Observations in VAEs

Oct 13th, 2024

Vedant Sachdeva, Ue-Yu Pen, Daniel Tan, Madhav Mani

Sequence-driven approaches for generative modeling in biology

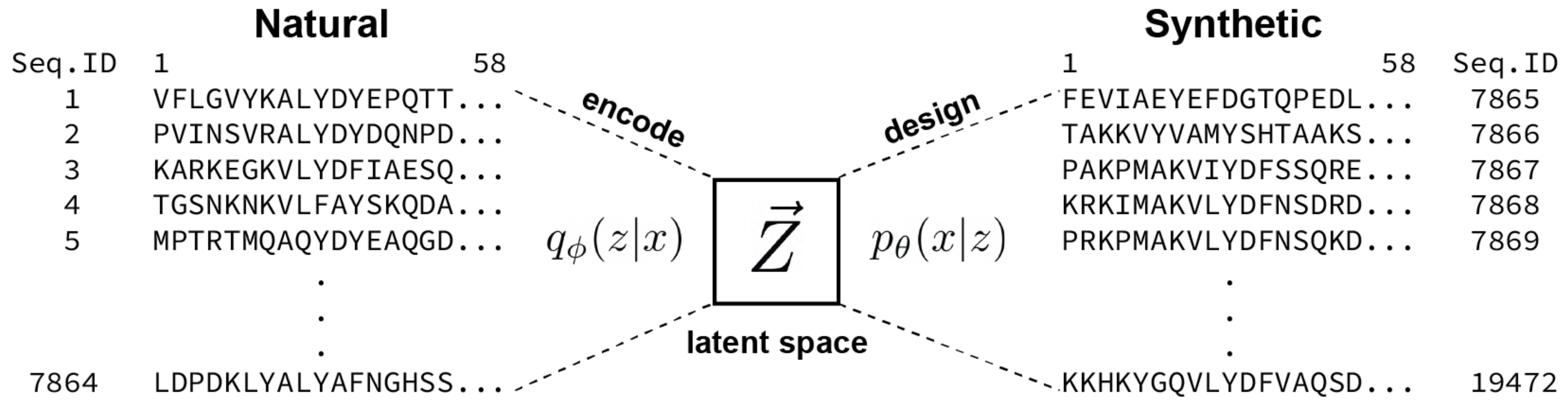
- Sequence-driven approaches leverage correlations from homologous sequences in order to generate novel sequences that obey the observed correlation structure, or predict function based on the sequence data
- Early work involved applying PCA to one-hot embedded sequence data or fitting data to maximum entropy models
- However, such work was largely constrained to considering second order moments

Lockless, 1999
Morcos, 2011

Sequence-driven approaches for generative modeling in biology

- Core principles to take away: Latent space-driven optimization and probability function estimation

- One modeling approach: VAEs (using MMD to enforce prior)



Natural

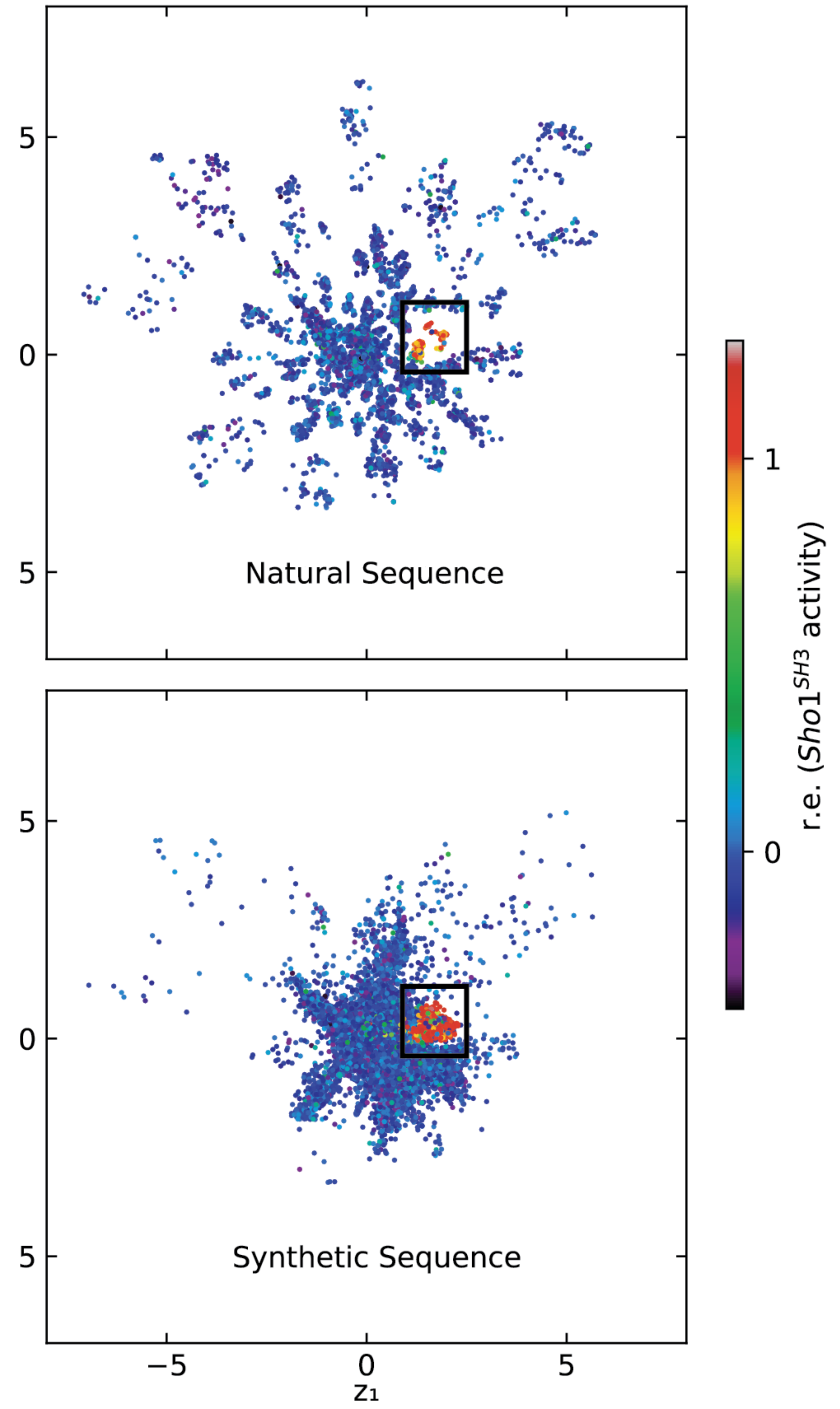
Seq.ID	1 … 58
1	VFLGVYKALYDYEPQTT...
2	PVINSVRALYDYDQNPD...
3	KARKEGKVLYDFIAESQ...
4	TGSNKNKVLFAYSKQDA...
5	MPTRTMQAQYDYEAQGD...
	.
	.
	.
7864	LDPDKLYALYAFNGHSS...

Synthetic

1 … 58	Seq.ID
FEVIAEYEFDGTQPEDL...	7865
TAKKVYVAMYSHTAAKS...	7866
PAKPMAKVIYDFSSQRE...	7867
KRKIMAKVLYDFNSDRD...	7868
PRKPMAKVLYDFNSQKD...	7869
.	
.	
.	
KKHKYGQVLYDFVAQSD...	19472

Variational Autoencoders on SH3 Domains

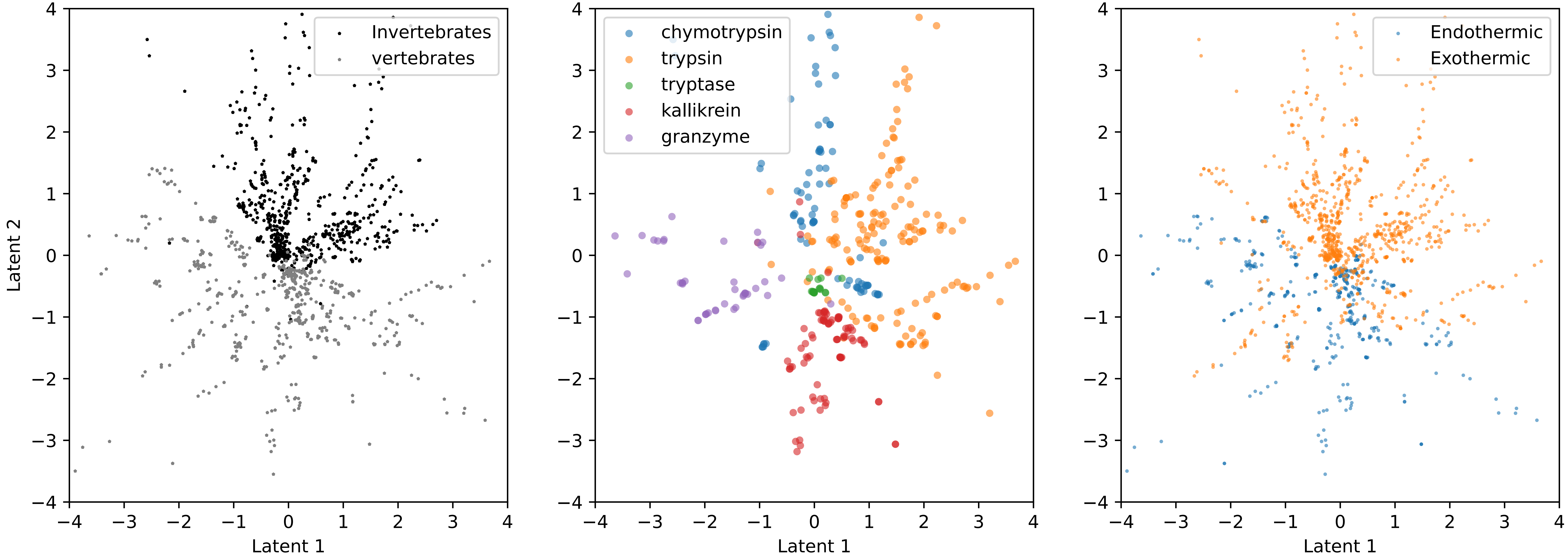
- Fit a VAE with a two-dimensional latent space
    - Colored by measurement of enrichment
    - For Sho1 ligand binding
- Sequences with high phenotype localize
    - Offers easy conditioning approach for design



Data from
Lian, 2021

Variational Autoencoders on Serine Proteases

Good separation across multiple characteristics of the data



Data from
Halabi, 2009

What do VAEs learn?

