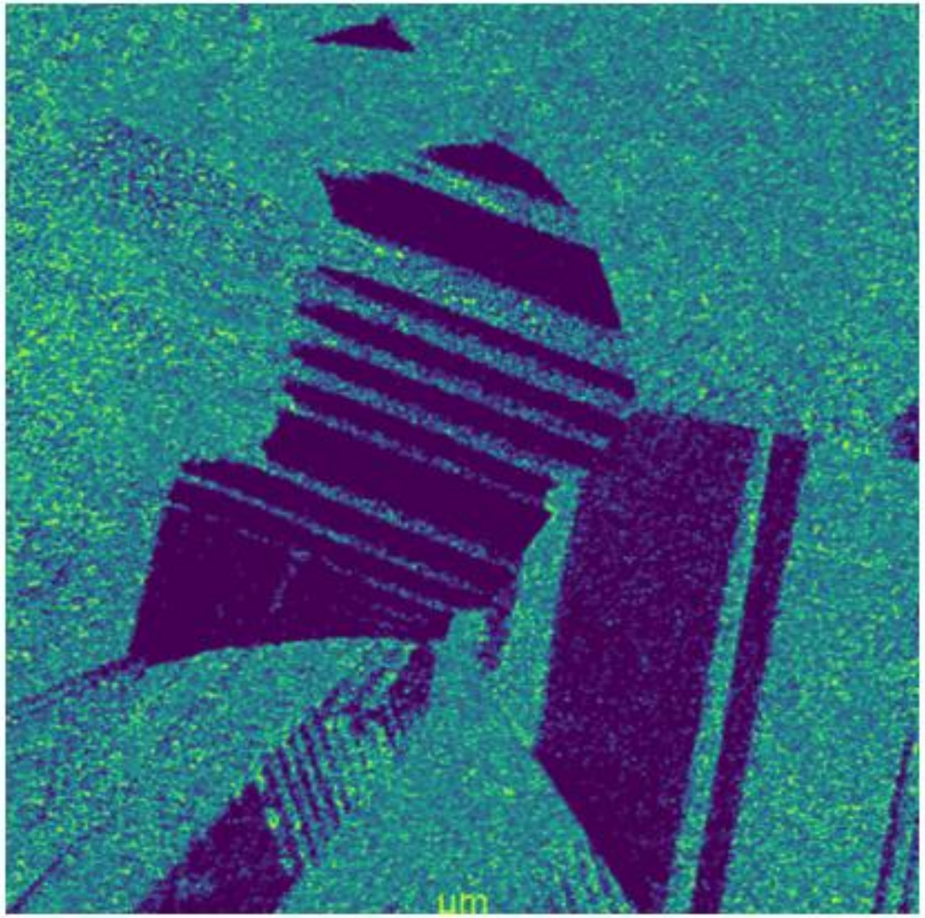


Ferroelectric domain and domain walls

Amplitude

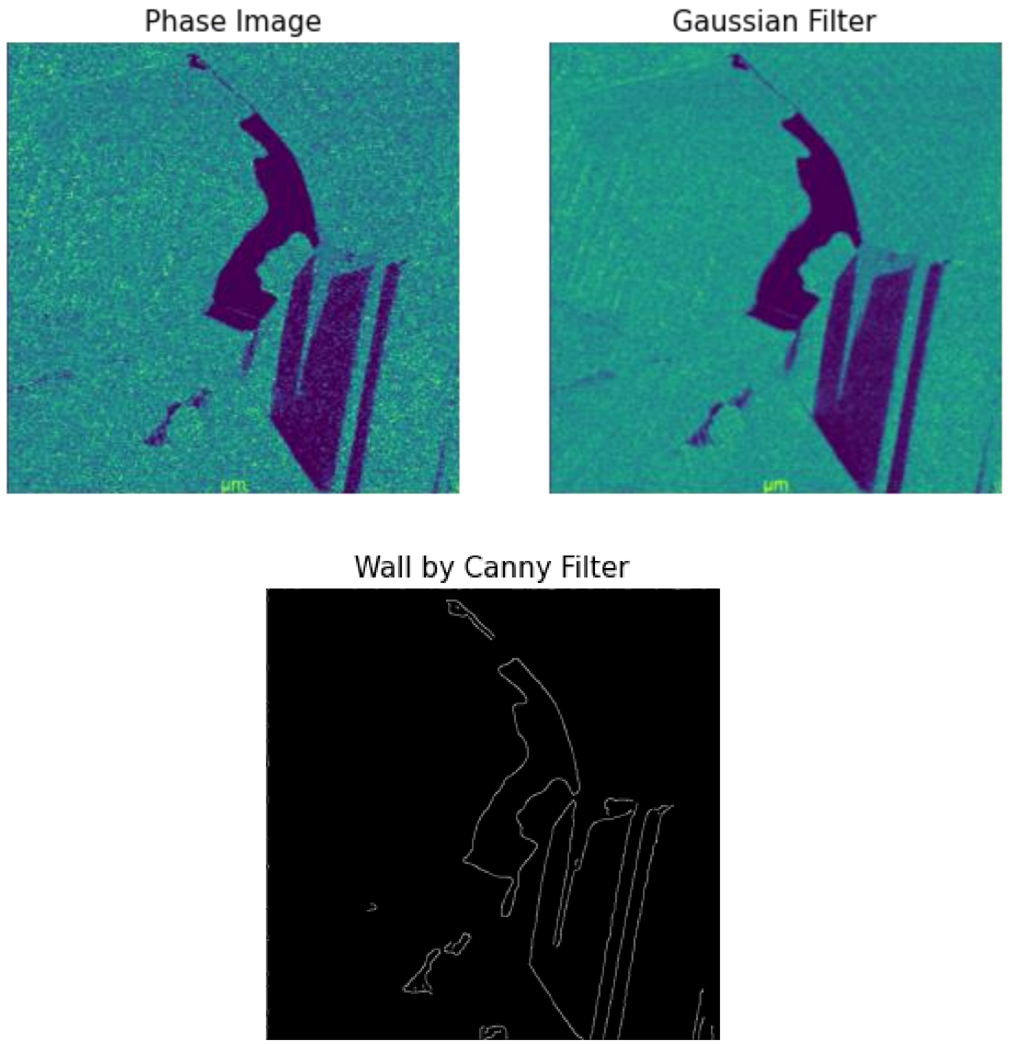


Phase

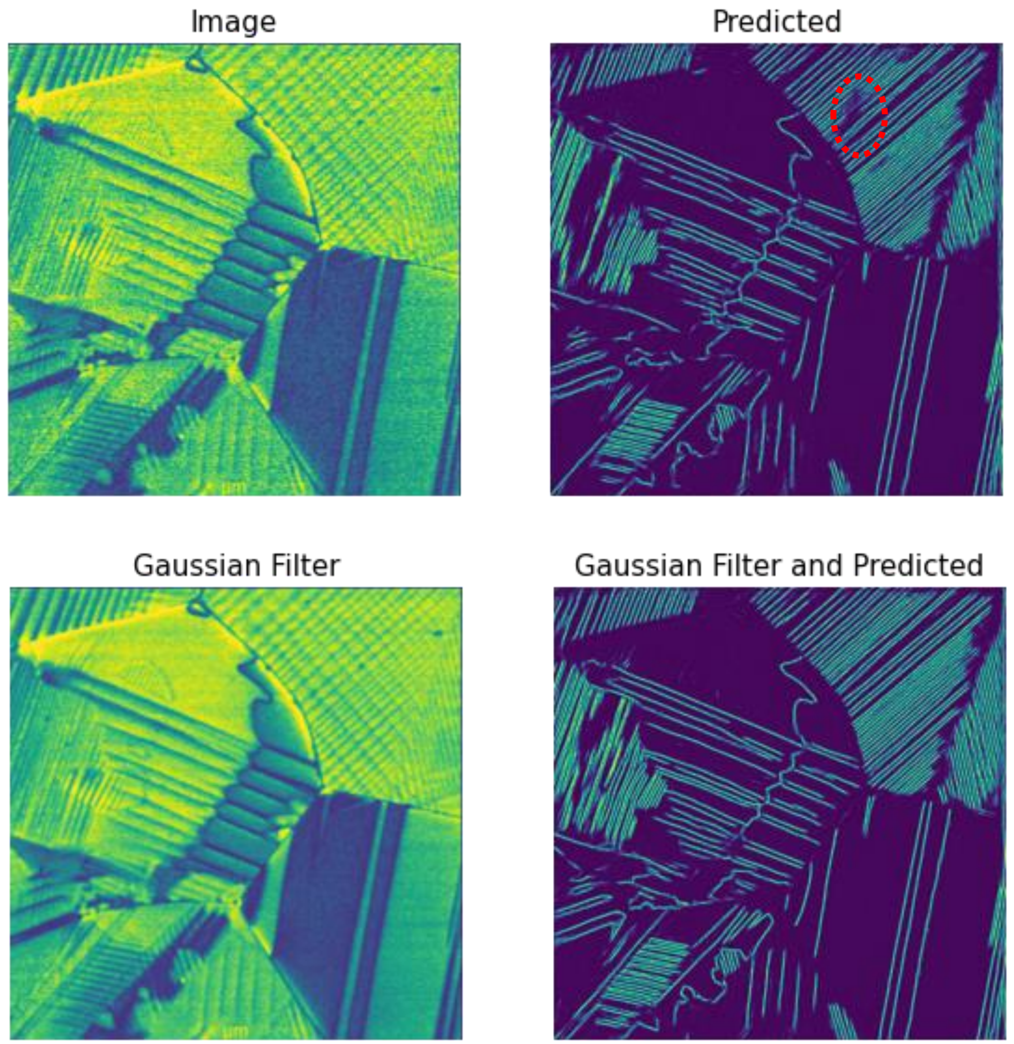


Detecting domain walls

Canny filter

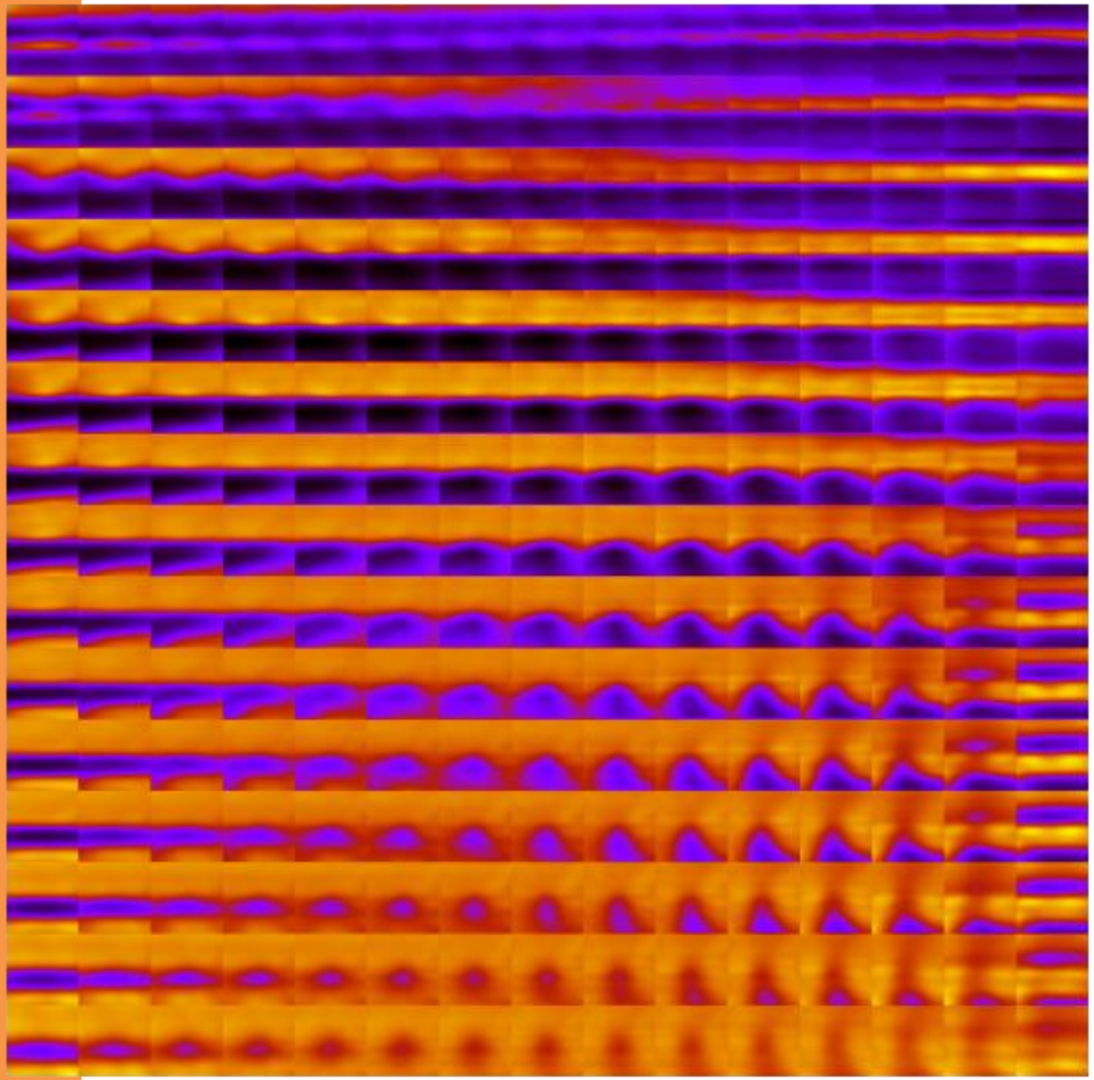


DCNN Prediction

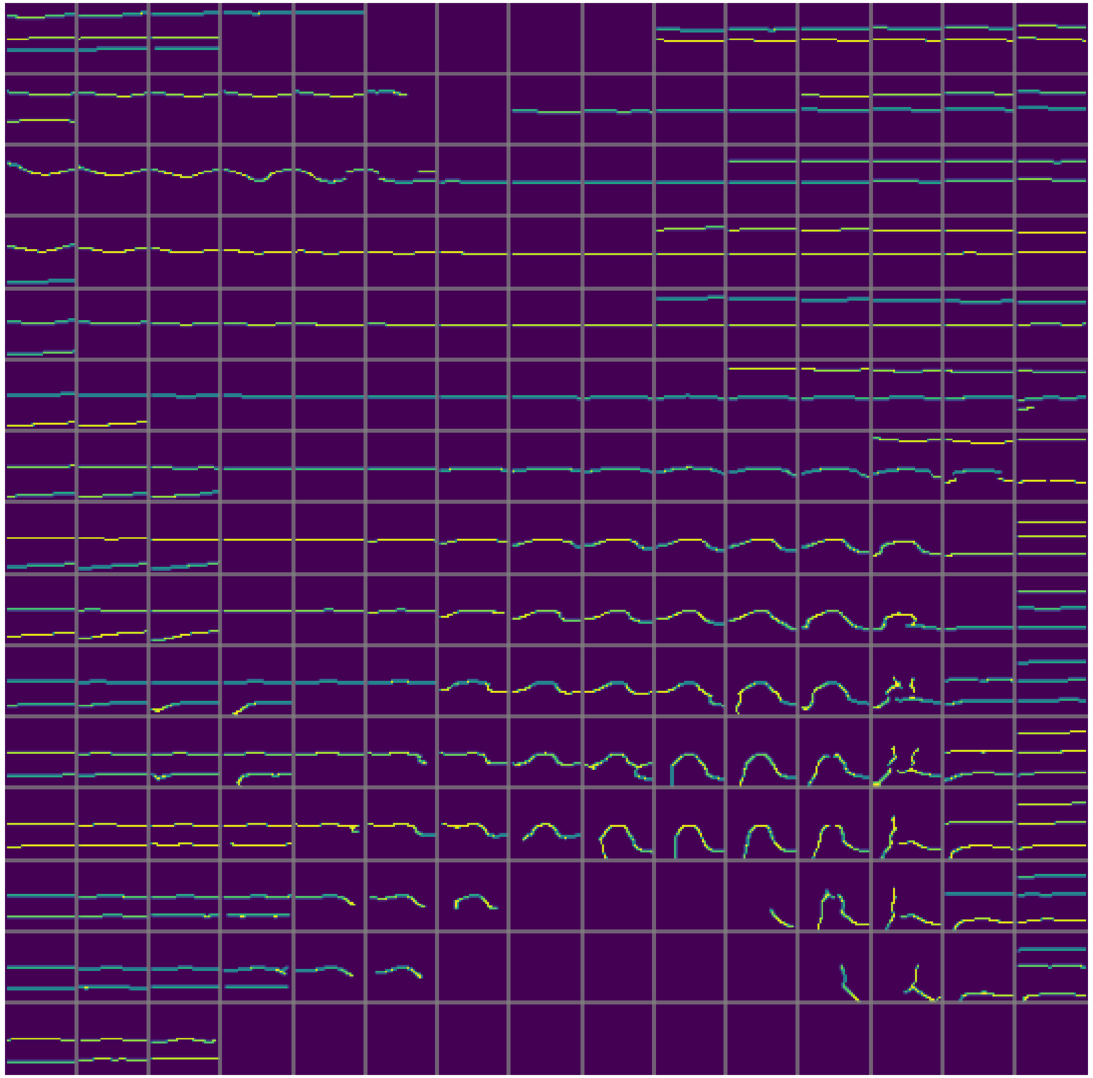


rVAE analysis

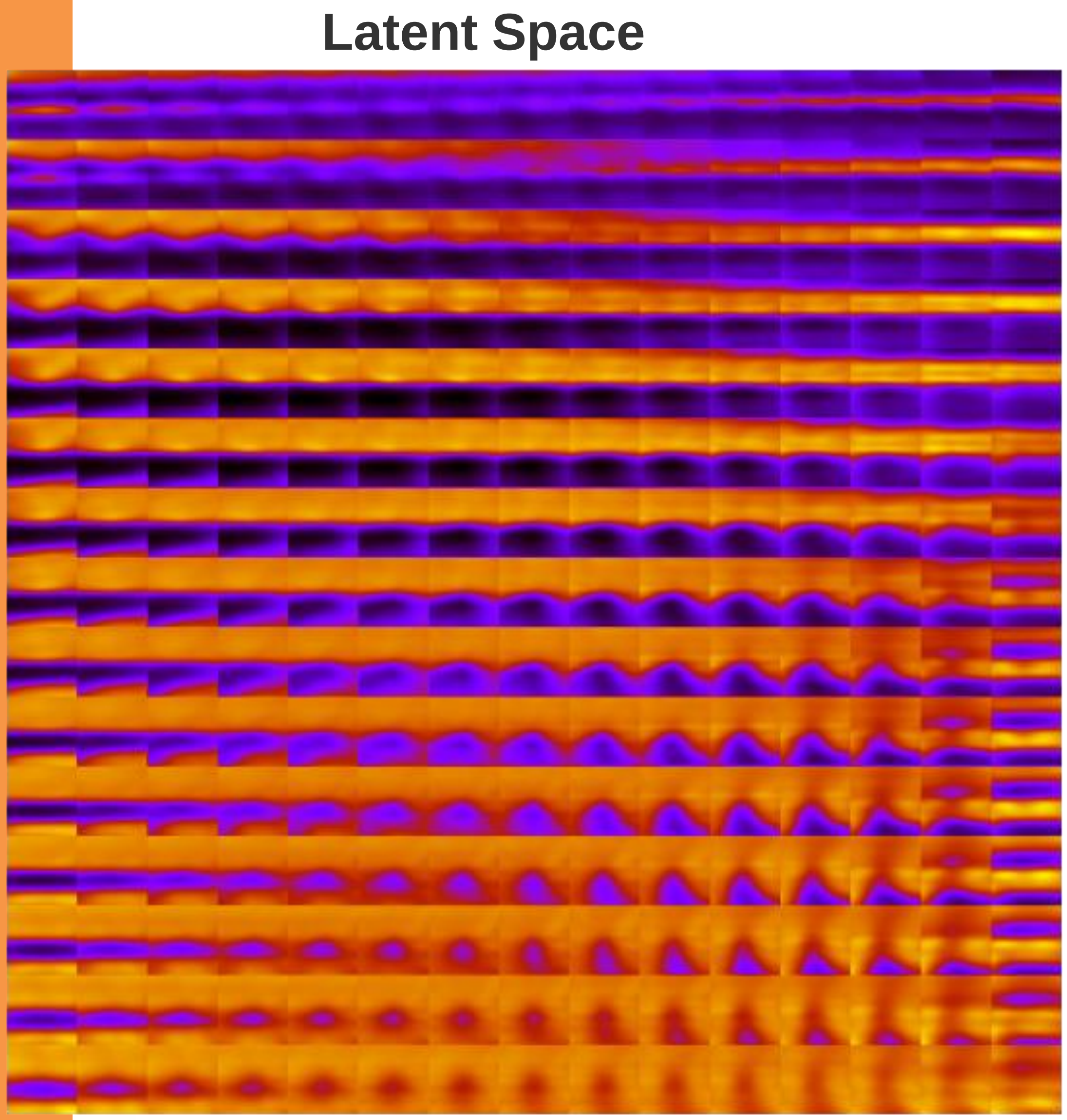
Latent Space

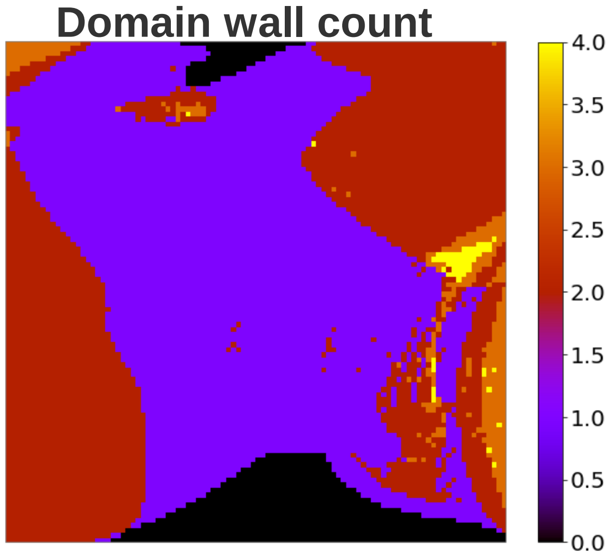


