A brief introduction to deep generative models

Clayton W. Seitz

January 27, 2022

Outline

The logic of generative modeling

Say we have a set of variables $\mathrm{x} = (x_1, x_2, ..., x_n)$ which might have some statistical dependence

The variable x might be an amino acid sequence, gene expression data, microscopy image, etc.

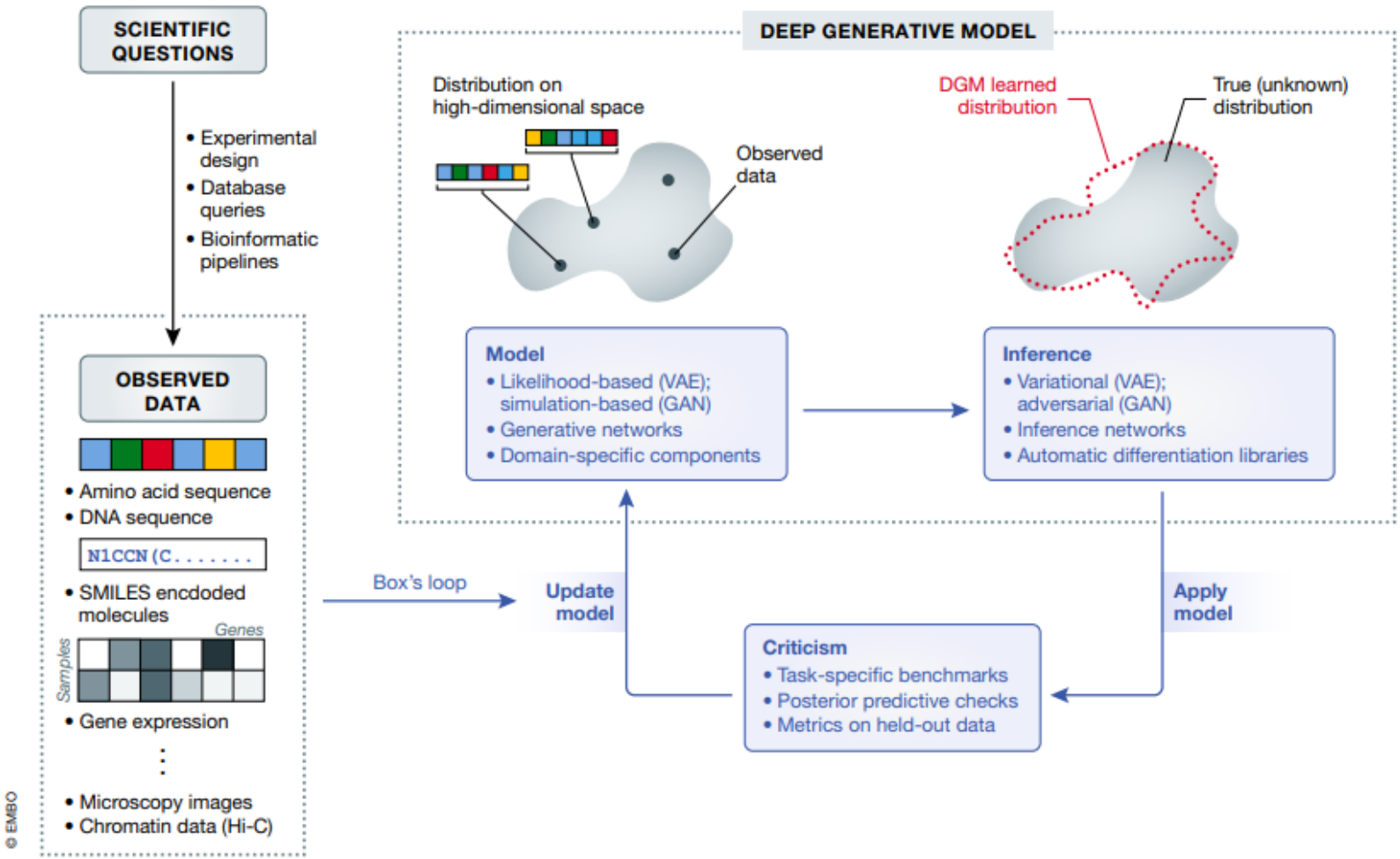
- Often we are handed a batch of empirical samples $\{\mathrm{x}_i\}_{i=1}^{N}$
- We want to know the generating distribution $p(\mathrm{x})$

In supervised generative learning, we try to explicity learn the joint distribution $p(\mathrm{x}) = \prod_{i=1}^{N-1} p(x_i|x_{i+1:N})p(x_N)$, which is generally more difficult than discriminative learning.