Extracting second-order moments from VAEs

Candidate Amino Acid Sequence: MAGICTHEGATHERING

Proposed Amino Acid Sequence: MAGICTHEMATHERING

$P(\text{Letter} \mid G \rightarrow M)$

Can infer this change in probability distribution using a VAE

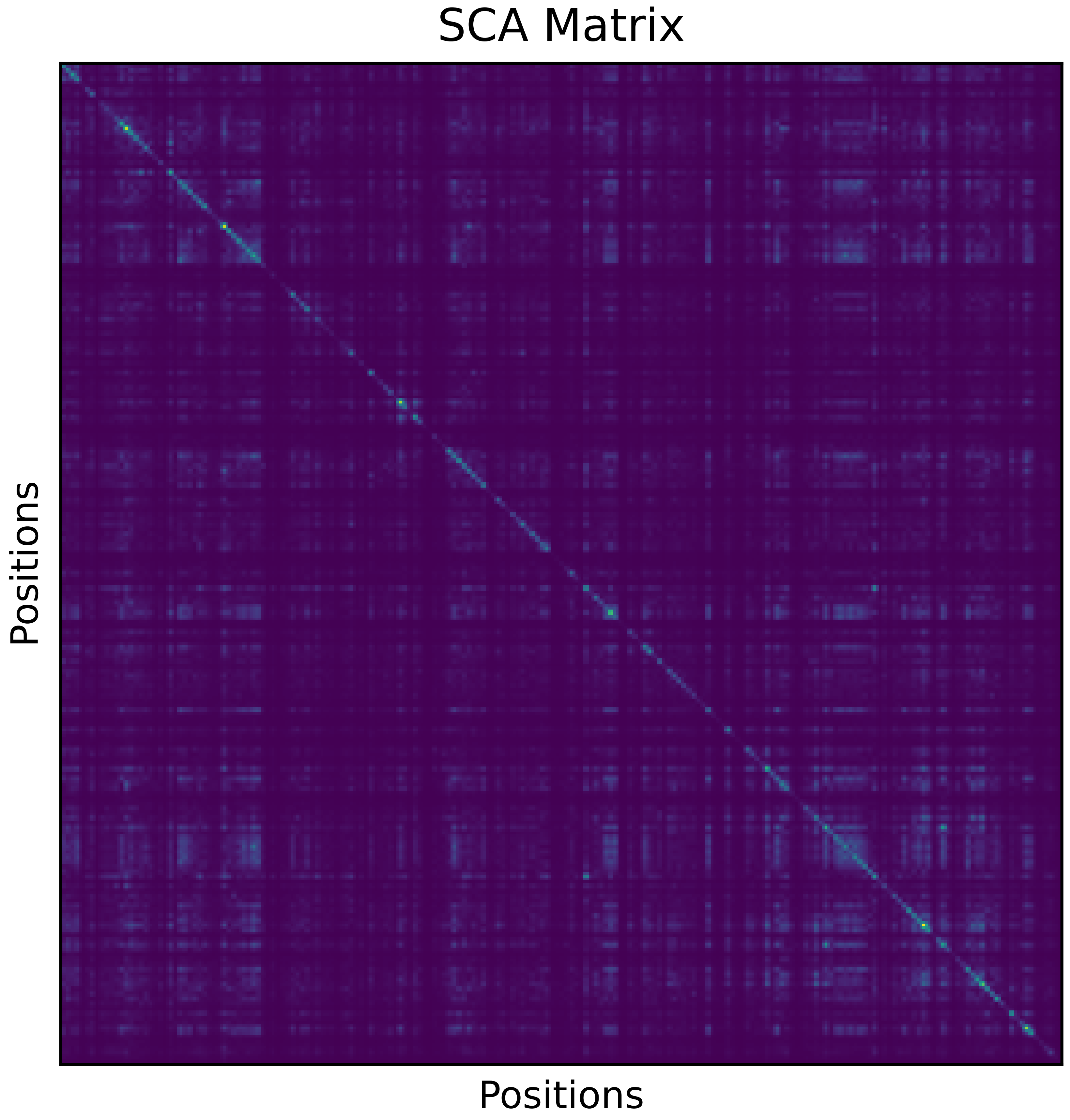
What do VAEs learn?

Extracting second-order moments from VAEs

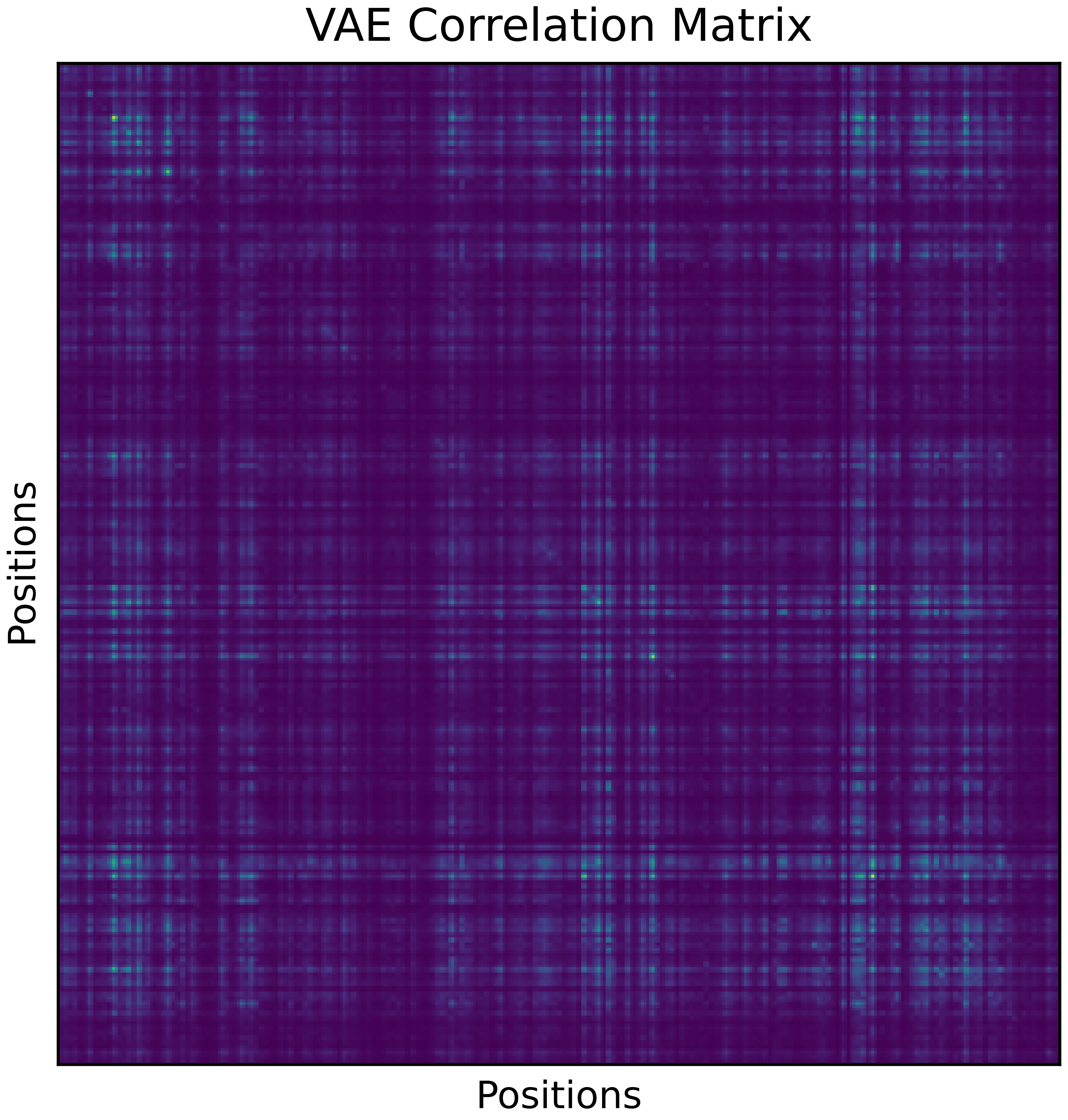
- Full prescription
  - Obtain all sequences 1 hamming unit away from training sequences
  - Encode and decode all perturbed sequences
  - Compute $\delta p(x_i \rightarrow x_i^{'} | x_j \rightarrow x_j^{'}) = \log\left(\frac{p(x_i^{'} | z(x_{syn}))}{p(x_i | z(x_{nat}))}\right)$
    - Where I’ve defined the latent coordinates as a function of the sequence

What do VAEs learn?