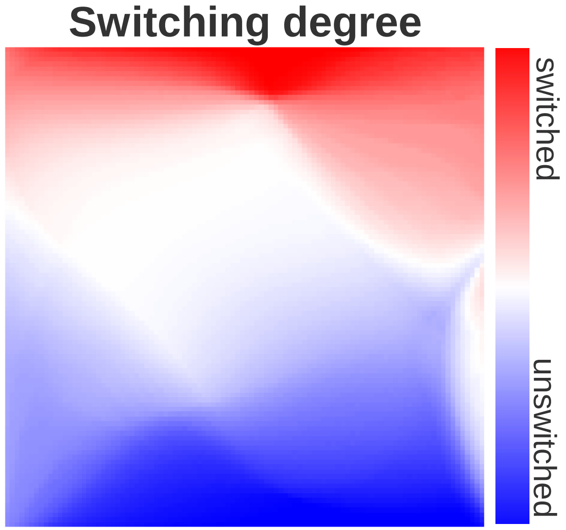
Domain Walls

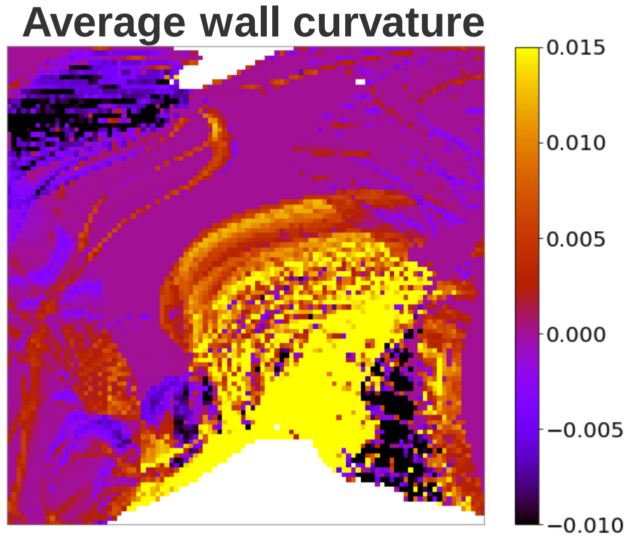


rVAE latent space

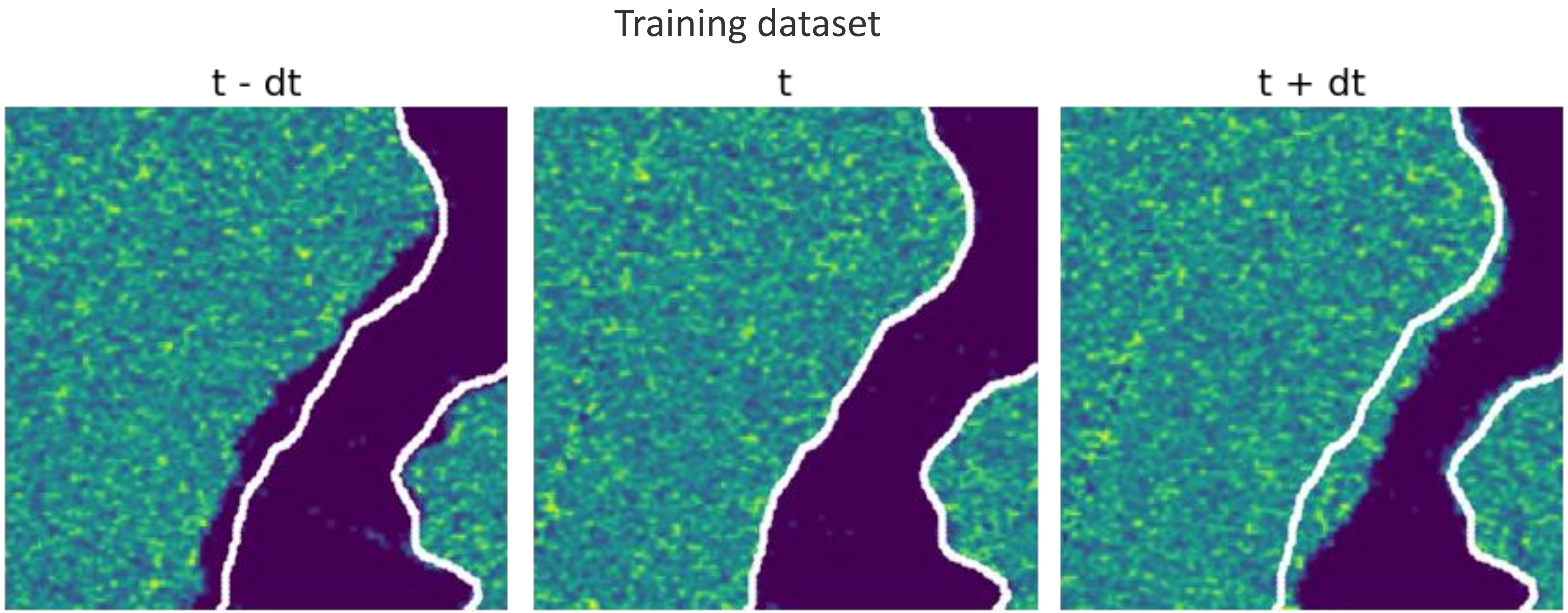






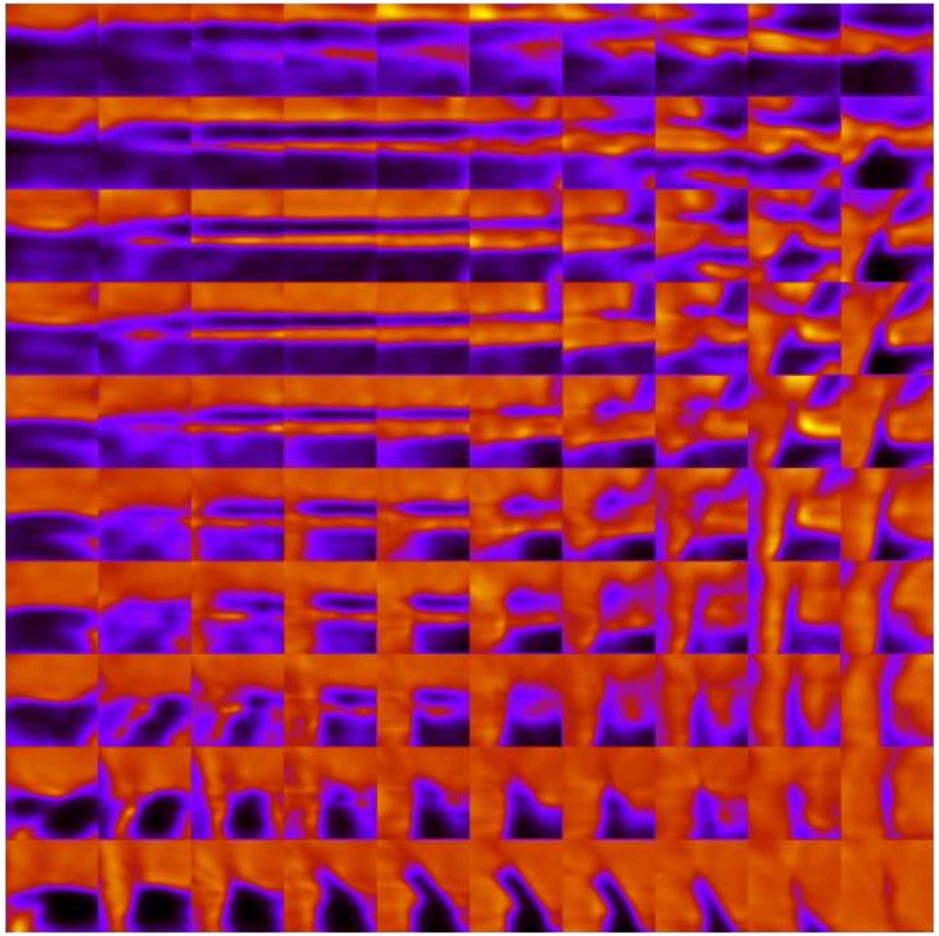


rVAE with time delay

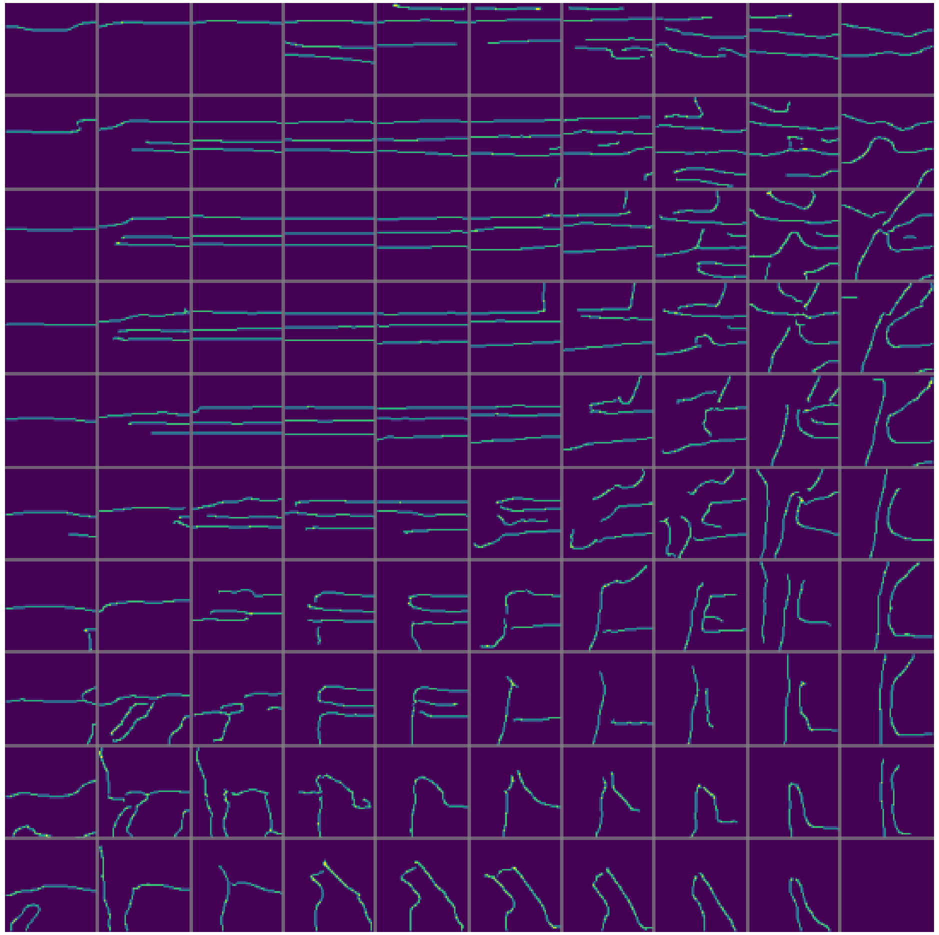


rVAE with time delay

Latent space

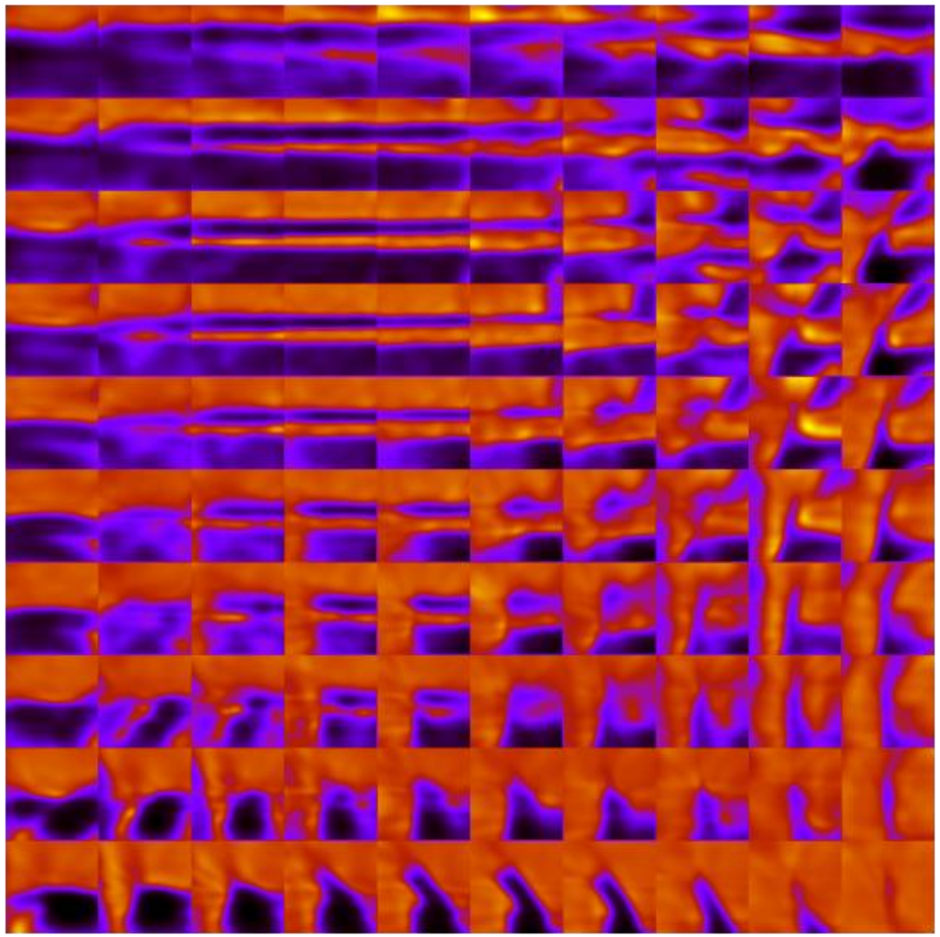


Domain wall

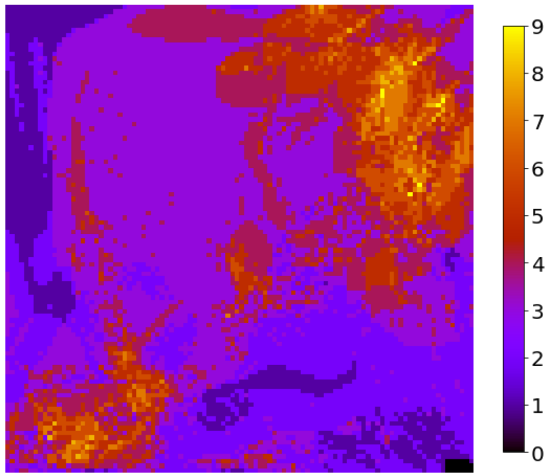


rVAE with time delay

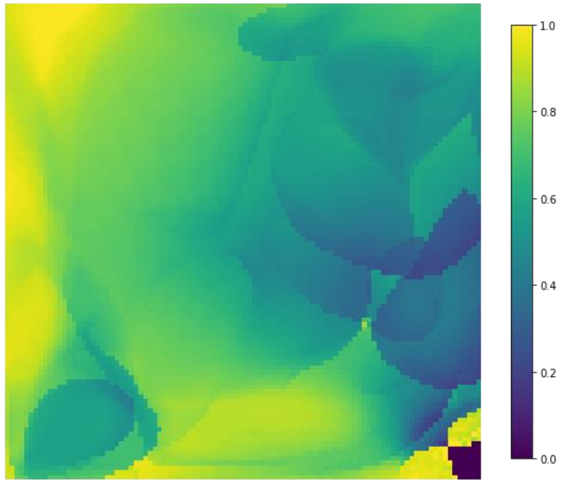
