c∗GMΔVÆs—q

Kolb, Yiftach

March 7, 2023

c∗GMΔVÆs—q

a Master Thesis in Bioinformatics

Advisor / Reviewer: Professor Martin Vingron
Reviewer: Professor Tim Conrad

Glossary

c∗GM∆VÆ (maybe Conditional) Gaussian Mixture with Dirichlet prior Variational Auto Encoder

c∗GM∆VÆs—q c∗GM∆VÆ (for) s(ingle–cell-RNA–Se)q.

Autoencoders

A "vanilla" autoencoder is a neural networks that "learns" the identity (subject to dimensional restriction).

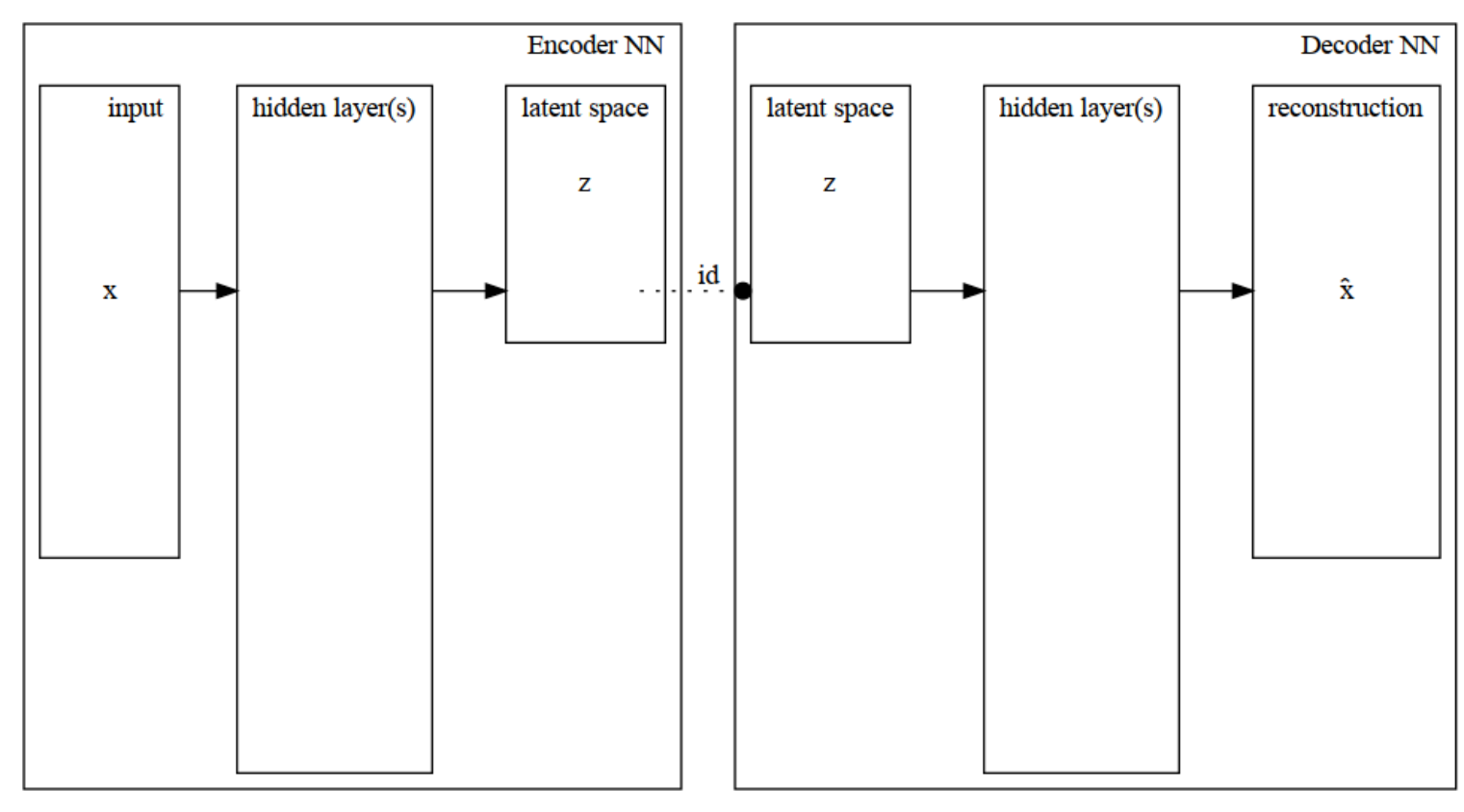


Figure: Autoencoder

Autoencoders and PCA

(On centered data[8])

PCA

$$\tilde{\boldsymbol{V}} = \text{argmin}_{\mathbf{W}}\{\|\mathbf{X} - \mathbf{X}\boldsymbol{W}\boldsymbol{W}^T\|_F^2 \quad : \quad \boldsymbol{W} \in \mathbb{R}^{n\times l}, \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_l\} \tag{1}$$

Linear AE

$$\text{argmin}_{\boldsymbol{E},\boldsymbol{D}}\{\|\mathbf{X} - \mathbf{X}\boldsymbol{E}\boldsymbol{D}\|_F^2 \quad : \quad \boldsymbol{E}, \boldsymbol{D}^{\boldsymbol{T}} \in \mathbb{R}^{n\times l}, \} \tag{2}$$

$$\tilde{\boldsymbol{W}} \in \text{argmin}_{\mathbf{W}}\{\|\mathbf{X} - \mathbf{X}\boldsymbol{W}\boldsymbol{W}^\dagger\|_F^2 \quad : \quad \boldsymbol{W} \in \mathbb{R}^{n\times l}, \} \tag{3}$$

$$\text{span}\{\tilde{\boldsymbol{W}}\} = \text{span}\{\tilde{\boldsymbol{V}}\}$$

VAEs

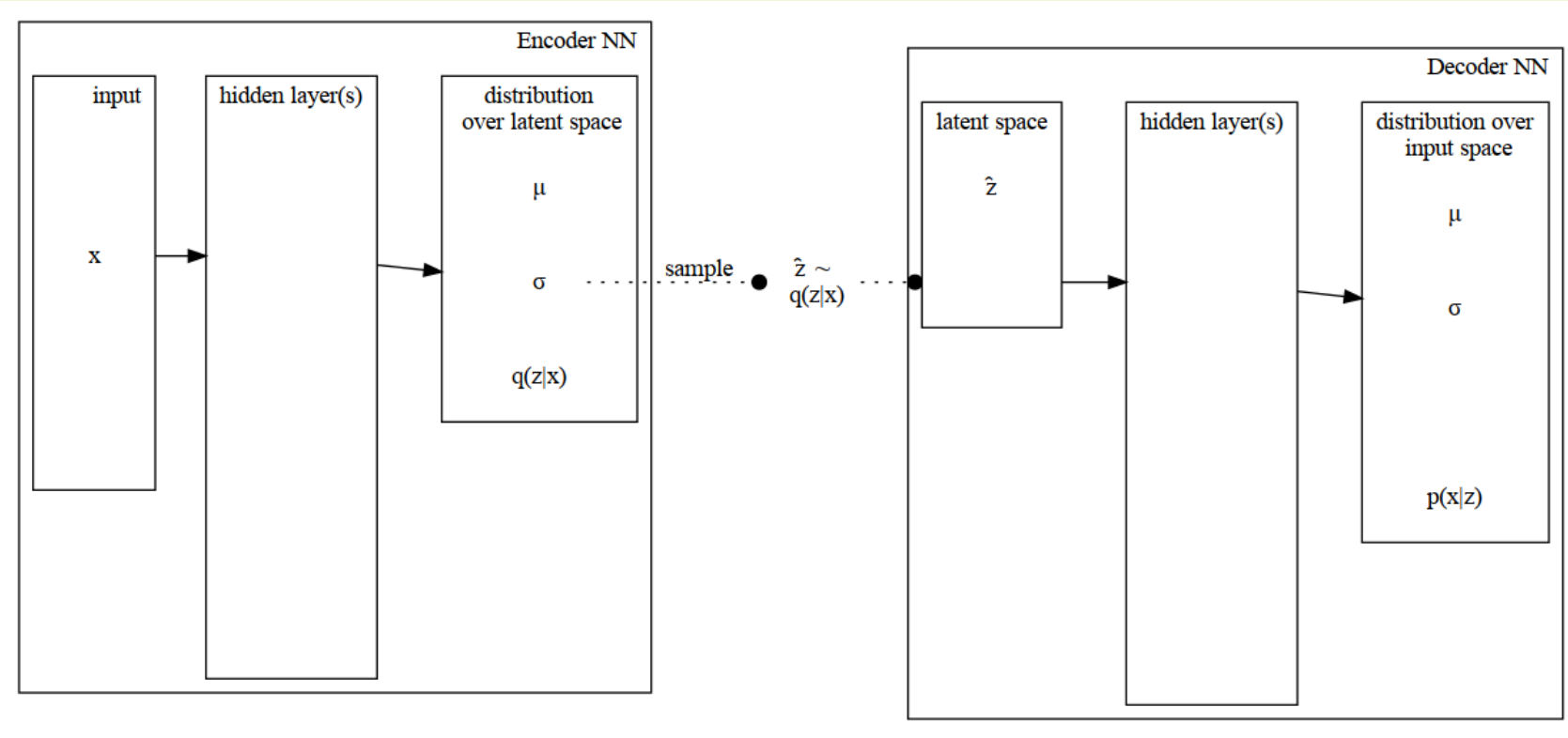


Figure: VAE

VAE can represent the true distribution of the data

VAE is a prbabilistic model. Theoretically, by choosing the correct type of distribution model, it can "learn" to represent the true distribution of the data.