Perks of generative modeling

- Fitting complete multivariate distributions $p(\mathrm{x})$ goes beyond correlation-based or clustering approaches
- Correlations cannot discover partial correlation in the context of other neighbors
- Fitting $p(\mathrm{x})$ permits sampling based inference

Applying deep generative models to biological data



Why generative modeling is difficult

When describing a distribution over multiple variables, we may not know the proper normalization Z. That is,

$$p(\mathrm{x}) = \frac{1}{Z}\tilde{p}(\mathrm{x})$$

This very important situation arises in several contexts:

1. In Bayesian inference where $p(x_1|x_2) = p(x_2|x_1)p(x_1)/p(x_2)$ is intractable due to $Z = p(x_2) = \int p(x_2|x_1)p(x_1)dx_1$. This integral can be very difficult or impossible to compute.

2. In models from statistical physics, e.g. the Ising model, we only know $\tilde{p}(\mathrm{x}) = e^{-H(\mathrm{x})}$ where $H(\mathrm{x})$ is the Hamiltonian

Variational autoencoders (VAEs)

A variational solution to generative modeling

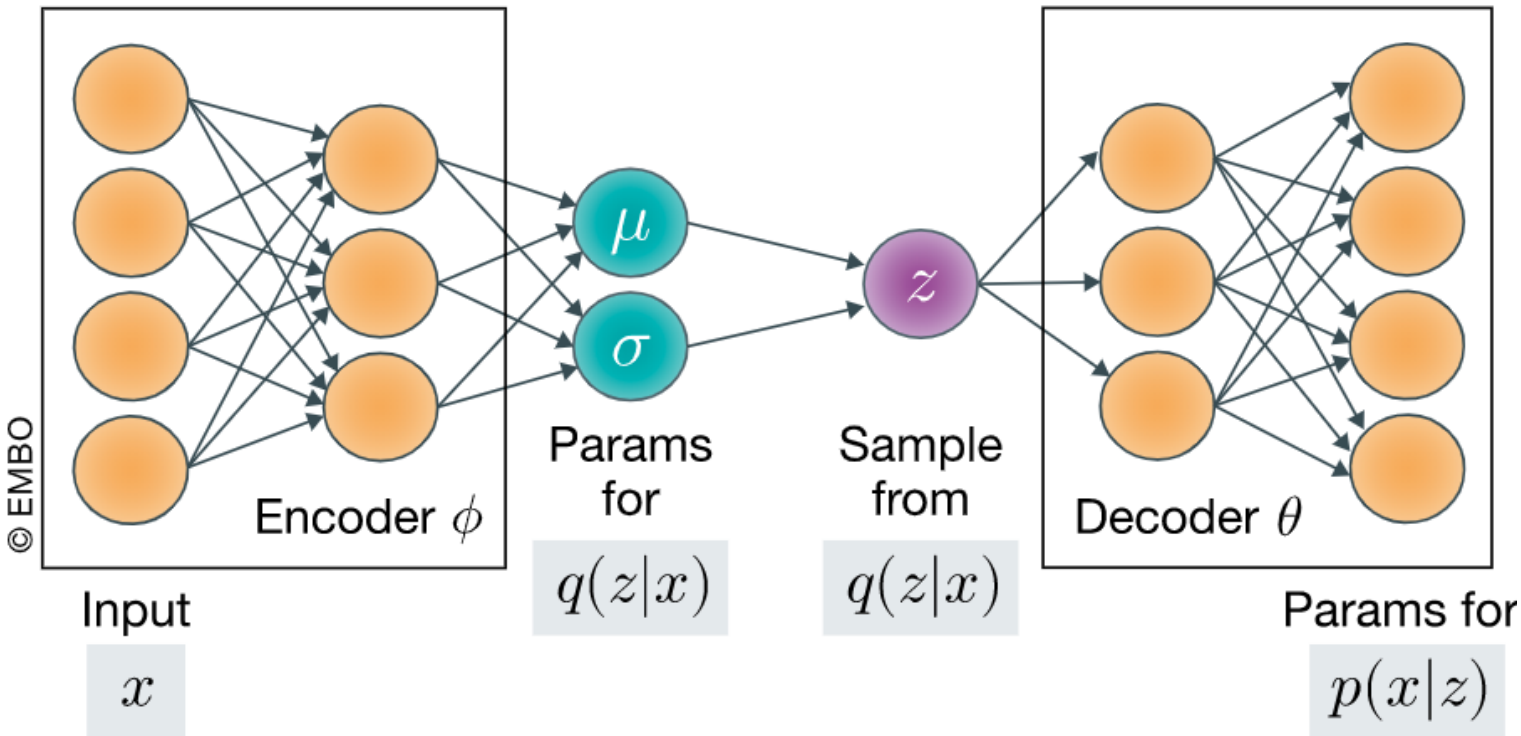


Figure 1: **Variational autoencoder architecture** Taken from Lopez 2020 in EMBO

Bayesian inference

The variable x has a latent representation or code z. We often say that z is the *causal source* of x. Ultimately, we would like to know the distribution $P_\Phi(\mathrm{x})$

$$P_\phi(\mathrm{x}) = \frac{P_\phi(\mathrm{x}|\mathrm{z})P_\phi(\mathrm{z})}{Q_\psi(\mathrm{z}|\mathrm{x})}$$