Latent space



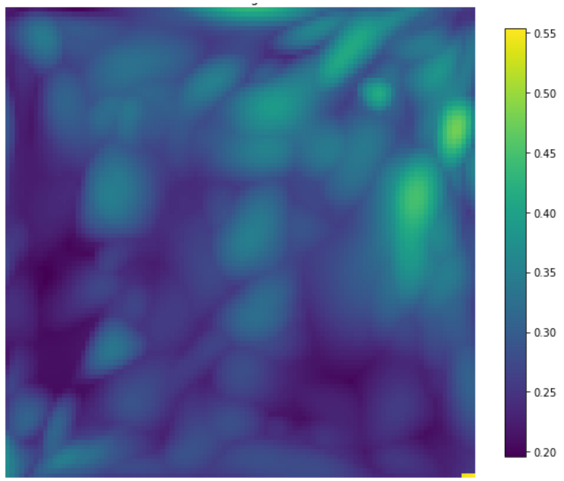
Wall count



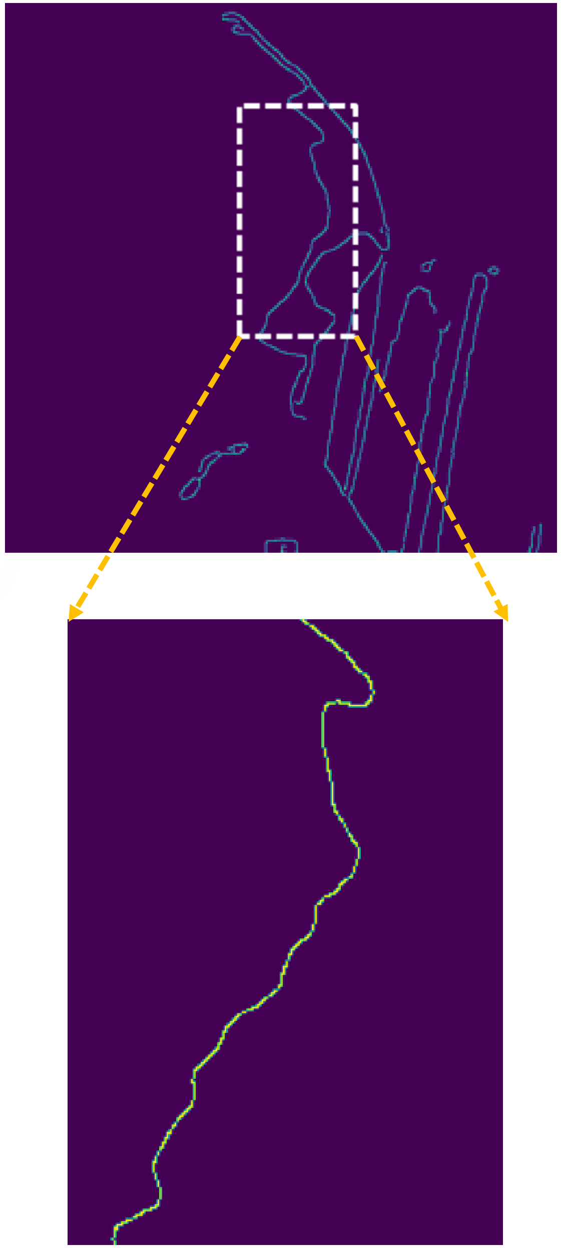
Domain convex

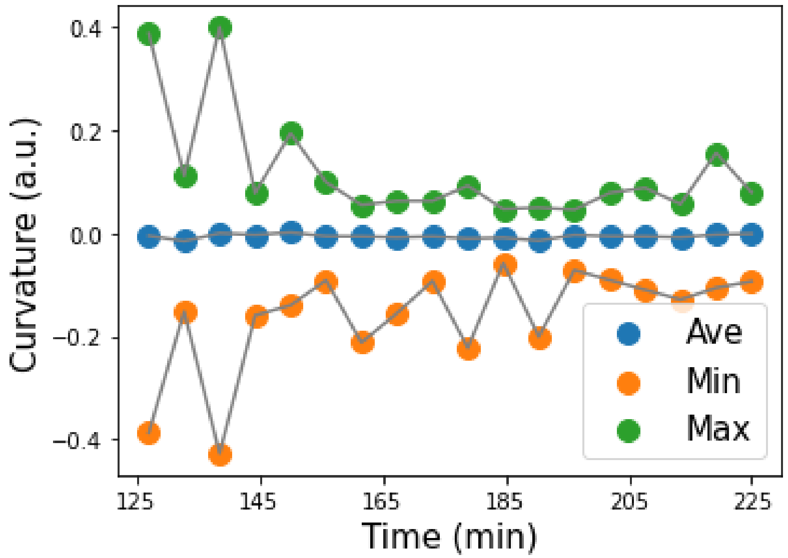


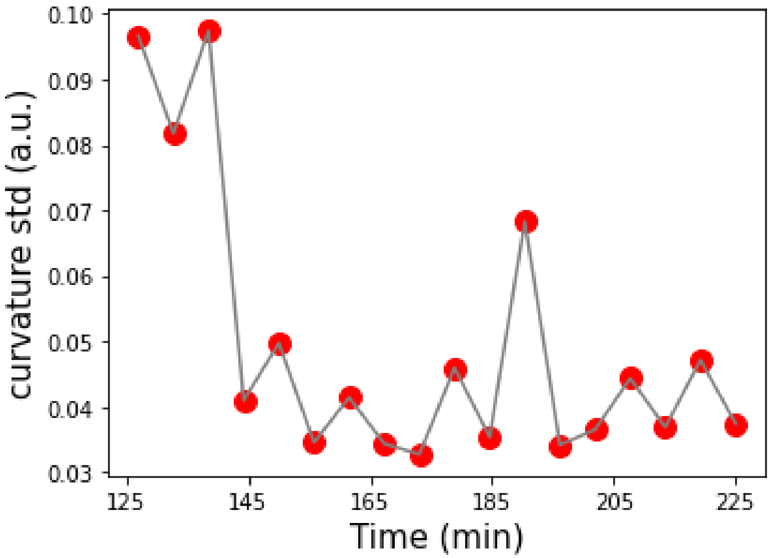
Switch significance



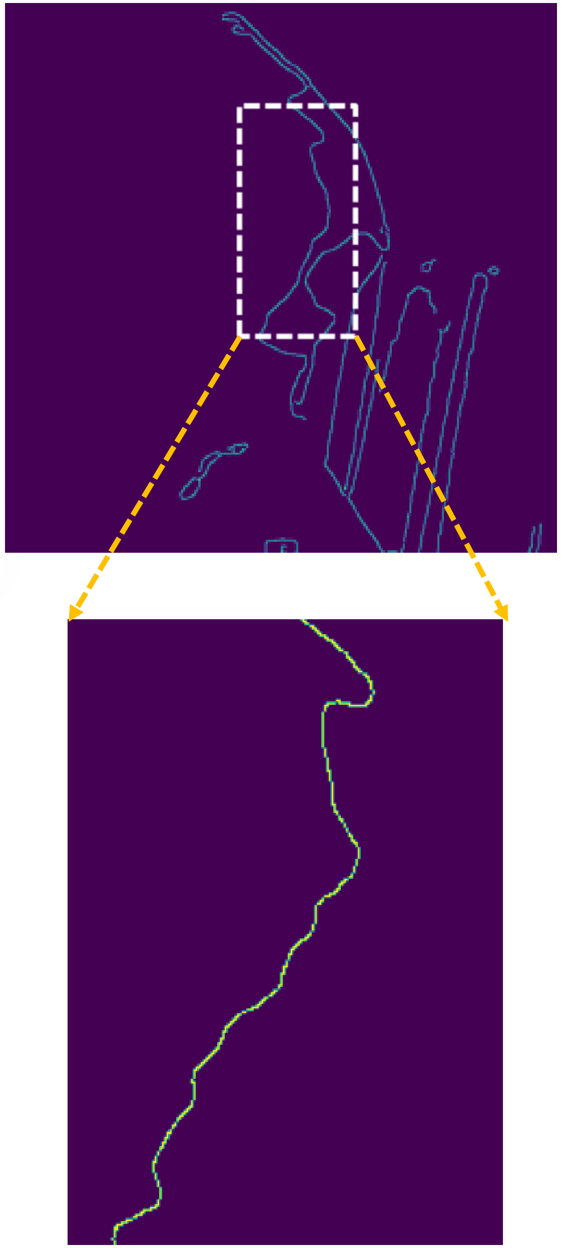
Domain wall evolution





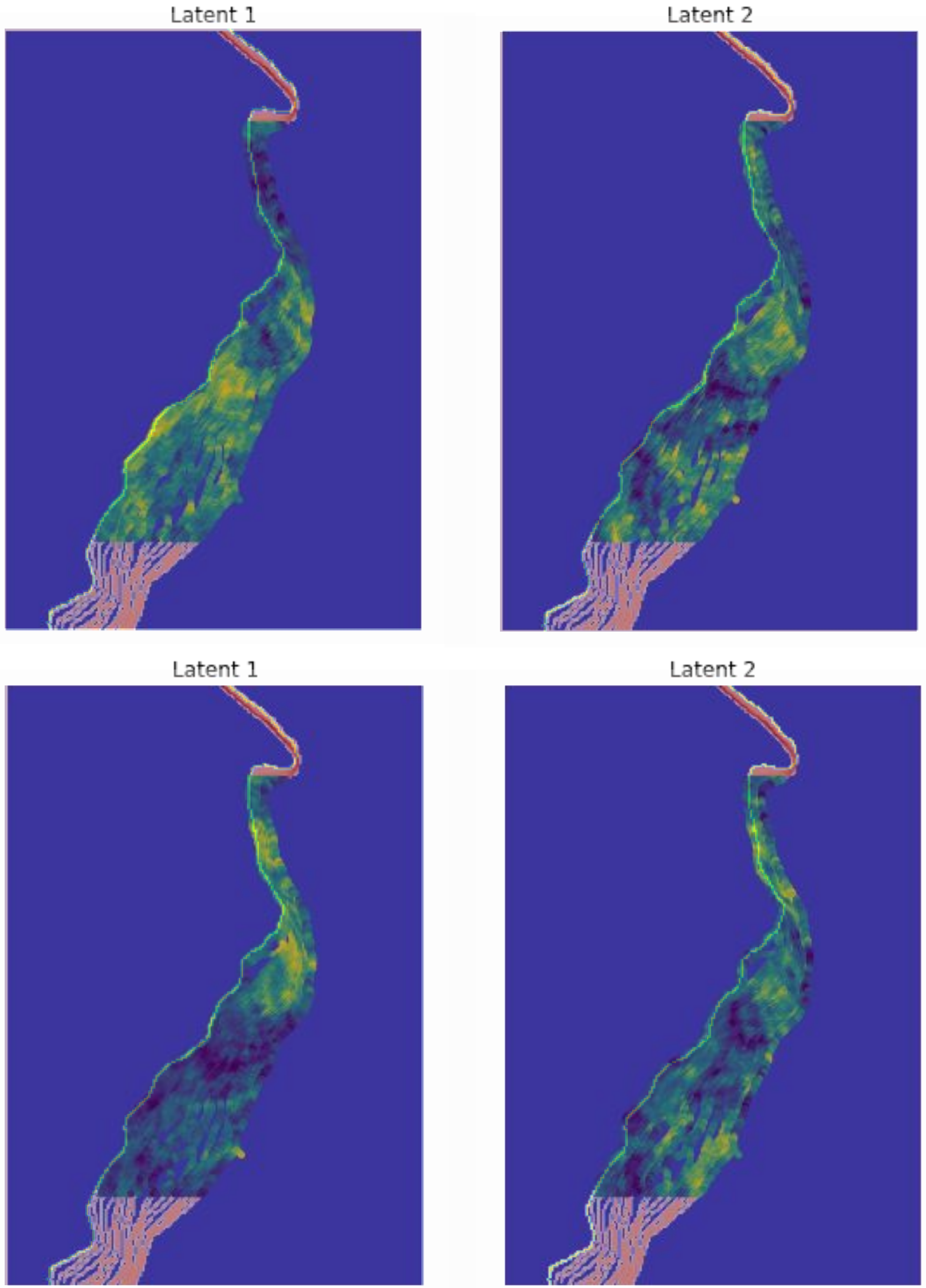


Domain wall evolution



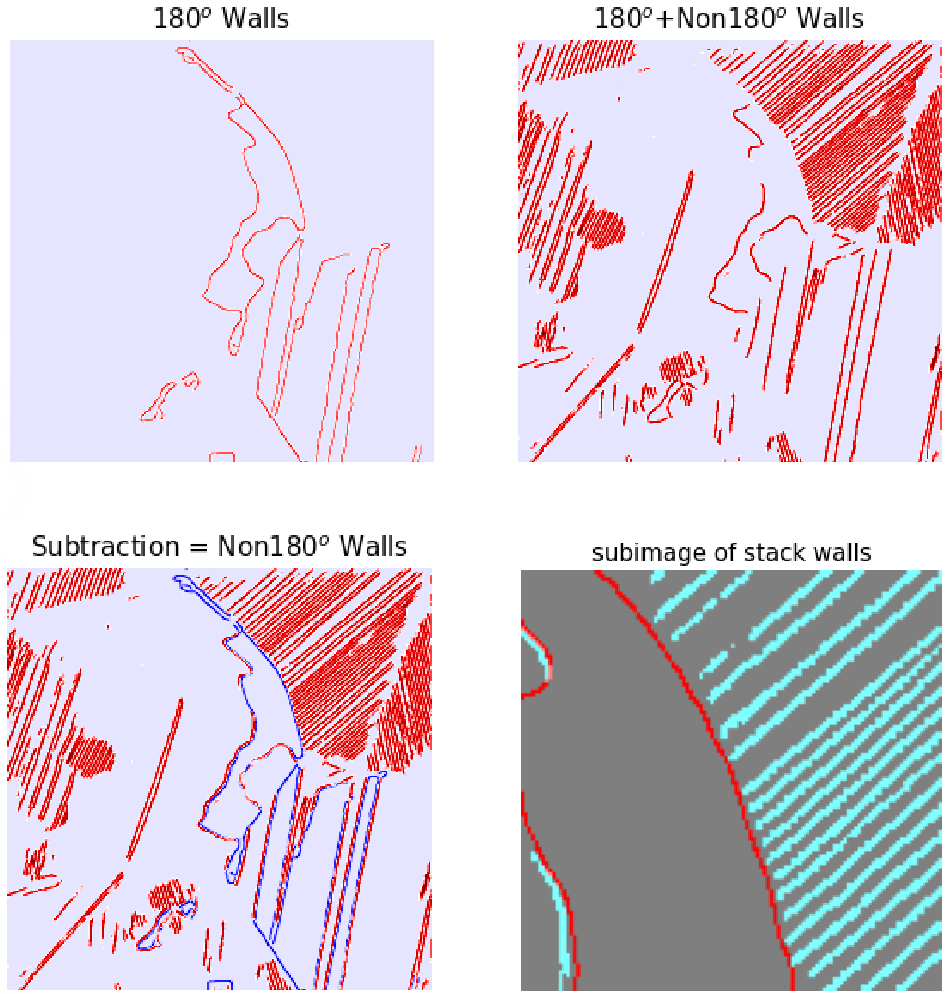
Forward:
t vs t+1