A comparison of methods



SCA Matrix

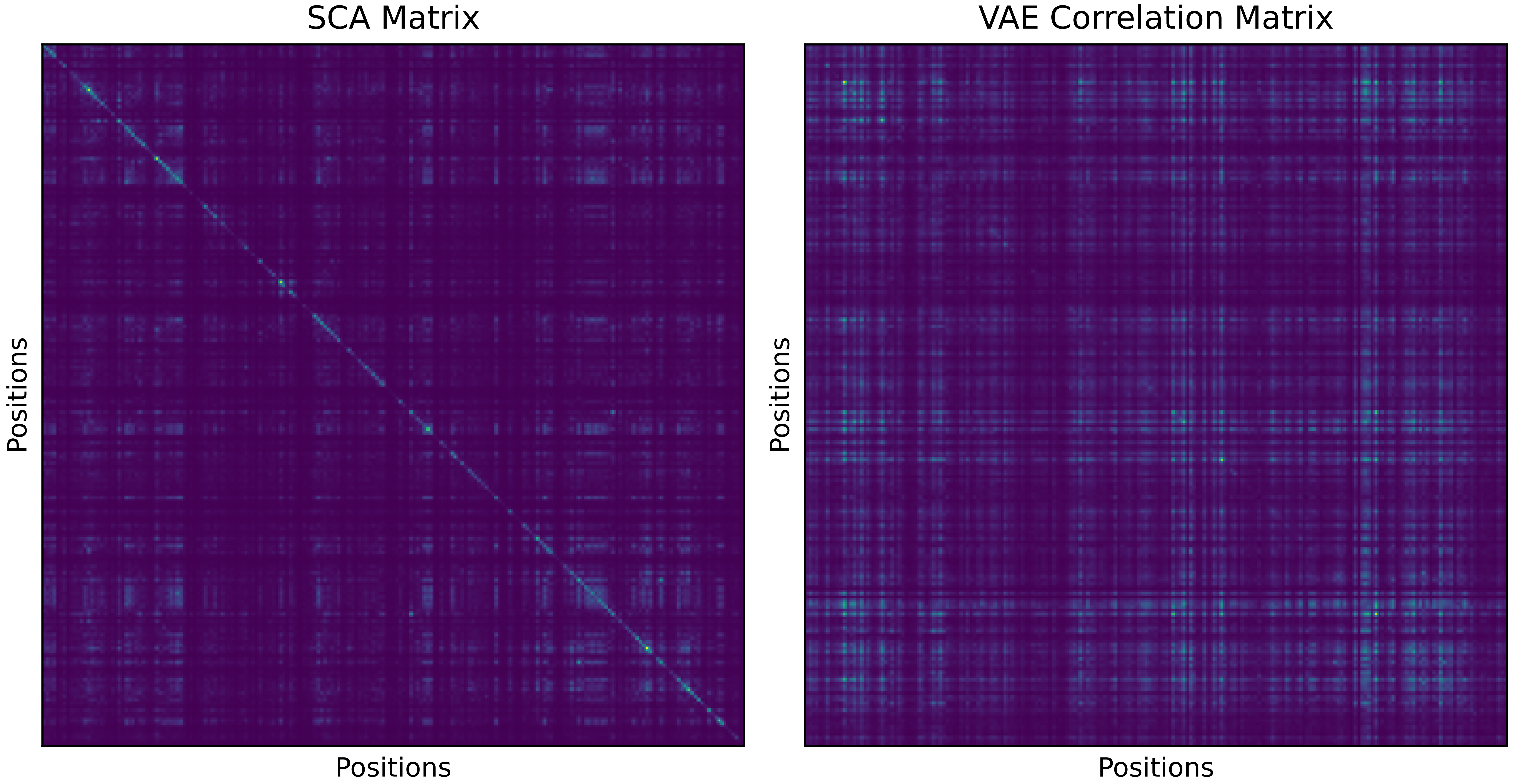


VAE Correlation Matrix

Halabi, 2009

What do VAEs learn?

A comparison of methods



Matrices look similar but not identical

Dot products of top three eigenmodes are greater than 0.5

This is high overlap, given the nature of the space

Halabi, 2009

What do VAEs learn

PCA directions, at the bare minimum

- Agreeing with older literature that VAEs seem to capture at least the information that emerges from PCA

What do VAEs learn

Coevolutionary signals beyond pairwise correlations

- Agreeing with older literature that VAEs seem to capture at least the information that emerges from PCA

- Better separation of the various phenotypes imply capturing coevolutionary signals

- How are such things represented?