in order to find the model parameters that maximize the likelihood of the observed data:

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} - \log P_\Phi(\mathrm{x})$$

but we generally do not know $P_\psi(\mathrm{z}|\mathrm{x})$ due to the intractable integral $Z = \int P_\phi(\mathrm{x}|\mathrm{z})P_\phi(\mathrm{z})d\mathrm{z}$ (see slide 5)

Computing the evidence

We can rewrite the evidence as

$$
\begin{aligned}
P_\phi(\mathrm{x}) &= \int P_\phi(\mathrm{z}) P_\phi(\mathrm{x}|\mathrm{z}) d\mathrm{z} \\
&= \int P_\phi(\mathrm{z}) P_\phi(\mathrm{x}|\mathrm{z}) \frac{P_\phi(\mathrm{z}|\mathrm{x})}{P_\phi(\mathrm{z}|\mathrm{x})} d\mathrm{z} \\
&= \mathbb{E}_{\mathrm{z} \sim P_\phi(\mathrm{z}|\mathrm{x})} \frac{P_\phi(\mathrm{z}) P_\phi(\mathrm{x}|\mathrm{z})}{P_\phi(\mathrm{z}|\mathrm{x})}
\end{aligned}
$$

where $P_\phi(\mathrm{z}|\mathrm{x})$ is our model "encoder"

The evidence lower bound (ELBO)

$$
\begin{aligned}
\log P_\phi(\mathrm{x}) &= \log \int_z P(x,z)dx \\
&= \log \int_z P(x,z)\frac{Q(z|x)}{Q(z|x)}dz \\
&= \log \mathbb{E}_{\mathrm{z}\sim P_\phi(\mathrm{z}|\mathrm{x})}\frac{P(x|z)P(z)}{Q(z|x)} \\
&\geq \mathbb{E}_{\mathrm{z}\sim P_\phi(\mathrm{z}|\mathrm{x})} \log \frac{Q(x|z)}{P(z)} + \log P(x|z)
\end{aligned}
$$

$$
-\log P_\phi(\mathrm{x}) \leq \mathbb{E}_{\mathrm{z}\sim P_\phi(\mathrm{z}|\mathrm{x})} \log \frac{Q(x|z)}{P(z)} - \log P(x|z)
$$