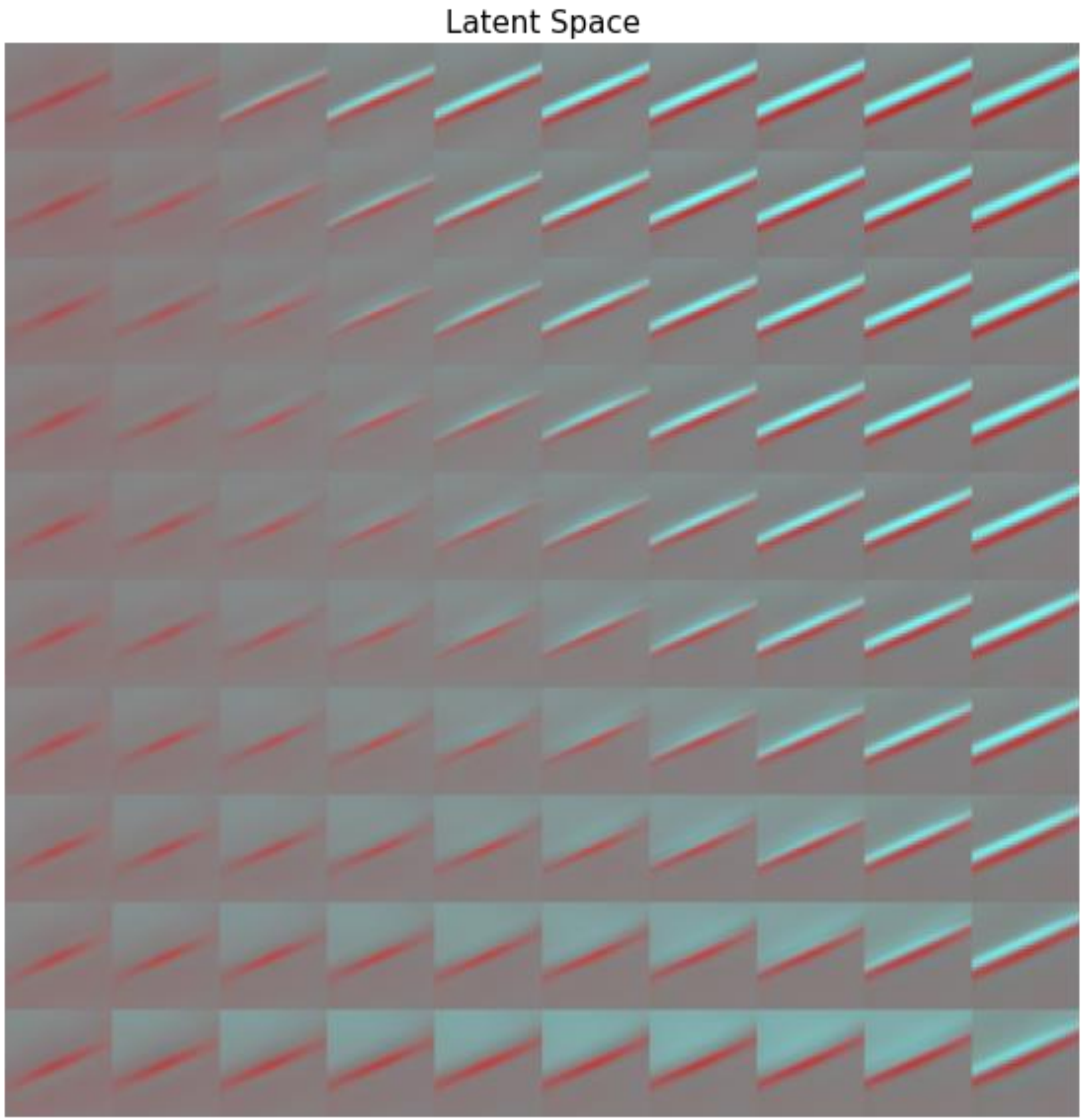
Reverse:
t vs t+1



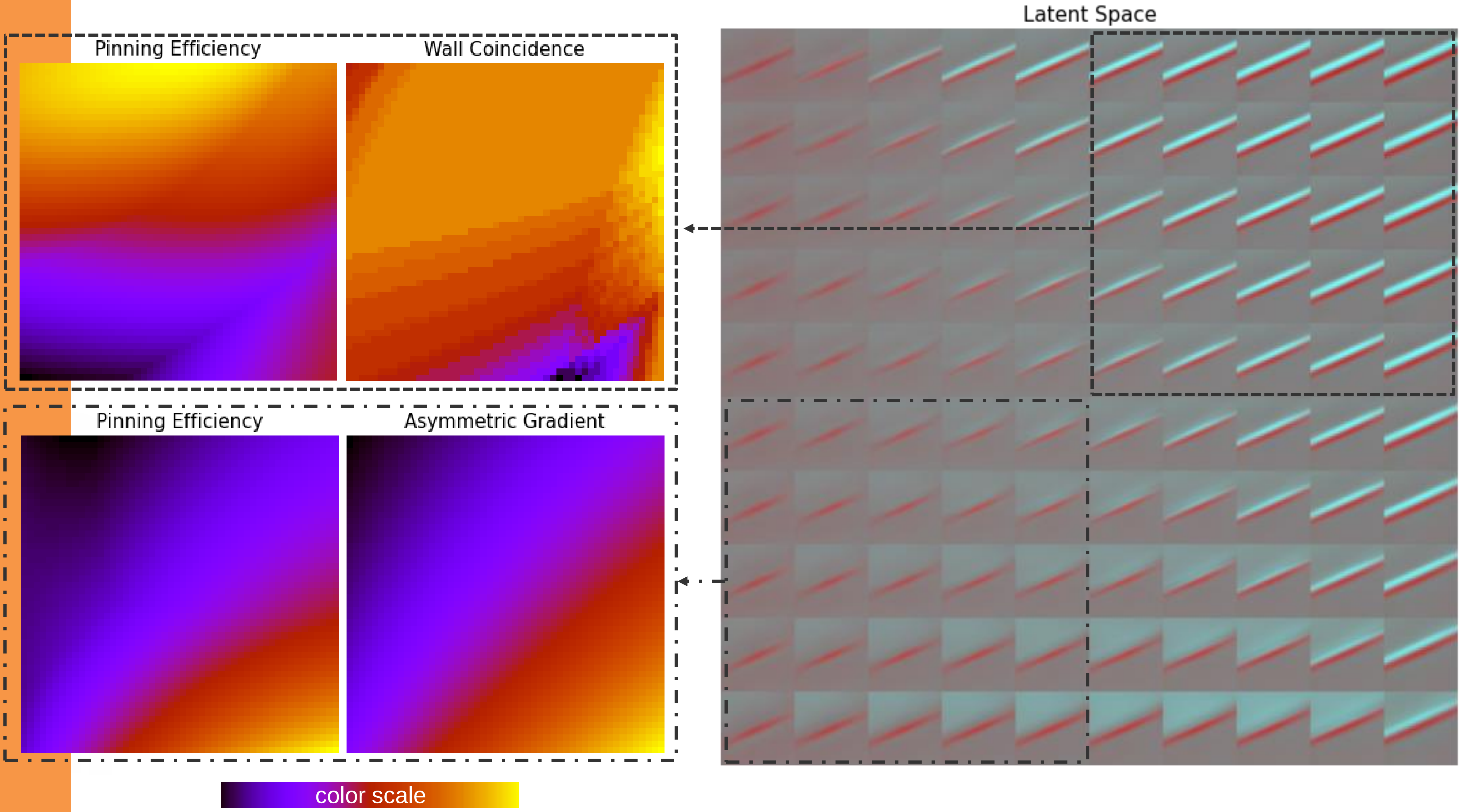
Multilayer rVAE

THE UNIVERSITY of TENNESSEE KNOXVILLE

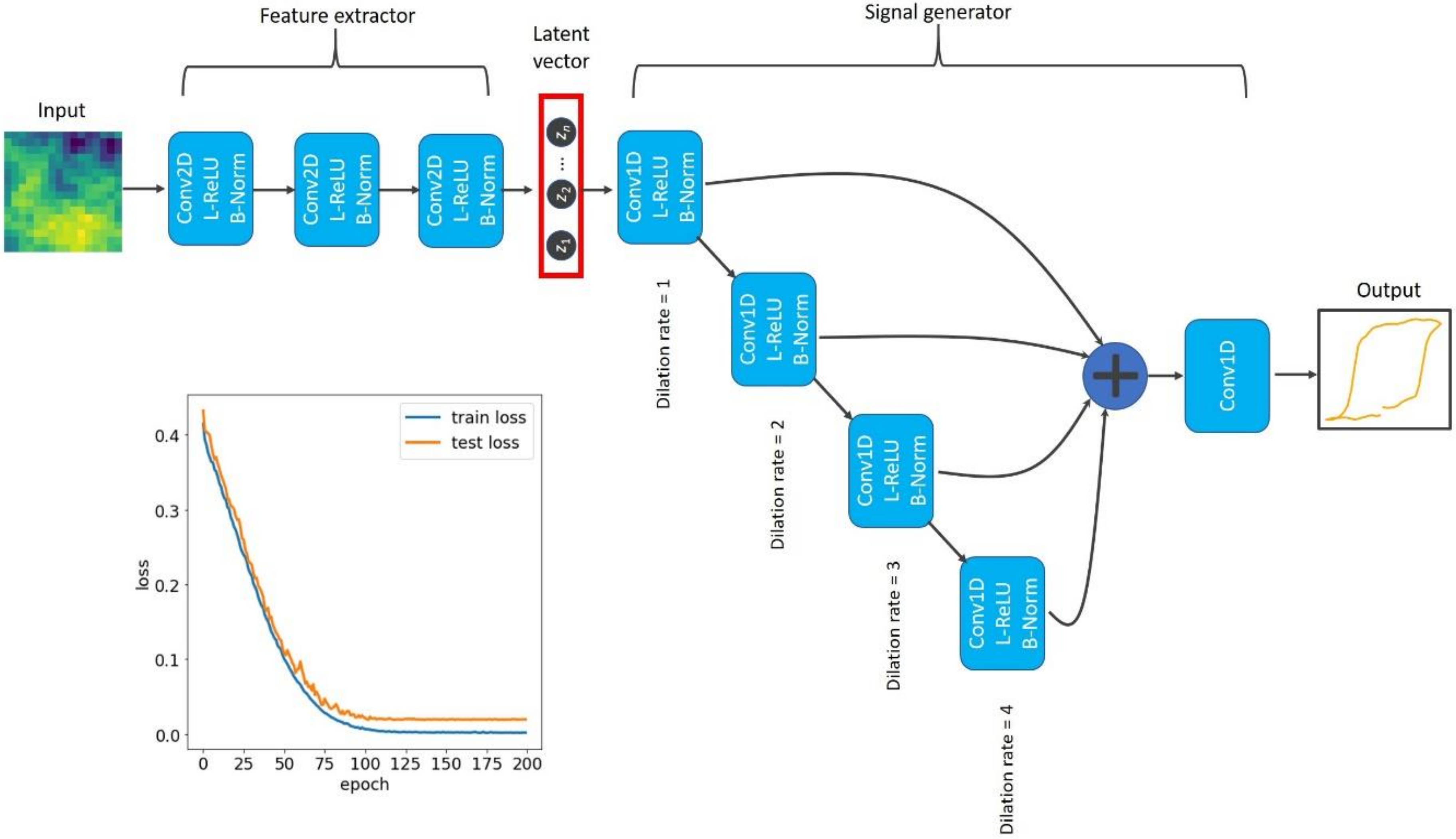




Pinning mechanism



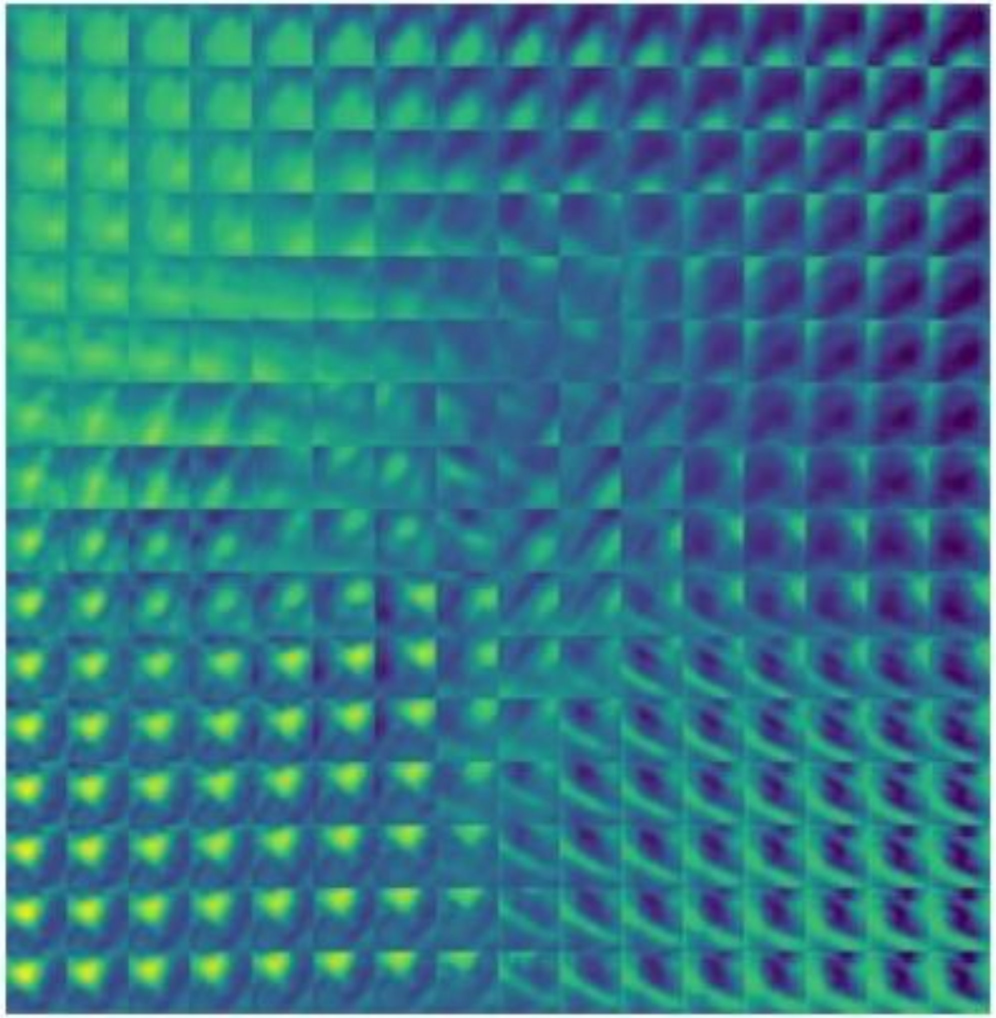
Encoders-Decoders

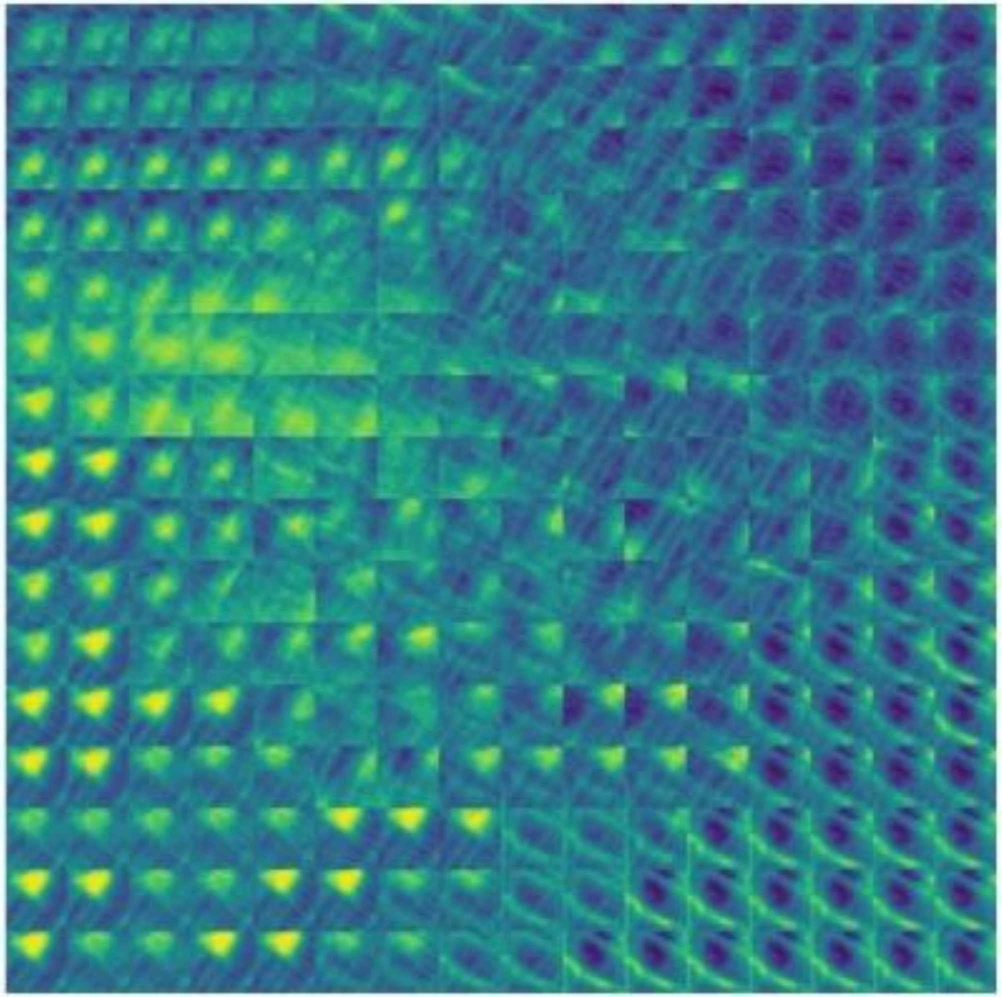


- Use encoder-decoder architecture to transform local structure to local spectra
- And spectra to images
- Predictive within the image