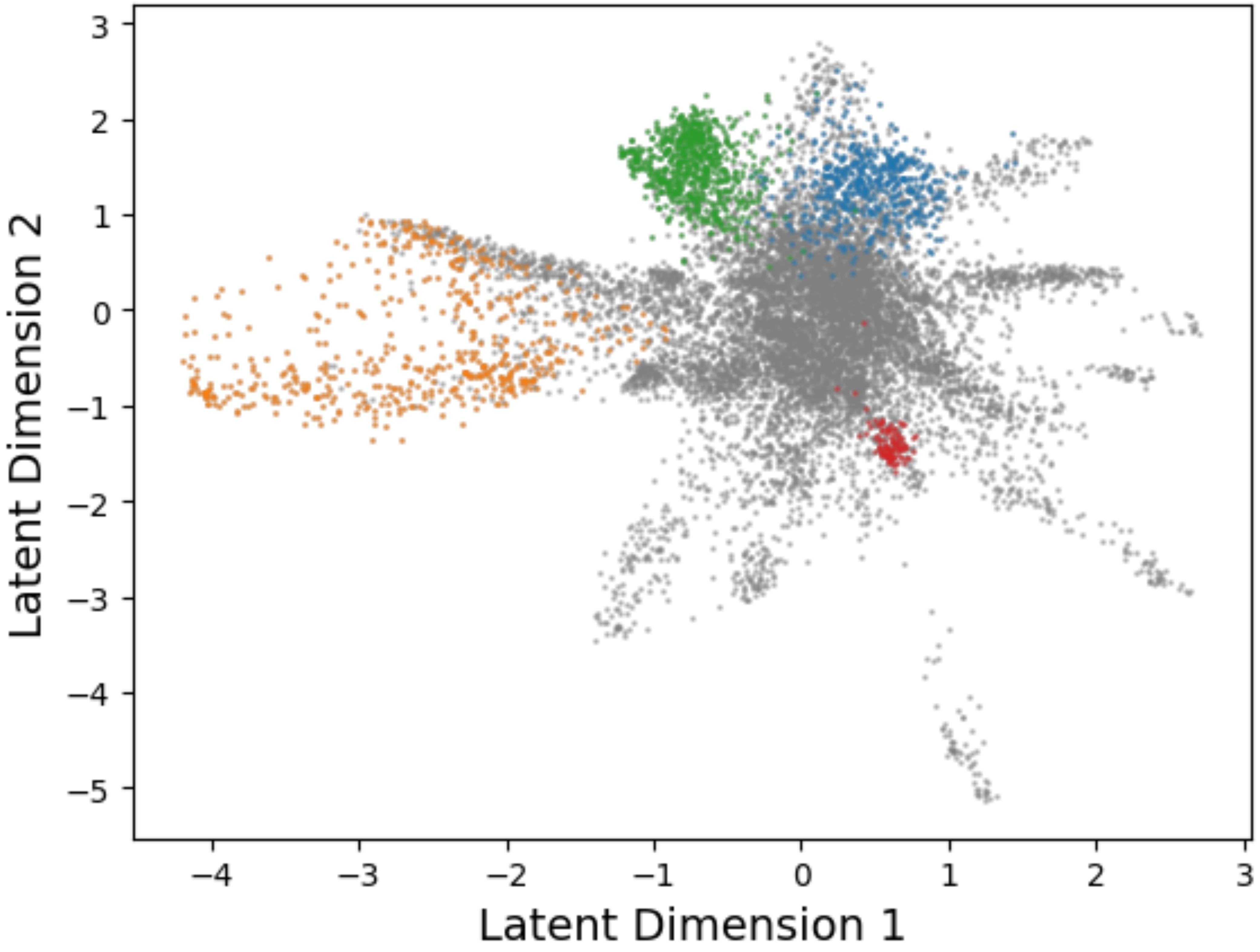
Training VAEs with synthetic data

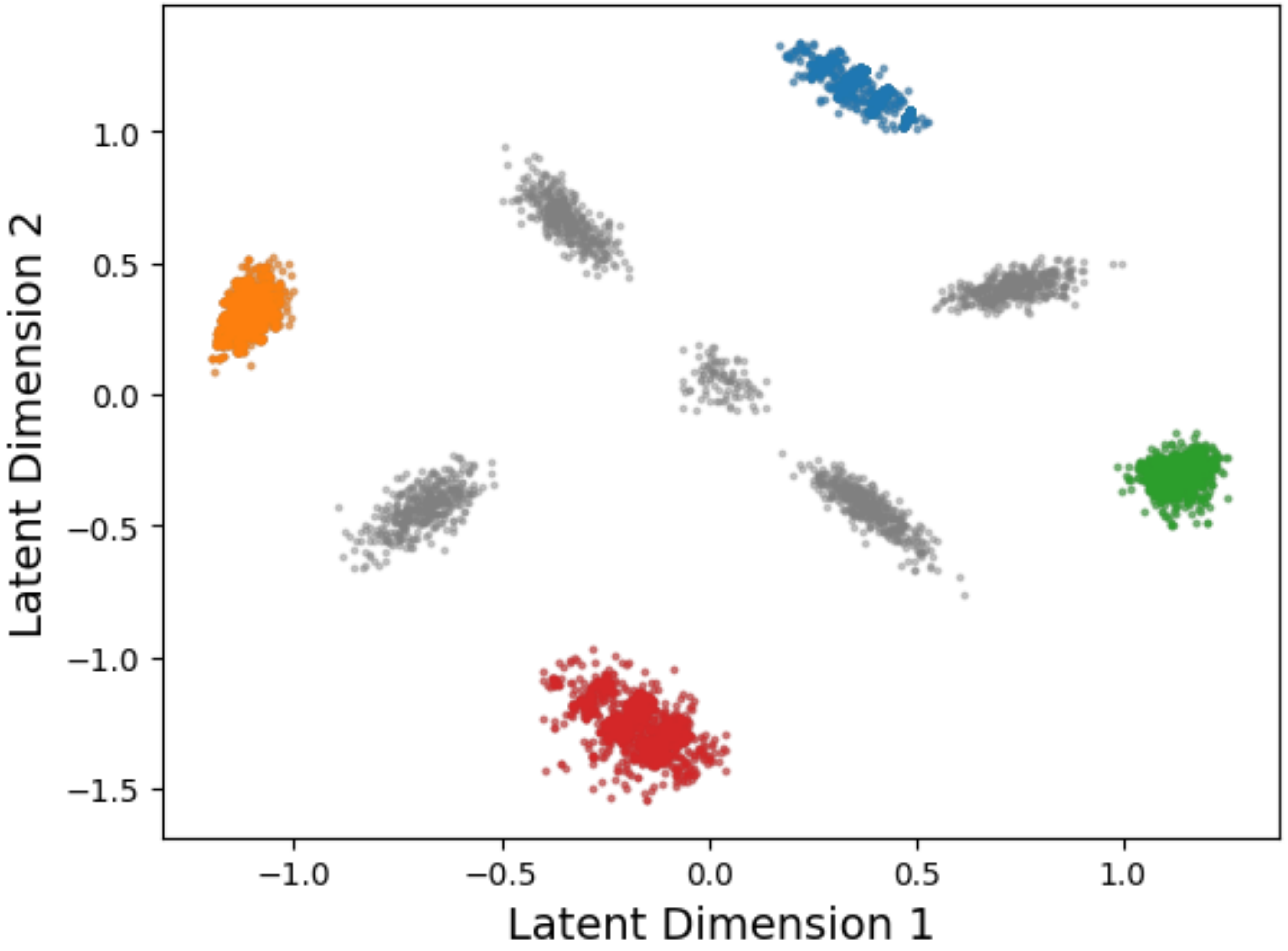
Drawing inspiration from associative memory networks

- Draw data from known probability distribution
- $$P(\vec{S}) \propto \exp(\beta(\sum_{\mu} \vec{S} \cdot \vec{\xi}^{\mu}_{\text{short}} + \sum_{\mu} (\vec{S} \cdot \vec{\xi}^{\mu}_{\text{long}})^2))$$
- Short motifs correspond to motifs on subsegments of sequences, while long motifs correspond to full length sequence motifs
- Inspired by observations in SH3 Domain data

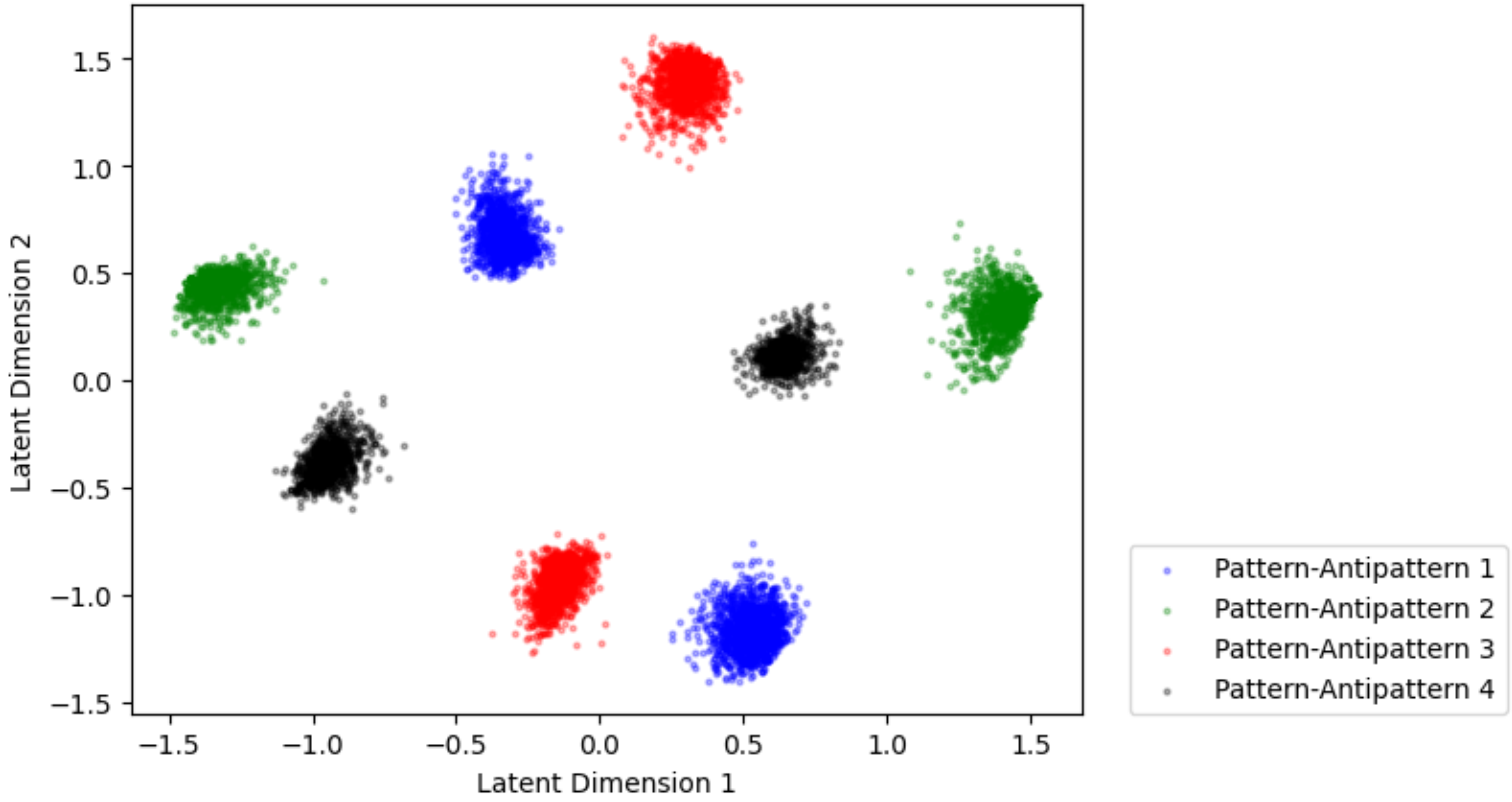
SH3 VAE vs Synthetic Data VAE

Model seems to capture some of the phenomenology observed in VAEs.
Notably, clustering of sequences in Hamming Space