Applying the VAE to transcriptomics

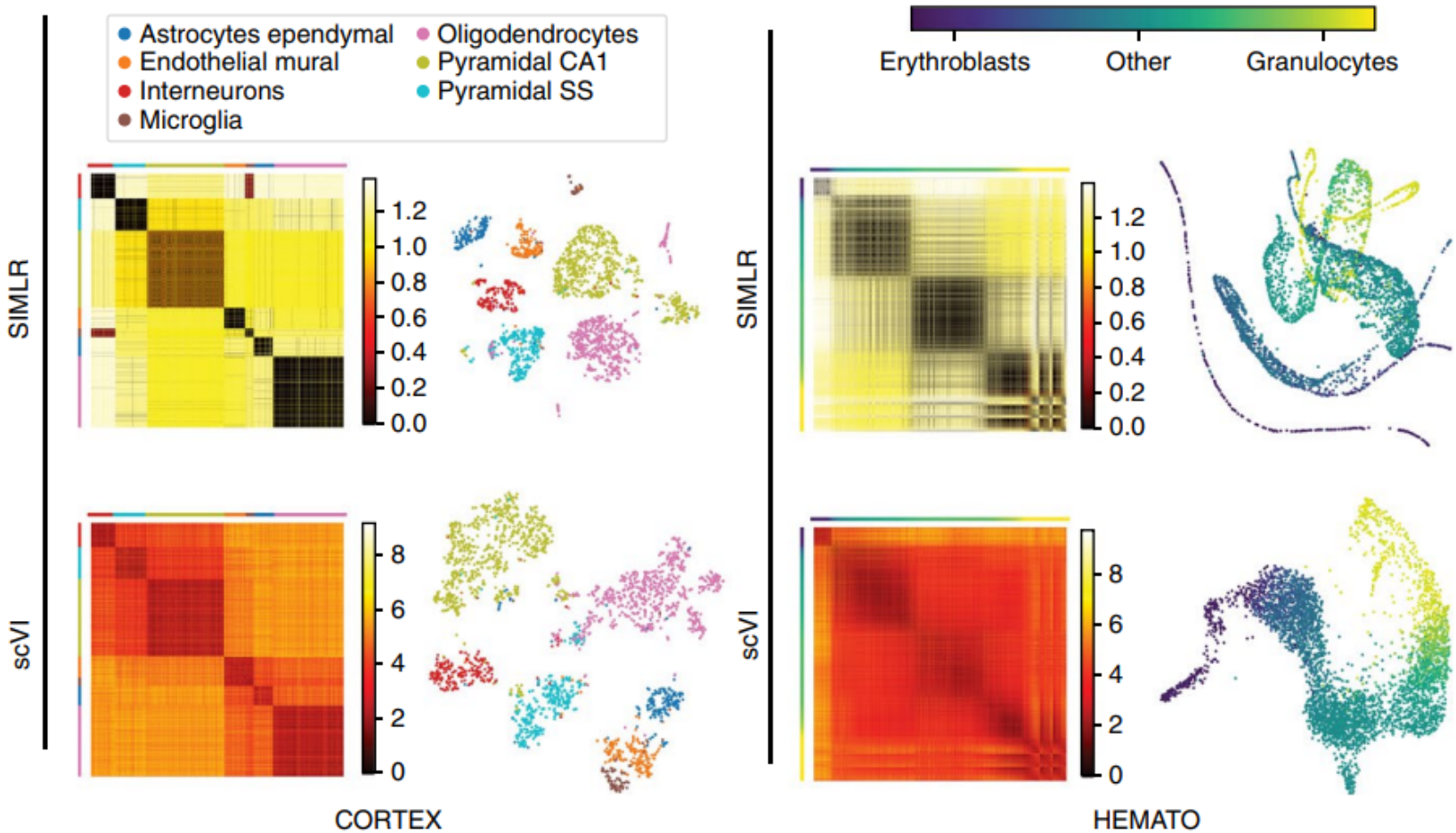
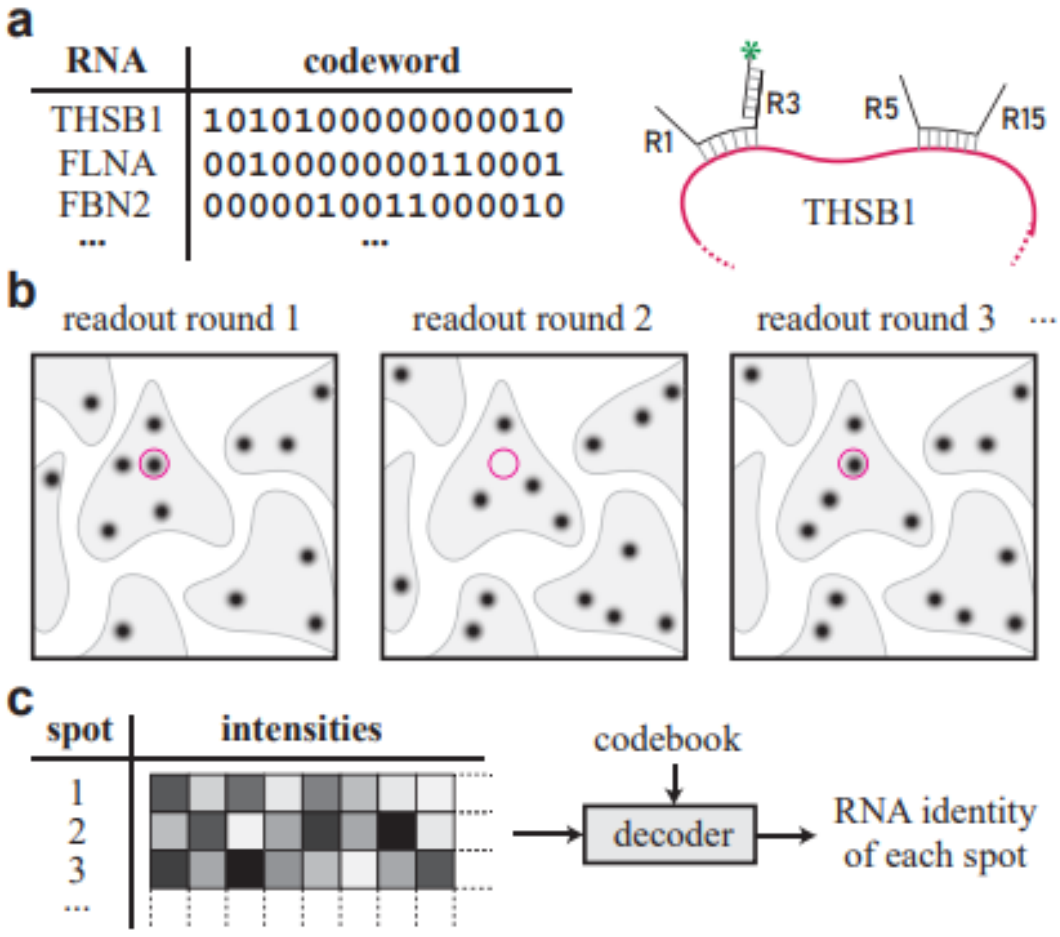


Figure 2: Distance matrices and 2D embedding of the latent space for CORTEX dataset ($n = 3005$ cells) and HEMATO ($n = 4016$ hematopoietic progenitor cells)

Previous studies lack interpretability and spatial information

- VAEs (and deep networks in general) are very flexible or "expressive"
- The latent distribution (which is typically multivariate normal) can be difficult to interpret
- But can still perform clustering, visualization, etc.
- Field is moving towards using spatial features as well

Multiplexed RNA imaging with MHD4 code



Open-source datasets from the Zhuang lab at Harvard

Multiplexed RNA imaging with MHD4 code

- Human genome codes nearly 30k non-redundant types of RNA molecules
- Some of those end up translated into protein
- Original study used 16-bit codes with minimum Hamming distance 4
- Code distance is chosen due to false positive and false negative rates

Using $L = 16$ bit codewords with distance 4, gives 140 unique RNAs. Other methods SlideSeq, 10x Visium can achieve whole-genome spatial transcriptomics

References I