Systematic characterization of generative models for de novo design of regulatory DNA

Nic Fishman, Avanti Shrikumar, Georgi Marinov, Anshul Kundaje
Stanford University

sequence generation: generate sequences which score highly on some property of interest (gene expression, chromatin accessibility, etc.)

This talk

1. Introducing a taxonomy
2. Develop an evaluation tool: 1NN accuracy
3. The apparent tradeoff between (1) biological realisticness and (2) maximizing the property of interest

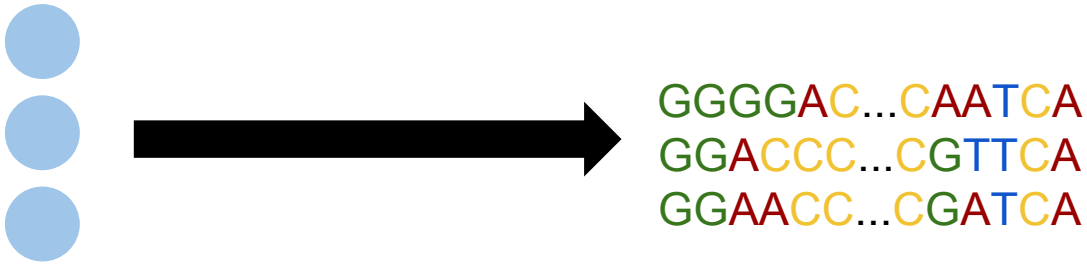
Methods for sequence generation

- VAE
- GAN
- supVAE
- FBGAN
- CEMPI
- RWR
- DbAS
- CbAS
- Differentiable Gradient Descent
- GPs and linear approximation
- Top percentile random sampling

The Taxonomy

Transducer (→ Tuner) → Sampler

transducers: maps random samples in a latent space Z to a generated regulatory sequence



Z (sampled)

Transducer

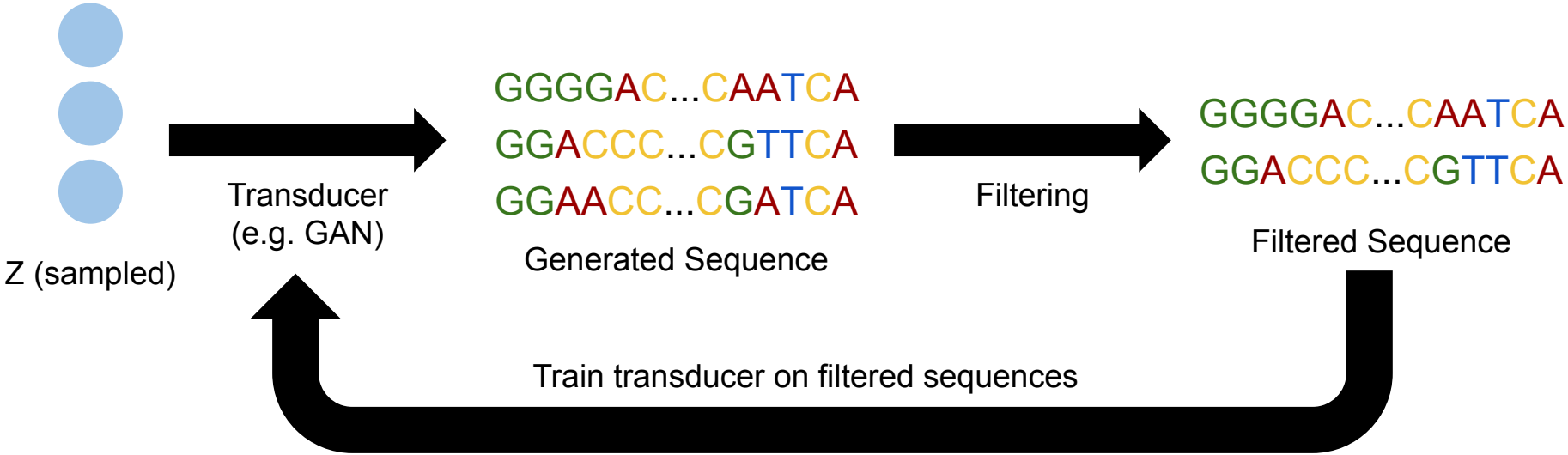
GANs = generator

VAEs = decoder

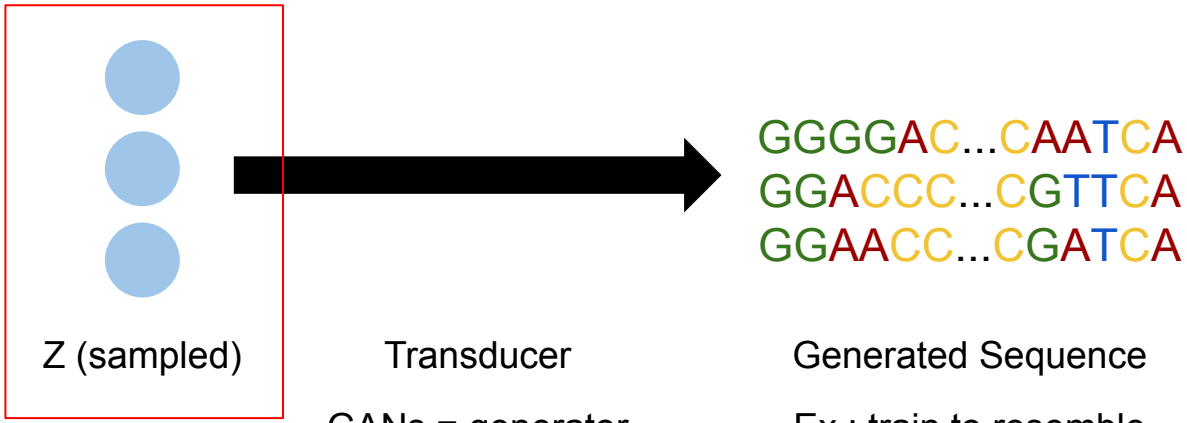
Generated Sequence

Ex.: train to resemble MPRA library

Transducer Tuners: tweak a transducer to generate high-expression regulatory sequences