Overcompression in VAEs



Anti-correlated data points are placed as far away from one another as possible

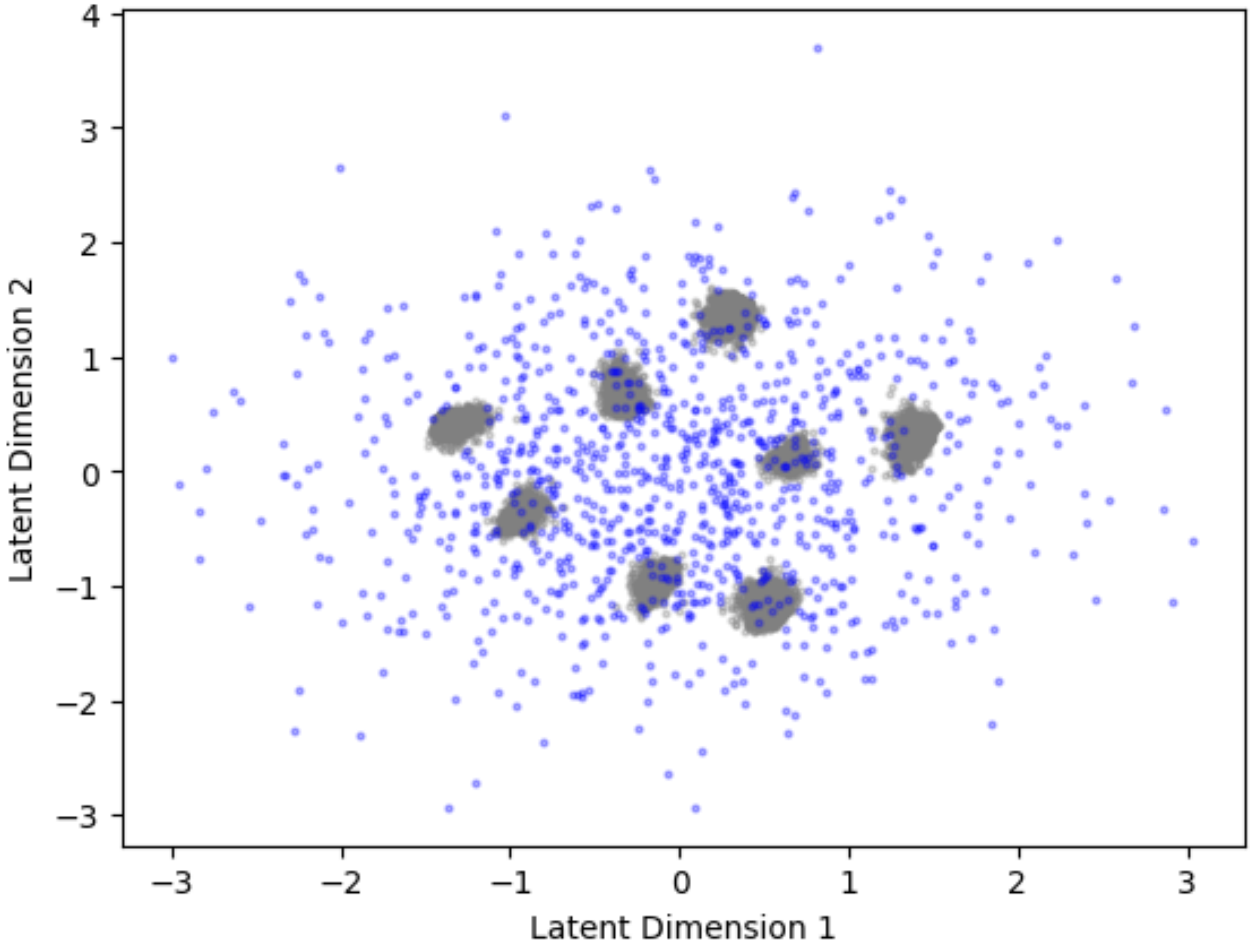
Seeming emergence of polysemanticity observed in Elhage et. al. 2022

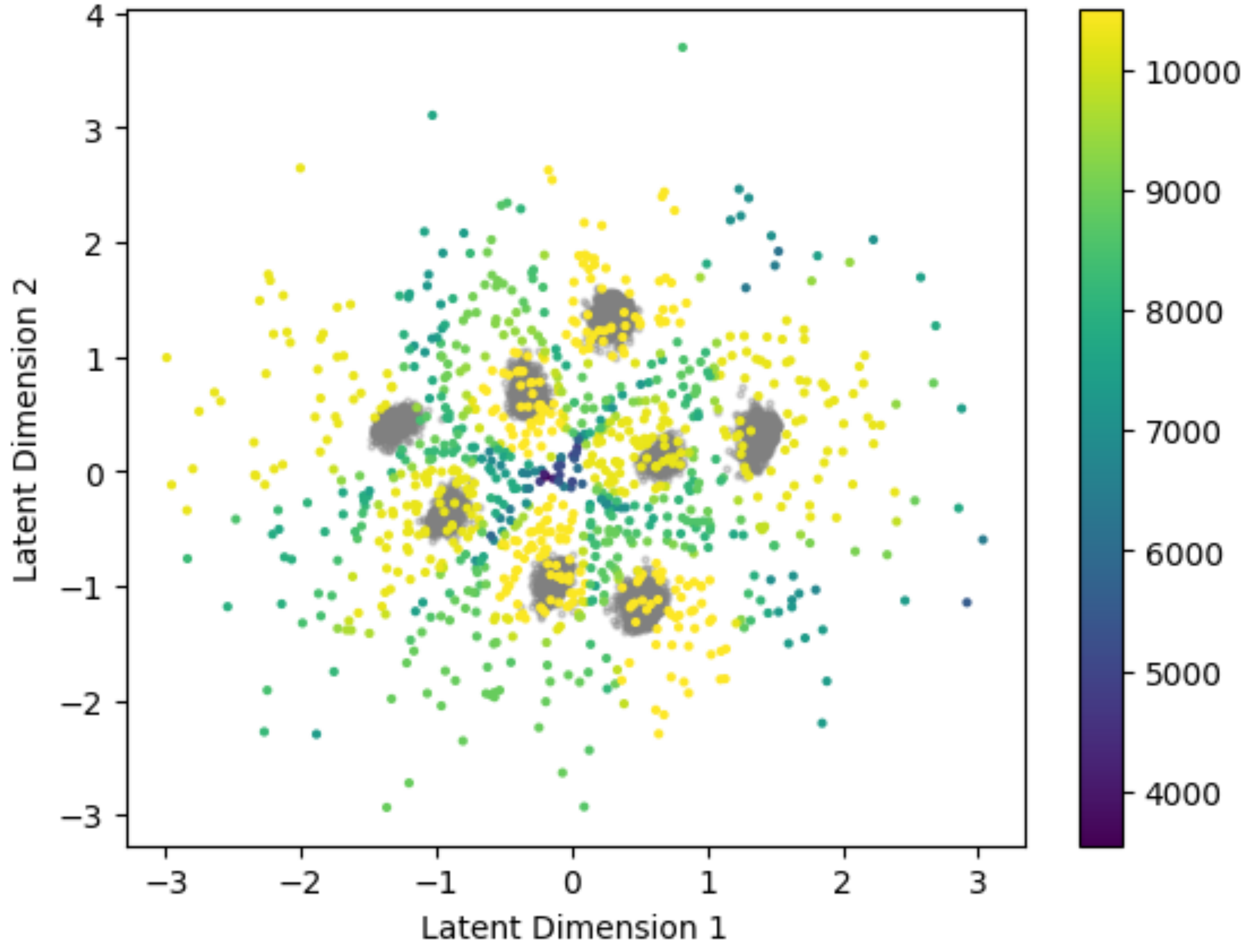
Overcompression in VAEs

Impact on generativity

$$P(\vec{S}) \propto \exp(\beta(\sum_{\mu} \vec{S} \cdot \vec{\xi}^{\mu}_{\text{short}} + \sum_{\mu} (\vec{S} \cdot \vec{\xi}^{\mu}_{\text{long}})^2))$$

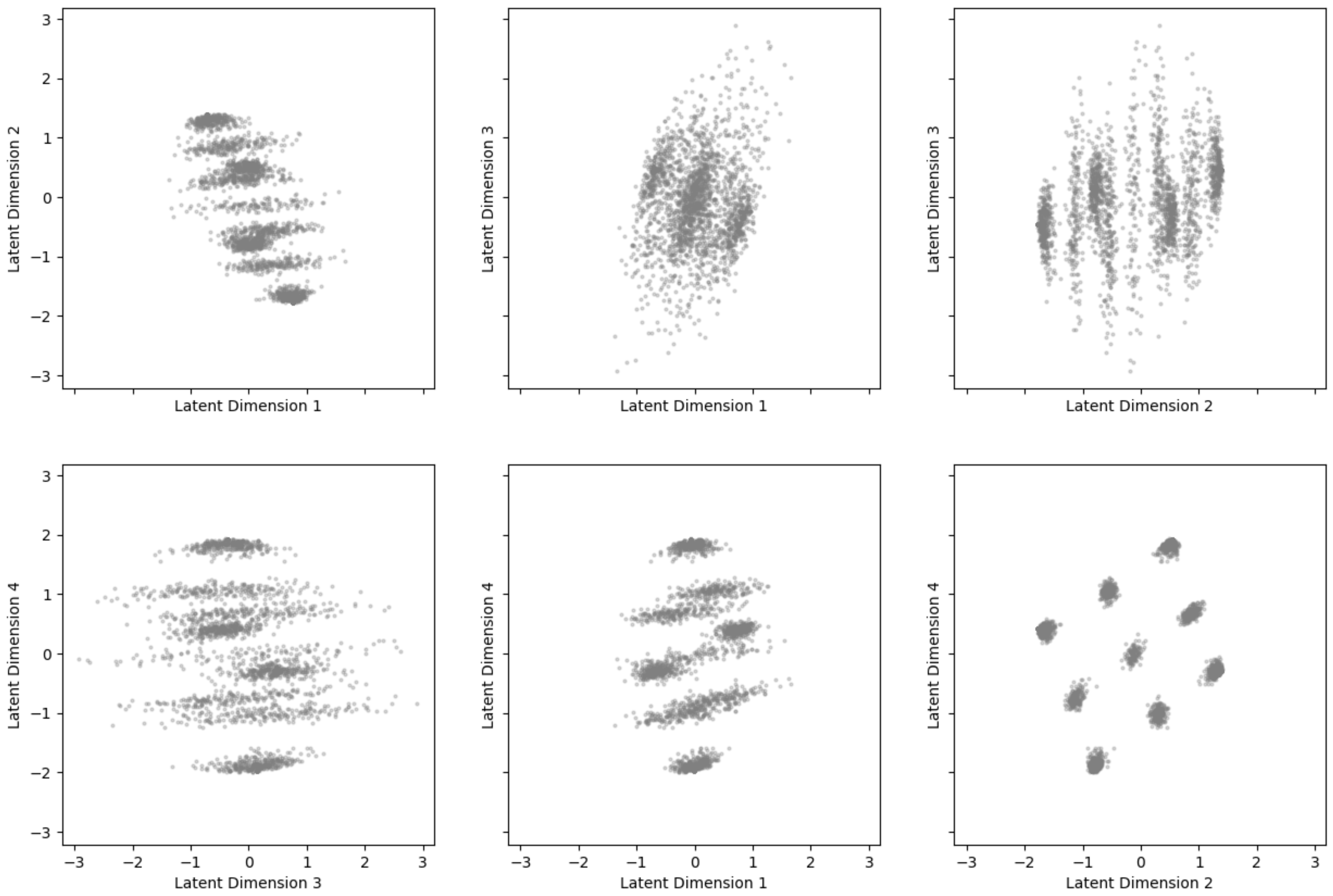
Model generates reasonable sequences, but are all reasonable sequences generated?





