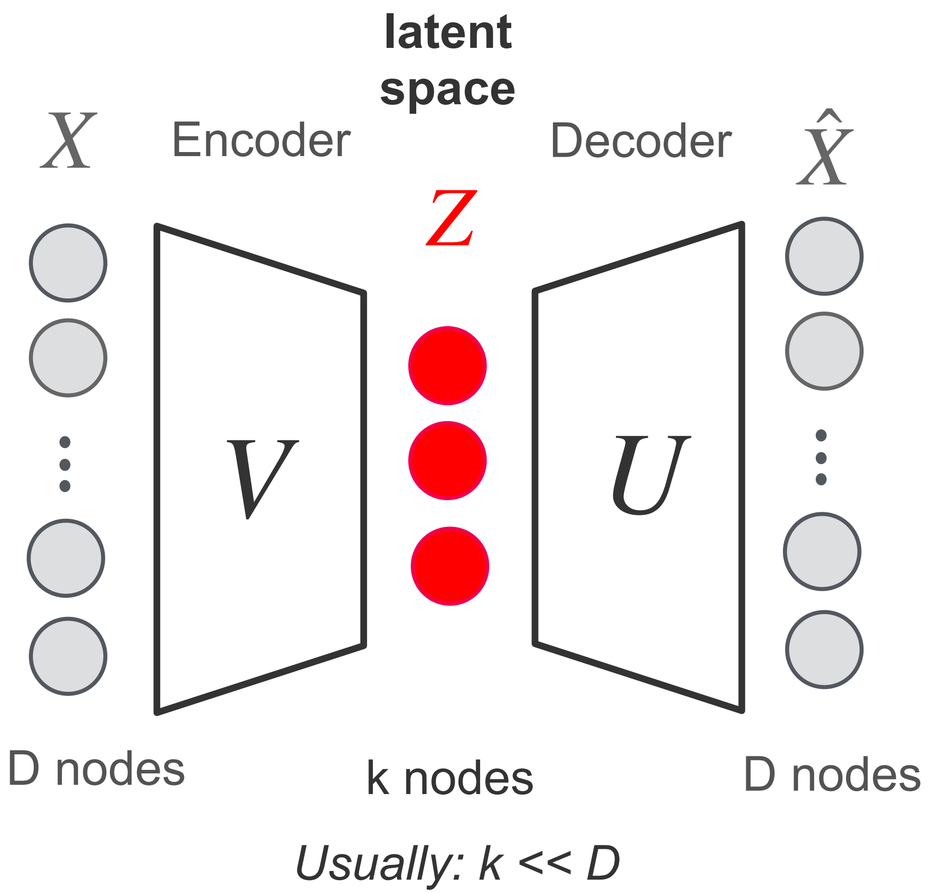
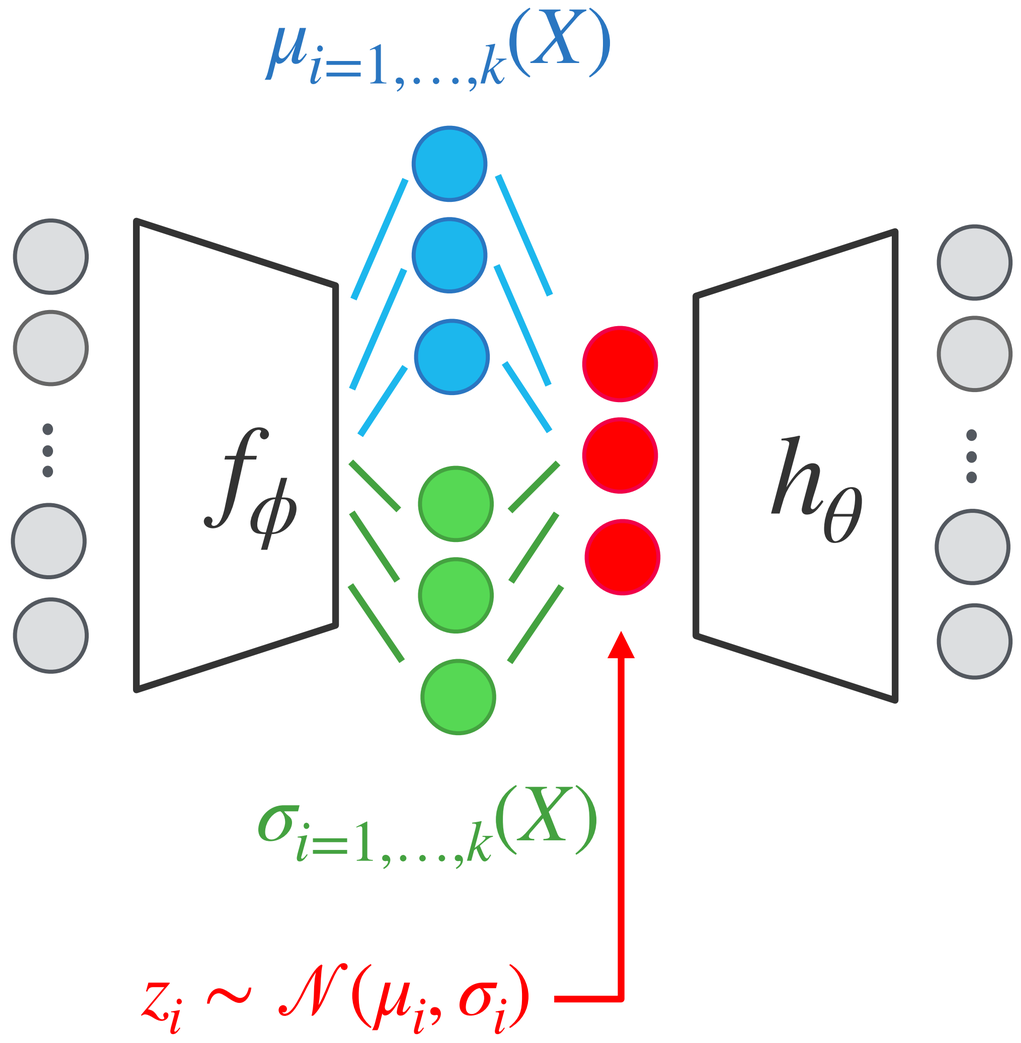
Deep Learning in Life Sciences