VAE: encoding

Instead of deterministic mapping ...define distribution over the laten space ($\mathbf{z}$) by mapping $\mathbf{x}$ into the distribution parameters e.g. $\mu(\mathbf{x}), \Sigma(\mathbf{x})$ when we use Gaussian $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu, \Sigma)$.

VAE: decoding

sample from the latent space $\mathbf{z} \sim \mathcal{N}(\cdot|\mu, \Sigma)$

map $\mathbf{z}$ to a distribution on the input space $p(\mathbf{x}|\mathbf{z})$

VAE: loss function

The *evidence lower bound (ELBO)* with respect to p, q is:

$$-\mathcal{L}(q, p, \mathbf{x}) \triangleq \int \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z})} dq(\boldsymbol{z}) \tag{4}$$

$$-\mathcal{L}(q, p) \triangleq -\mathcal{L}(q, p, \mathbf{X}) = \frac{1}{N} \sum_{1}^{N} (-\mathcal{L}(q, p, \mathbf{x}_i) \tag{5}$$

$$\approx \mathbf{E}_{\mathbf{x}}[-\mathcal{L}(q, p, \mathbf{x})] \tag{6}$$

(we want to minimize the minus ELBO function)

VAE: log evidence

It can be shown that maximizing the ELBO is equivalent to maximiminizing the "log evidence" $\log p(\mathbf{x})$

$$\log p(\boldsymbol{x}) = -\mathcal{L}(q, p, \mathbf{x}) - KL(q(\boldsymbol{z}) \| p(\boldsymbol{z}|\boldsymbol{x})) \tag{7}$$

Monte Carlo integration

As you can see we the loss function requires us to compute an integral, which usually cannot be done analytically.

Instead the integral is approximated by Monte Carlo integration.

$$\text{sample } \mathbf{z}_i \sim q(\mathbf{z}|\mathbf{x}) \text{ then :}$$

$$\begin{aligned}
\mathcal{L}(p, q, \mathbf{x}) &= \int -\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} dq(\mathbf{z}|\mathbf{x}) \\
&= \int -\log p(\mathbf{x}|\mathbf{z}) dq(\mathbf{z}|\mathbf{x}) + \int \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} dq(\mathbf{z}|\mathbf{x}) \qquad (8) \\
&\approx \frac{1}{k} \sum_{i=1}^{k} [-\log p(\mathbf{x}|\mathbf{z}_i) + \log \frac{q(\mathbf{z}_i|\mathbf{x})}{p(\mathbf{z}_i)}]
\end{aligned}$$

VAE: compounding the latent distribution

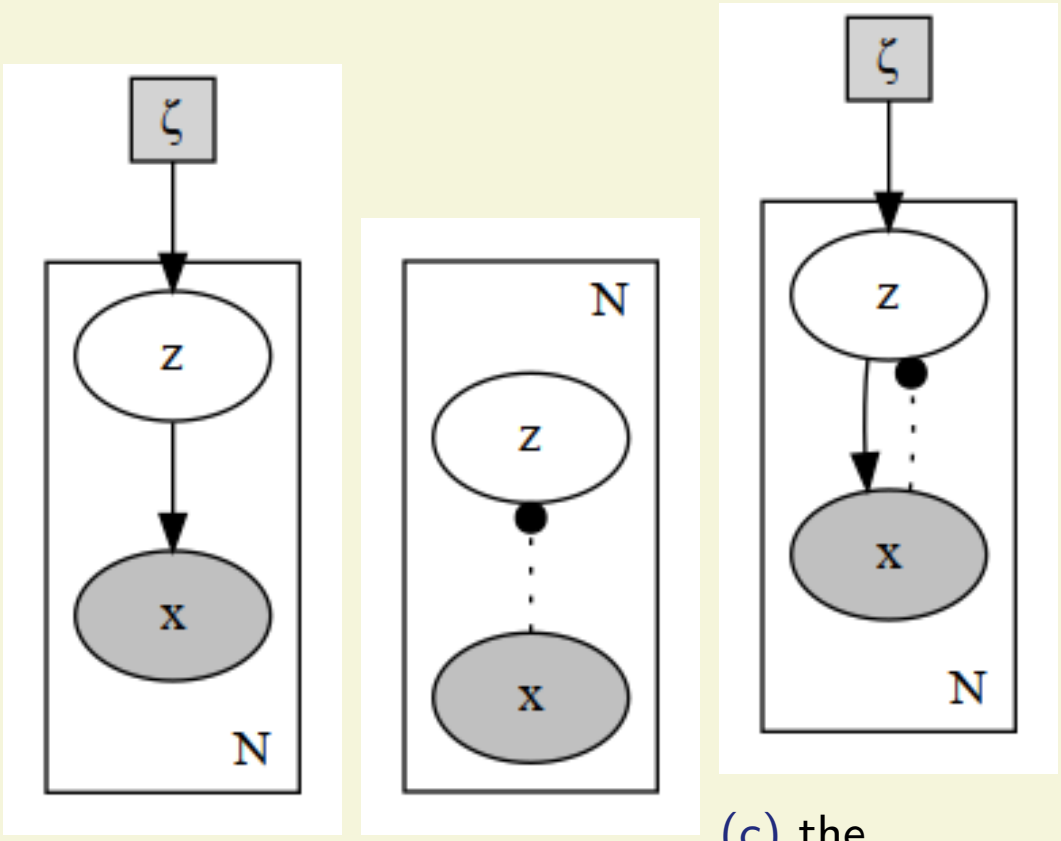
More complicated distributions such as mixture distribution can be modelled by "unpacking" the latent $\mathbf{z}$ and the observed $\mathbf{x}$

1. Define the set of observed random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k$, and the set of latent random vectors and stochastic parameters $\mathbf{z}_1, \ldots \mathbf{z}_l$.
2. Specify how to factor the generative model $p(\mathbf{x}_1, \ldots, \mathbf{x}_k | \mathbf{z}_1 \ldots, \mathbf{z}_l)$
3. Specify how to factor the inference model $q(\mathbf{z}_1 \ldots \mathbf{z}_l | \mathbf{x}1, \ldots \mathbf{x}_k)$
4. Choose appropriate priors $p(\mathbf{z}_i)$ and
5. Choose appropriate distribution families for the $\mathbf{x}_i$ and $\mathbf{z}_i$, and choose priors $p(\mathbf{z}_i)$.

VAE: Graphical representation

Evry distribution can be represented by a DAG. Nodes represent random variables (and also priors), and directed arrows represent conditional dependency.