Undercompression in VAEs

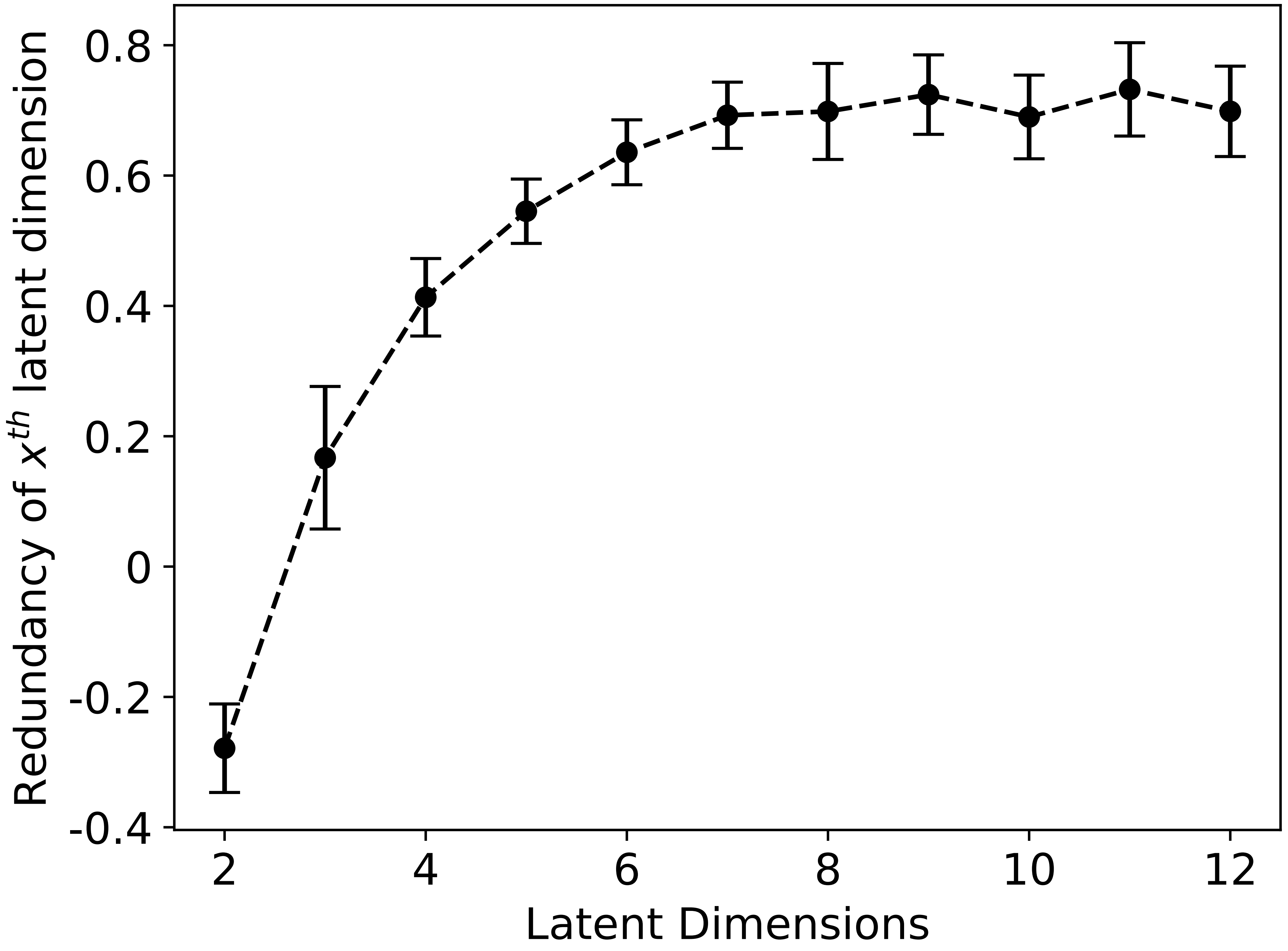
Undercompression: 2 patterns, 4 latent dimensions



Clearly, there exists some subspace where the relevant dimensions are

So what do the other dimensions encode?

Undercompression leads to redundancy



In SH3, we observe additional dimensions are redundant

Not shown: higher latent dimensions leads to instability in sequences generated from ‘empty’ parts of latent space

Work done with Ue-Yu Pen and Madhav Mani

Why is redundancy emerging?

- MMD only constrains on first two moments distribution to have pairwise distances mirroring that of a multivariate gaussian
    - Is there a way to shove information into higher order moments without changing the first and second order moments?
- Is this bad?
    - Reconstruction of generated sequences worsens as redundancy rises
    - “Patchy” latent space

Beta-VAEs for Feature Engineering: The Impact of Over- and Under-Compressing with VAEs

Project pursued by intern: Daniel Tan

- Let’s strengthen this claim: use beta-VAE on controlled synthetic dataset
- Suppose we have $u_1,\ u_2, \ldots, u_i \in \mathbb{R}^{100}$ where $u_i \bullet u_j \cong 0 : i \neq j$
- Construct a dataset of linear combinations of these *i* basis vectors. Our case: $i = 6$
- $D = \{ \sum_i \alpha_i u_i : \alpha_i \sim N(0,1) \}$
- Data has six independent factors: though the dataset is 100-dimensional, it spans only six dimensions
- We can tune beta to select out 6 out of 10 dimensions…
- On real data, we don’t know underlying dimensionality