Sample Optimizers: keep the transducer as-is and SEARCH the latent space for points that give high expression sequences.



Transducers	Tuners	Samplers
• VAE • GAN • supVAE • supGAN	• FBGAN • CEMPI • RWR • DbAS • CbAS	• Differentiable Gradient Descent • GPs and linear approximation • Top percentile random sampling

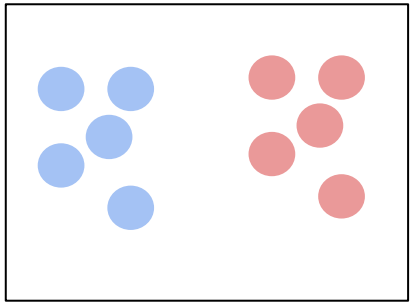
Unified codebase: https://github.com/kundajelab/seq_gen

Quantifying biological realisticness

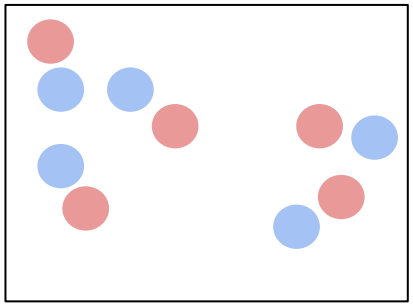
Xu et al:

- Embed generated & real sequences → 1-NN classification

High 1-NN accuracy
(Real sequences separated from generated sequences)



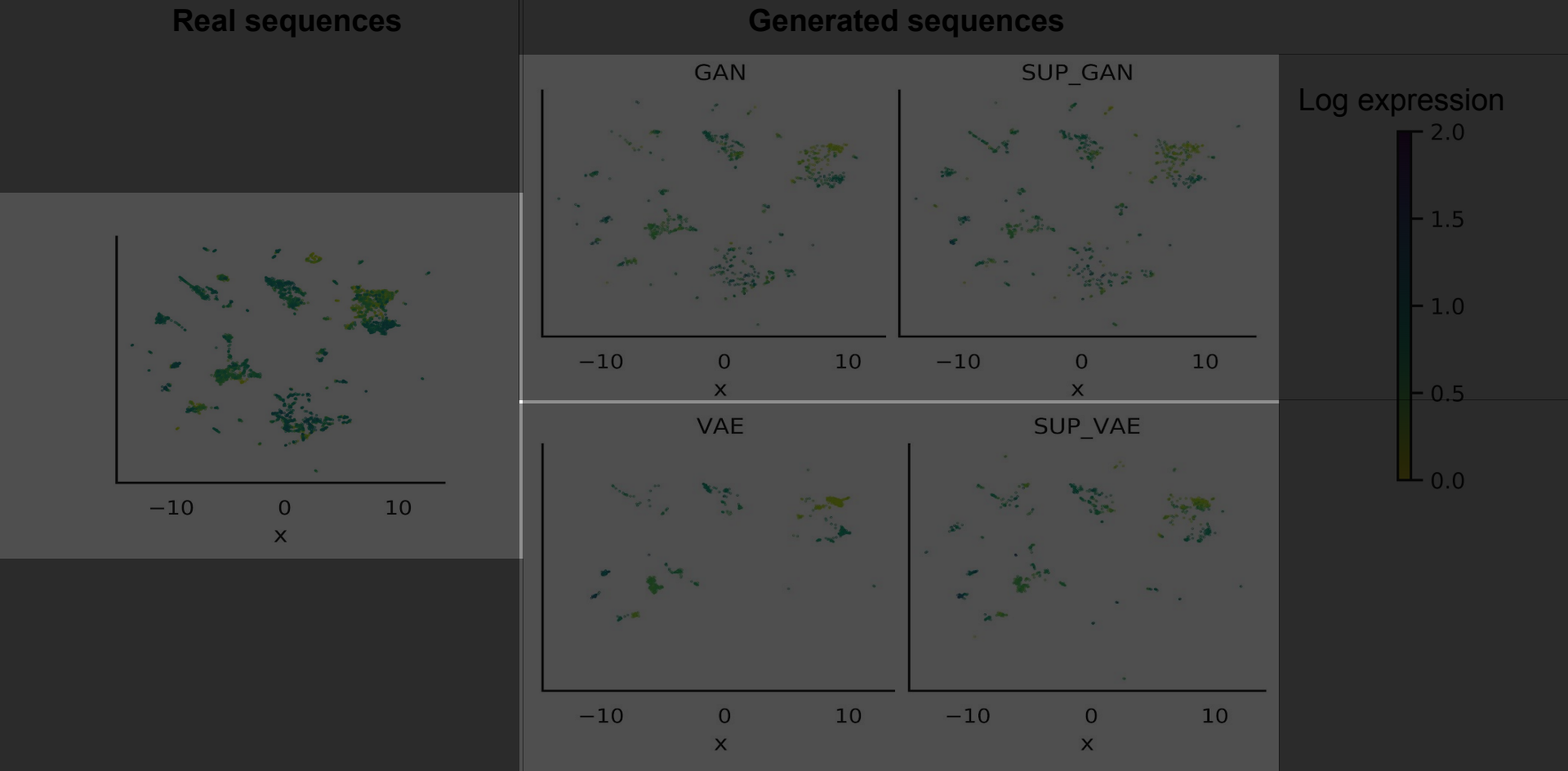
Low 1-NN accuracy
(Real sequences similar to generated sequences)



