Explicitly disentangling image content from translation and rotation with spatial-VAE

Tristan Bepler[1,2], Ellen Zhong[1,2], Kotaro Kelley[3], Ed Brignole[4], and Bonnie Berger[2,5]
[1]Computational and Systems Biology, MIT, Cambridge MA
[2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge MA
[3]Simons Electron Microscopy Center, New York Structural Biology Center, New York NY
[4]Department of Biology, MIT, Cambridge MA
[5]Department of Mathematics, MIT, Cambridge MA

Code
https://github.com/tbepler/spatial-VAE

Paper
https://arxiv.org/abs/1909.11663

Background

- In single particle electron microscopy, we are interested in experimentally determining the structure of a protein
- Statistical methods assume that the many observations (2D projections) of the protein arise from either a single structure or from a discrete mixture of structures
- In 2D classification, the protein views are also grouped into discrete clusters
- These conformations are confounded by orientation in the collected images

Motivation: learn generative models that are orientation invariant. That is, we wish to structure our generative model to enable structured inference on the unobserved orientations.

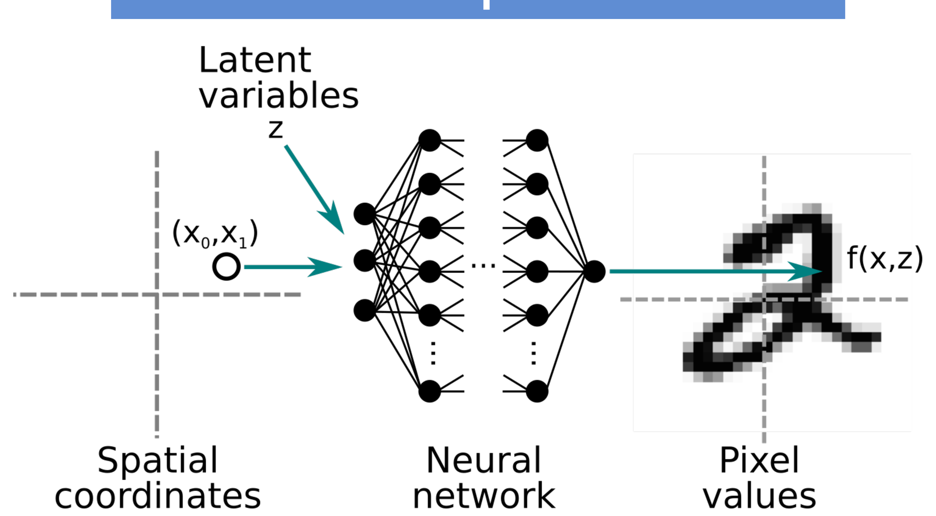
Problem

Learn a generative model of 2-D signals (y) in which the content is described by a set of unstructured, continuous latent variables, *z*. That is

$$z \sim N(0, I),$$
$$y \sim f(z).$$

However, we observe noisy images that are corrupted with *translation* and *rotation (in EM, particles are randomly oriented)*, but we do not wish to encode these features into z.

Solution: spatial-VAE



Parameterize the generative model as a function of the spatial coordinates, i.e. the log probability of an image, y, given latent variables, z, and spatial coordinates, x, is