(Variational) Autoencoders in genomics

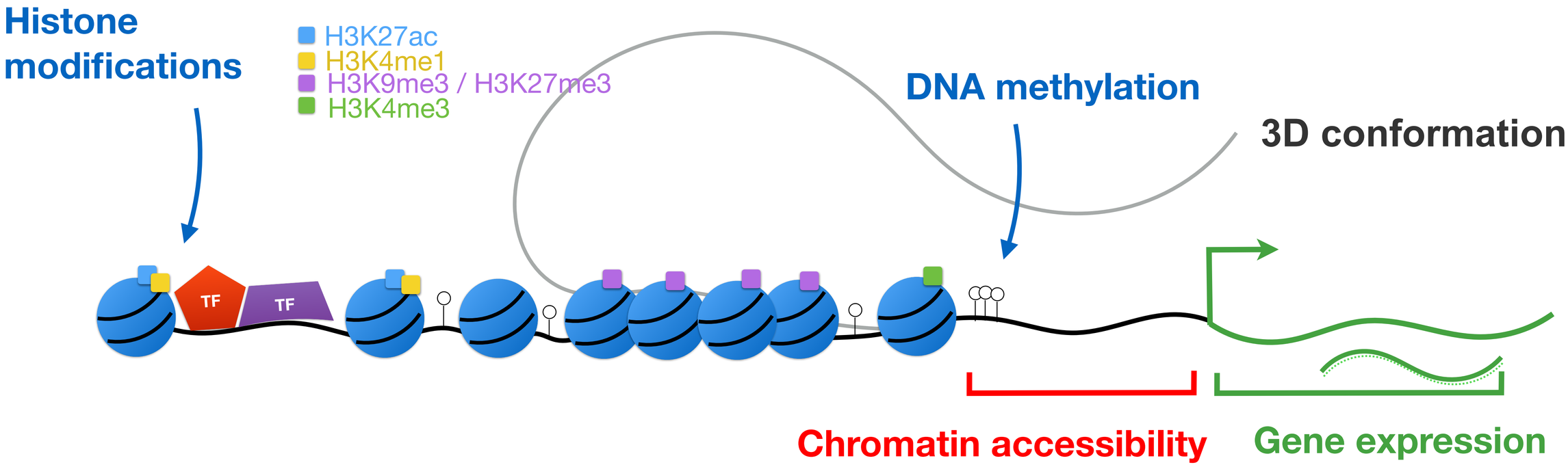
Carl Herrmann
Heidelberg University

Recap from week 3: AE / VAE



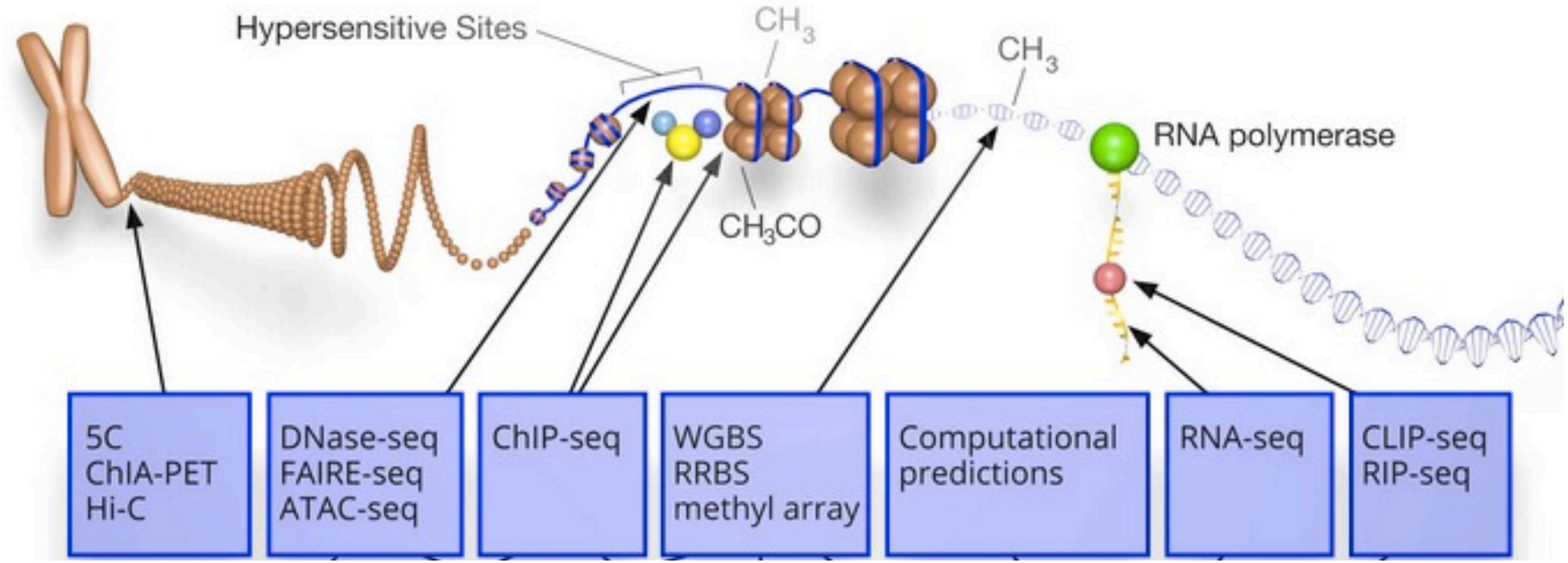


(Epi)genomic data



→ *high-throughput sequencing*

(Epi)genomic data

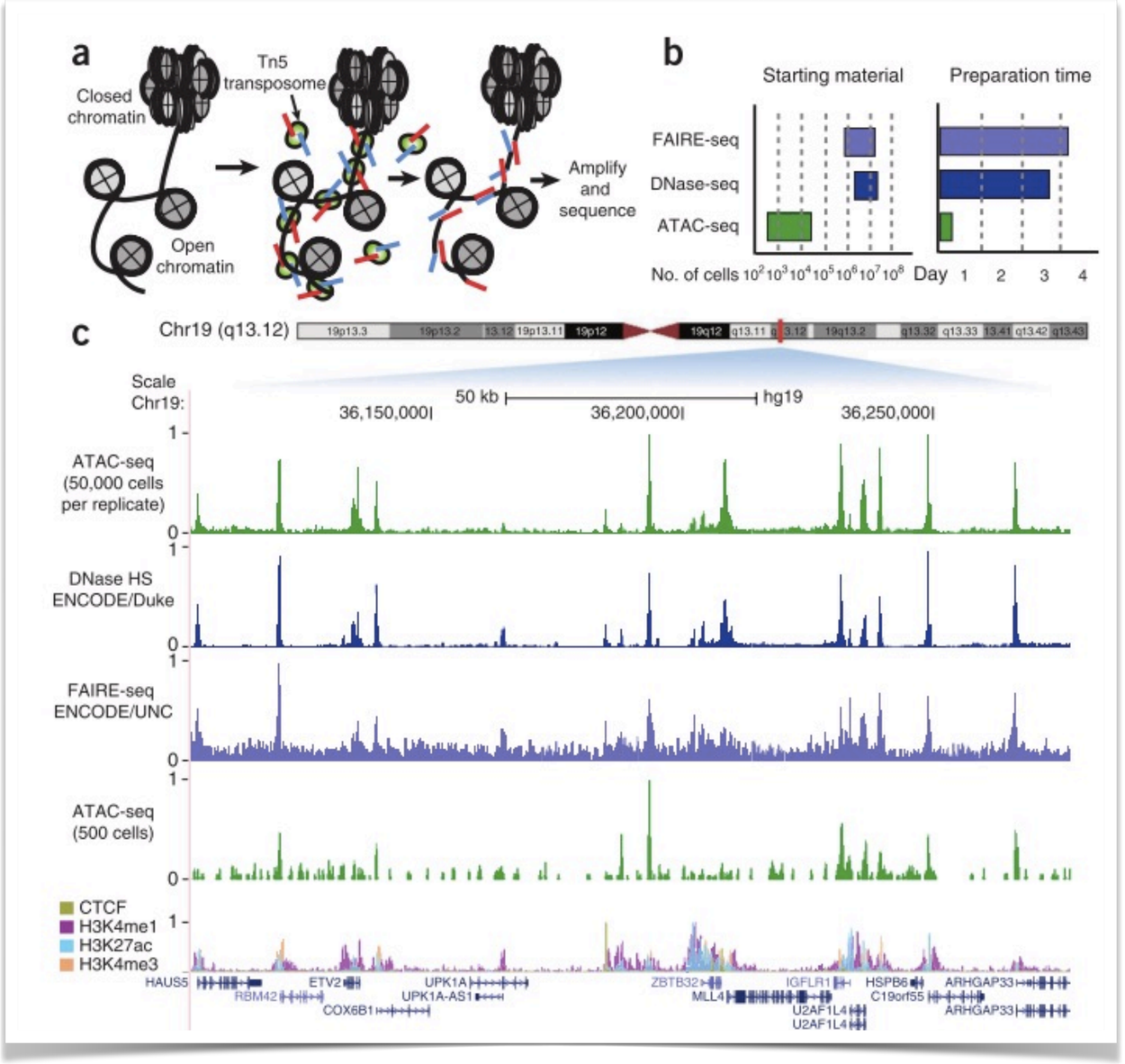


Most of these assays are
now available for single-cell genomics!

[ENCODE Project]

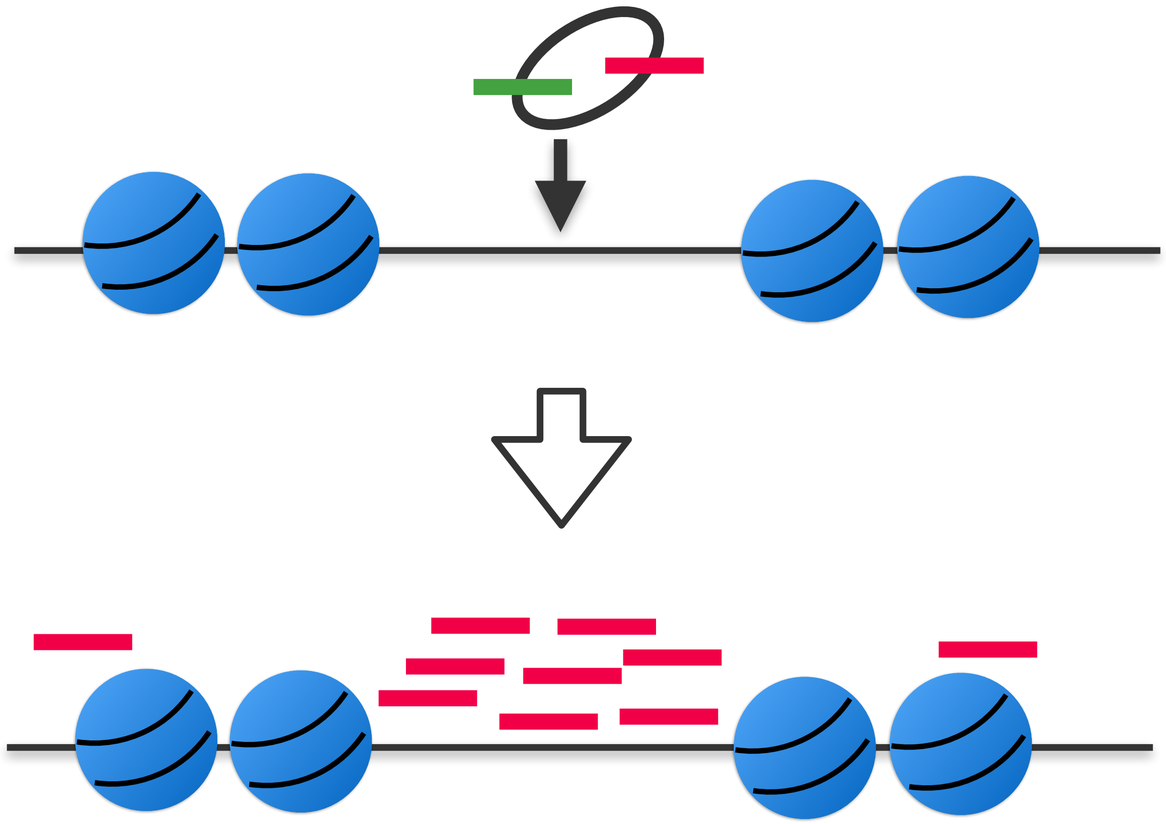
Chromatin accessibility

- **ATAC-seq**: using Tn5 transposase prepared with sequencing primers
- requires a small number of input material (~10,000 cells)
- easily adapter to single-cell sequencing
- identification of open chromatin regions (peaks)



[Greenleaf (2013)]

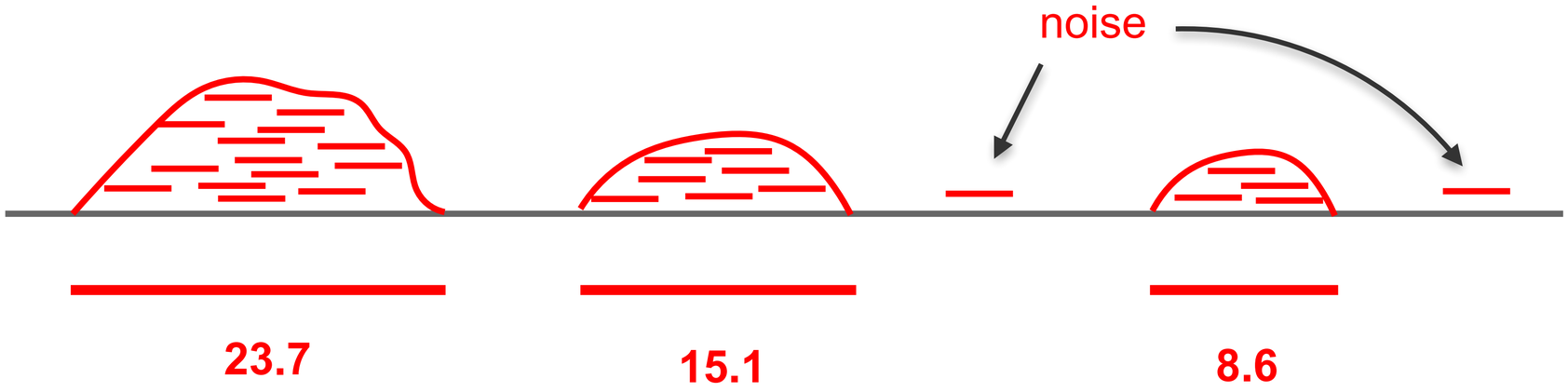
Chromatin accessibility (ATAC-seq)