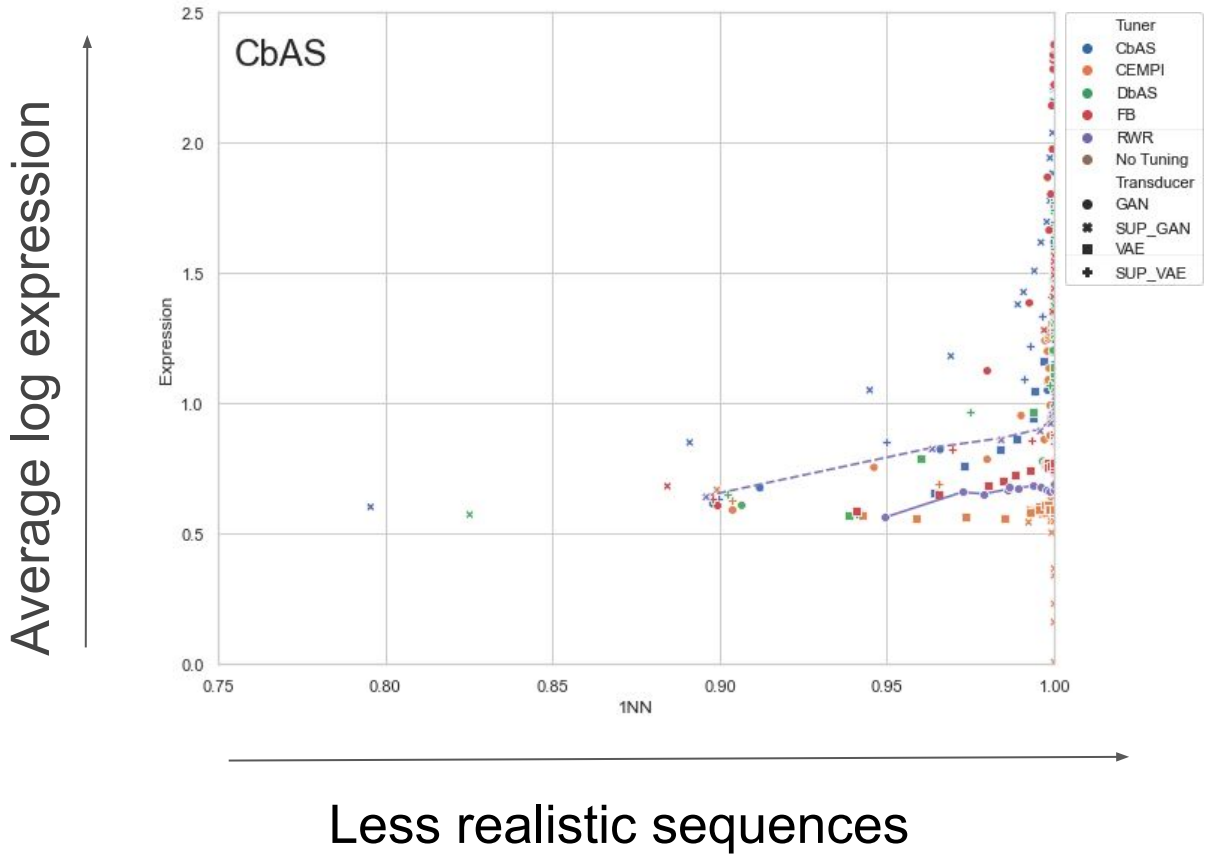
For embedding:

Activations of intermediate layer in DeepSEA model (Zhou & Troyanskaya, 2015)