VAE: base case



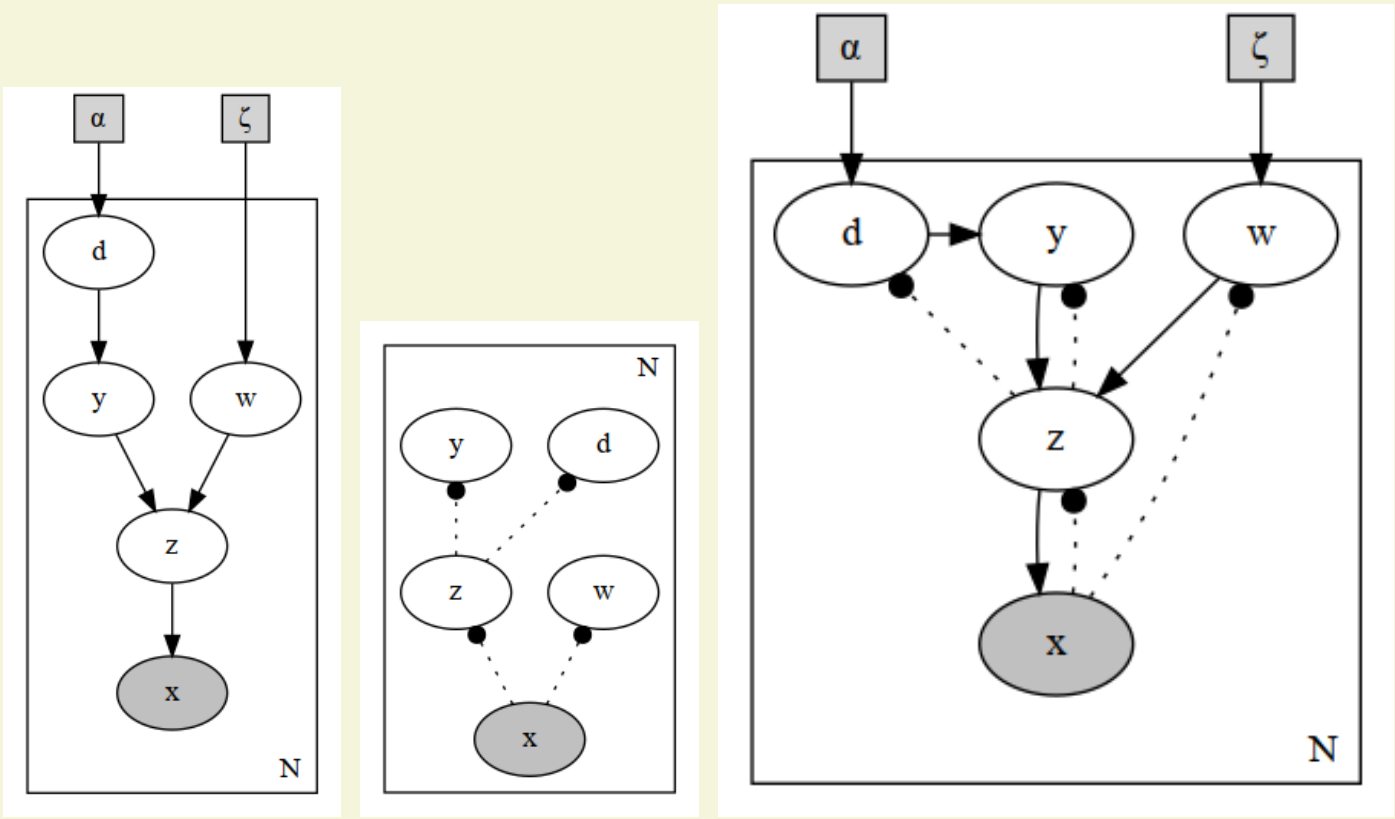
(a) generative model $p(x, z)$

(b) inference model $q(z|x)$

(c) the combined graphical model

Figure: VAE graphical model

VAE: patholocigacl case



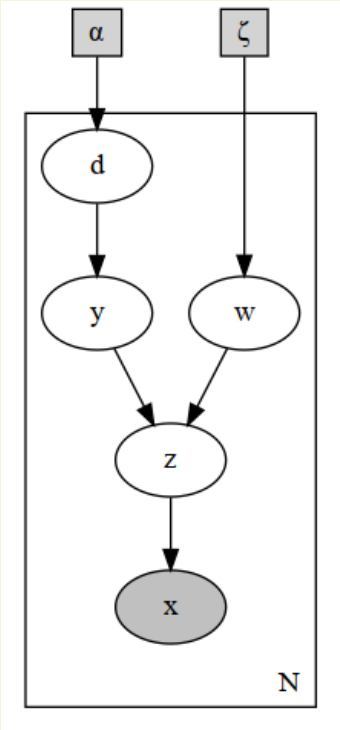
(a) generative model $p(x, z)$

(b) inference model $q(z|x)$

(c) the combined graphical model

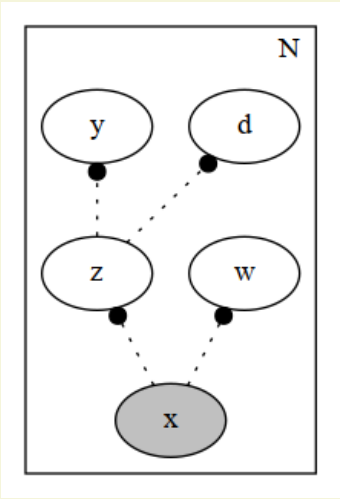
Figure: c∗GMΔVÆ graphical model

c∗GMΔVÆ generative model



$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}, \mathbf{d}) &= p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{w}, \mathbf{y})p(\mathbf{y}|\mathbf{d})p(\mathbf{d})p(\mathbf{w}) \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{1}) \\
p(\mathbf{d}) &= \mathrm{Dir}(\mathbf{d}|\alpha) \\
p(\mathbf{y}|\mathbf{d}) &= \mathrm{Cat}(\mathbf{y}|\mathbf{d}) \\
p(\mathbf{z}|\mathbf{w}, \mathbf{y}) &= \mathcal{N}(\mathbf{z}|\mu(\mathbf{w})_{\mathbf{y}}, \sigma(\mathbf{w})_{\mathbf{y}})) \\
p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mu(\mathbf{z}), \sigma(\mathbf{z}))
\end{aligned}
\tag{9}
$$

c∗GMΔVÆ inference model



$$
\begin{aligned}
q(\mathbf{y}, \mathbf{z}, \mathbf{w}, \mathbf{d}|\mathbf{x}) &= q(\mathbf{z}|\mathbf{x})q(\mathbf{w}|\mathbf{x})q(\mathbf{y}|\mathbf{z})q(\mathbf{d}|\mathbf{z}) \\
q(\mathbf{z}|\mathbf{x}) &= \mathcal{N}(\mathbf{z}|\mu_z(\mathbf{x}), \sigma_z(\mathbf{x})) \\
q(\mathbf{w}|\mathbf{x}) &= \mathcal{N}(\mathbf{w}|\mu_w(\mathbf{x}), \sigma_w(\mathbf{x})) \quad (10) \\
q(\mathbf{y}|\mathbf{z}) &= \mathrm{Cat}(\mathbf{y}|f(\mathbf{z})) \\
q(\mathbf{d}|\mathbf{z}) &= \mathrm{Dir}(\mathbf{d}|g(\mathbf{z}))
\end{aligned}
$$

c∗GMΔVÆ loss function

The loss function remains the -ELBO and we can break it into different terms:

$$\mathcal{L}(p, q, \mathbf{x}) = \int -\log \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}, \mathbf{d})}{q(\mathbf{z}, \mathbf{y}, \mathbf{w}, \mathbf{d}|\mathbf{x})} dq(\mathbf{z}, \mathbf{y}, \mathbf{w}, \mathbf{d}|\mathbf{x}) \tag{11}$$

$$= \int -\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{w}, \mathbf{y})p(\mathbf{y}|\mathbf{d})p(\mathbf{w})p(\mathbf{d})}{q(\mathbf{z}|\mathbf{x})q(\mathbf{w}|\mathbf{x})q(\mathbf{y}|\mathbf{z})q(\mathbf{d}|\mathbf{z})} dq \tag{12}$$

$$= \int -\log p(\mathbf{x}|\mathbf{z}) dq \tag{13}$$

$$+ \int \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{w}, \mathbf{y})} dq \tag{14}$$

$$+ \int \log \frac{q(\mathbf{w}|\mathbf{x})}{p(\mathbf{w})} dq \tag{15}$$

$$+ \int \log \frac{q(\mathbf{y}|\mathbf{z})}{p(\mathbf{y}|\mathbf{d})} dq \tag{16}$$