Latent space

THE UNIVERSITY of TENNESSEE KNOXVILLE





Prediction

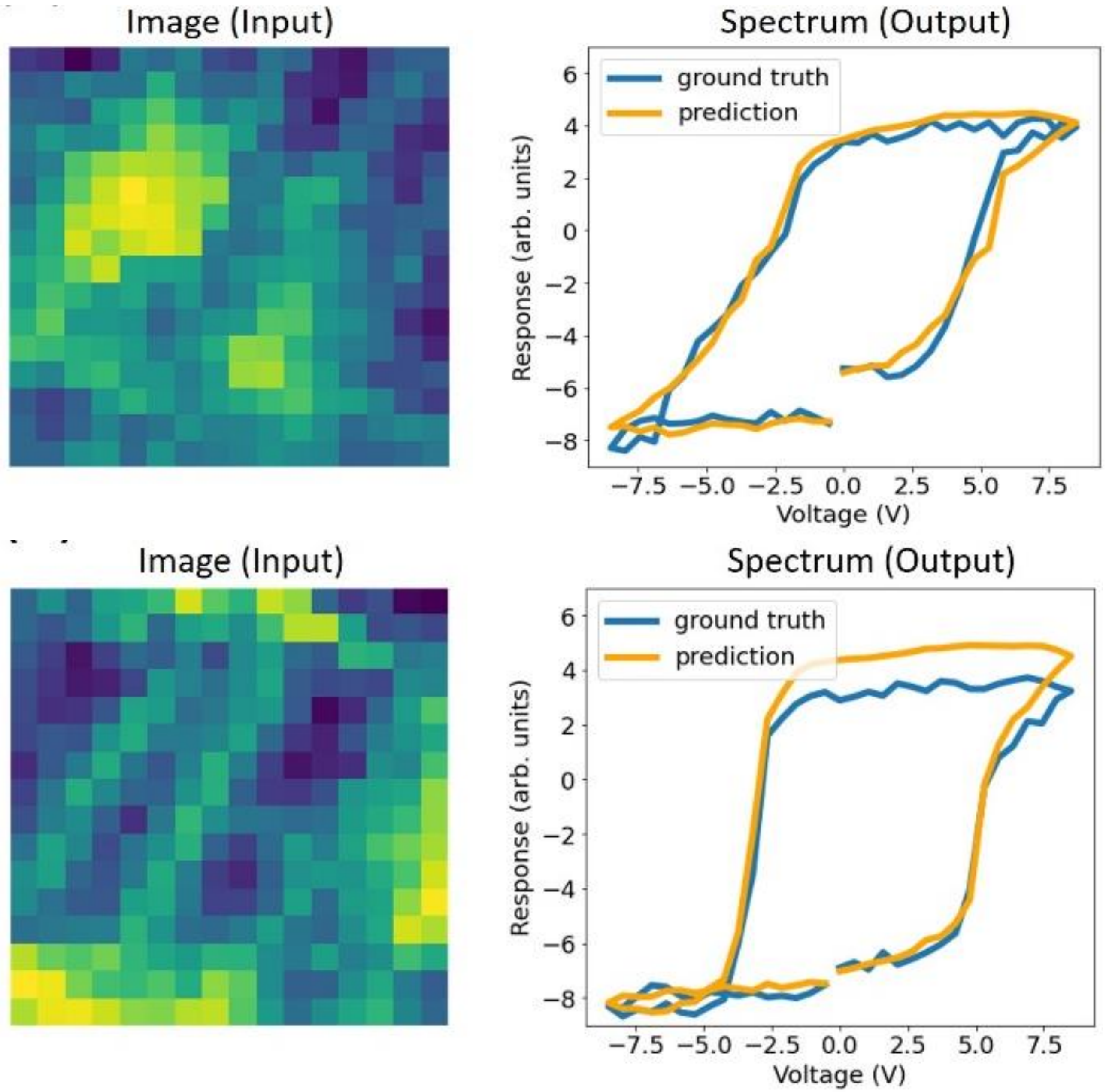
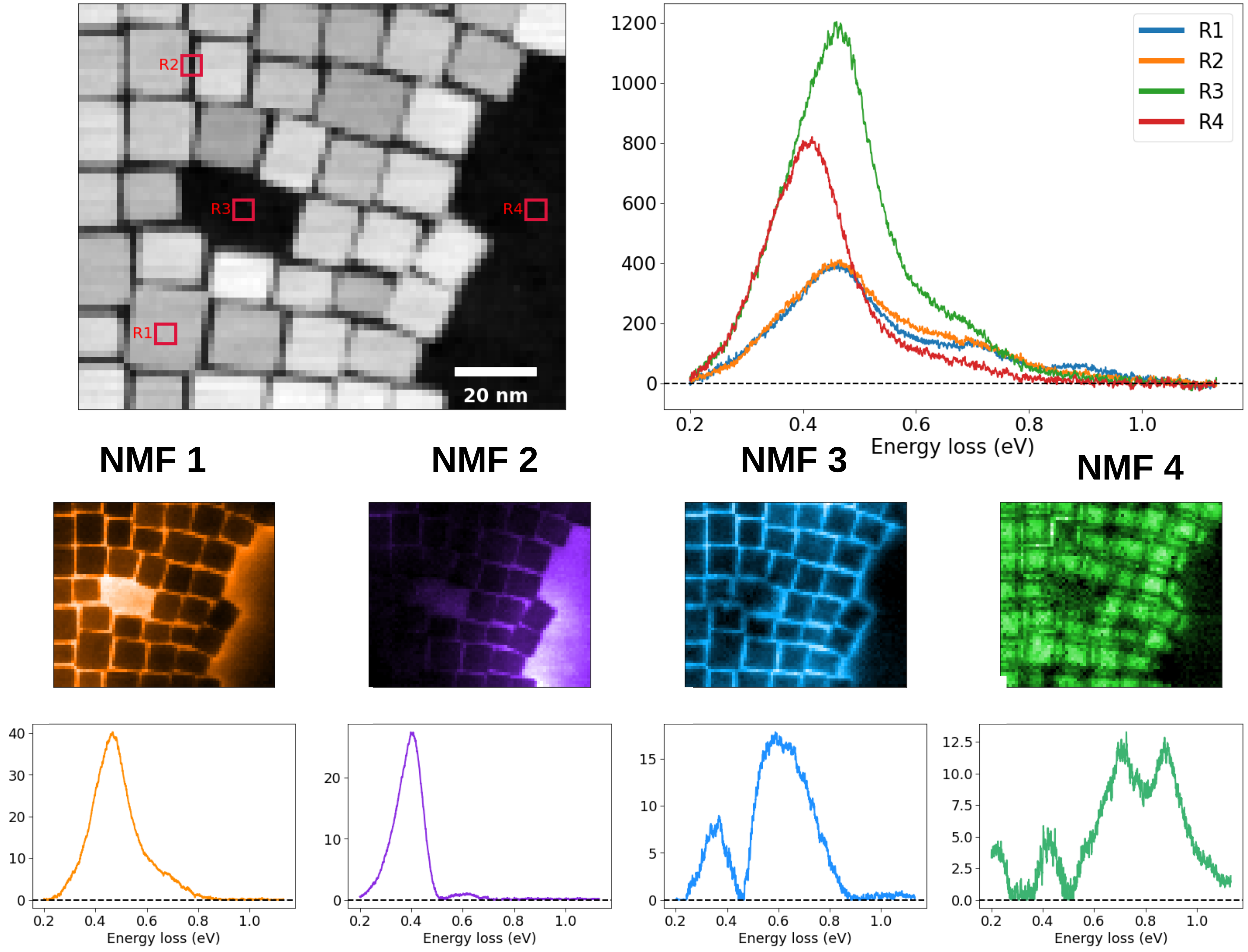


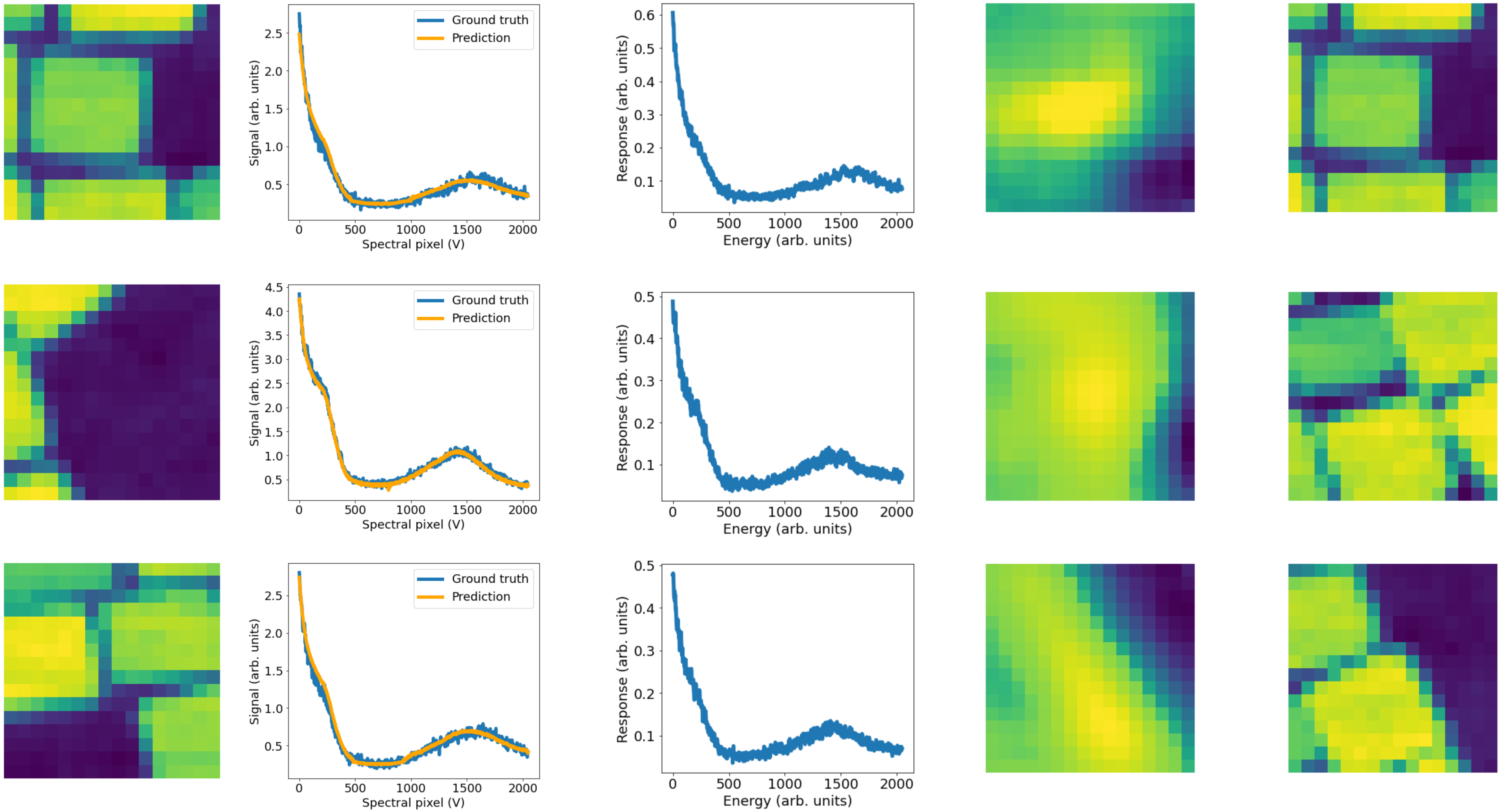
Image (Input)
Spectrum (Output)
ground truth
prediction
Response (arb. units)
Voltage (V)
6
4
2
0
−2
−4
−6
−8
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
Image (Input)
Spectrum (Output)
ground truth
prediction
Response (arb. units)
Voltage (V)

Plasmonic nanoparticles



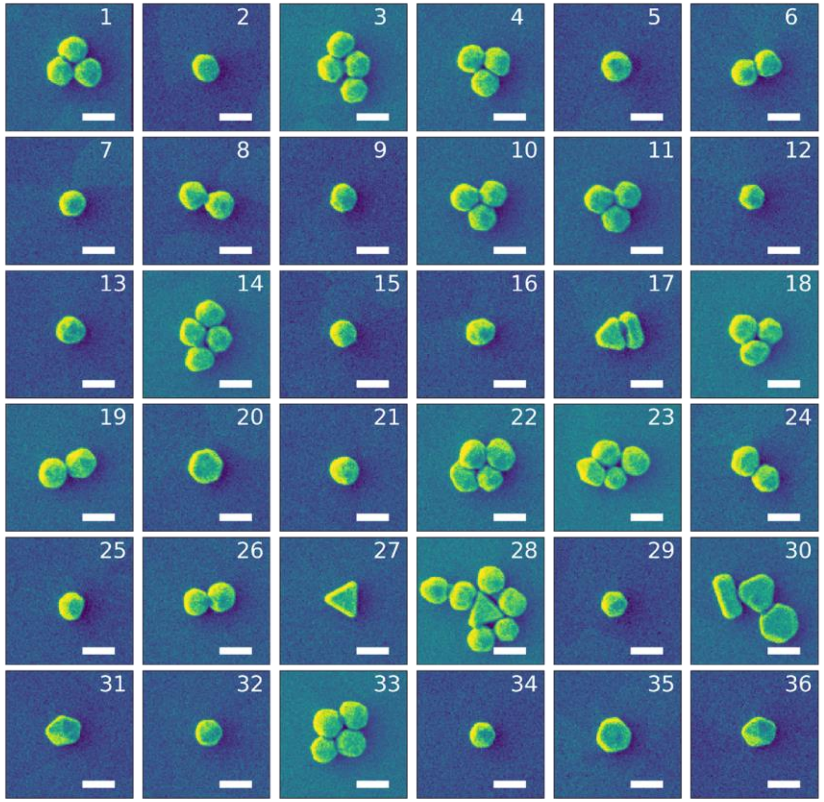
Encoders-Decoders

Sub-image | **Prediction vs. truth** | **Spectrum** | **Prediction** | **Ground truth**

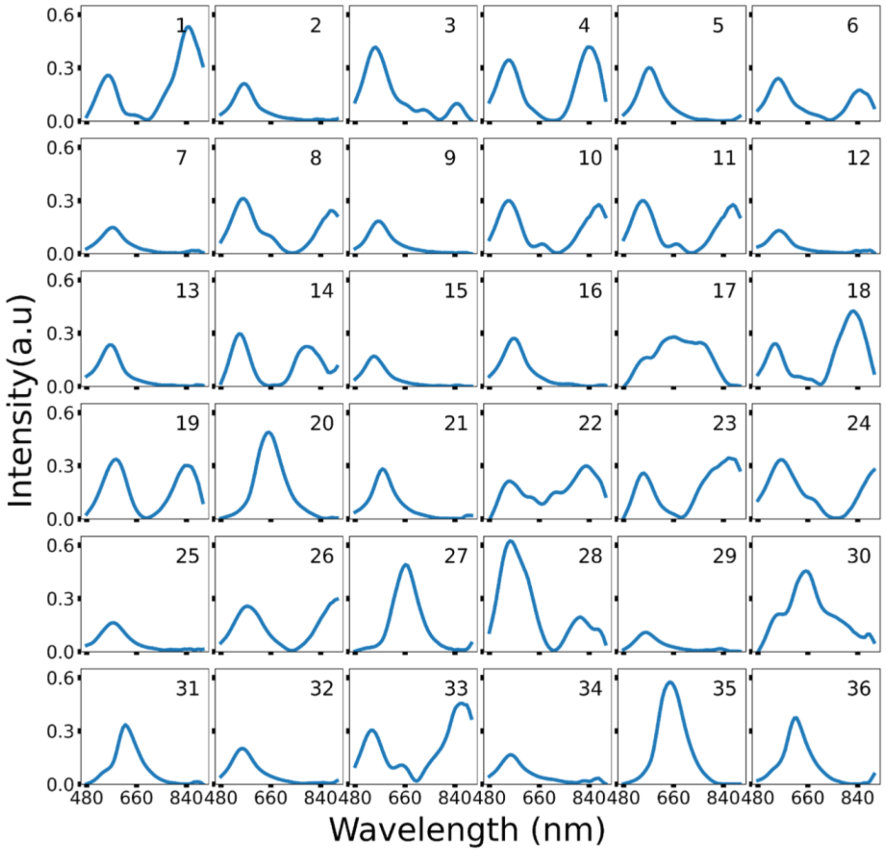


Dual VAE: structure-property relationships

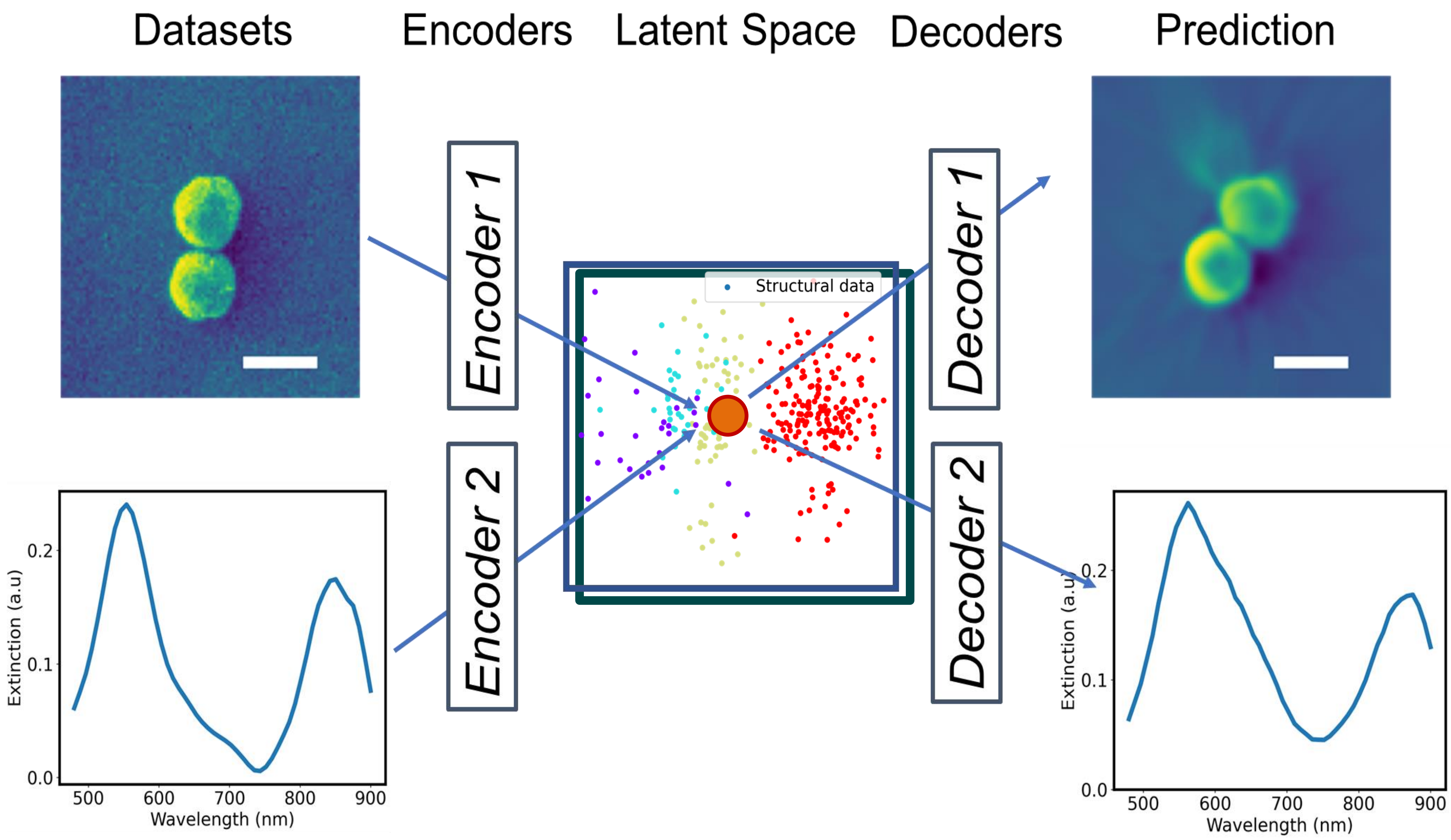
SEM images: “Structure Information”