insertion of Tn5 in multiple cells
at open regions

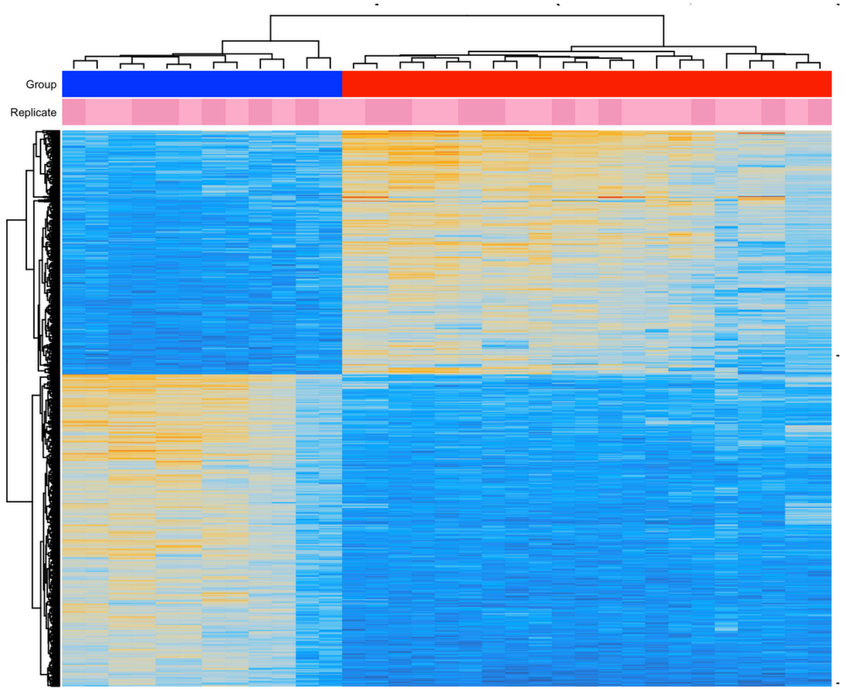
accumulation of sequencing reads
in open chromatin regions!

Chromatin accessibility



peaks
~ 50,000 - 100,000

samples



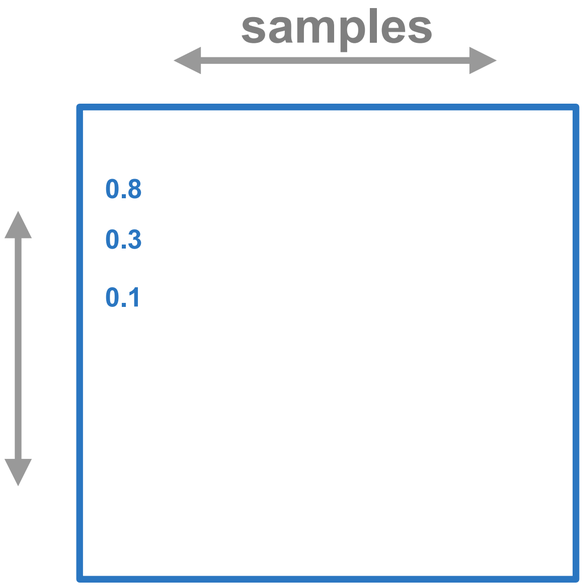
DNA methylation

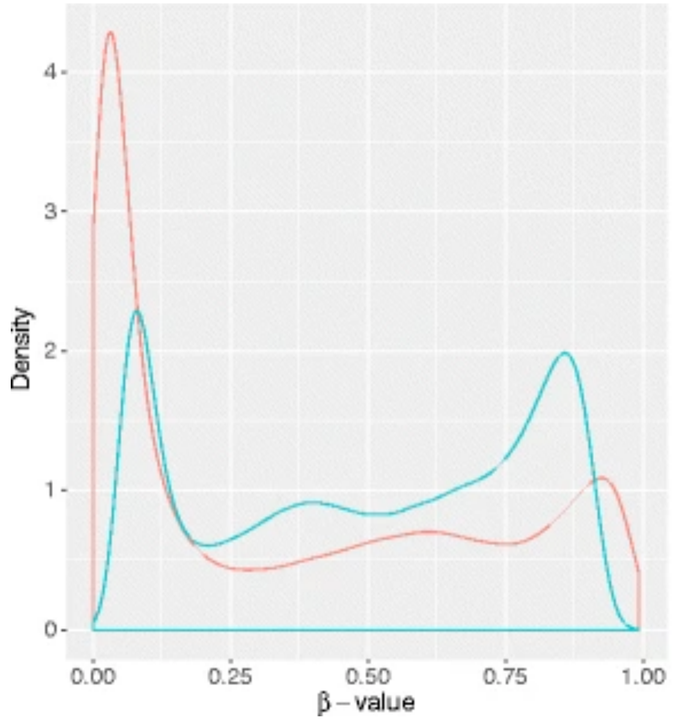
DeepLearning inLifeSciences

CpG dinucleotides

DNA methylation
$0 < \beta < 1$

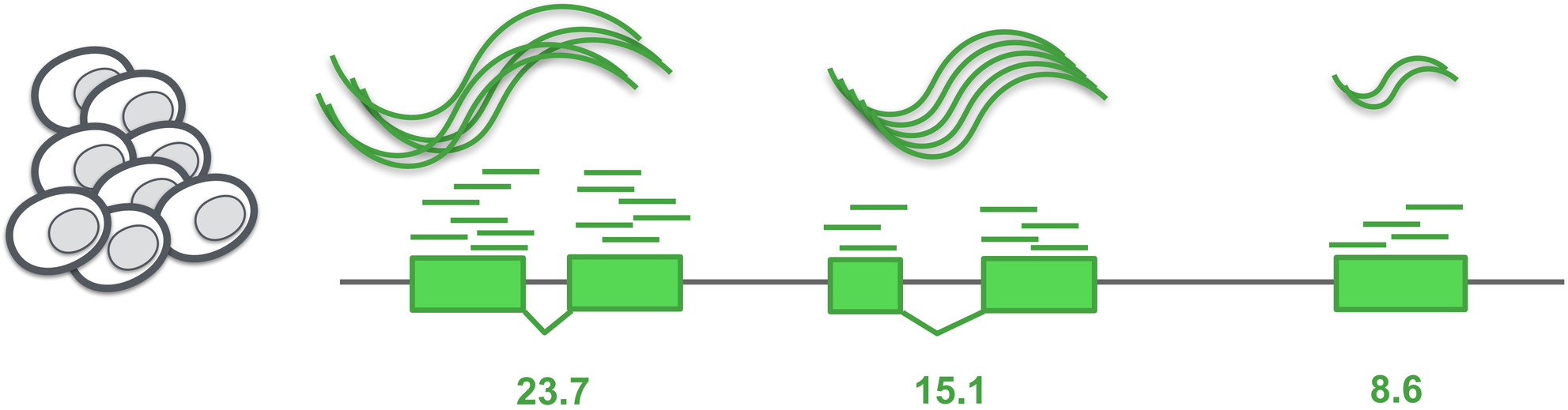
CpG dinucleotides
~ 100,000 - 16^6



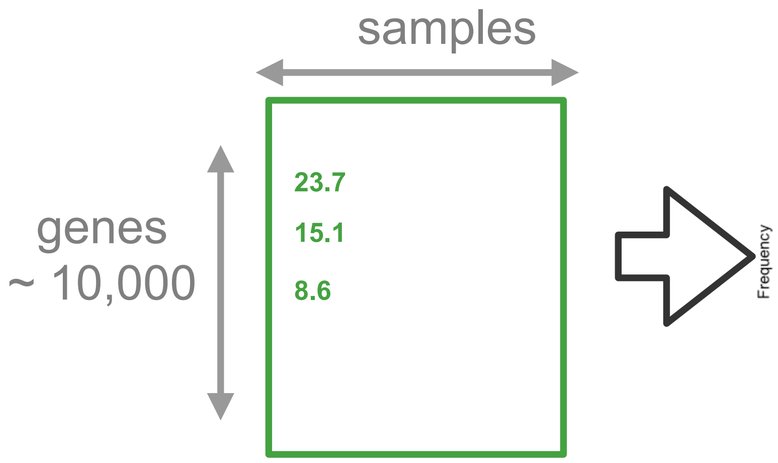


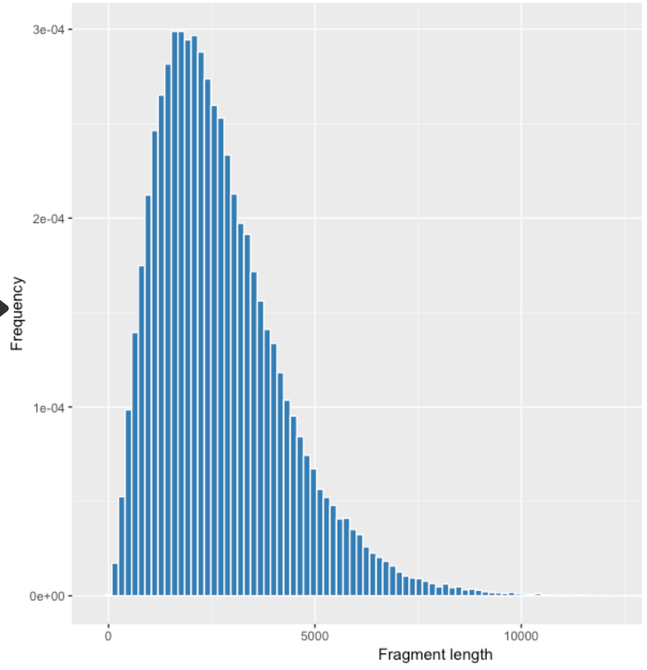
Bimodal distribution

Gene expression



gene expression quantification





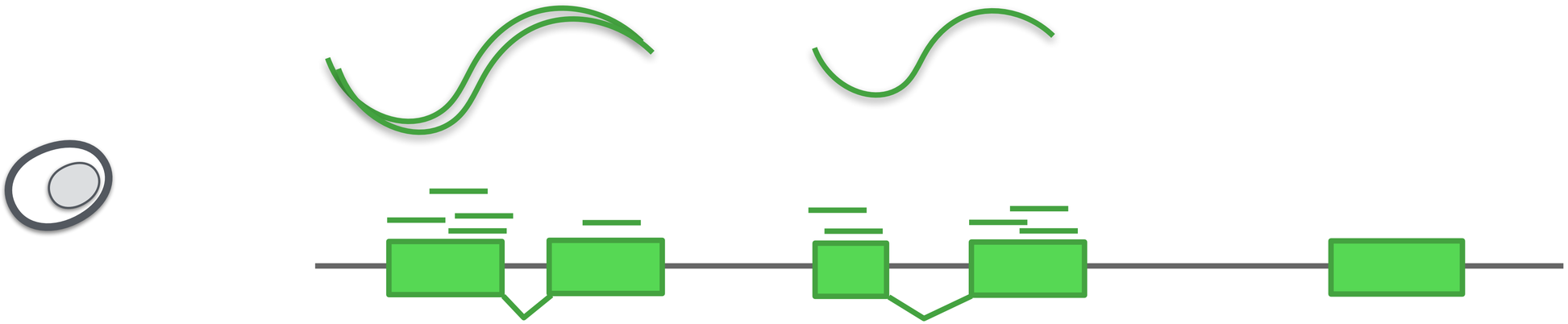
negative binomial distribution

$$NB(x;\theta,p) = \frac{\Gamma(\theta+x)}{x!\Gamma(\theta)}p^{\theta}(1-p)^{x}$$