Selective Posterior Collapse as a useful regularizer

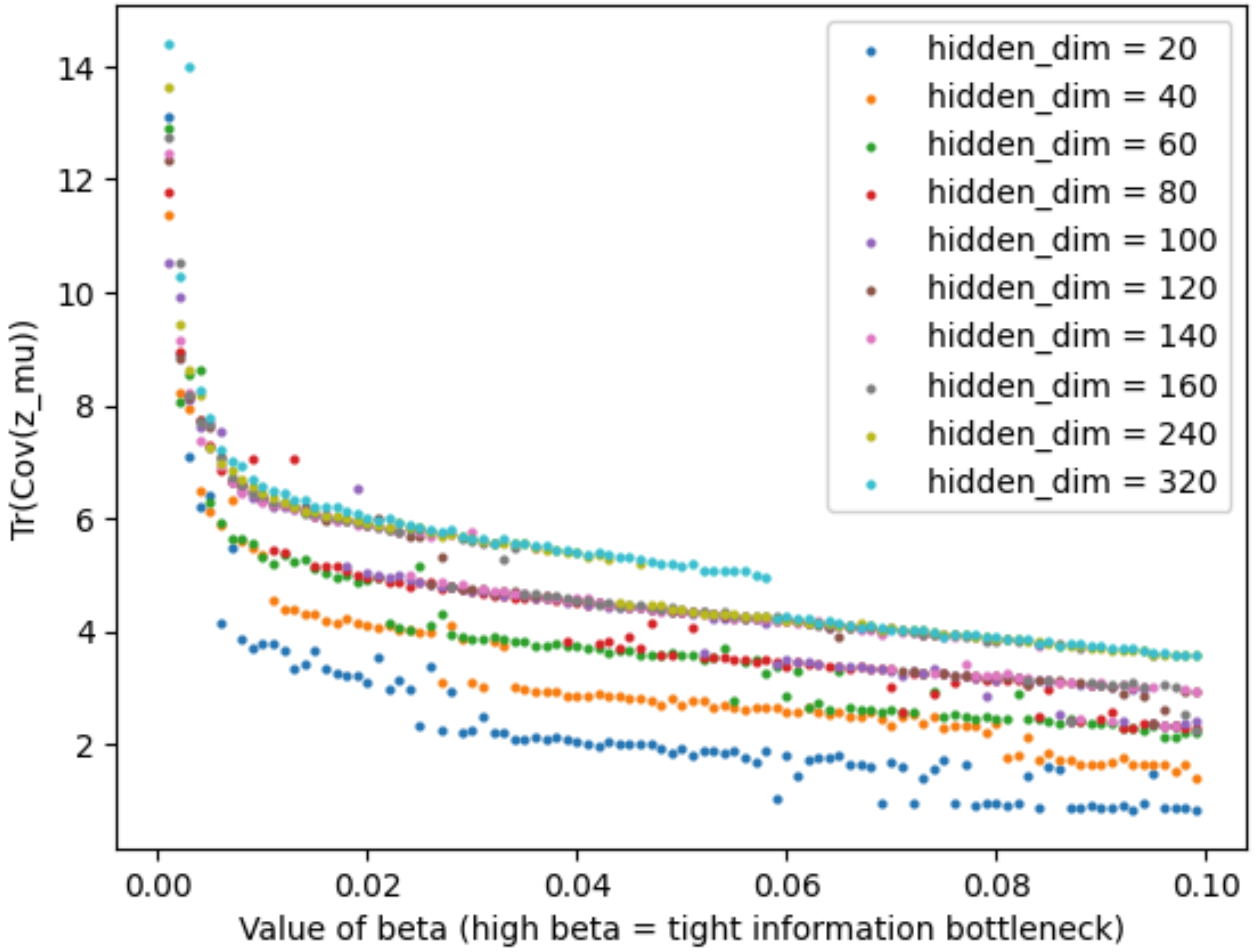
Project pursued by intern: Daniel Tan

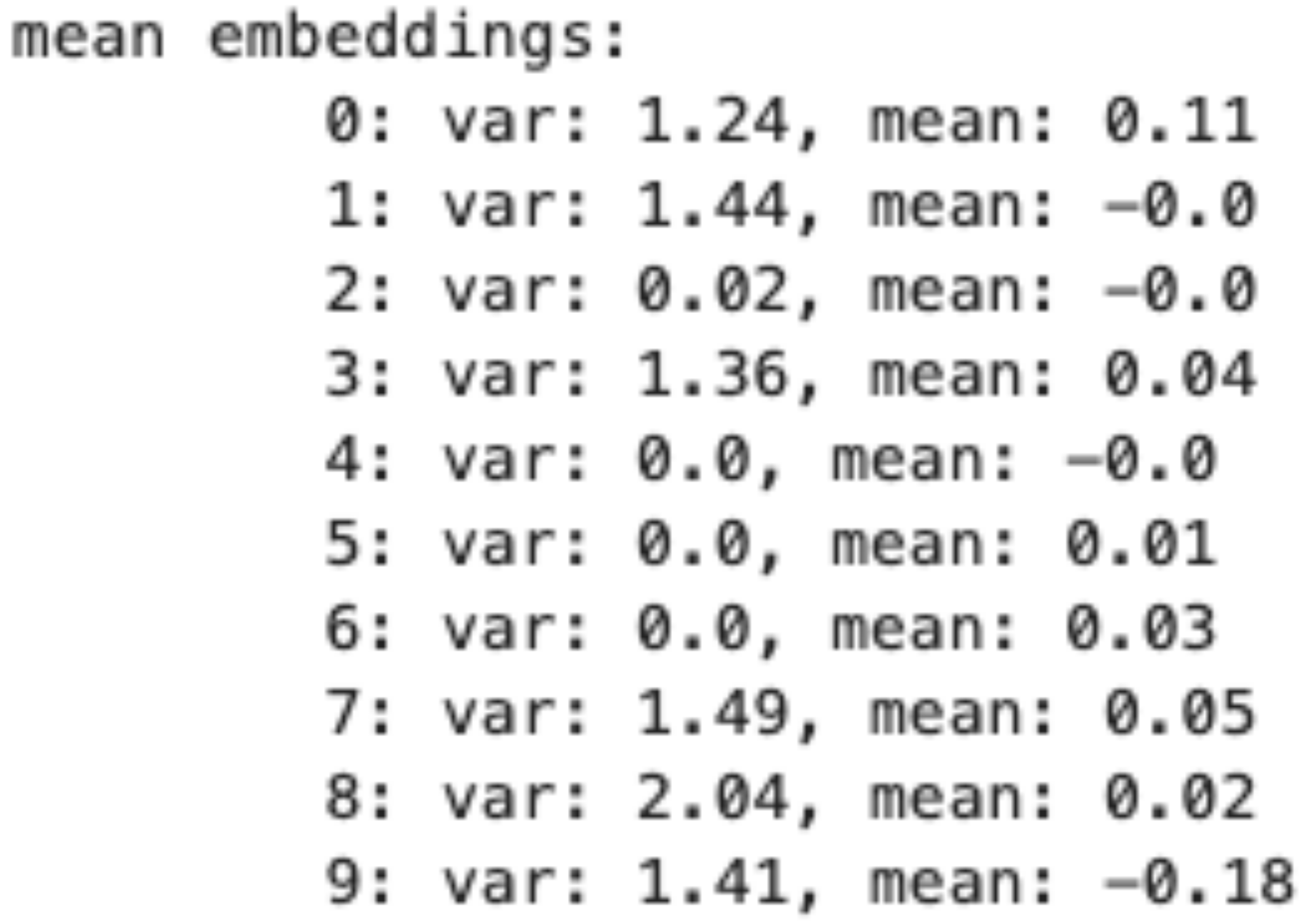
- KL beta-VAE does not continuously decrease variance of dimension with increasing beta; it is either ~1 or ~0

$$Tr\left(Cov\left(z_\mu\right)\right) = \sum_i Var\left(z_{\mu,i}\right)$$

Selective posterior collapse: trace of covariance matrix of coordinates of mean embeddings vs. size of info bottleneck

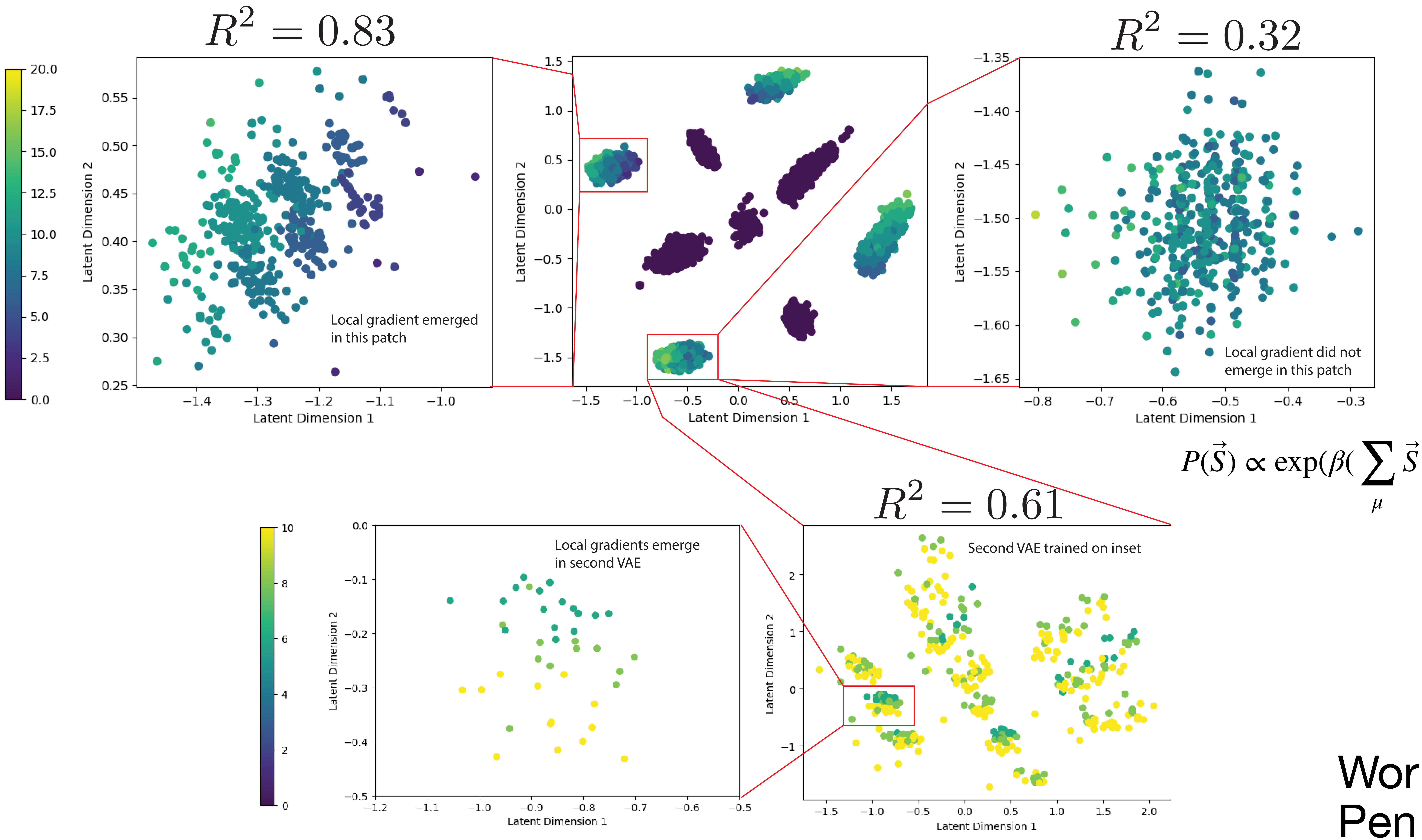
Discretized changes in trace, showing sudden dimension drop out





```
mean embeddings:
        0: var: 1.24, mean: 0.11
        1: var: 1.44, mean: -0.0
        2: var: 0.02, mean: -0.0
        3: var: 1.36, mean: 0.04
        4: var: 0.0, mean: -0.0
        5: var: 0.0, mean: 0.01
        6: var: 0.0, mean: 0.03
        7: var: 1.49, mean: 0.05
        8: var: 2.04, mean: 0.02
        9: var: 1.41, mean: -0.18
```