$$+ \int \log \frac{q(\mathbf{d}|\mathbf{z})}{p(\mathbf{d})} dq \tag{17}$$

c∗GMΔVÆ : supervised case

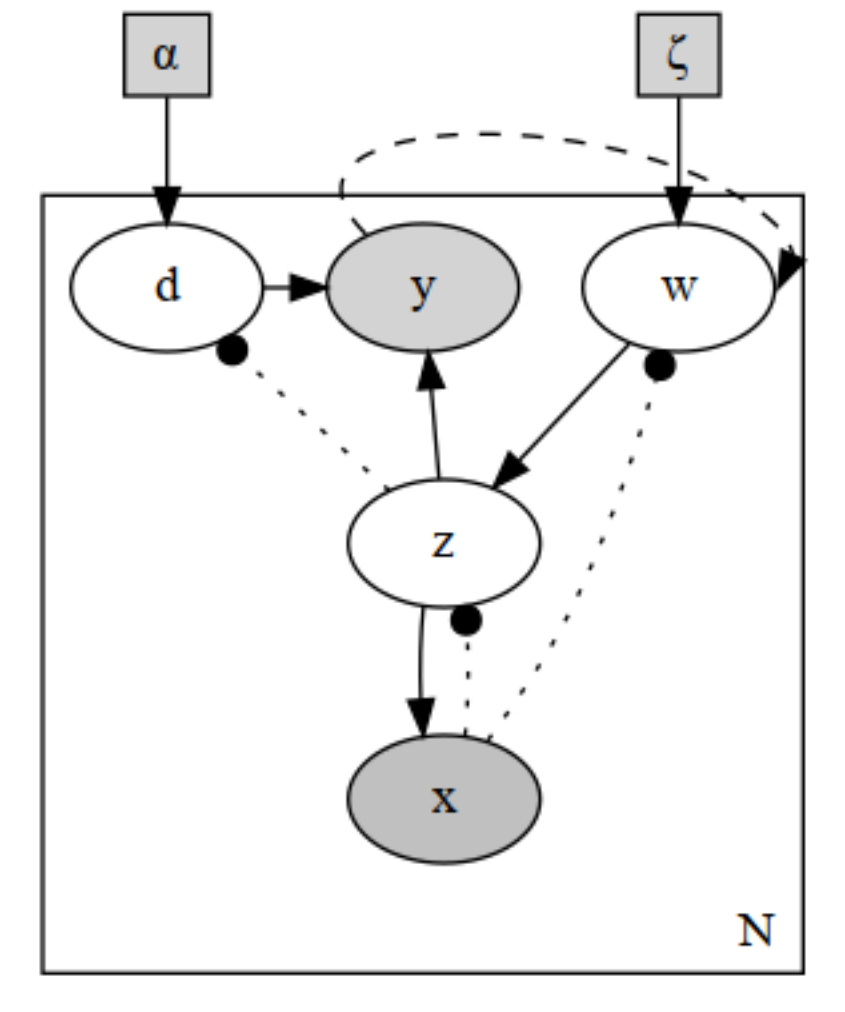
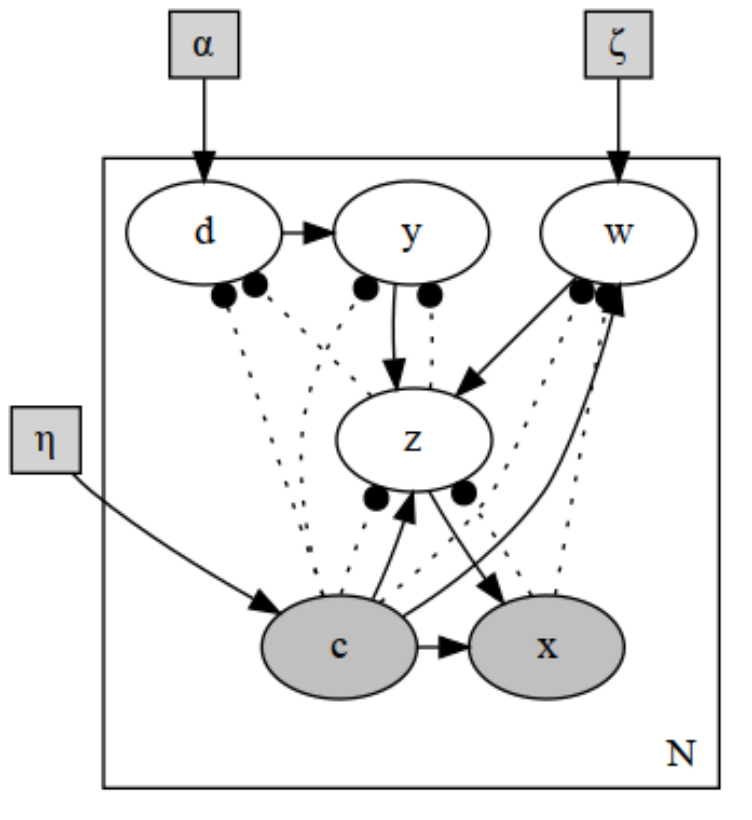
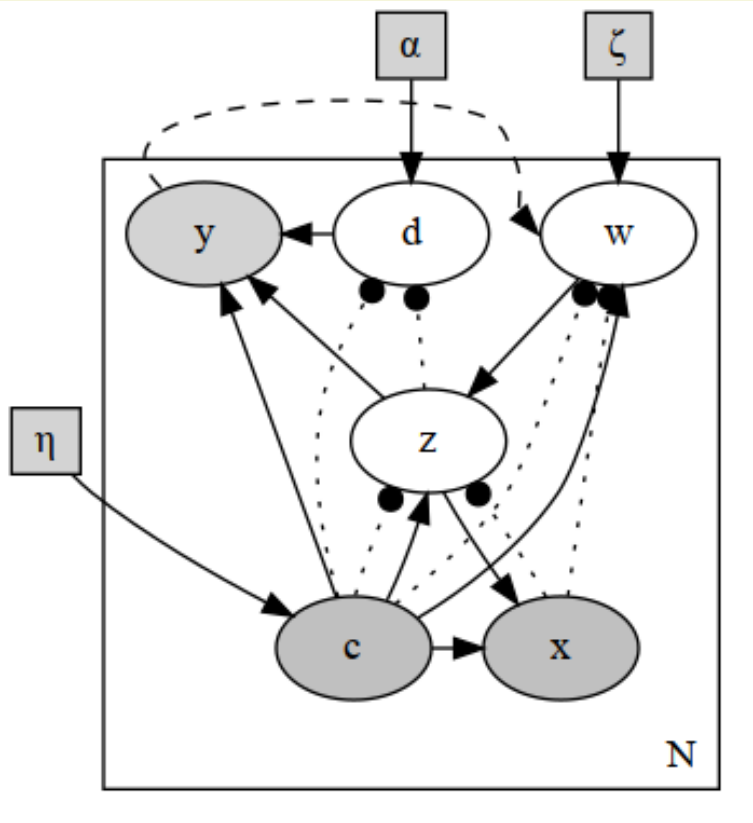


Figure: c∗GMΔVÆ : the supervised case, where **y** is an observed variable.