$$E(X) = \frac{\theta(1-p)}{p} \qquad Var(X) = \frac{\theta(1-p)}{p^2} = \frac{E(X)}{p}$$

single-cell Gene expression

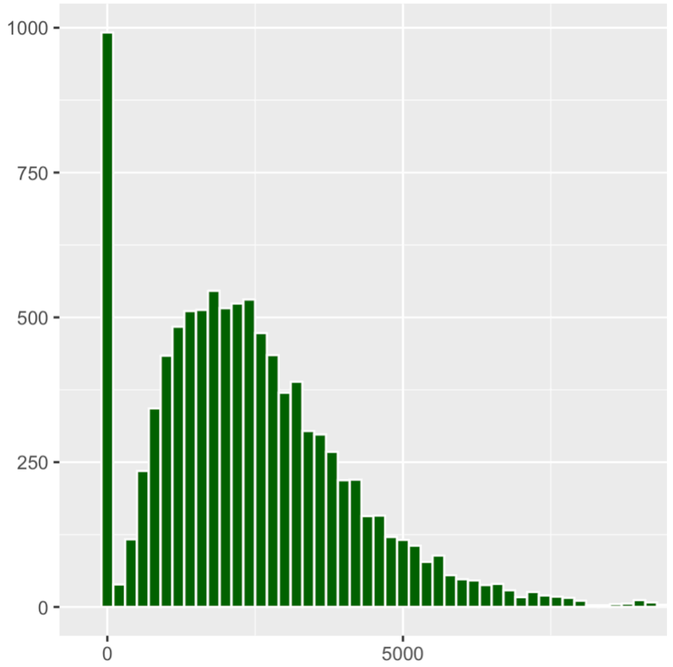
DeepLearning inLifeSciences



mRNA

4.1 2.3 0 (?)

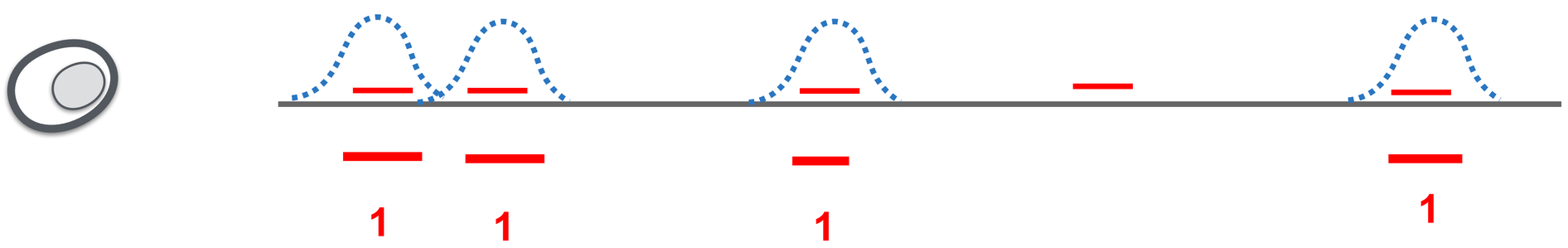
gene expression quantification



<u>sparse</u> data !
zero-inflated negative binomial distribution (ZINB)

$$ZINB(x;\pi,\theta,p) = \pi\delta(x) + (1-\pi)\,NB(x;\theta,p)$$

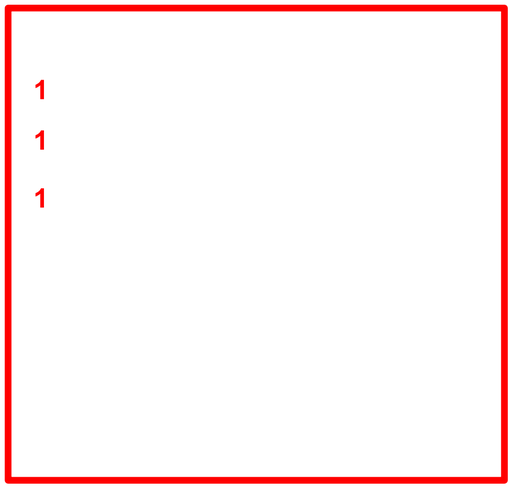
single-cell chromatin accessibility



peak regions
from pseudo-bulk

single-cells

fragments
~ 50,000 - 100,000

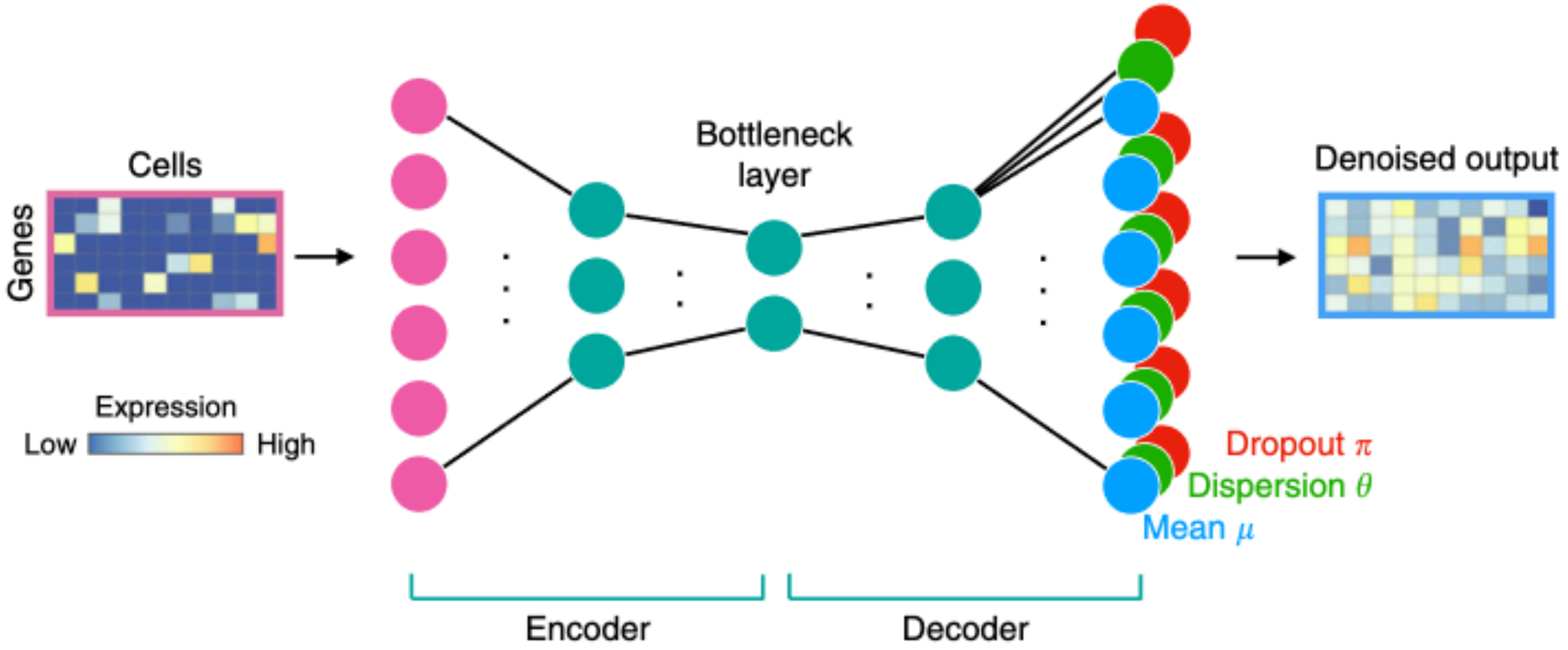


single-cell chromatin accessibility can be represented as a sparse binary matrix

1 - denoising model for scRNA-seq

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).

"deep count autoencoder"



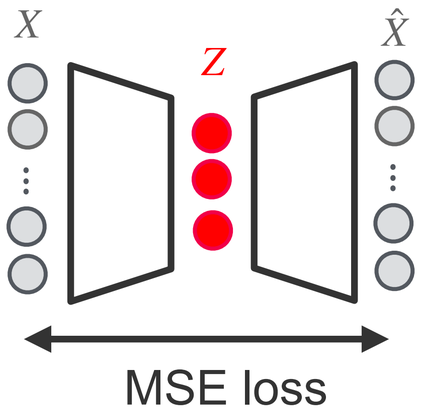
$$ZINB(x;\pi,\theta,\mu) = \pi\delta(x) + (1-\pi)\,NB(x;\theta,\mu)$$

learn parameters π, θ, μ
using likelihood maximization

$$\hat{\pi},\hat{\theta},\hat{\mu} = \text{argmax}\; ZINB(x;\pi,\theta,\mu)$$

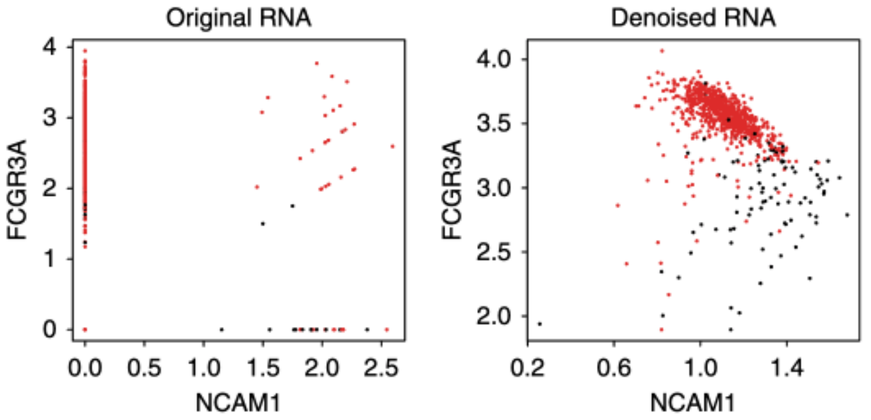
"deep count autoencoder"

$$ZINB(x;\hat{\pi},\hat{\theta},\hat{\mu}) = \hat{\pi}\delta(x) + (1-\hat{\pi})\,NB(x;\hat{\mu},\hat{\theta})$$