Hyperspectral microscope: “Property Information”



Dual VAE

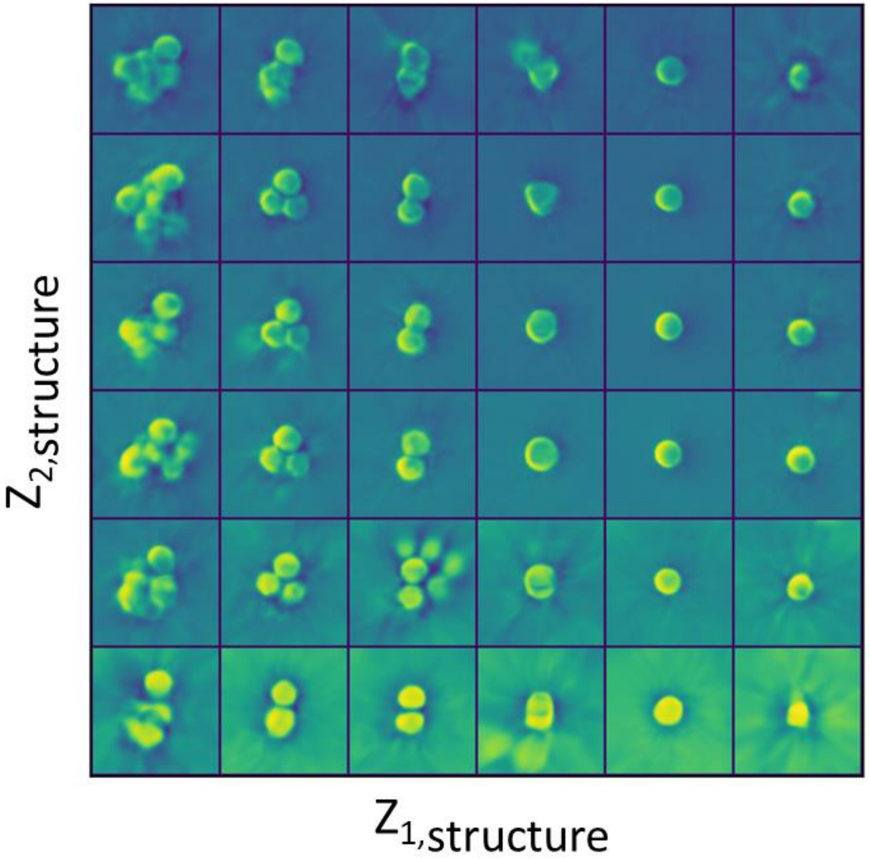


Datasets
Encoders
Latent Space
Decoders
Prediction
Encoder 1
Encoder 2
Decoder 1
Decoder 2
Structural data
Extinction (a.u)
Wavelength (nm)
0.0
0.1
0.2
500
600
700
800
900

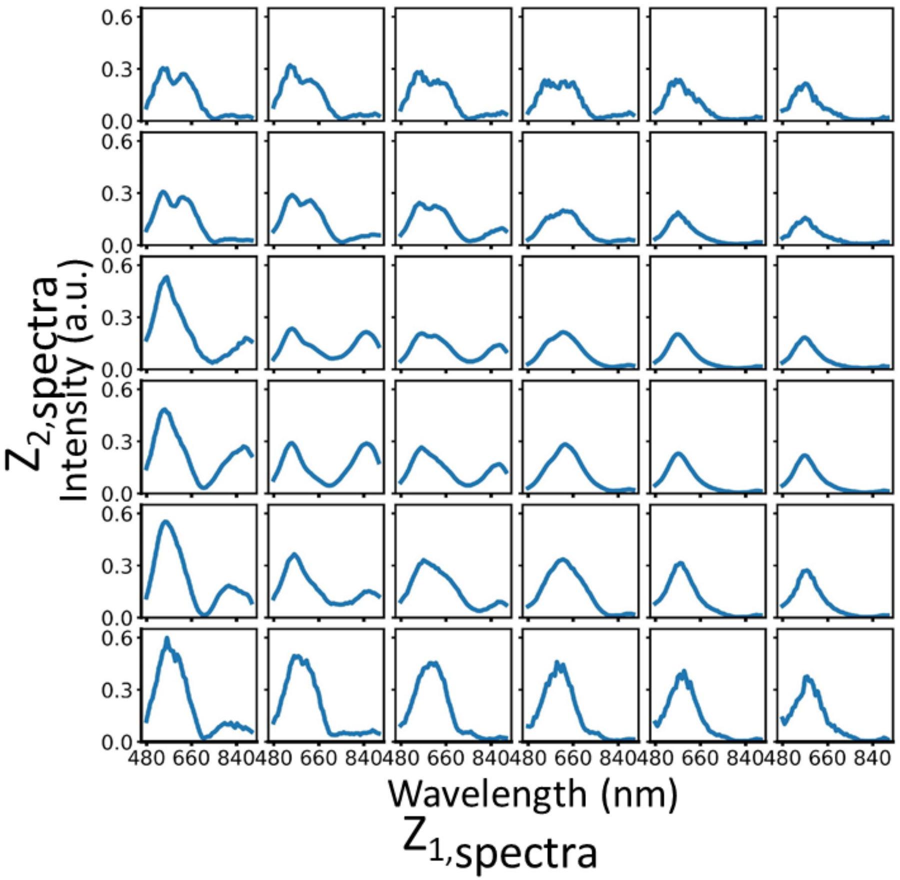
Dual VAE: Latent Representations

Manifold Representation

Changing in particle shape



increasing in particle number in a cluster

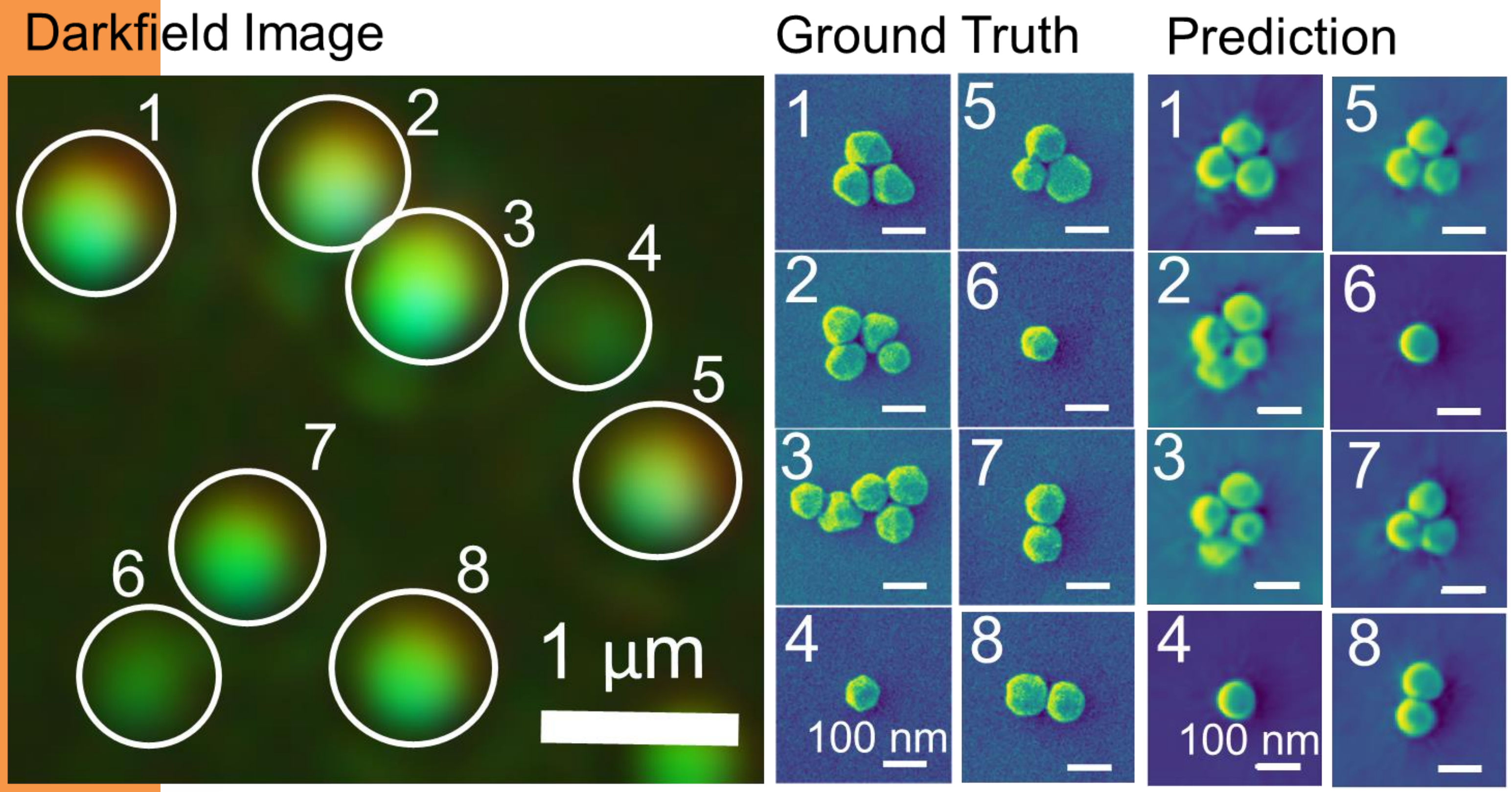


Changing in spectrum shape

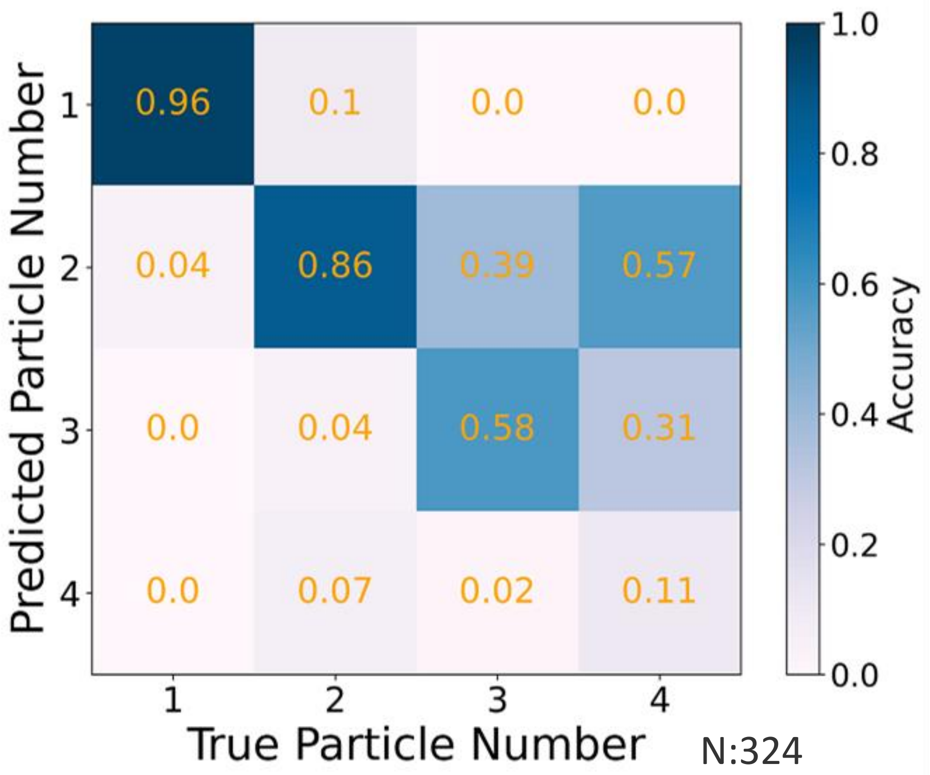
increasing in the peak intensity

Dual VAE: Predictions

Example



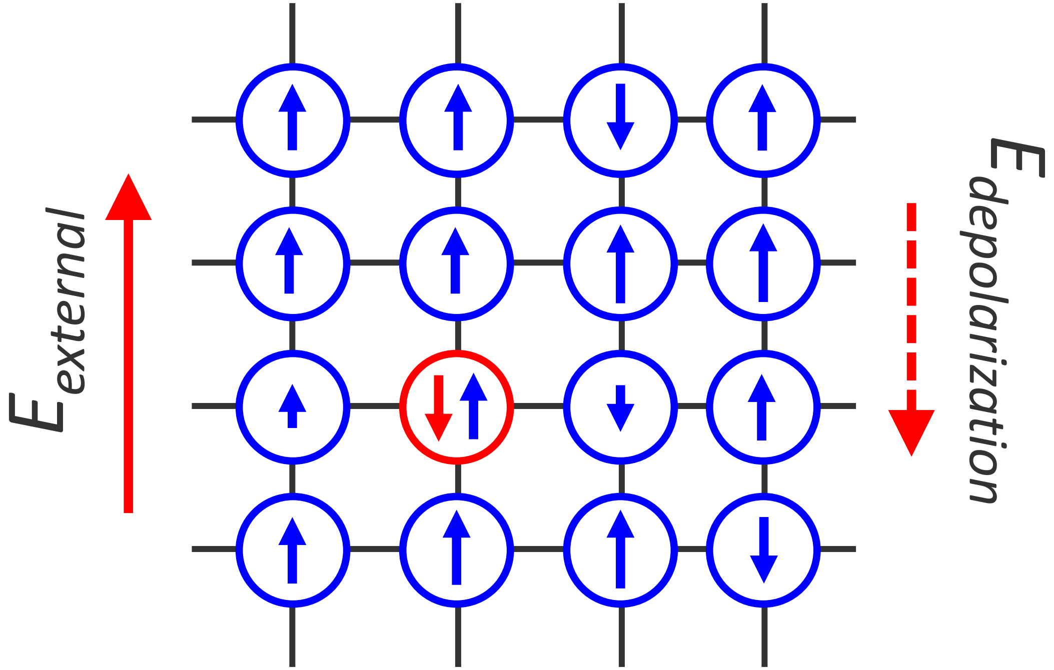
Overall Particles

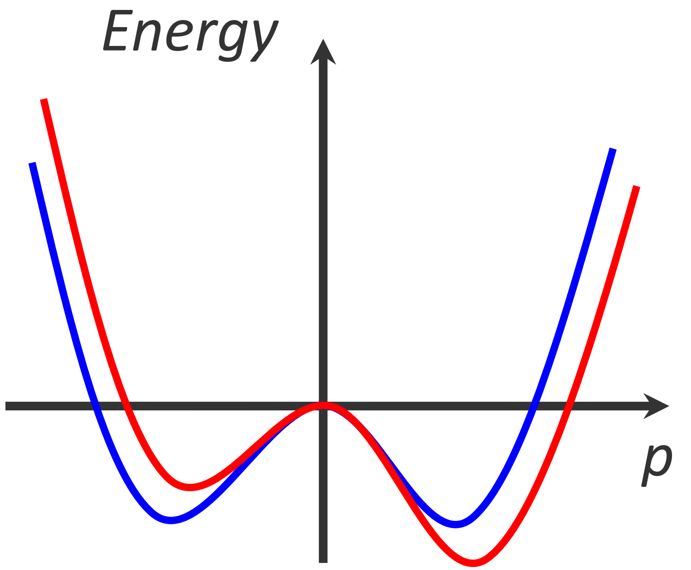


2303.18236.pdf (arxiv.org)

https://github.com/saimani5/VAE-tutorials

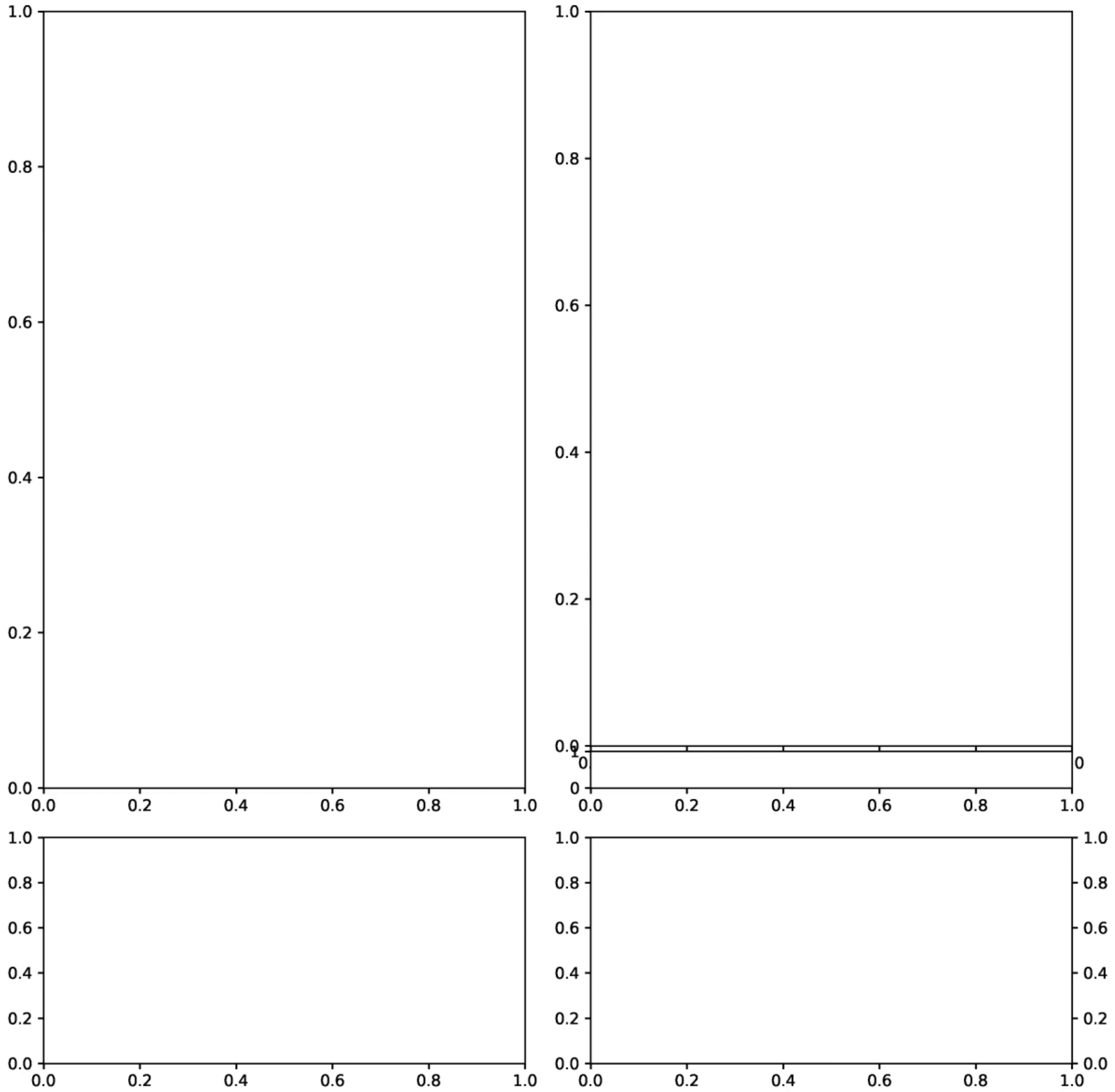
FerroSIM: the simplest interesting ferroelectric





- A discrete square lattice where a continuous polarization vector resides at each lattice site
- The local free energy at each site takes the GLD form:
  - $F_{ij} = \alpha_1 \left(p_{x_{ij}}^2 + p_{y_{ij}}^2\right) + \alpha_2 \left(p_{x_{ij}}^4 + p_{y_{ij}}^4\right) + \alpha_3 p_{x_{ij}}^2 p_{y_{ij}}^2 - E_{loc_{x_{ij}}} p_{x_{ij}} - E_{loc_{y_{ij}}} p_{y_{ij}}$
  - Where, $E_{loc} = E_{ext} + E_{dep} + E_d(i,j)$ and $E_d = -\alpha_{dep} < p >$
- The total free energy is the sum of local free energies and coupling terms:
  - $F = \sum_{i,j}^{N} F_{ij} + K \sum_{k,l} (p_{x_{ij}} - p_{x_{i+k,j+l}})^2 + K \sum_{k,l} (p_{y_{ij}} - p_{y_{i+k,j+l}})^2$
- Polarization at each lattice site is updated to decrease the free energy using $\frac{\gamma d p_{i,j}}{dt} = -\frac{\partial F}{\partial p_{i,j}}$

But what about trajectories?