$$\log p(y|z) = \sum_{i=1}^{n} \log p_g(y^i|x^i, z)$$

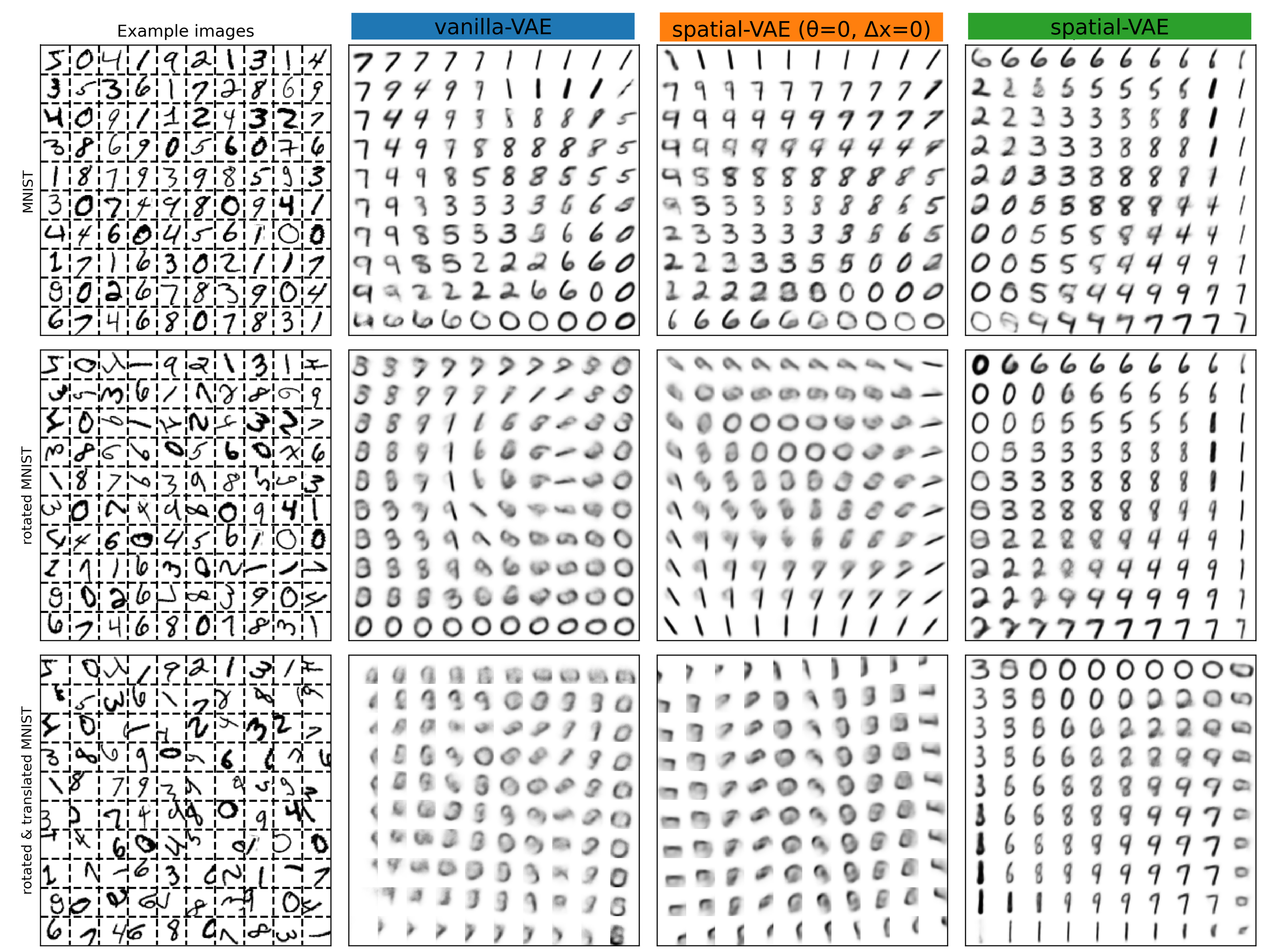
This allows us to naturally model rotation and translation as transformations of the spatial coordinates. Let R(θ) be the rotation matrix for angle θ and Δx be the translation, then the log probability of the image is given by