Takeaway 1: GANs generate greater diversity than VAEs



Takeaway 2: trade-off between realism & maximization



Takeaway 3: Too much tuning produces adversarial sequences

Reasonable generated sequence

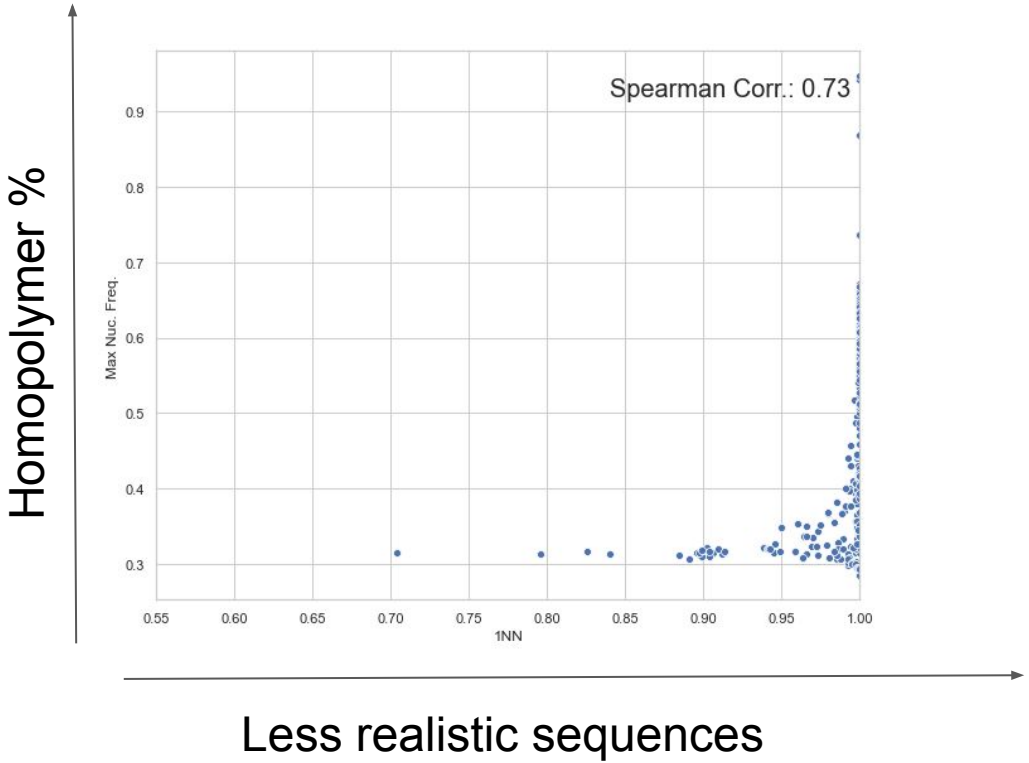
GGGGAC...CAATCA

Mean predicted expression, but biologically realistic

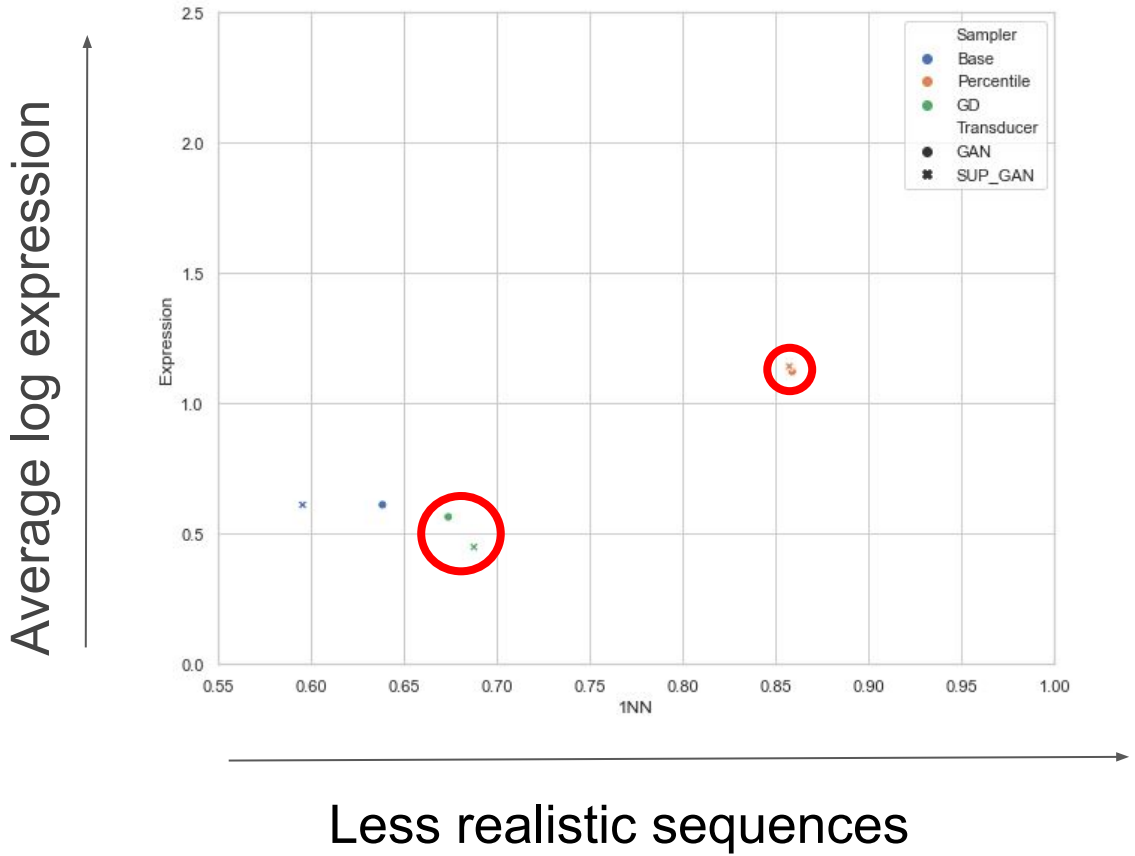
Adversarial sequence

TTTTTT...TTTTCA

High predicted expression, not biologically realistic



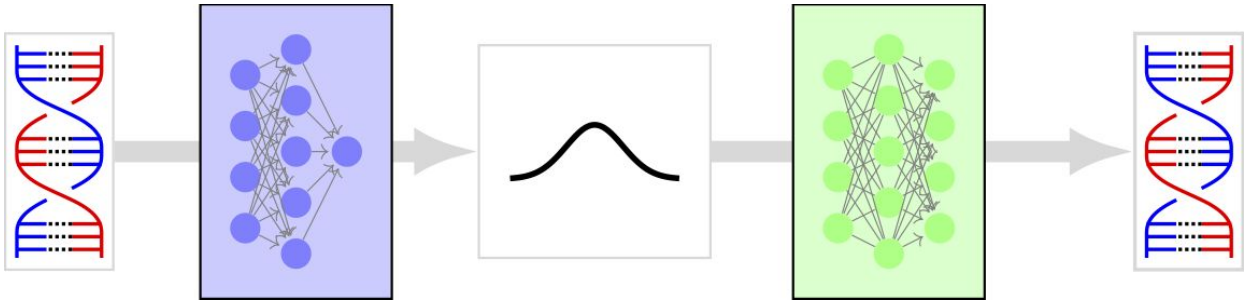