the conditional flavor of c∗GMΔVÆ

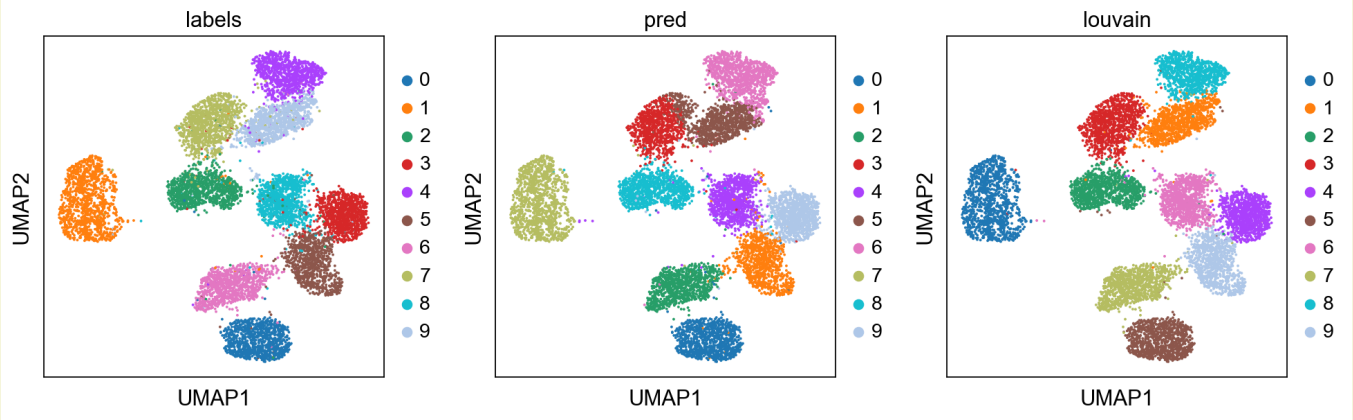


(a) c∗GMΔVÆ (cond.), unsupervised case. Sorry for the arrow clutter

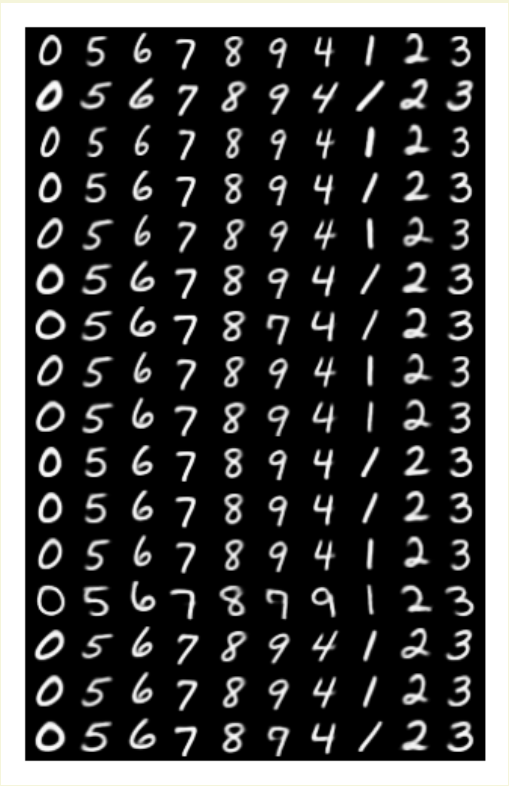


(b) c∗GMΔVÆ (cond.), supervised case

Tests on MNIST



(a) UMAP of the latent space



Testing on synthetic conditional-categorical "blobs" dataset

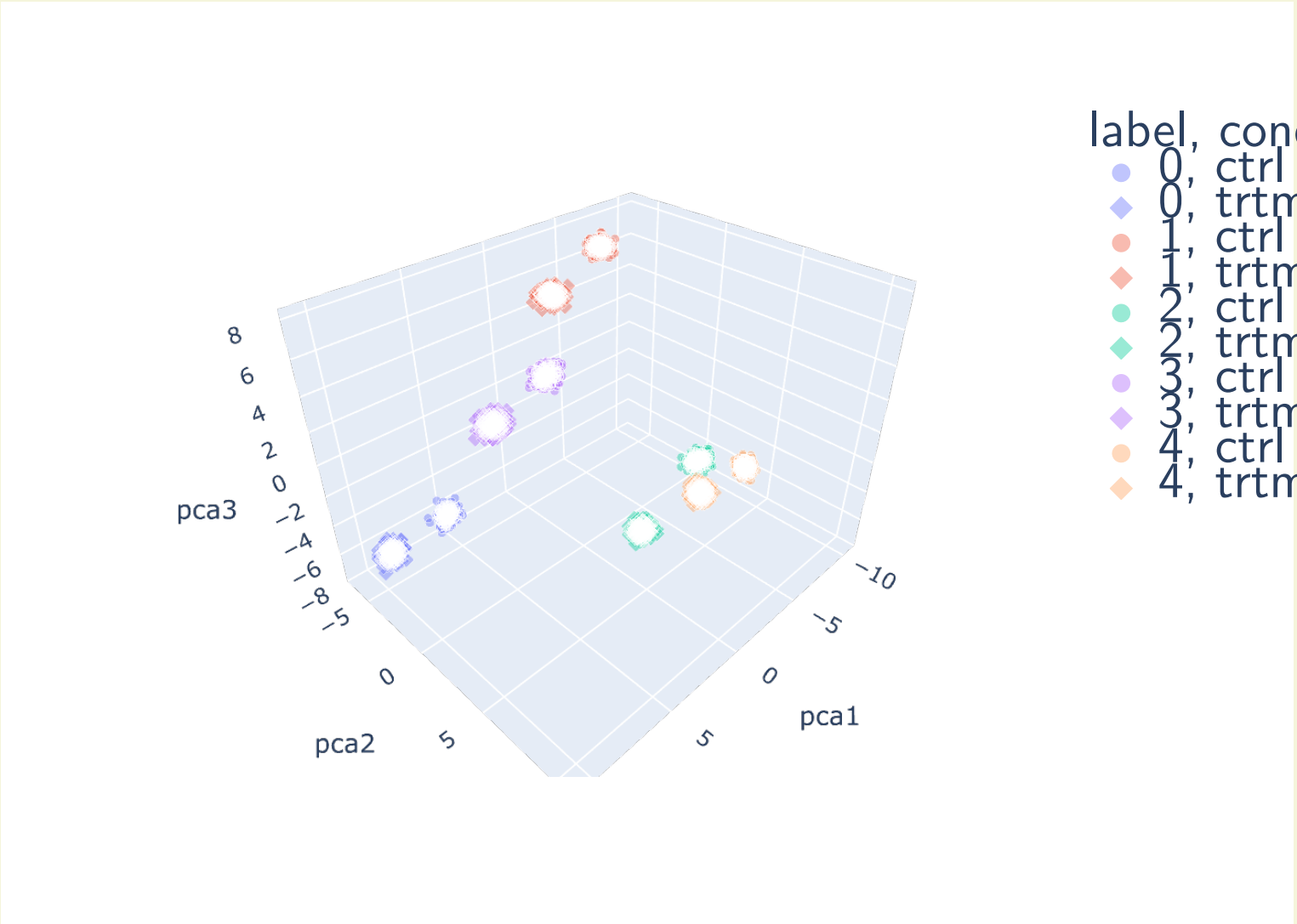


Figure: The toy dataset, showing a 3d plot of its 3 major PCA components.

w embeding

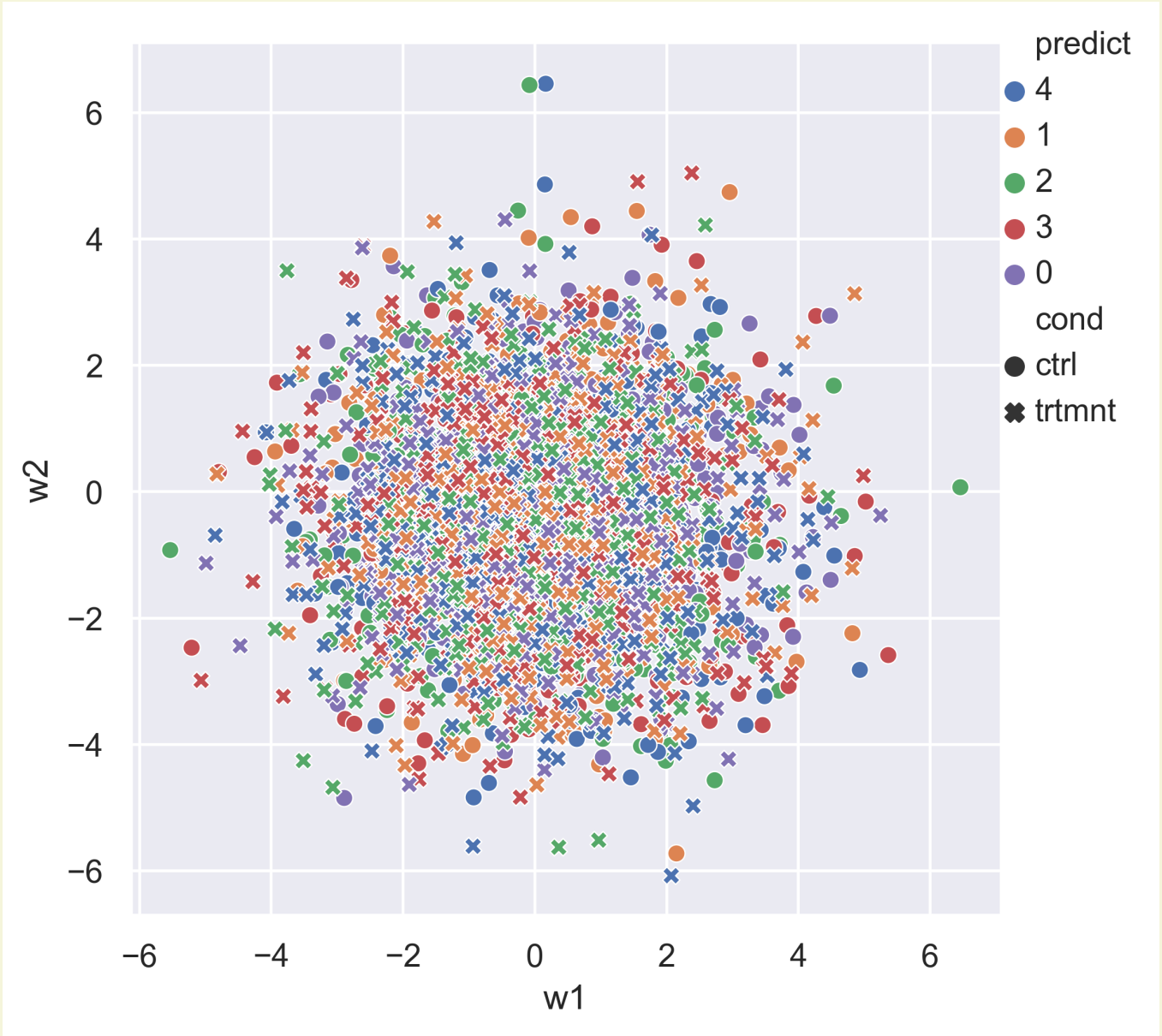


Figure: Random sampling from the encoded distribution of the **w** space.

z embeding

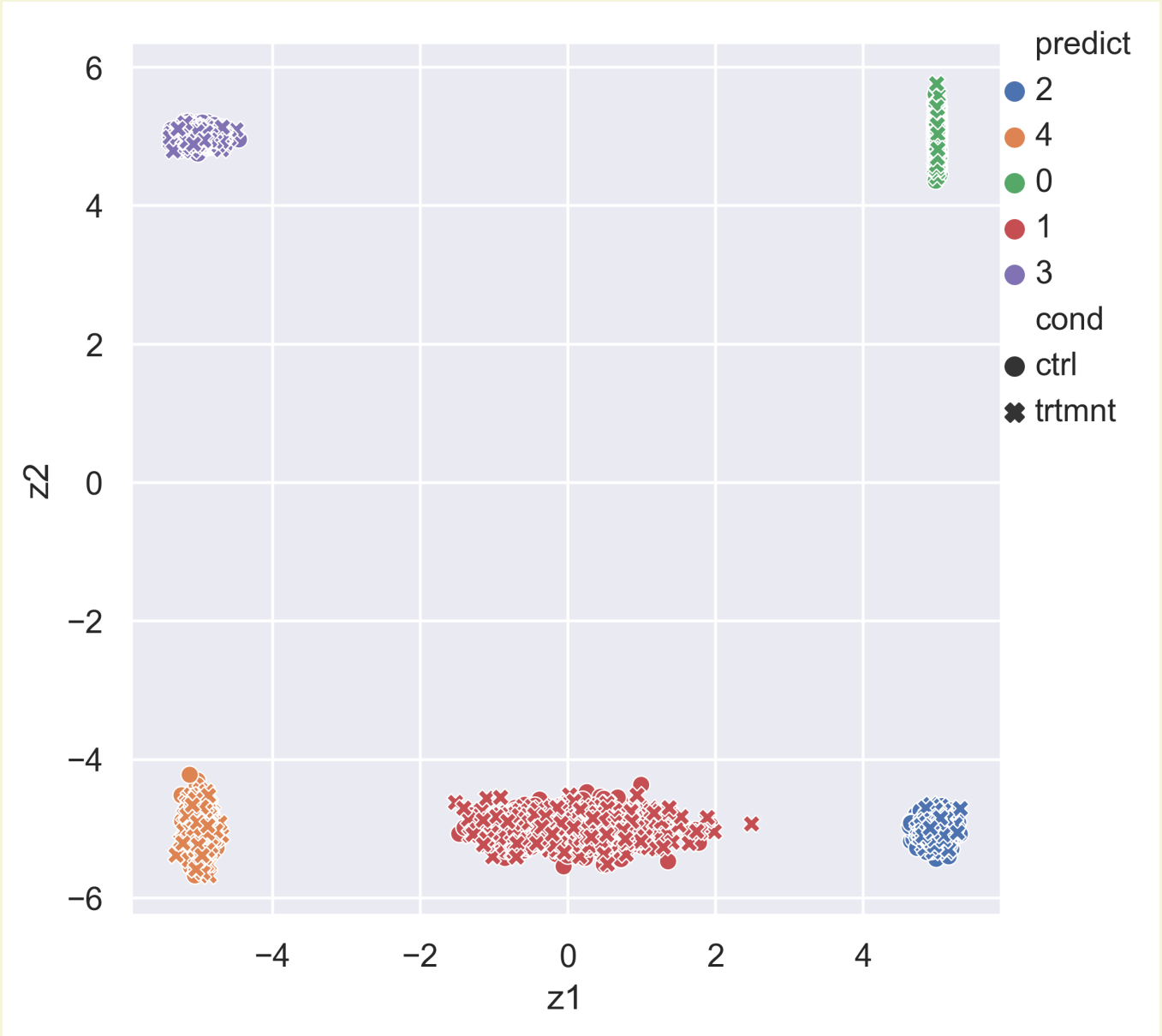


Figure: Sampling from encoded **z** space, fixed standard normal prior case.