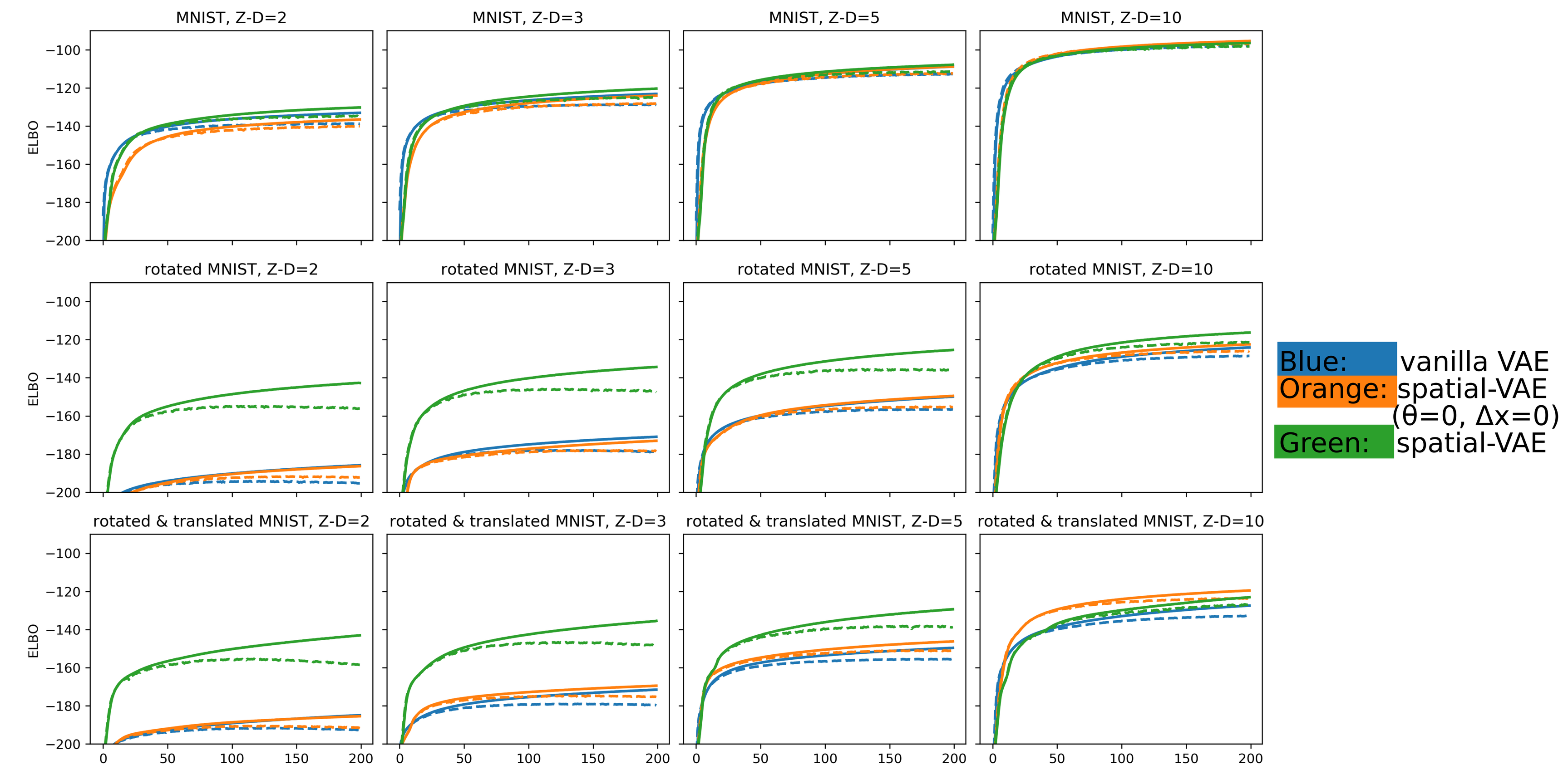
$$\log p_g(y|z, \theta, \Delta x) = \sum_{i=1}^{n} \log p_g(y^i|x^i R(\theta) + \Delta x, z)$$

We then perform approximate inference on z, θ, and Δx using a variational autoencoder framework.