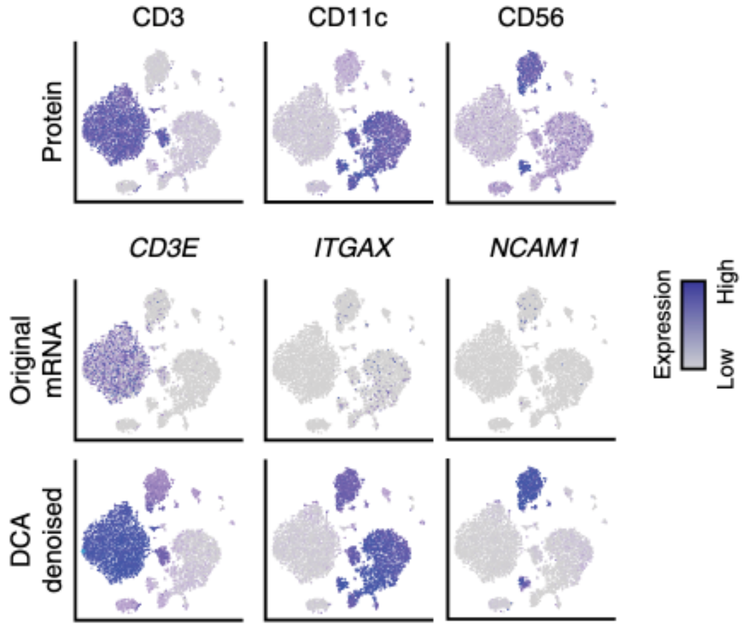


Advantages of denoising

- better separation of cell sub-populations
  - **identification of small populations** with subtle phenotypic differences (e.g. NK cells)



- better correlation between **protein and gene expression**



2 - VAE for data integration

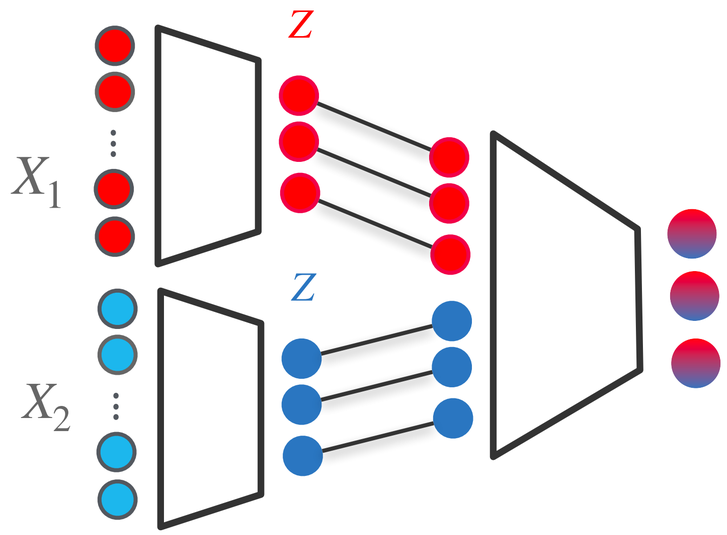
[1] Zhang, X. et al. Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification. in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 765–769 (IEEE, 2019). doi:10.1109/bibm47256.2019.8983228.

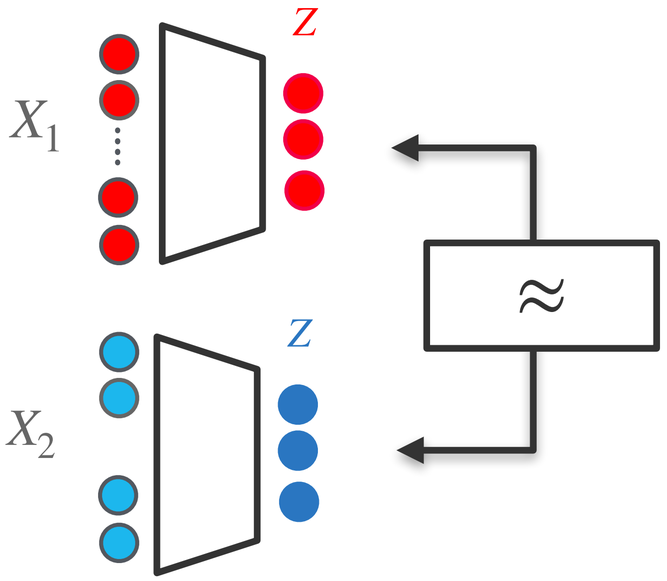
[2] Yang, K. D. et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat Commun 12, 31 (2021).

VAE for data integration

- We often have multiple data modalities for the same samples / cells / ...
  - gene expression
  - epigenomic data
  - proteomics
  - image data

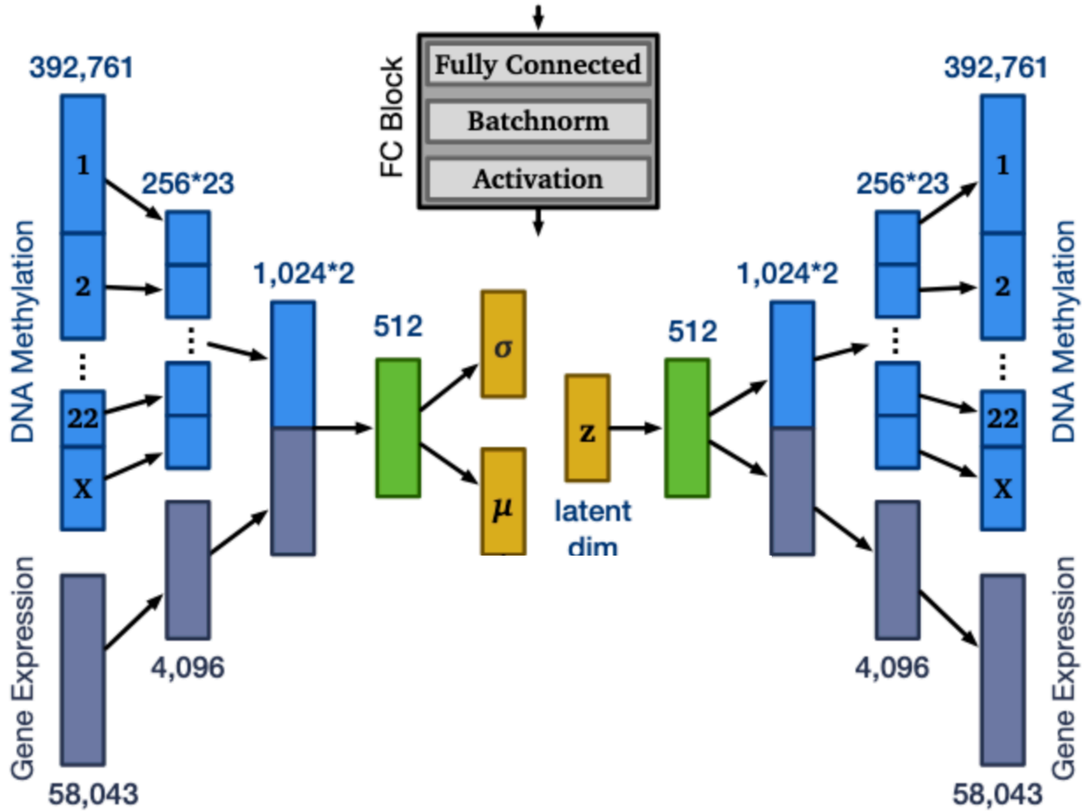
- We can use autoencoders to project these modalities into a **shared latent space**
  - simple **concatenation** of embeddings
  - latent space **coupling**
  - latent space **translation** → allows translating one modality into another!





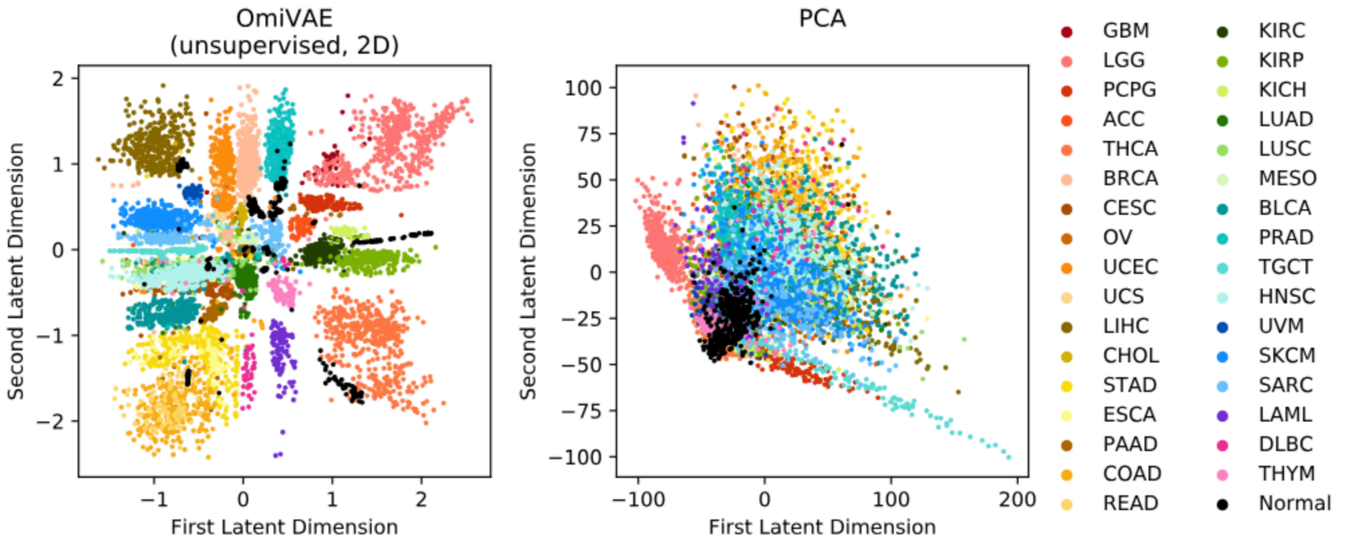
OmiVAE model

- VAE to integrate bulk
  - **gene expression** (RNA-seq)
  - **DNA methylation** data (array)
- trained on bulk data from cancer patients (TCGA)
- Embedding + concatenation of the embeddings
- 2 models
  - unsupervised
  - supervised (with classifier on latent space)