Transfer learning

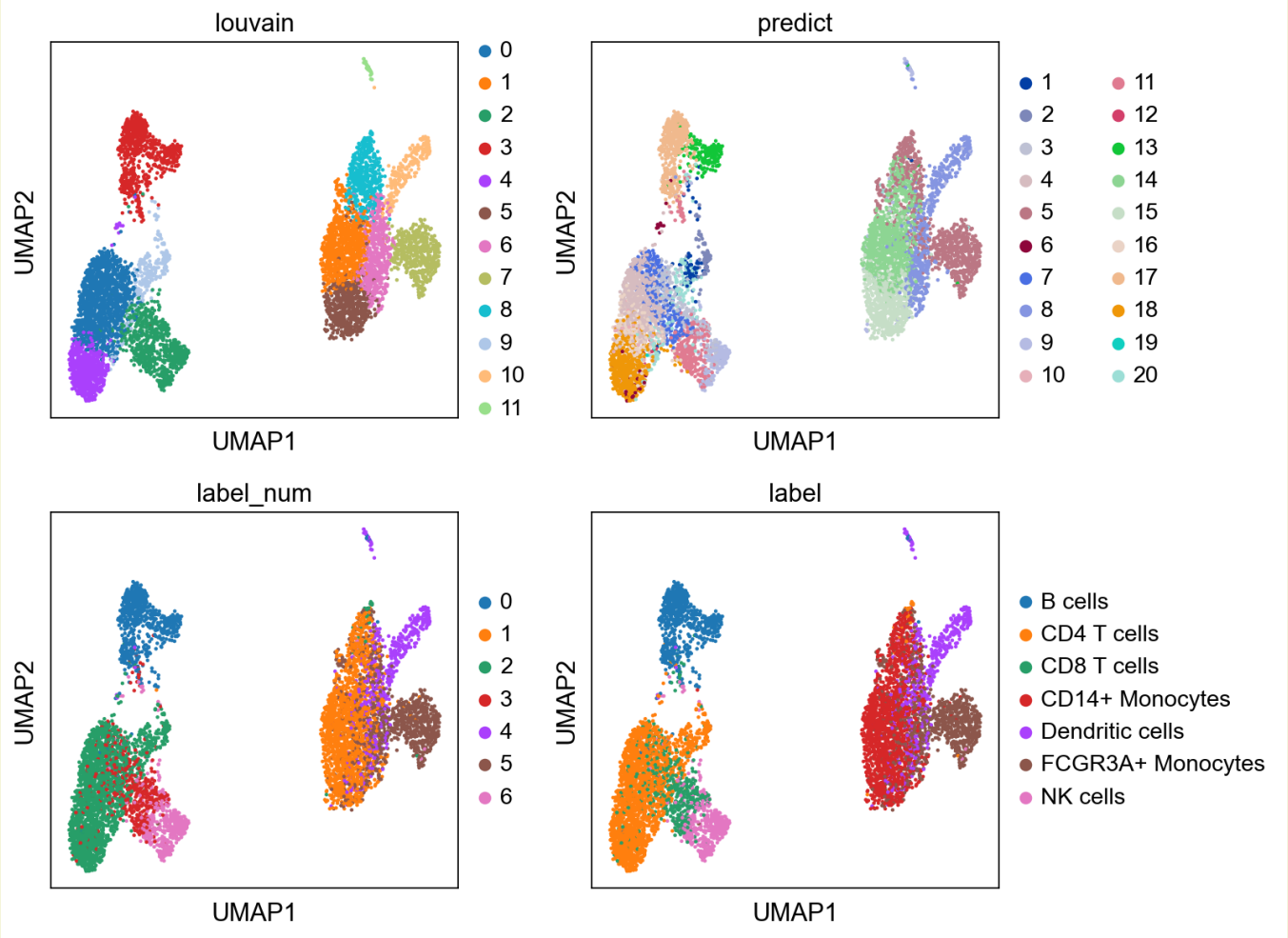


Figure: UMAP of PCA space of the Kang train data.

Transfer learning

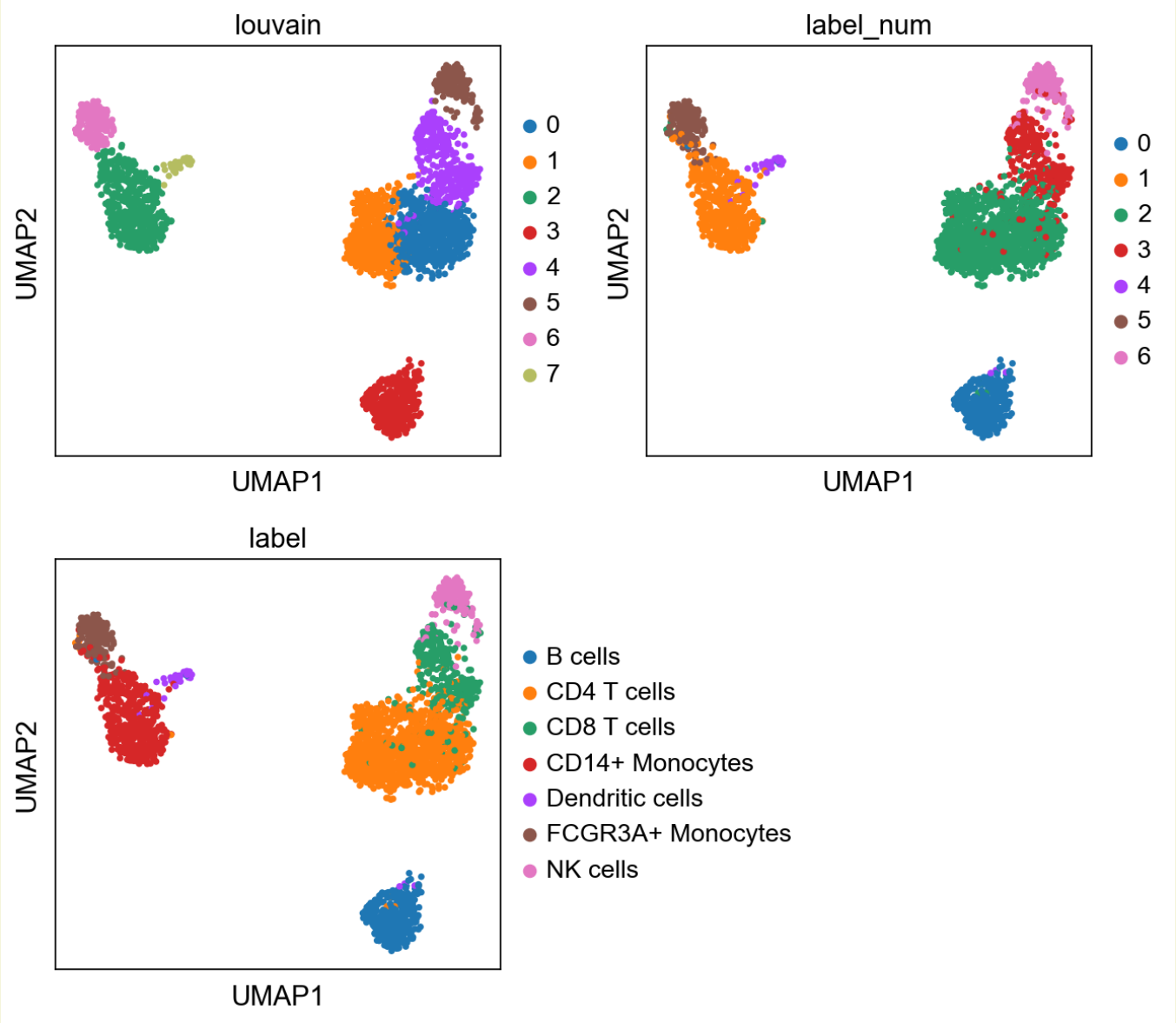


Figure: UMAP of the Zheng dataset.