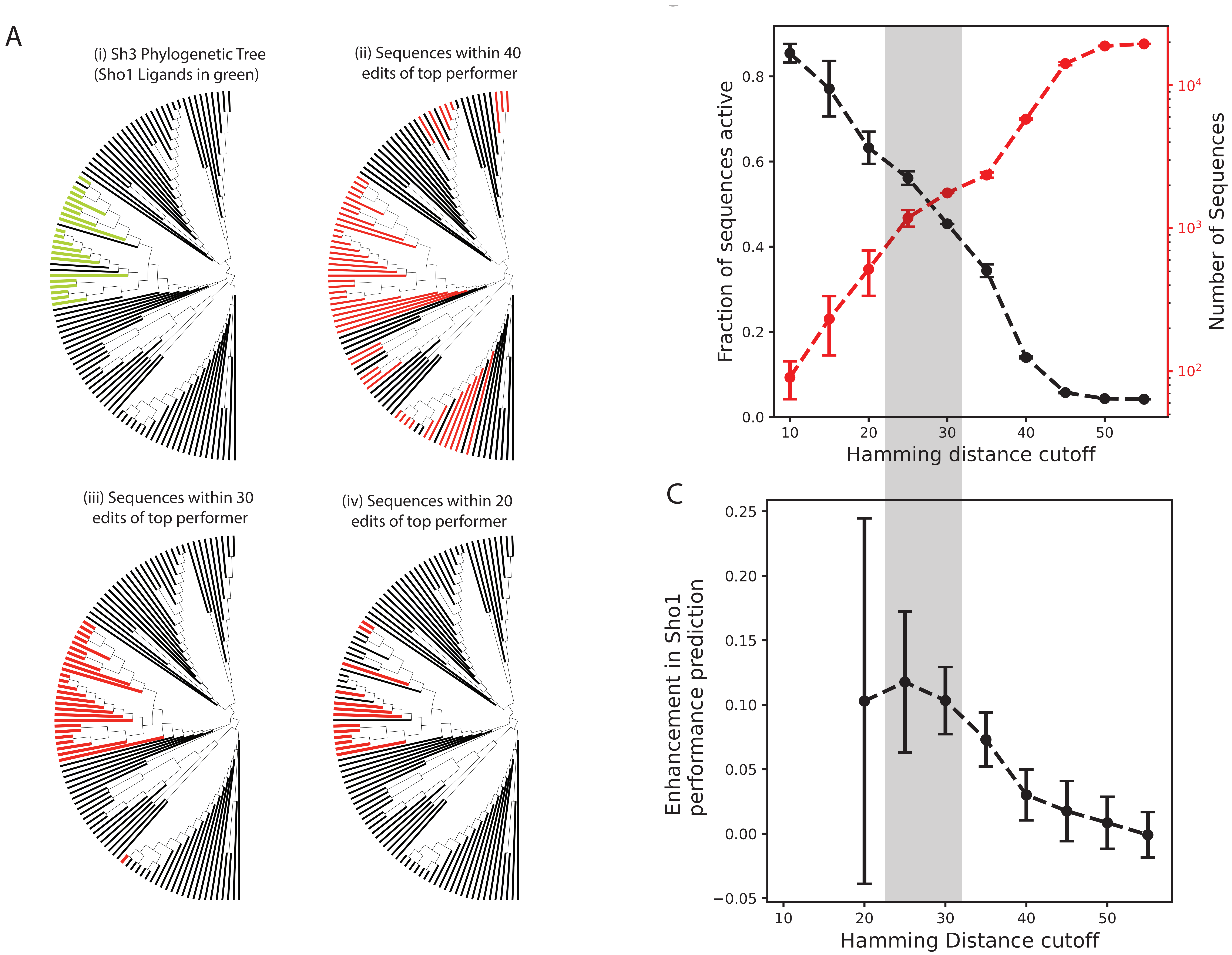
Simpson’s Paradox in VAEs



$$P(\vec{S}) \propto \exp(\beta(\sum_{\mu} \vec{S}\cdot\vec{\xi}^{\mu}_{\text{short}} + \sum_{\mu}(\vec{S}\cdot\vec{\xi}^{\mu}_{\text{long}})^2))$$

Work done with Ue-Yu Pen and Madhav Mani

Observed in real proteins as well

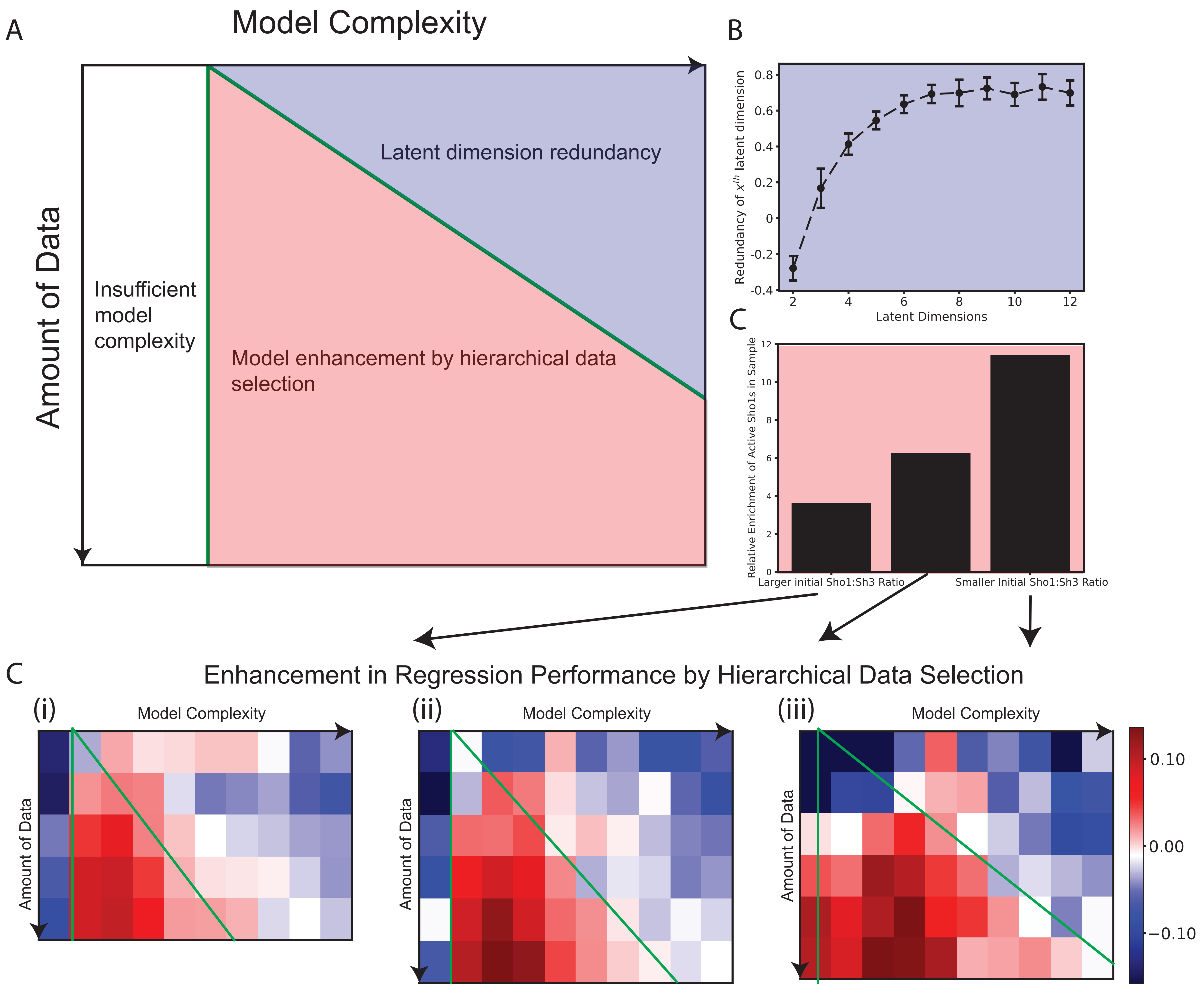


By subselecting sequences, we can focus very specifically on a part of a phylogenetic tree of interest

This is relevant, where phylogenetic signal might dominate the variation in the sequence signal

Work done with Ue-Yu Pen and Madhav Mani

A phase diagram where this effect emerges



Region of interest gets larger relative enrichment of proteins of interest increases

Work done with Ue-Yu Pen and Madhav Mani

Forward directions

- What order moments are important for learning on protein sequence data?
  - Write activation functions using Taylor expansion, and determine how performance of model depends on the order of the Taylor expansion in one layer networks
- Contrastive learning to control what’s learned in unsupervised settings
  - Use unsupervised latent space methods to define sequences that have background variation and sequences that have variation of interest to learn specifically what defines variation amongst a small subfamily of proteins
- Linking what ML models learn to physical properties of proteins or inducing inductive biases therein
  - Do models learn energy landscapes that have physical relevance? Could a model be constrained to learn the energetic landscape that explains conformational changes in a protein?
- Enhancing generative models through the use of associative memory networks
  - “Energy Transformer” as a new paradigm for studying the landscape of proteins