OmiVAE

- **unsupervised** analysis
- better multi-class classification performance compared with other methods



	Accuracy	Precision	Recall	F1 Score
PCA+SVM	30.13±1.62%	0.26±0.02	0.30±0.02	0.26±0.02
KPCA+SVM	30.16±1.65%	0.26±0.02	0.30±0.02	0.26±0.02
t-SNE+SVM	82.94±0.87%	0.80±0.01	0.83±0.01	0.80±0.01
UMAP+SVM	80.39±0.96%	0.73±0.01	0.80±0.01	0.76±0.01
OmiVAE+SVM (1st phase, 2D)	**84.40±0.75%**	**0.83±0.01**	**0.84±0.01**	**0.82±0.01**

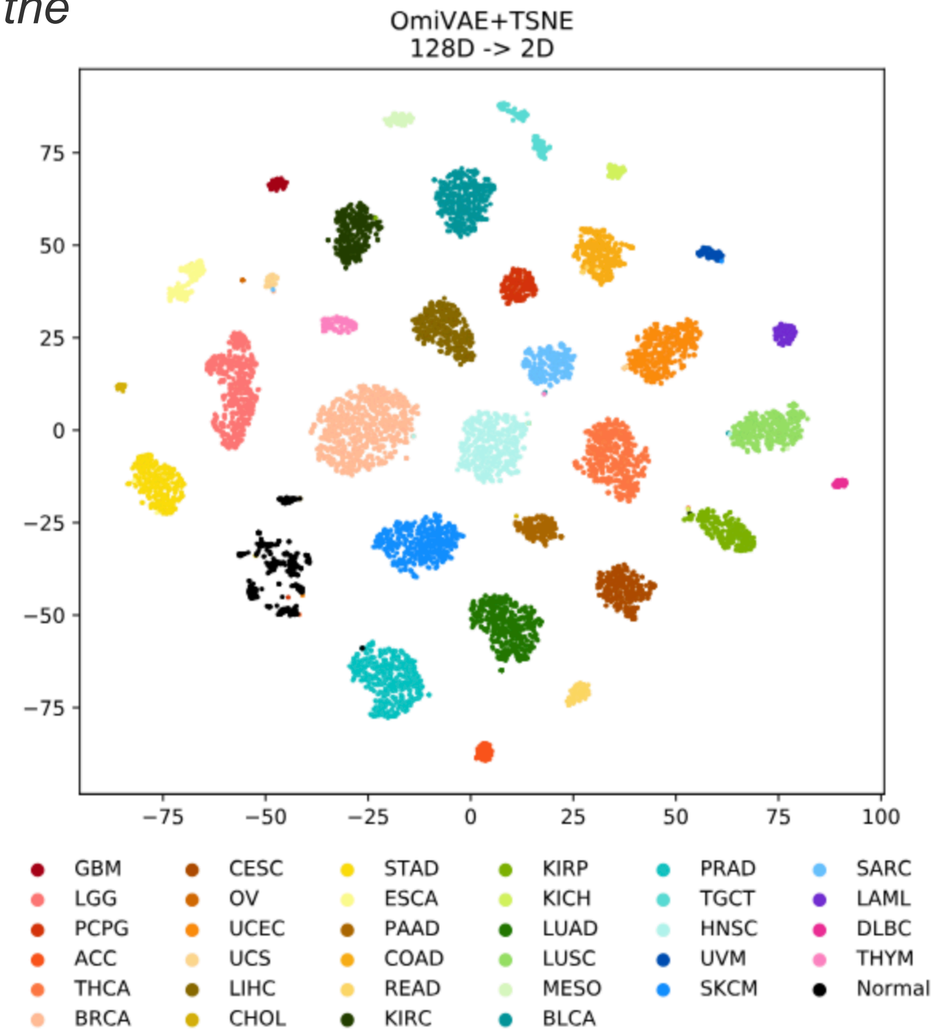
OmiVAE

- **supervised** analysis
  - unsupervised training
  - fine-tuning with classifier attached to latent space

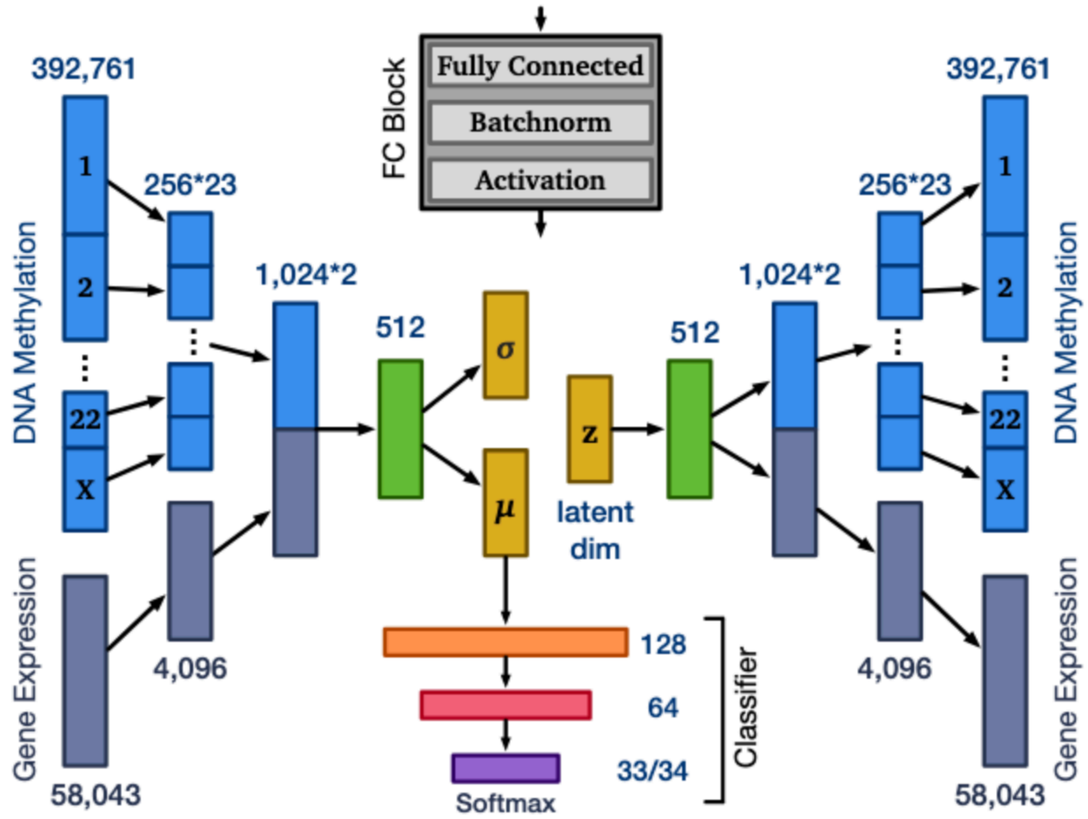
$$\mathscr{L}_{total} = \mathscr{L}_{VAE} + \beta \mathscr{L}_{class}$$

performance on the testing set

	Application Mode	Accuracy	F1 Score
Only Gene Expression	OmiVAE+SVM (unsupervised phase)	93.12±0.54%	0.926±0.006
	OmiVAE (end-to-end model)	**96.37±0.46%**	0.963±0.005
Only DNA Methylation	OmiVAE+SVM (unsupervised phase)	91.10±0.92%	0.899±0.010
	OmiVAE (end-to-end model)	**96.37±0.70%**	0.963±0.007
Multi-Omics Data	OmiVAE+SVM (unsupervised phase)	94.16±0.74%	0.937±0.008
	OmiVAE (end-to-end model)	**97.49±0.45%**	0.975±0.005

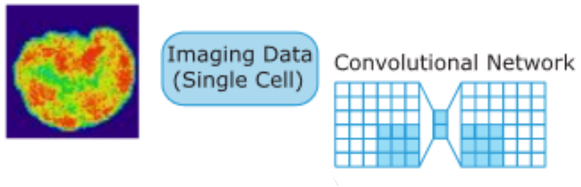


Caveat

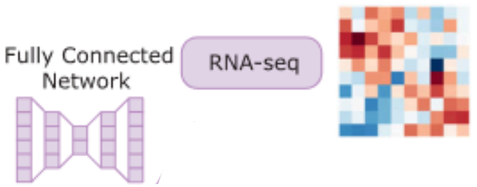


Model cannot be used if you have just one of the 2 data modalities ...