Transfer learning

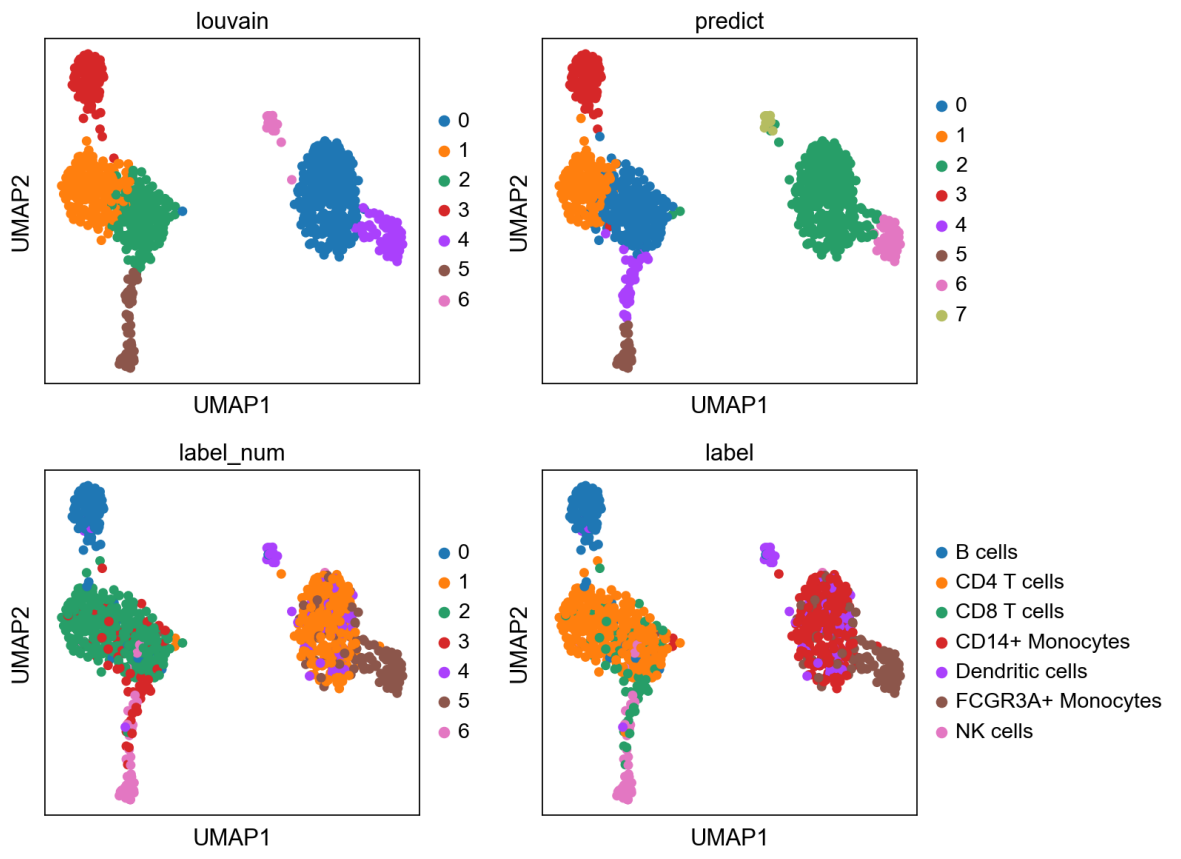


Figure: UMAP of the latent space of the Kang validation subset (the holdout).