Takeaway 4: random sampling beats gradient descent



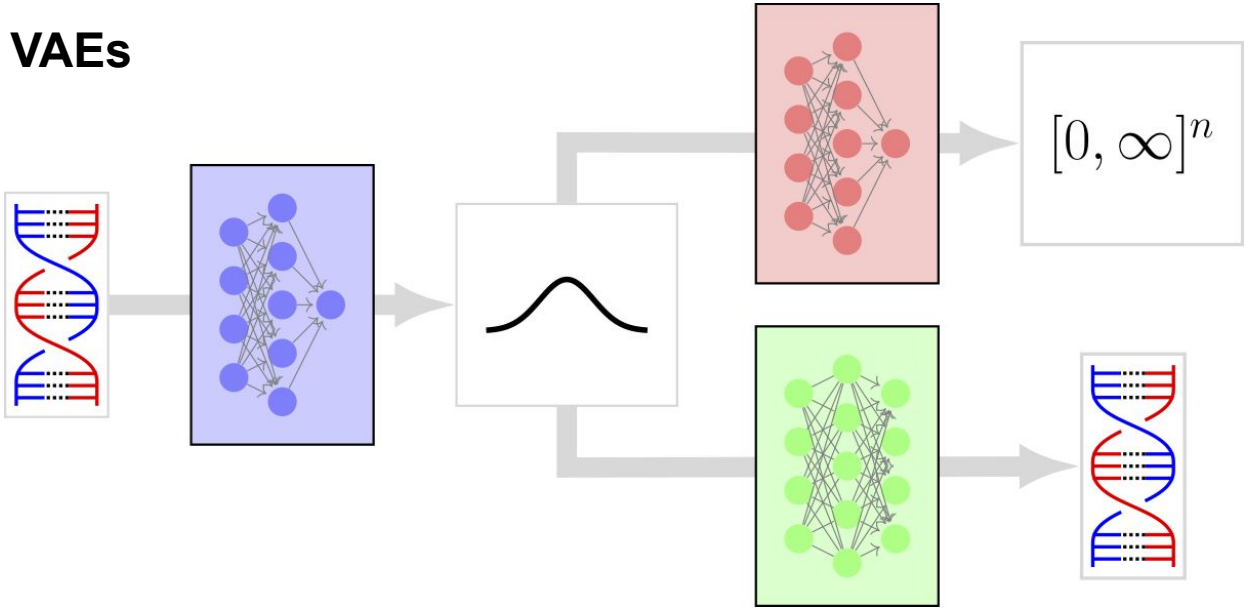
Takeaways

1. Transducer → Tuner → Sampler
2. 1NN accuracy as an evaluation tool
3. Takeaways:
    a. GANs > VAEs
    b. The apparent tradeoff between (1) biological realisticness and (2) maximizing the property of interest
    c. Adversarial examples with continued tuning
    d. Random sampling with a high threshold beats gradient descent in the latentspace