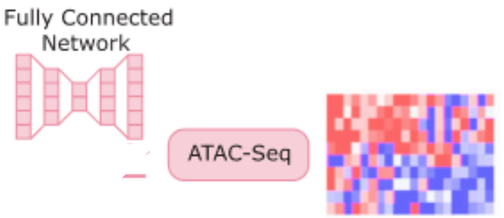
Co-embeddings



cellular imaging data



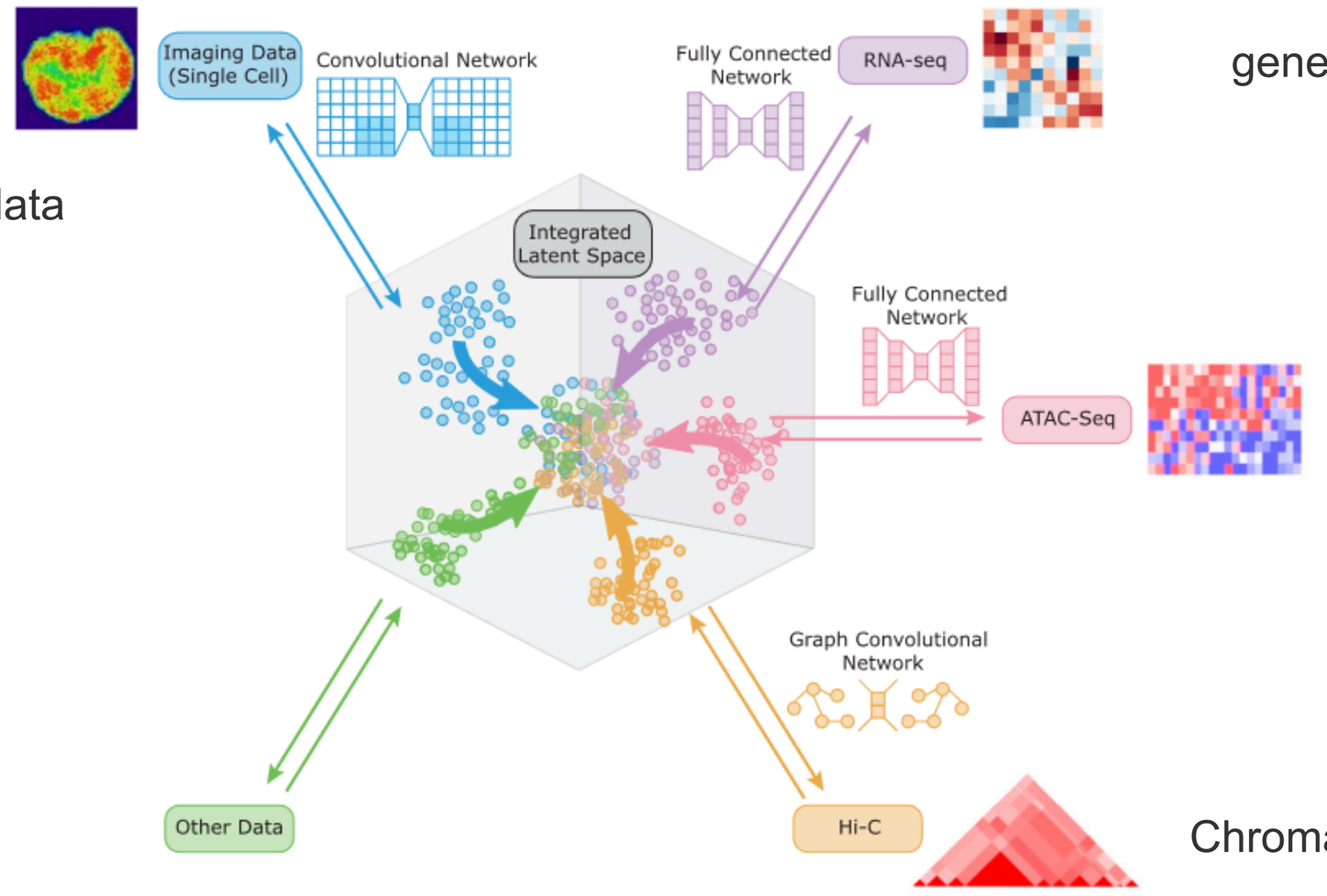
gene expression



Chromatin accessibility

Chromatin interactions

Co-embeddings



cellular imaging data

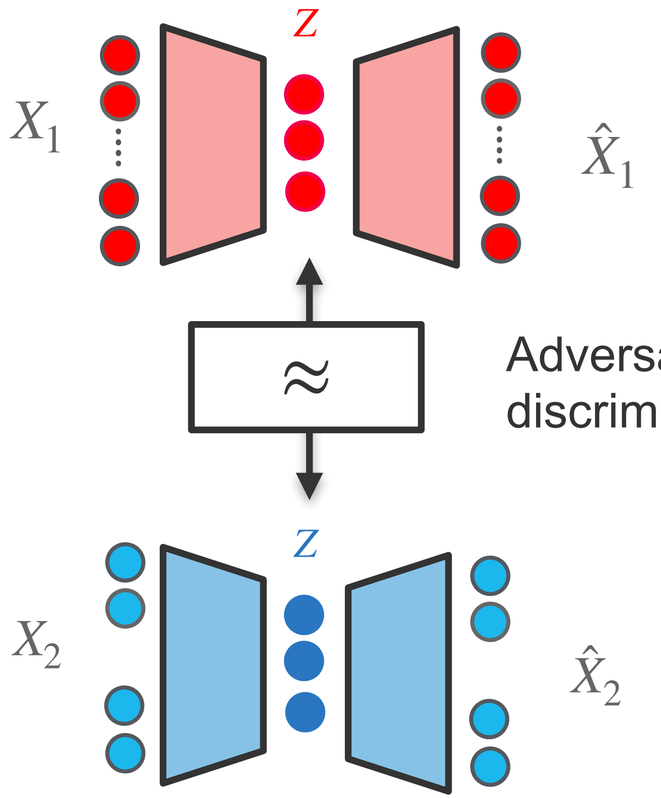
gene expression

Chromatin accessibility

Chromatin interactions

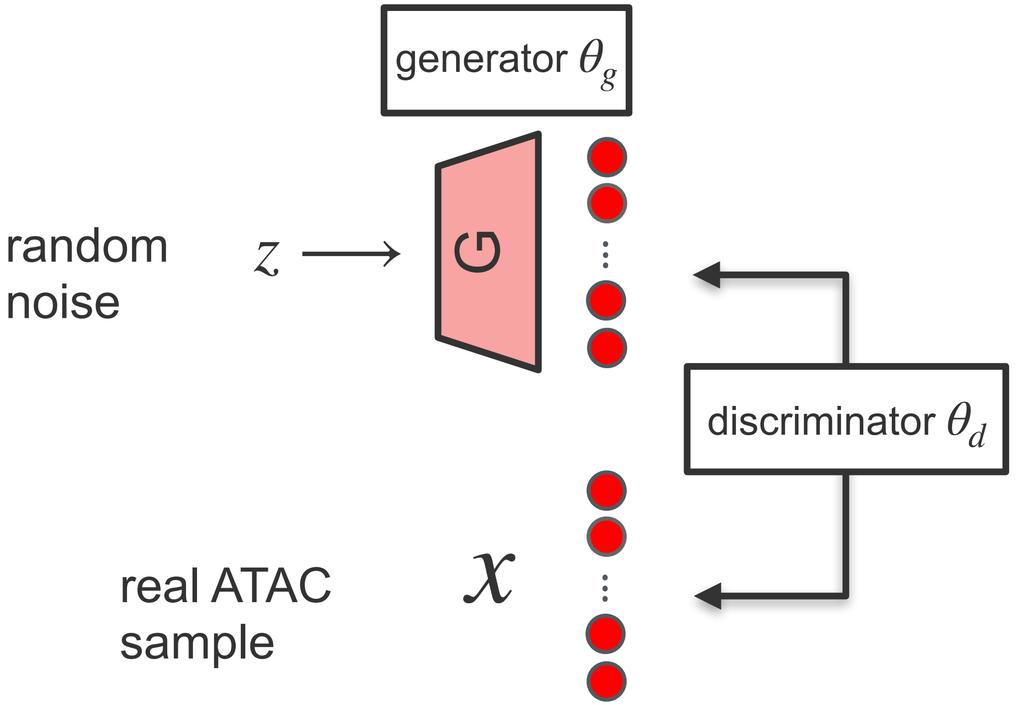
Co-embeddings

- Latent spaces of different modalities are aligned using a **General Adversarial network (GAN)**
  → trained until the initial data-type is no longer recognizable
- **Advantages**
  - model can be trained on multiple modalities and applied on a test data with one single modality!
  - Can be used to **translate** modalities



Adversarial networks (GAN)

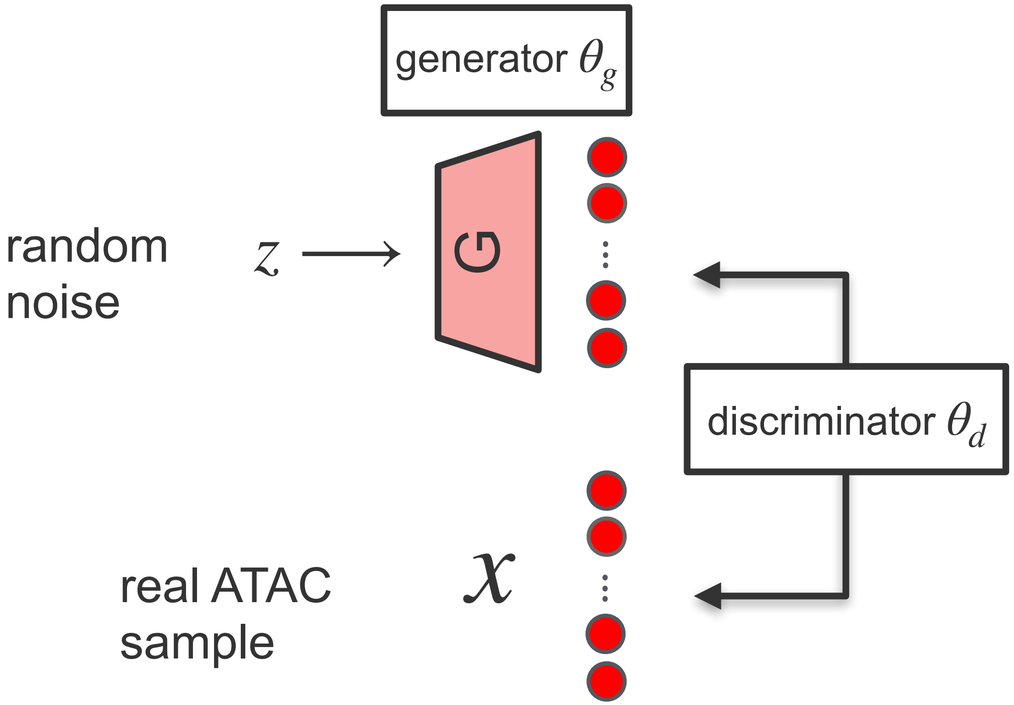
- Game theory scenario in which 2 components compete
  - **generator**: generates samples
  - **discriminator**: classifies samples into genuine samples or samples produced by the generator



1. Sample random noise.
2. Produce **generator** output from sampled random noise.
3. Get **discriminator** "Real" or "Fake" classification for generator output.
4. Calculate **loss from discriminator** classification.
5. Backpropagate through both the discriminator and generator to obtain gradients.
6. Use gradients to change only the **generator weights**.