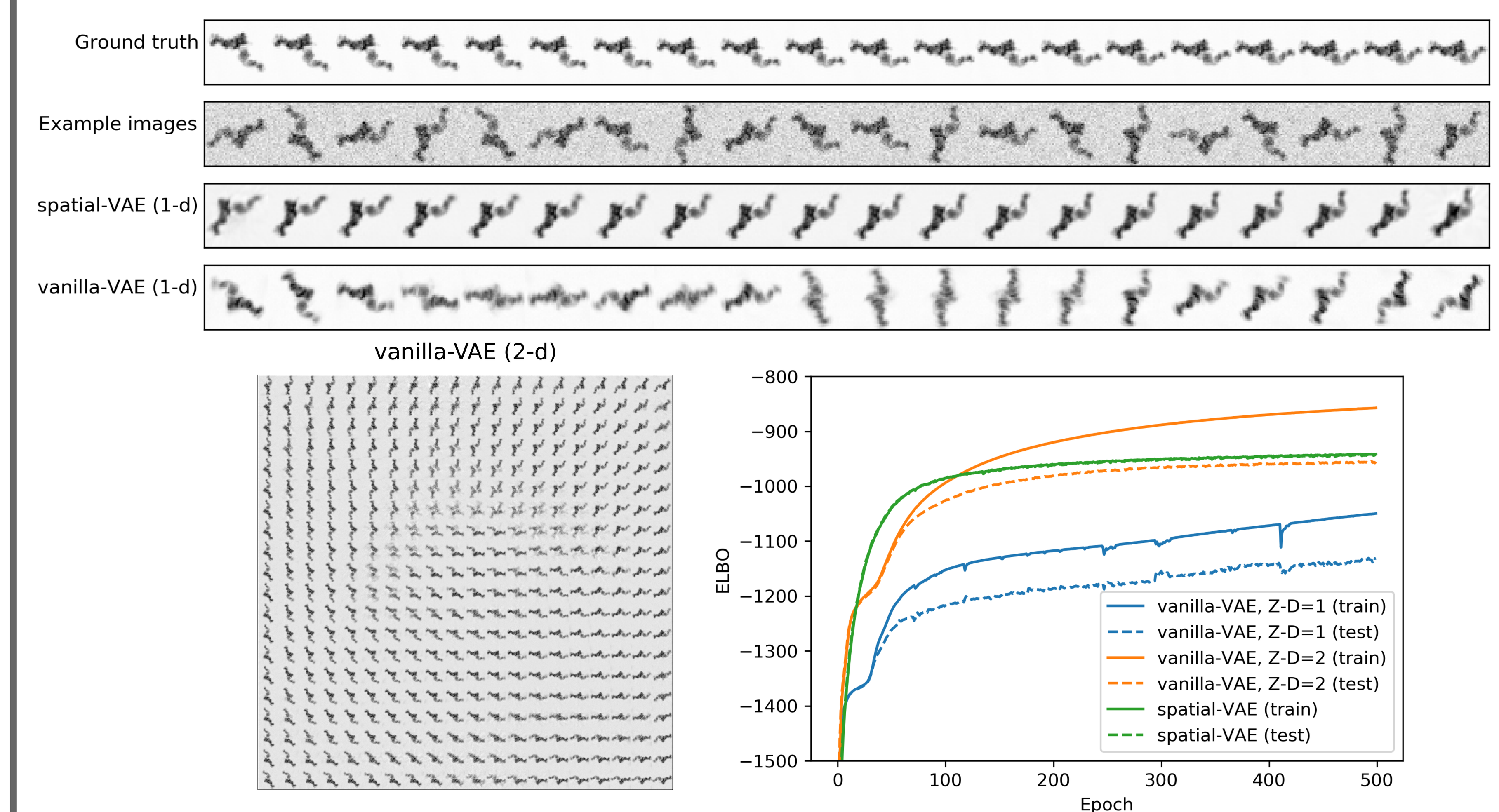
Results

Learning to generate digits from MNIST

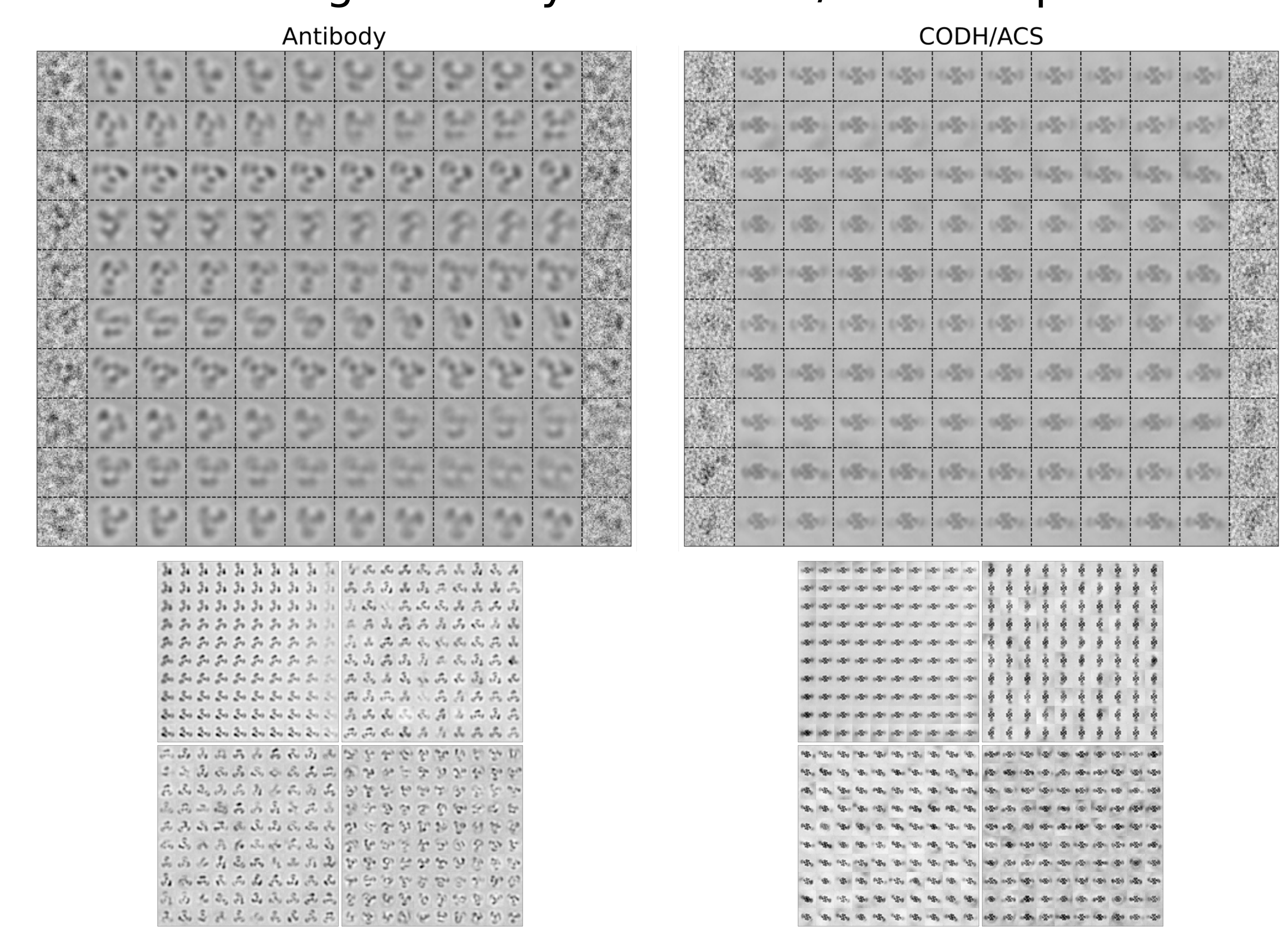




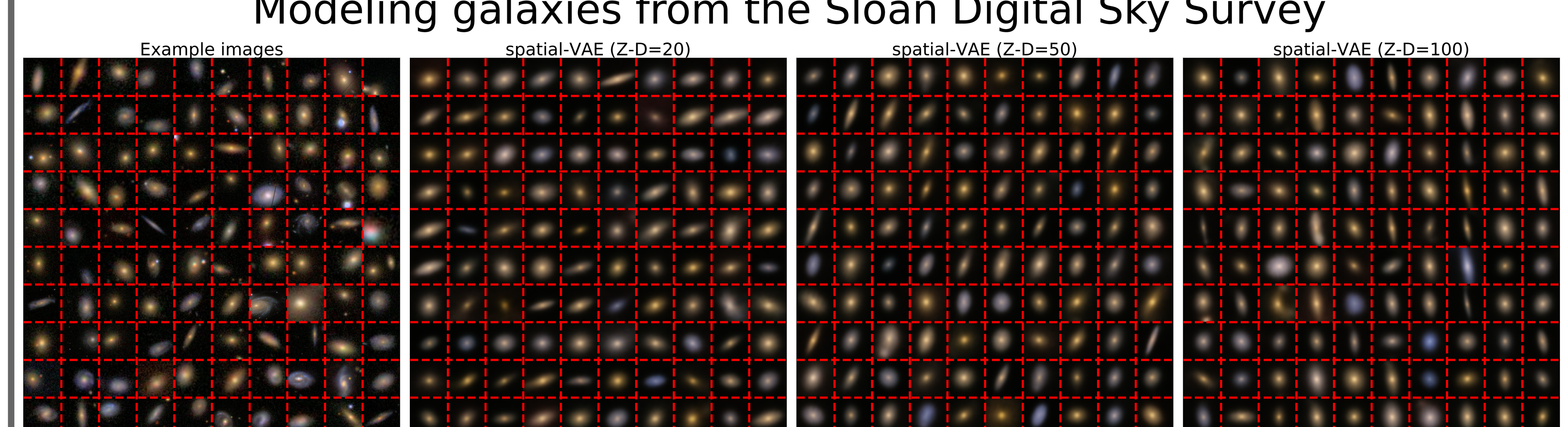
Simulated particle with hinge (5HDB)



Modeling antibody and CODH/ACS complex



Modeling galaxies from the Sloan Digital Sky Survey



Conclusions

- Spatial-VAE enables learning generative models of images that separate latent variables encoding content from rotation and translation
- The framework learns continuous conformations of proteins and galaxies in 2D
- The general framework extends to higher dimension signals and other coordinate transformations