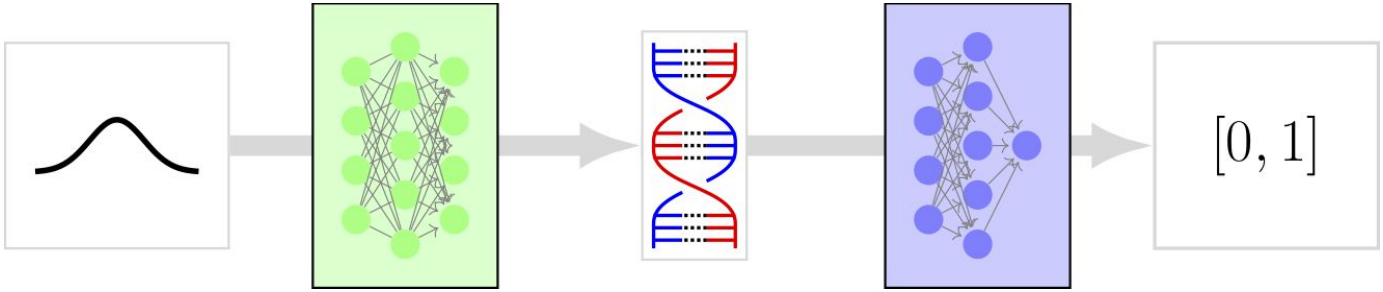
VAEs



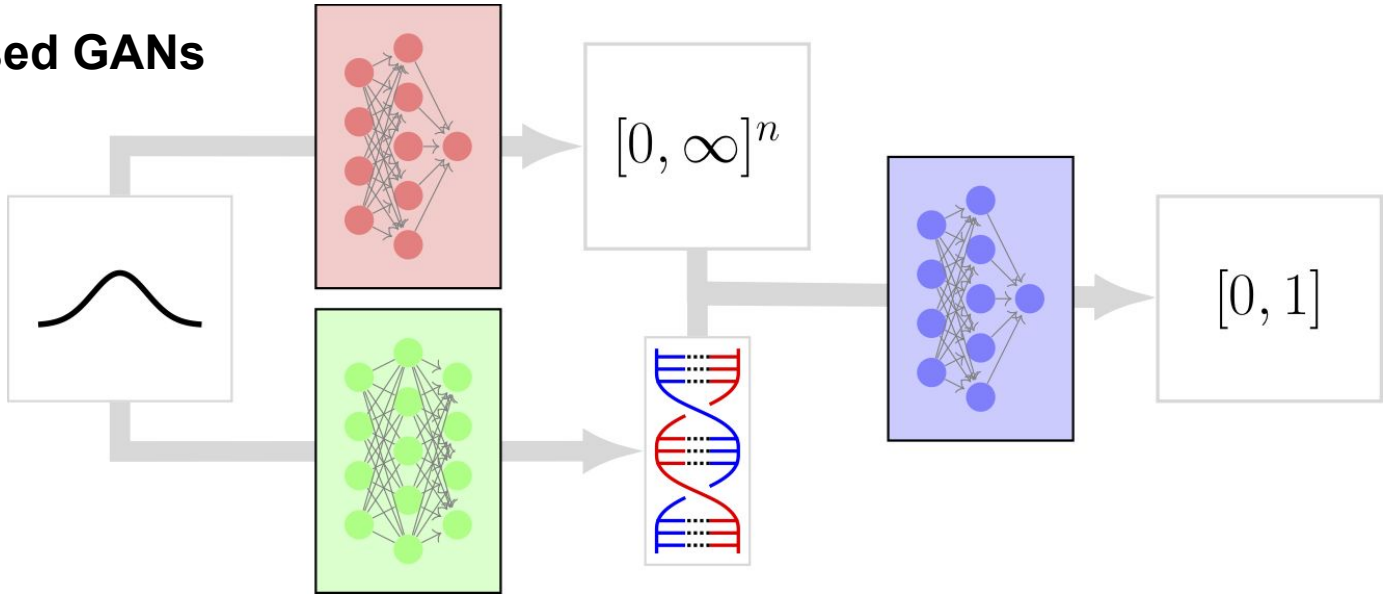
Supervised VAEs

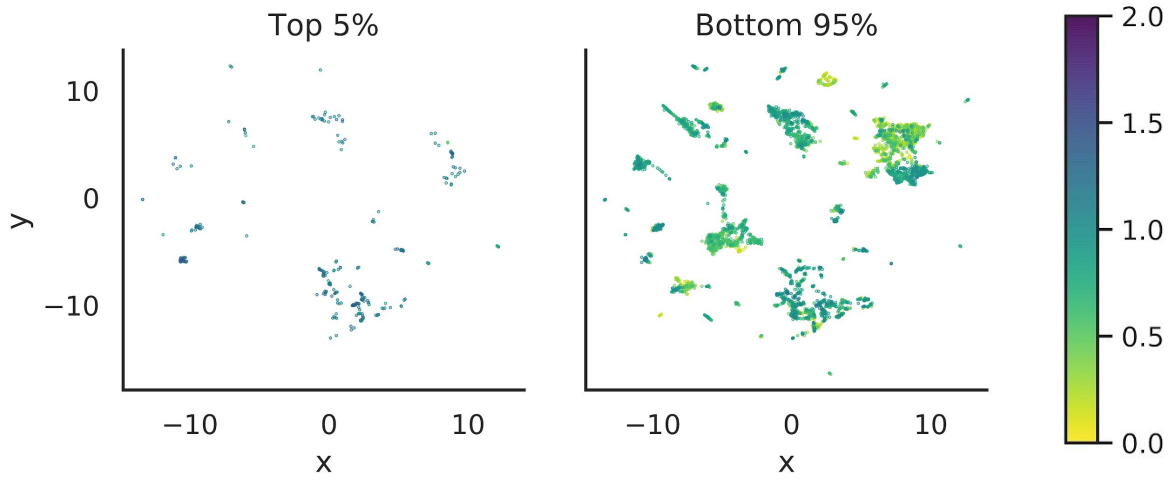


GANs

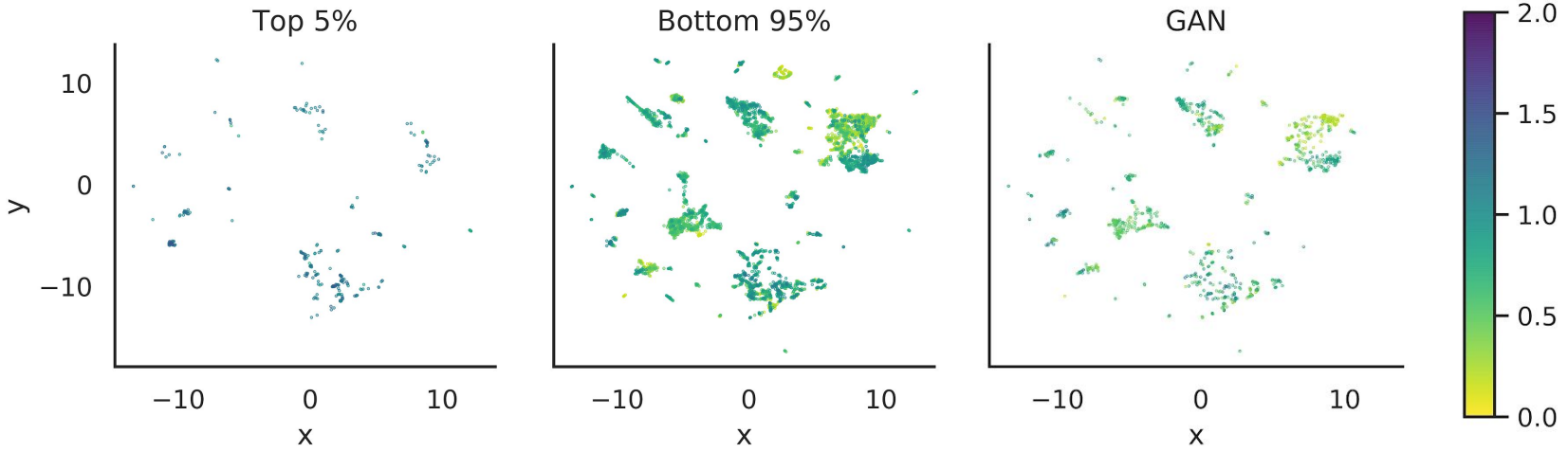


Supervised GANs

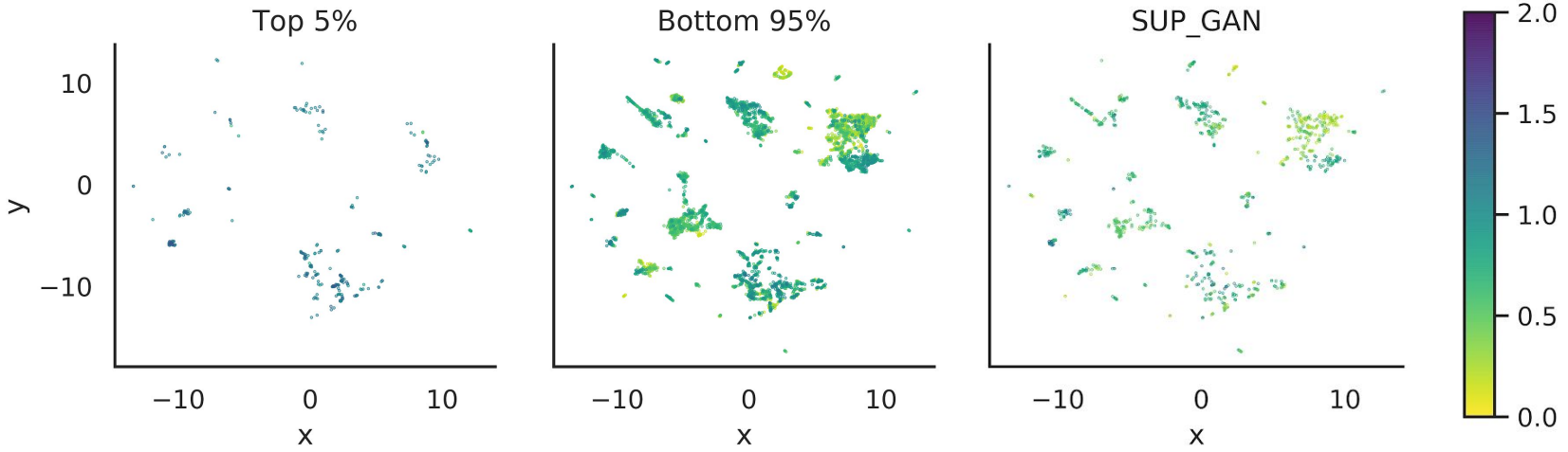




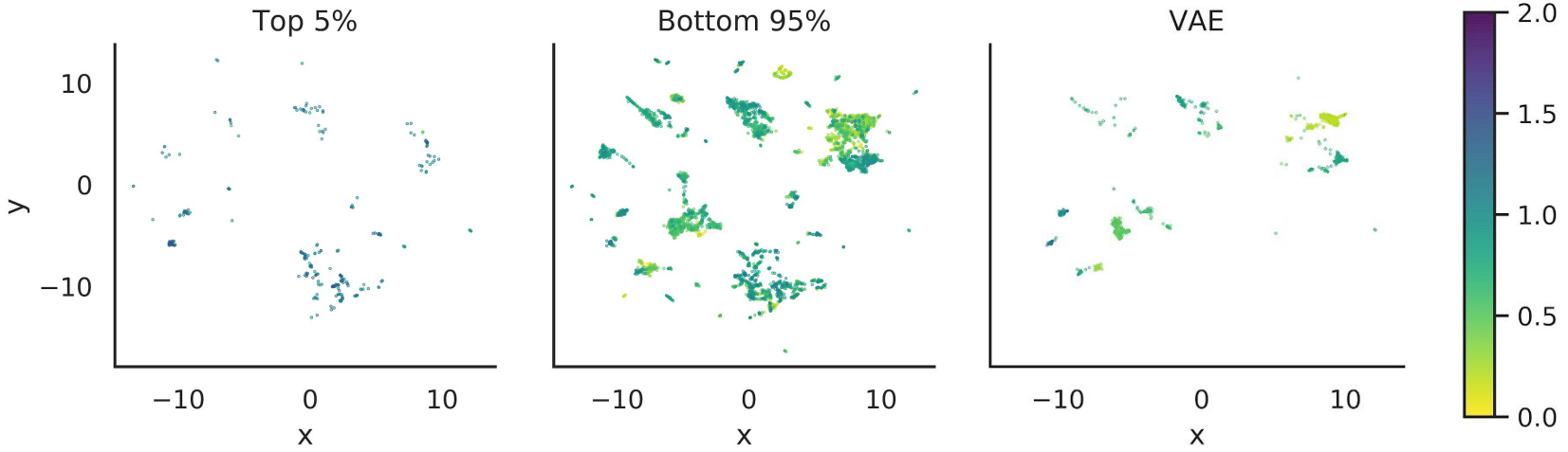
Visualizing GAN/VAE quality. A common 2D projection for the sequences was created as follows: all generated and real sequences were scanned with the first five convolutional layers of the DeepSEA network, and the output of the last convolutional layer was flattened to derive a sequence embedding. This embedding was projected into 2D space using UMAP. The two panels here are the real sequences.



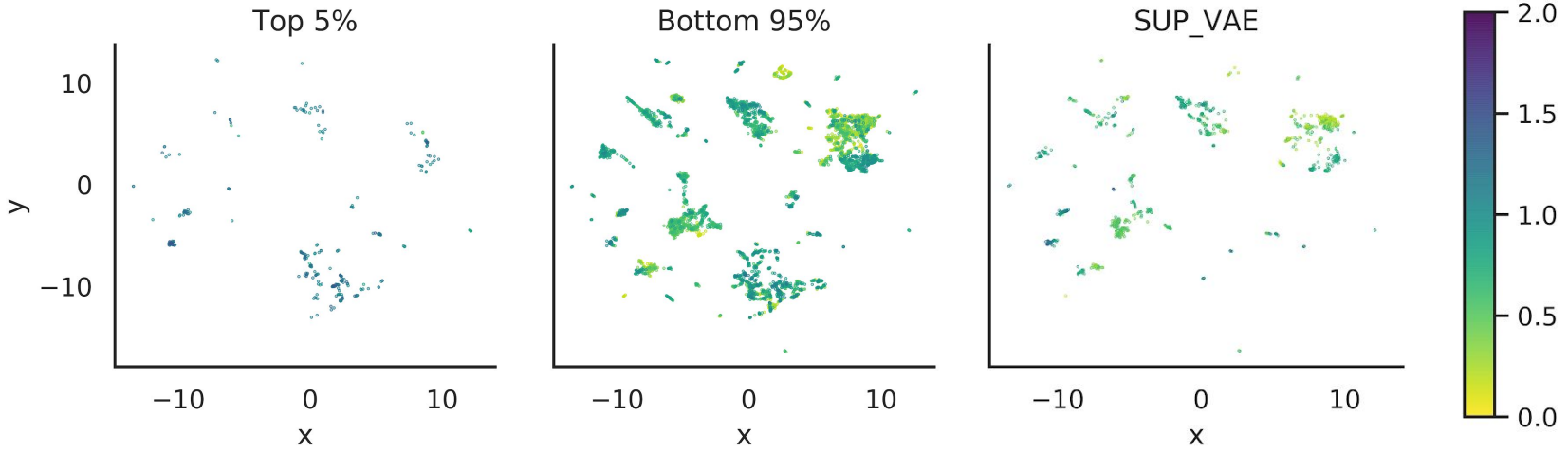