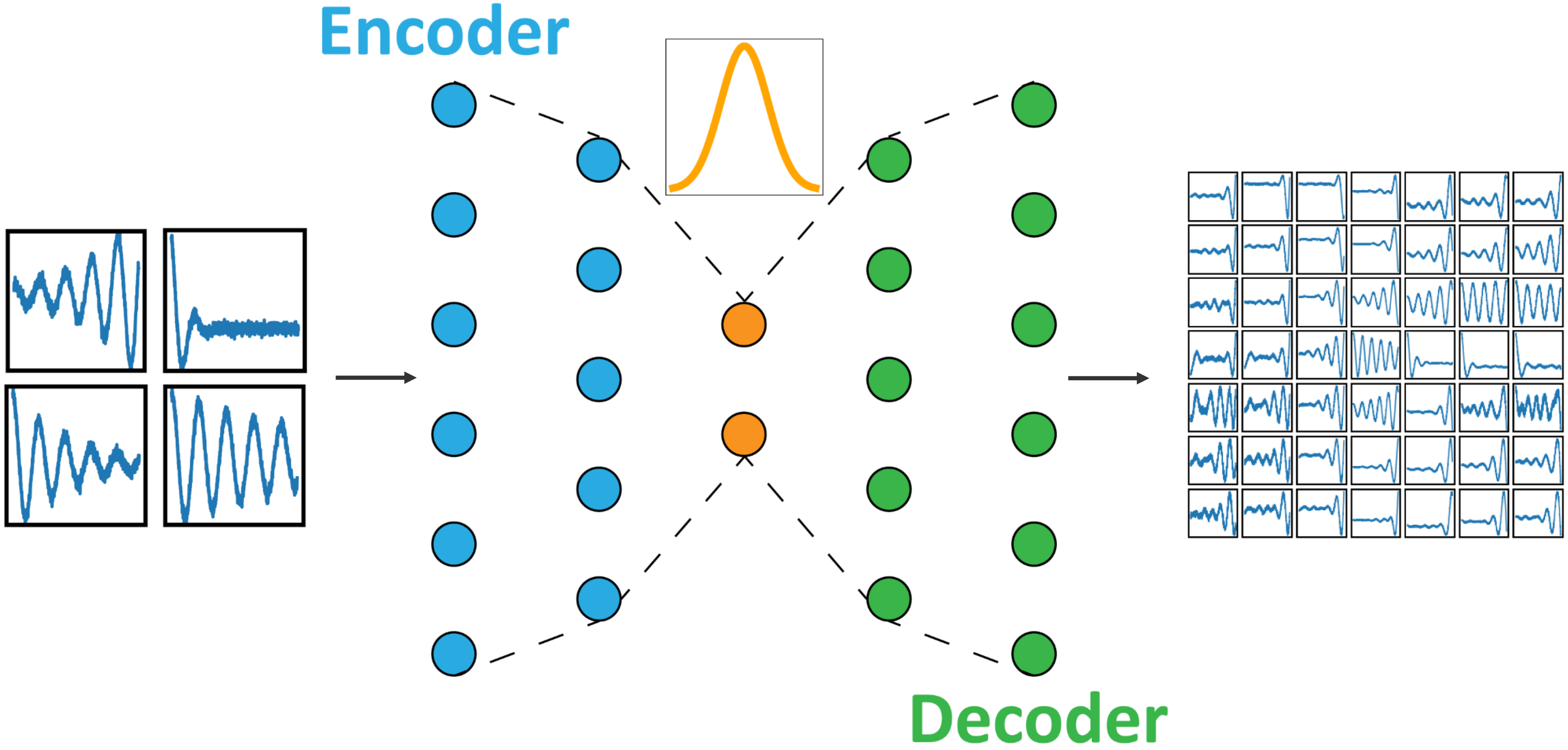
- The model has large number of microstates
- The global state depends on history, i.e. dependence of field vs. time
- Can we somehow optimize the chosen global state in the space of possible histories?

- This space is obviously intractable…

- … however, we are not interested in ALL possible histories. We are interested in relatively simple histories

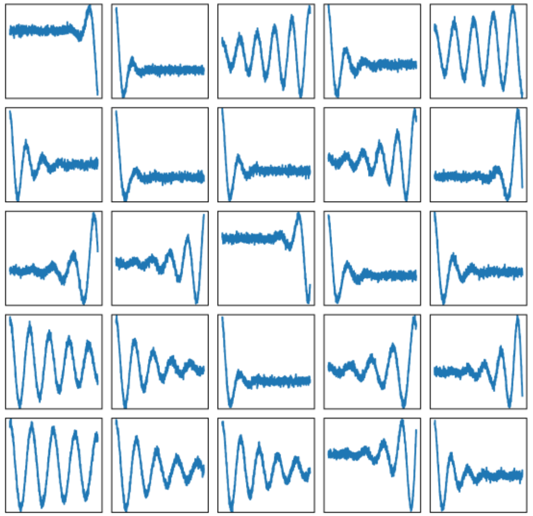
- **Thought:** what if we start with the histories that make sense from domain perspective, and look for way to simplify them?

Can VAE help?



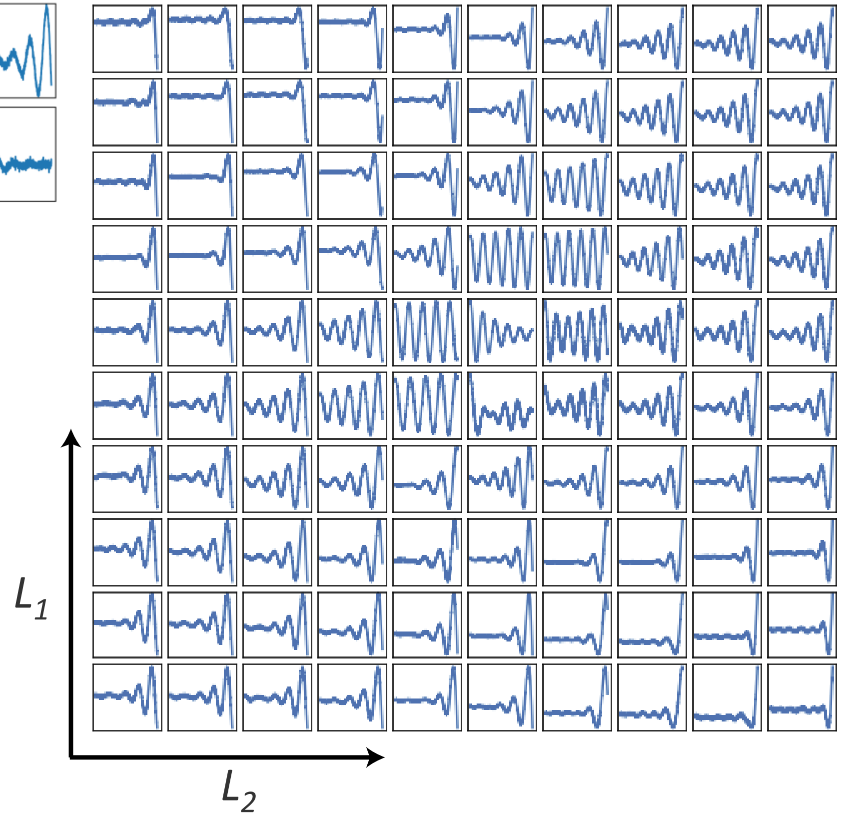
VAE encoding of domain trajectories

- Sinusoidal trajectories with exponential functions as amplitude modulators
  - $A exp(\alpha t) \sin(\omega t) + B$
- A: [0, 0.75],
- α: [-2.75, 2.75],
- ω: [2π, $\frac{8}{3}\pi$],
- B: [-0.5, 0.5]
- These electric fields are divided into 900 discrete time steps.
- 7500 of these curves are then used to create a smooth latent space using a Variational Autoencoder (VAE)

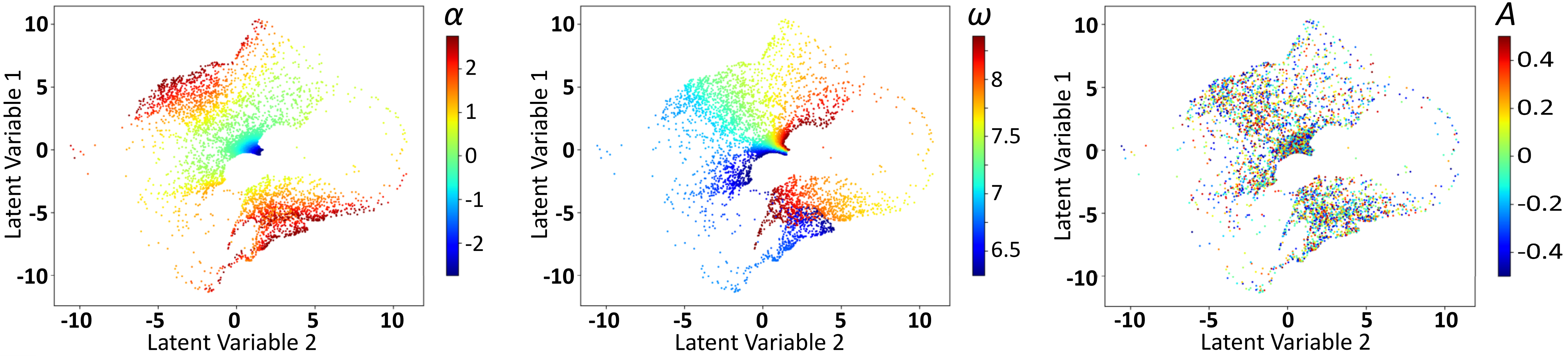


Example Trajectories

Decoded Latent Space

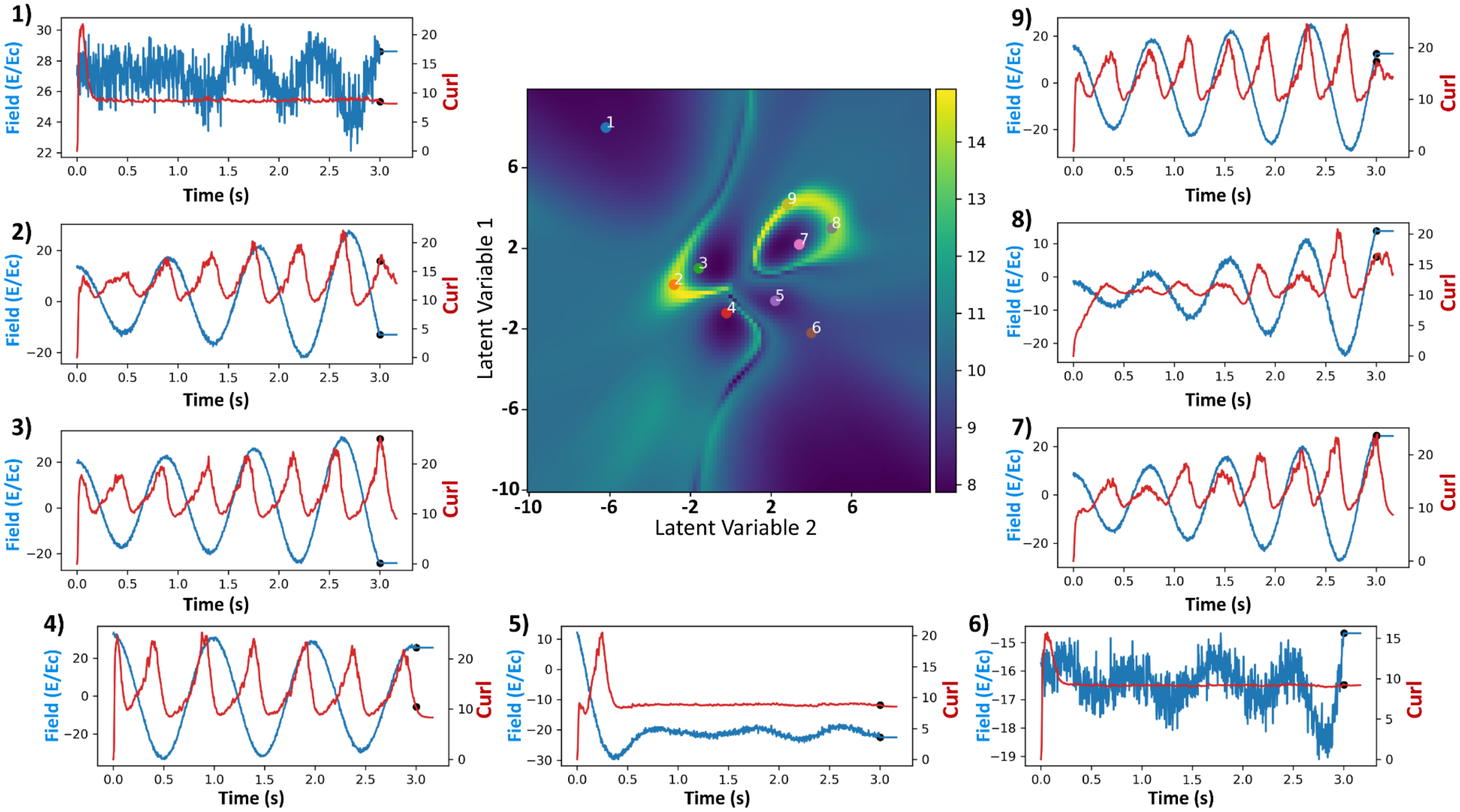


Latent space distributions



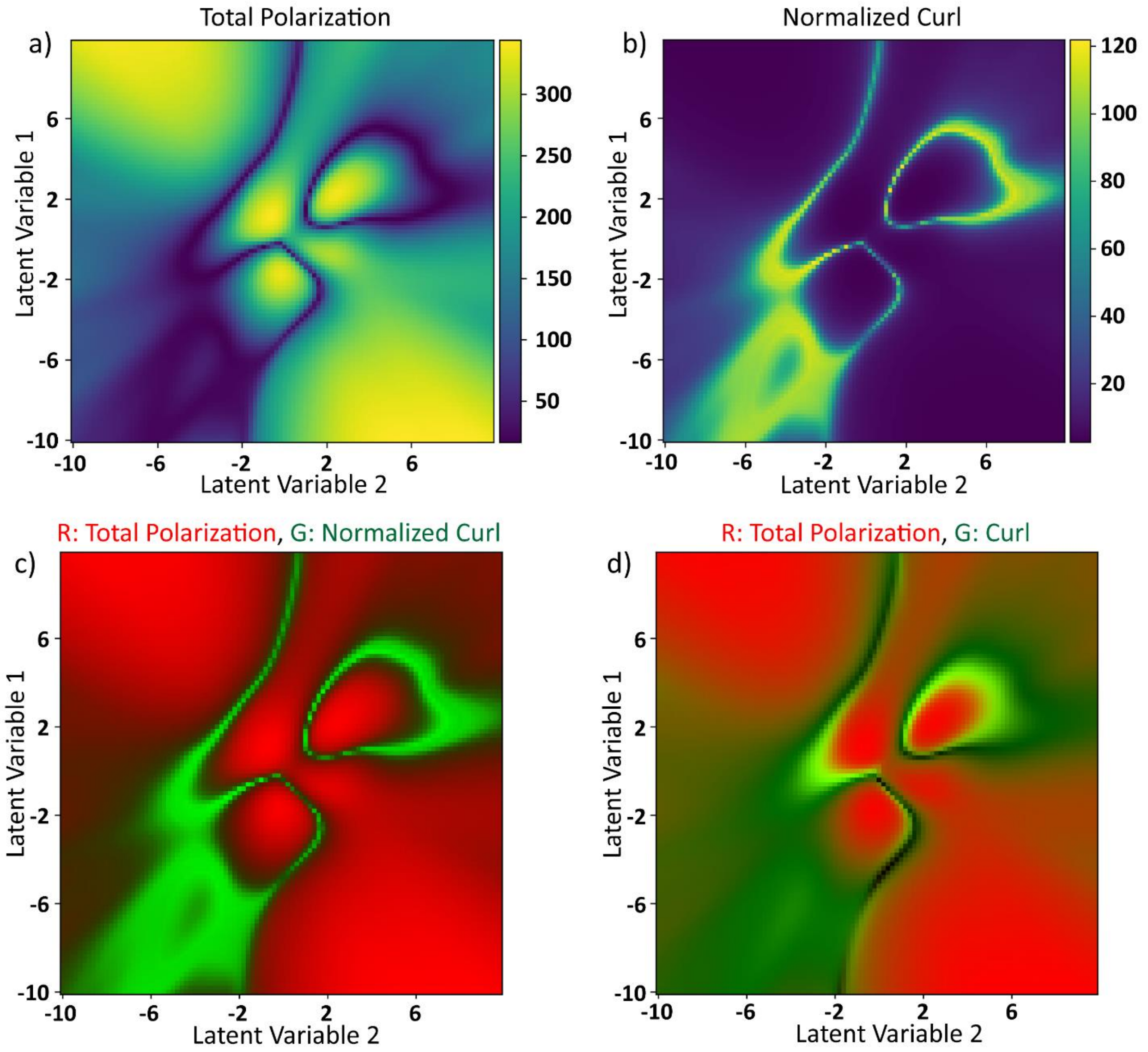
Ground truth target function

- Latent space is sampled and then decoded back into the space of electric field of 900-dimensions
- An equilibration region of 50-time steps is then added where the electric field is held constant at the final value of the decoded electric field.
- The sum of absolute value of curl at each lattice site at the end of the simulation is the target value to be optimized



- Curl decays in the equilibration region
- The rate of decay of the curl is proportional to the curl at the onset of the equilibration region
- The local maxima of the curl seemingly coincides with the local optima of the electric field.