converting treatment effect

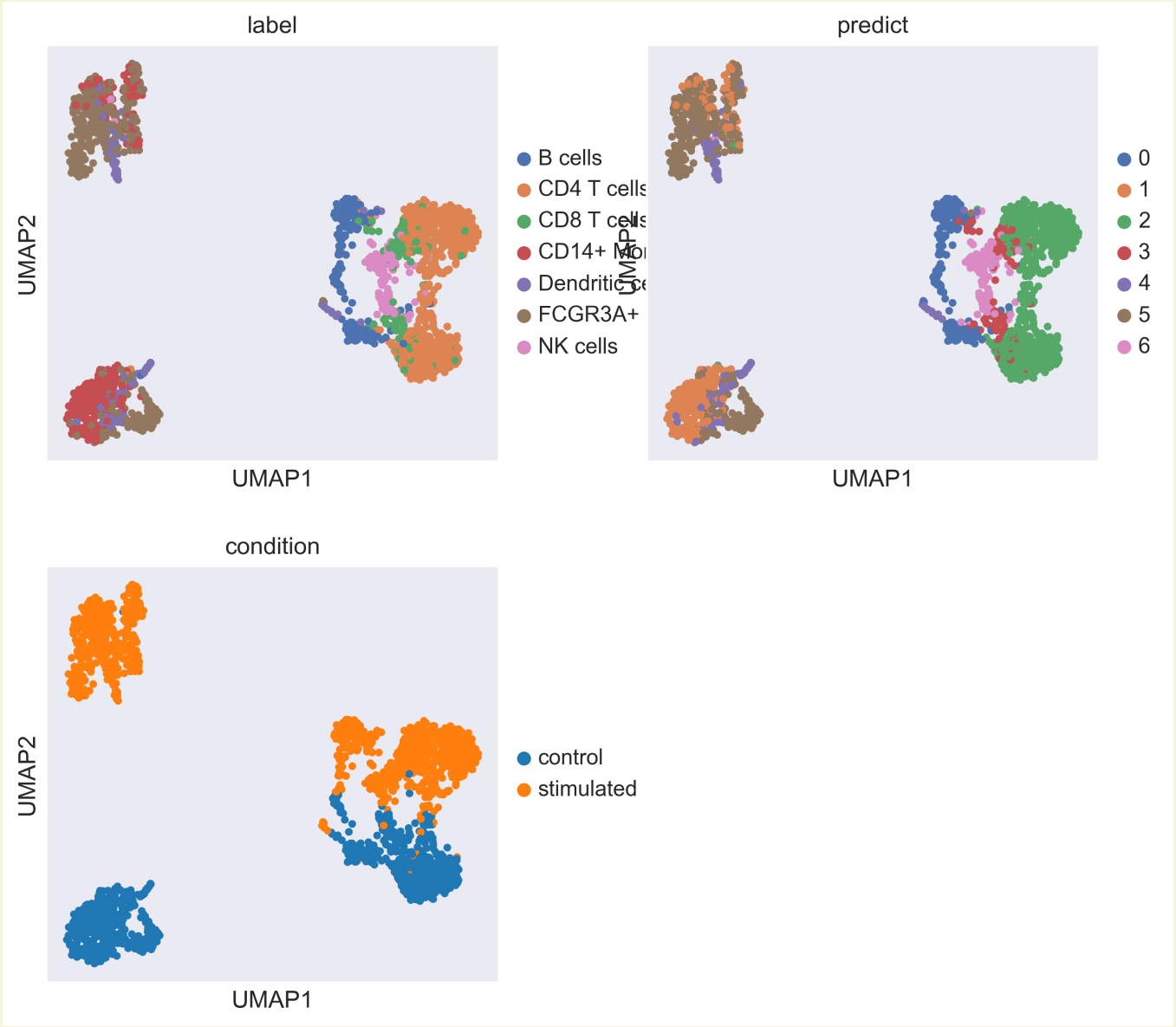


Figure: UMAP of the Kang validation dataset

converting treatment effect

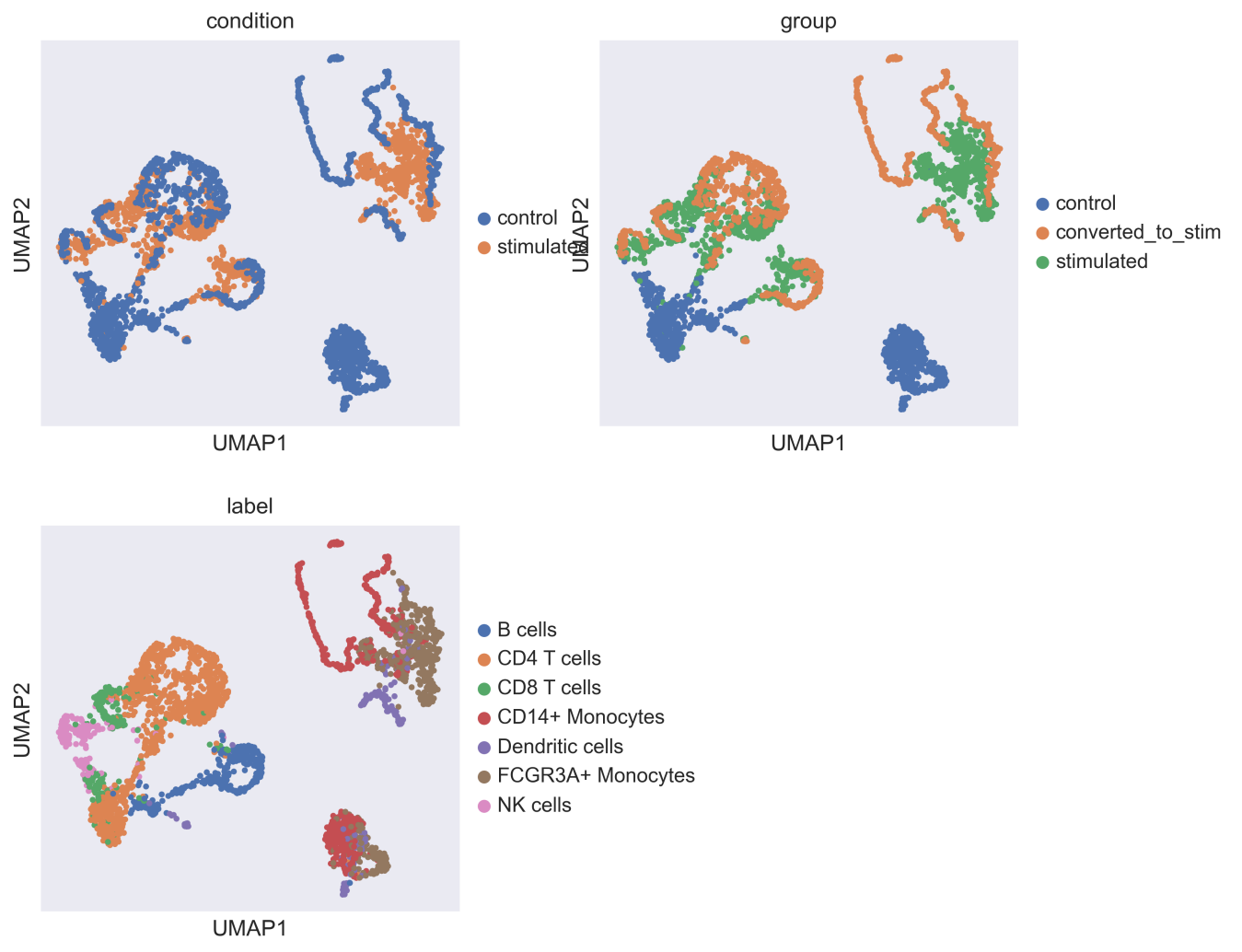
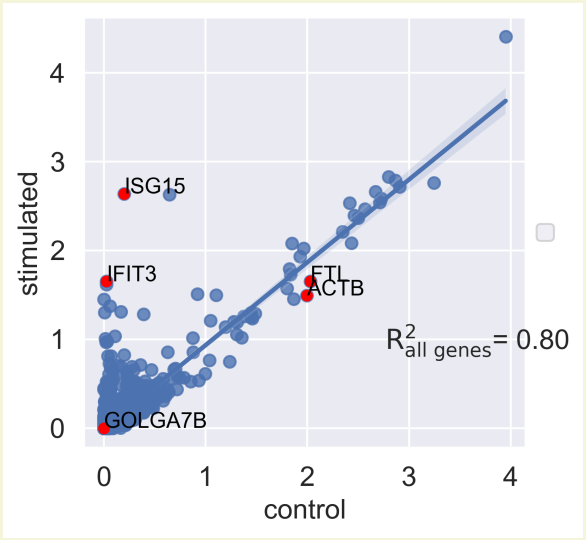
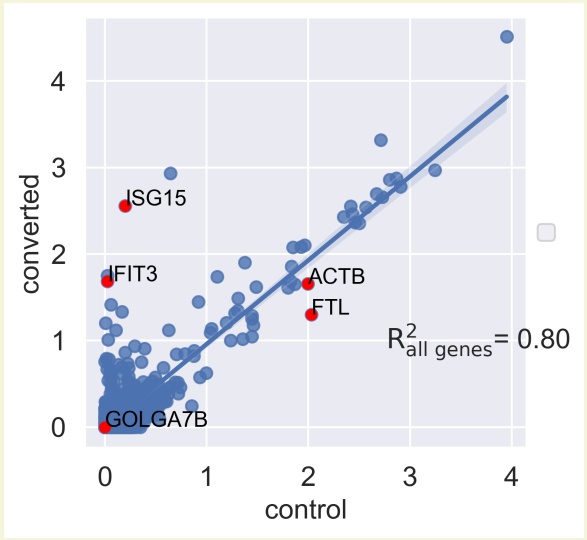
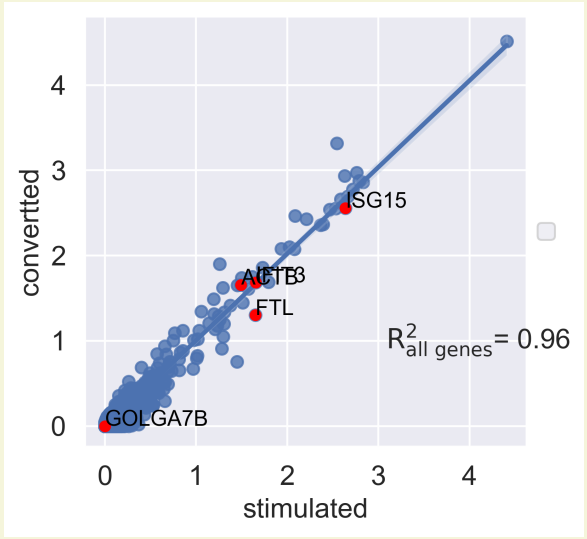


Figure: UMAP of the combined validation subset with its control group remapped into stimulated state.

converting treatment effect







Generating data

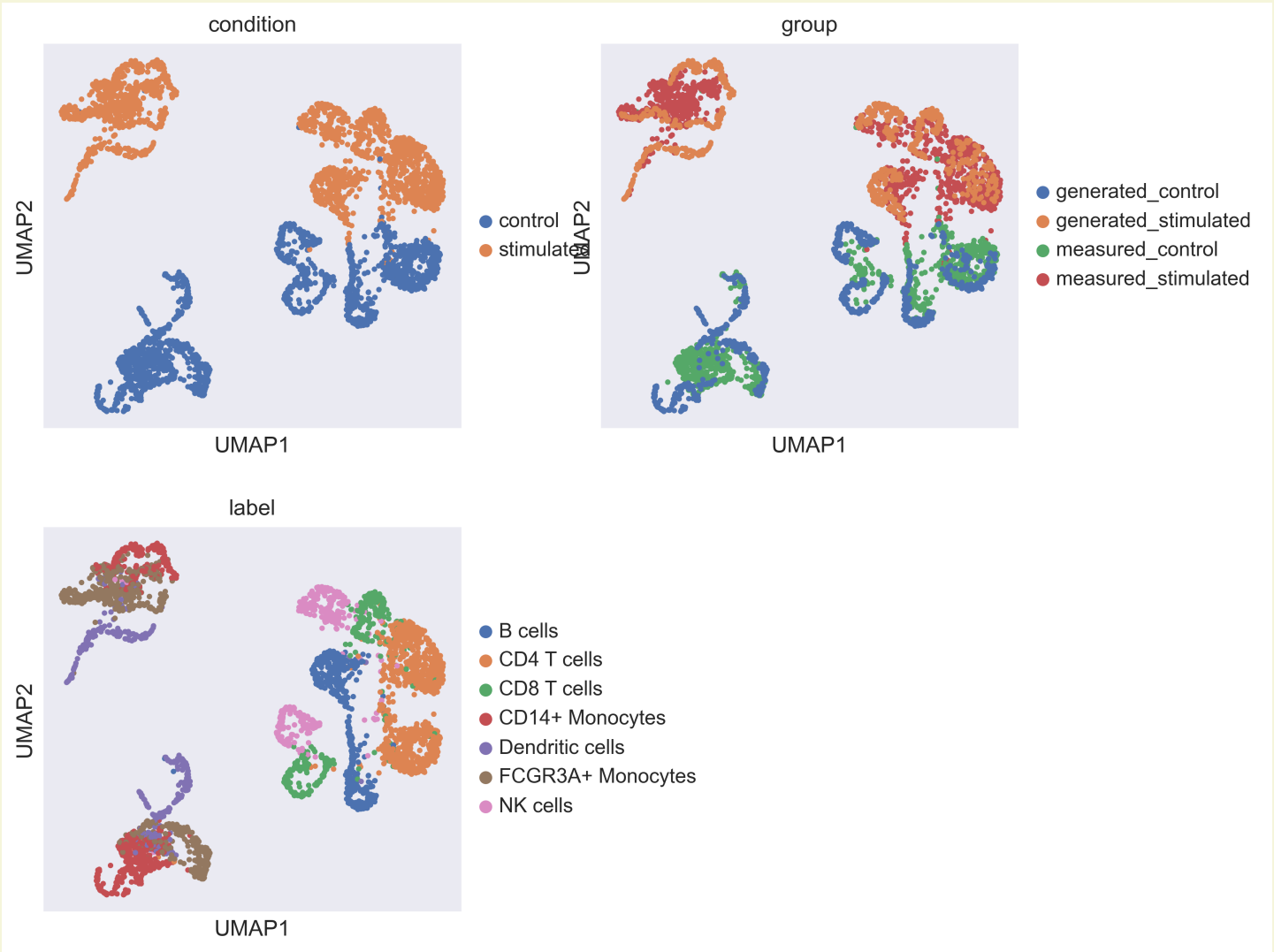


Figure: The Kang UMAP of real data combined with randomly generated data.

if you like to learn more

The code used for this thesis is available as a python module. The thesis, the python module, and a short tutorial are on this git [5]. The entire work (with extra mess and junk) is here [6].

Acknowledgement

- Martin Vingron
- MPG and Group V
- Tim Conrad

big warm thank you!

¯_(ツ)_/¯

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[2] Nat Dilokthanakul et al. "Deep unsupervised clustering with gaussian mixture variational autoencoders". In: *arXiv preprint arXiv:1611.02648* (2016).

[3] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[4] Durk P Kingma et al. "Semi-supervised learning with deep generative models". In: *Advances in neural information processing systems* 27 (2014).

[5] Yiftach Kolb. *The "official" c*GMΔVAE project git*. URL: https://github.com/zelhar/mg22.

[6] Yiftach Kolb. *The Github project housing this thesis*. URL: https://github.com/zelhar/mg22.

[7] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. "scGen predicts single-cell perturbation responses". In: *Nature methods* 16.8 (2019), pp. 715–721.

[8] Elad Plaut. “From principal subspaces to principal components with linear autoencoders”. In: *arXiv preprint arXiv:1804.10253* (2018).