Visualizing GAN/VAE quality. A common 2D projection for the sequences was created as follows: all generated and real sequences were scanned with the first five convolutional layers of the DeepSEA network, and the output of the last convolutional layer was flattened to derive a sequence embedding. This embedding was projected into 2D space using UMAP. The first two panels are the real sequences, the last is generated sequences from the GAN.



Visualizing GAN/VAE quality. A common 2D projection for the sequences was created as follows: all generated and real sequences were scanned with the first five convolutional layers of the DeepSEA network, and the output of the last convolutional layer was flattened to derive a sequence embedding. This embedding was projected into 2D space using UMAP. The first two panels are the real sequences, the last is generated sequences from the supGAN.



Visualizing GAN/VAE quality. A common 2D projection for the sequences was created as follows: all generated and real sequences were scanned with the first five convolutional layers of the DeepSEA network, and the output of the last convolutional layer was flattened to derive a sequence embedding. This embedding was projected into 2D space using UMAP. The first two panels are the real sequences, the last is generated sequences from the VAE.



Visualizing GAN/VAE quality. A common 2D projection for the sequences was created as follows: all generated and real sequences were scanned with the first five convolutional layers of the DeepSEA network, and the output of the last convolutional layer was flattened to derive a sequence embedding. This embedding was projected into 2D space using UMAP. The first two panels are the real sequences, the last is generated sequences from the supVAE.