Exploring the curl surface



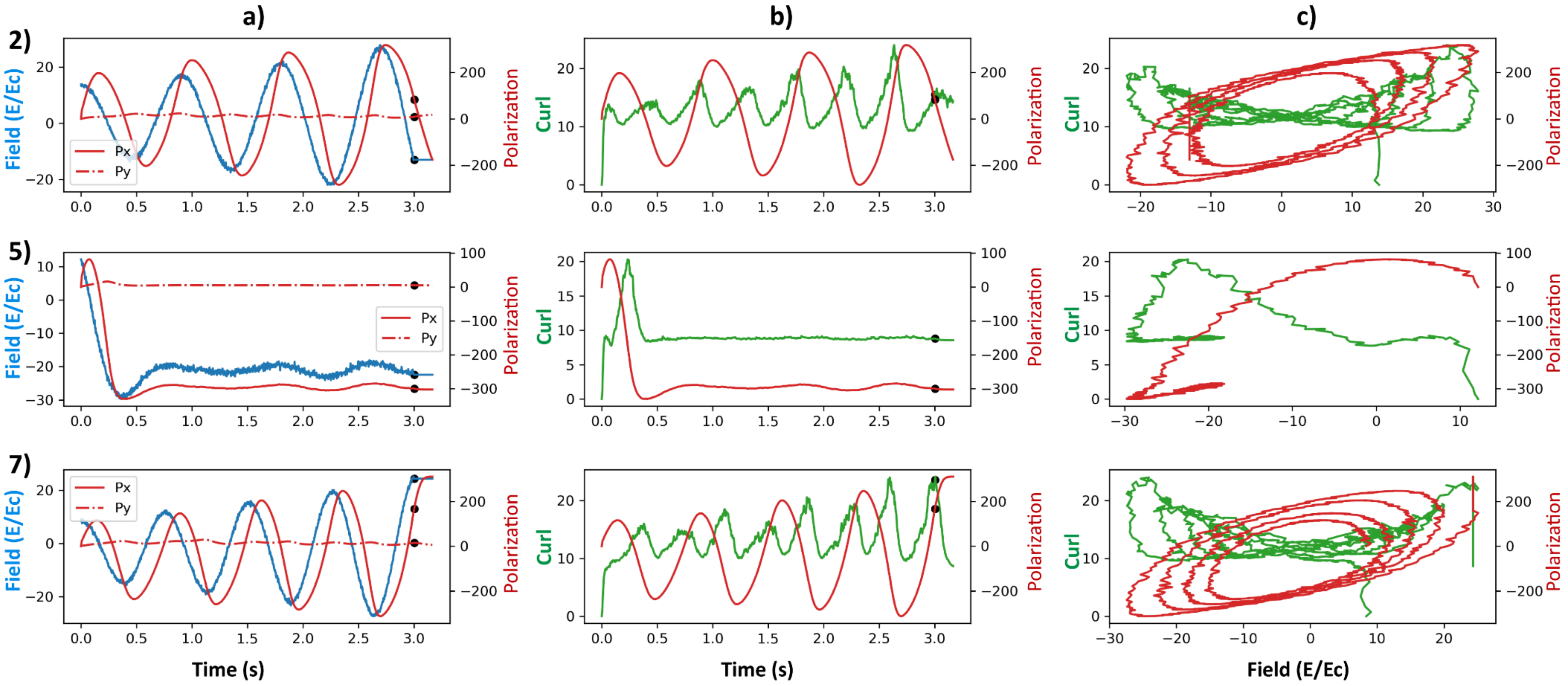
Normalized Curl

- At the end of the simulations, the polarization vector at each lattice site is normalized
- These normalized polarization vectors are then used to recalculate curl
- We will refer to it as the normalized curl
- It is supposed to estimate how much the vector field rotates without considering the magnitude of the field

Observations

- Normalized curl is inversely proportional to the magnitude of polarization
- The system is allowed to be in the most chaotic state when the polarization is the least as the effect of coupling terms is low
- The system's polarization is at the lowest the coercive field

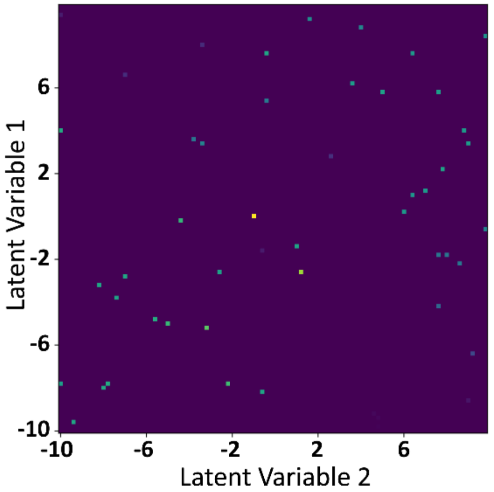
Exploring the curl surface



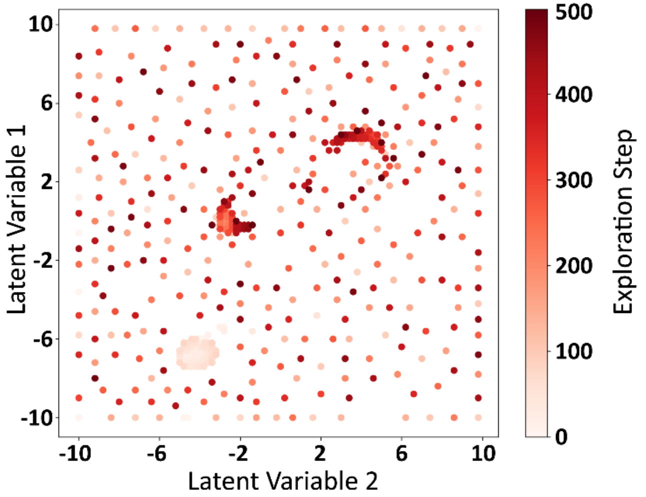
- The system's polarization is at the lowest the coercive field (A state of maximum normalized curl)
- But the curl is also a function of magnitude of the polarization
- Hence, the magnitude of the curl is maximum a few steps after the coercive field where the polarization grows in magnitude just enough that the coupling terms do not take over to kill the curl in the system
- This time coincides with the time it takes the electric field to reach the maximum from the coercive field, hence the overlap of the local maxima of curl and electric field

Bayesian Optimization in the Latent Space

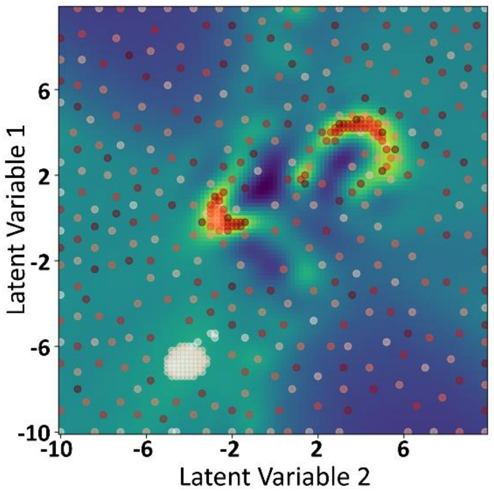
100 initial points



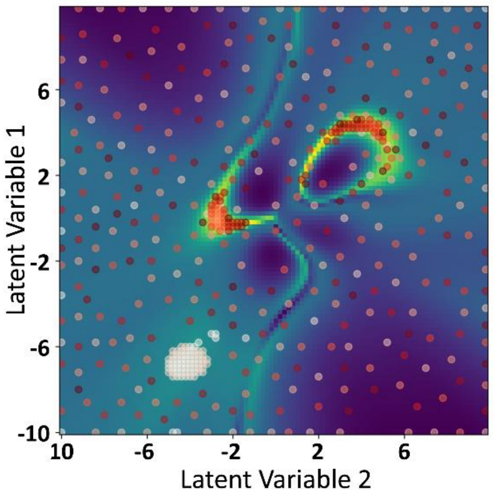
Explored points



Reconstructed curl surface

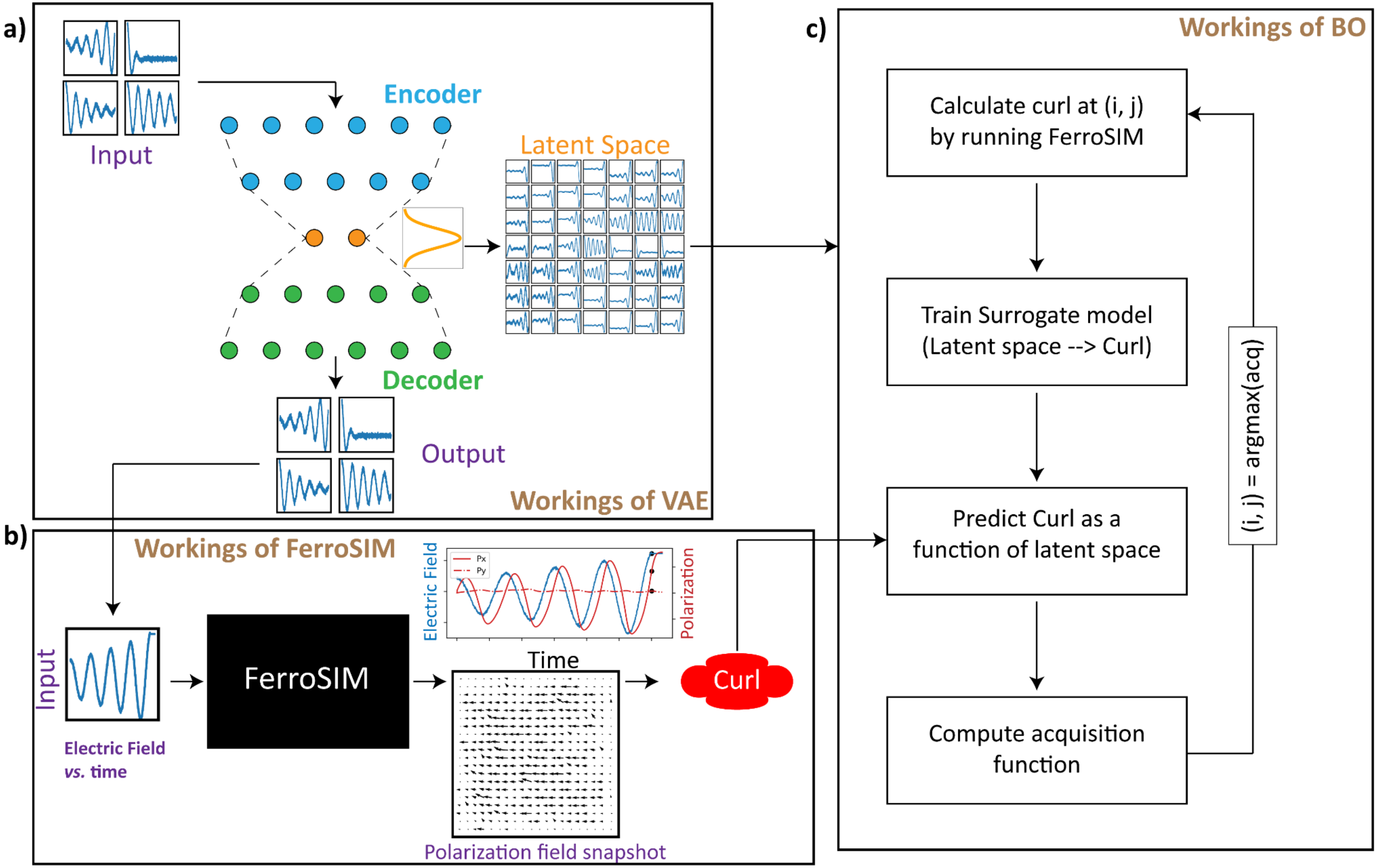


Original curl surface



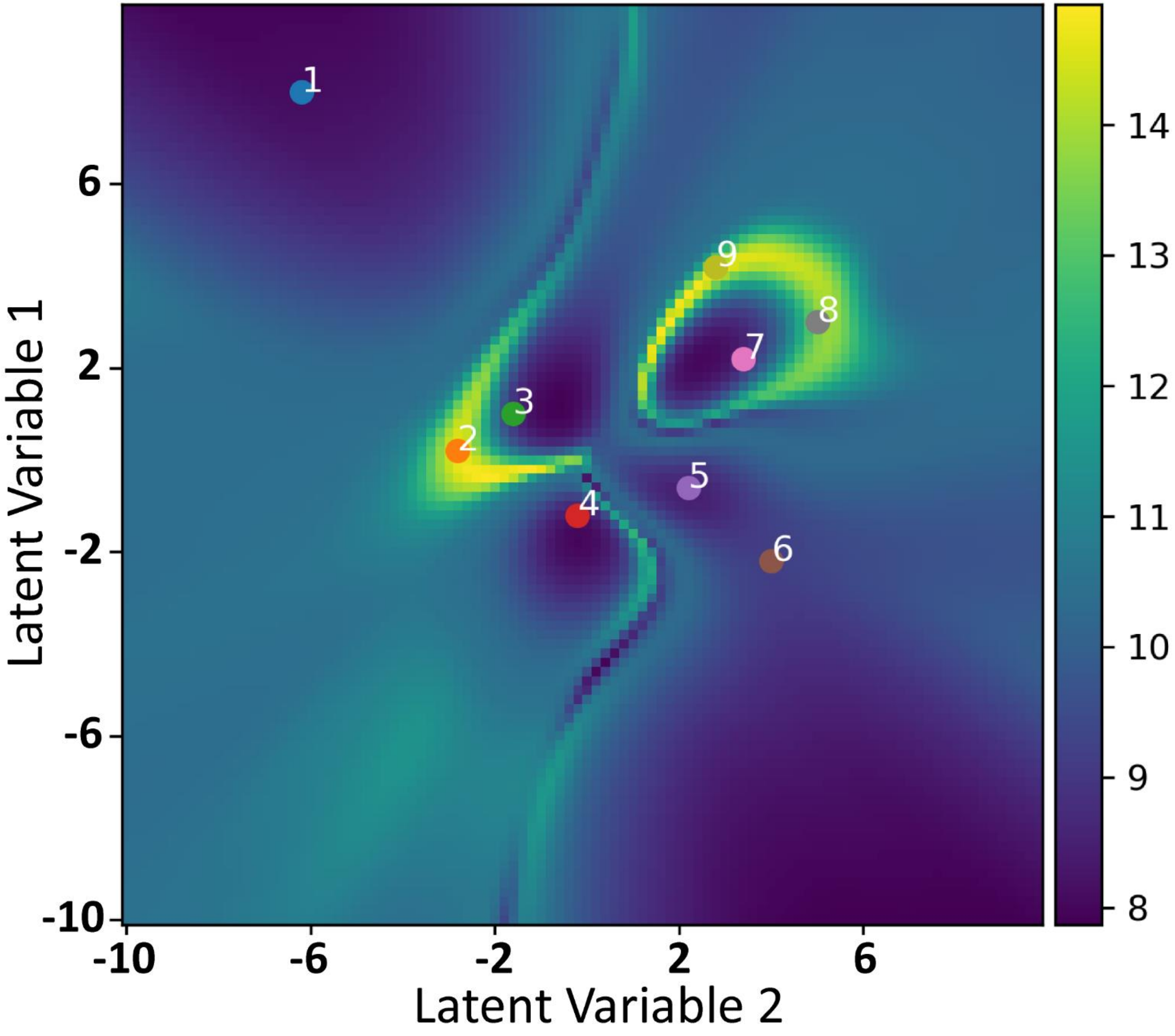
- 100 initialization points and the BO explored the latent space for the next 500 points
- Acq function: μ + 10σ
- So, at the end BO only explored a total of 600 points out of 10,000 points the latent space is divided into
- Caveat: we had to tune the Acq with the ground truth data known

Putting everything together



Same approaches are used for molecular discovery, polymers, and biomolecules

What determines success?



The success of the BO in the latent space clearly depends on the shape on the manifold that points of interest form.

For VAE, the shape of the manifold is determined by the properties of the data only, including

(a) how strong correlations in data reflect in correlation in properties and
(b) weight of the "good" trajectories