Adversarial networks (GAN)

- Game theory scenario in which 2 components compete
  - **generator**: generates samples
  - **discriminator**: classifies samples into genuine samples or samples produced by the generator



Minmax Loss

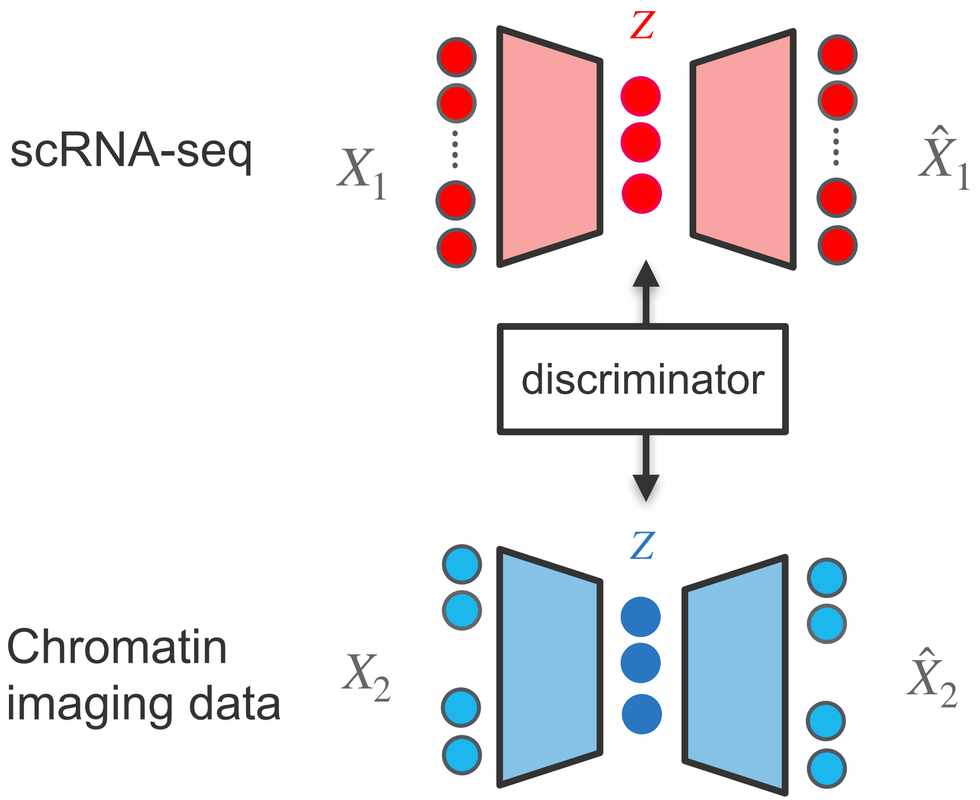
$$v(\theta_d, \theta_g) = \mathbb{E}_x(\log D(x)) + \mathbb{E}_z(1 - \log(D(G(z)))$$

$D(x)$ = prob. that x is real

- **discriminator** tries to **maximize** v
- **generator** tries to **minimize** v

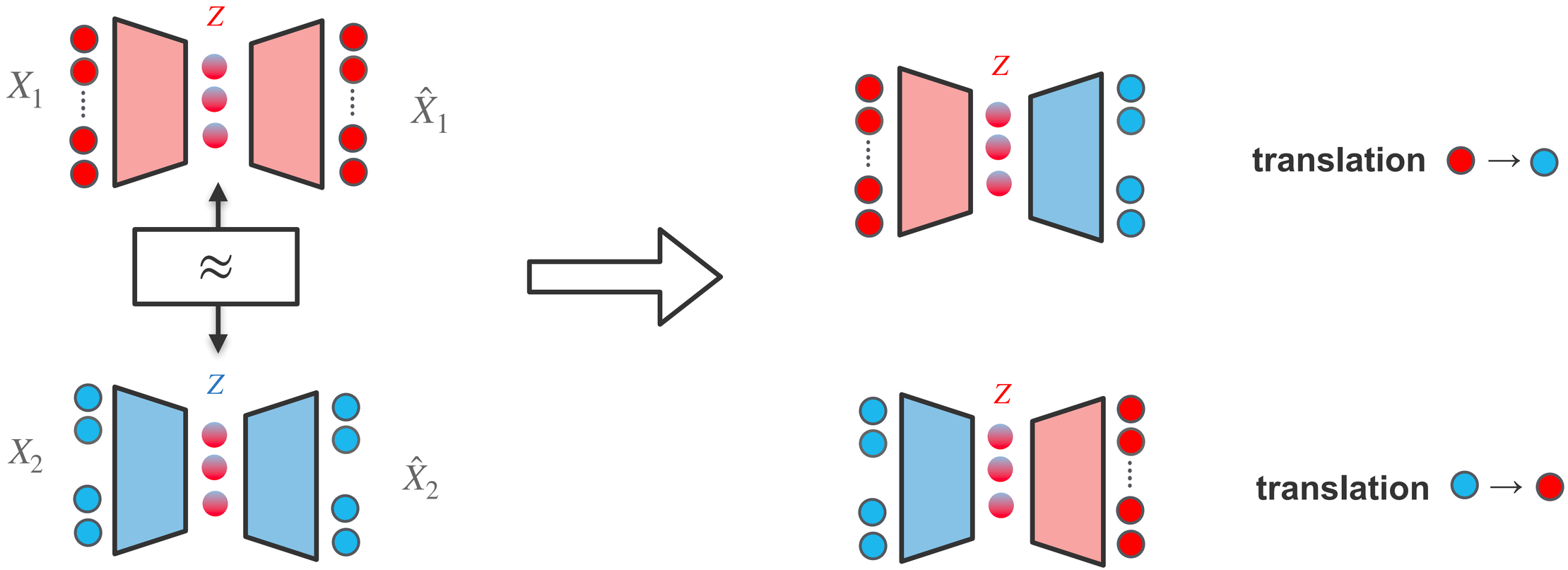
$$G^* = \arg\min_G \max_d v(\theta_d, \theta_g)$$

Co-embeddings



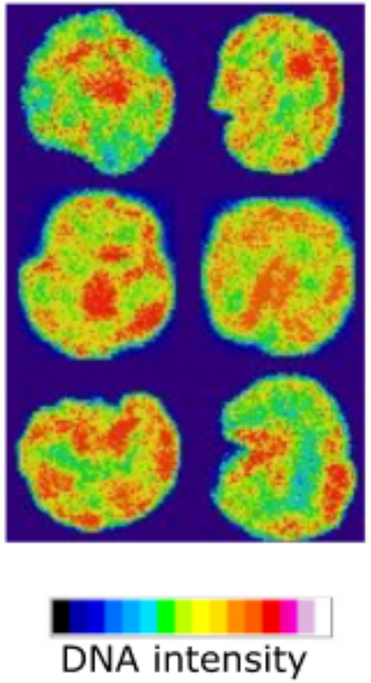
- **Generators** are here the **encoders**
- Generate latent spaces Z_1, Z_2
- Model is trained until the discriminator can no longer discriminate between Z_1, Z_2

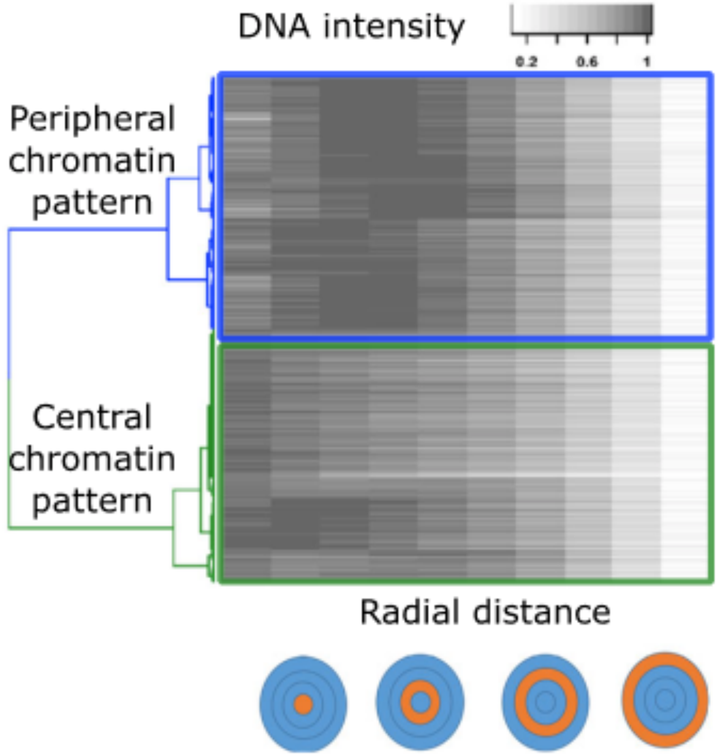
Multi-domain translation



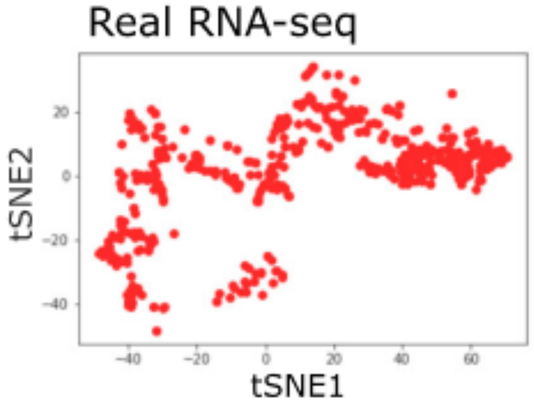
Translating data modalities

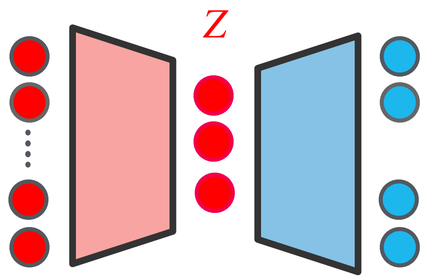
- DAPI staining of cell nuclei reveals chromatin density within the nucleus
- Identification of 2 distinct sub-populations
  - central
  - peripheral

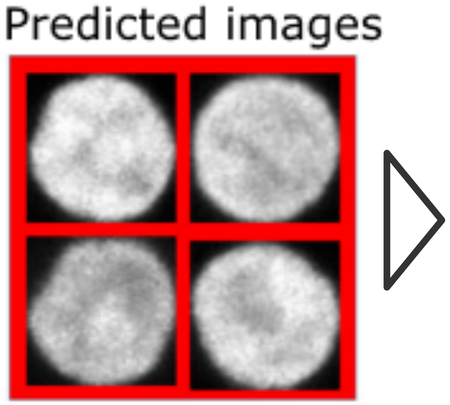


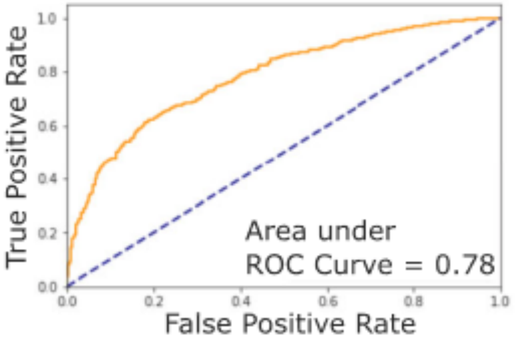


Translating data modalities

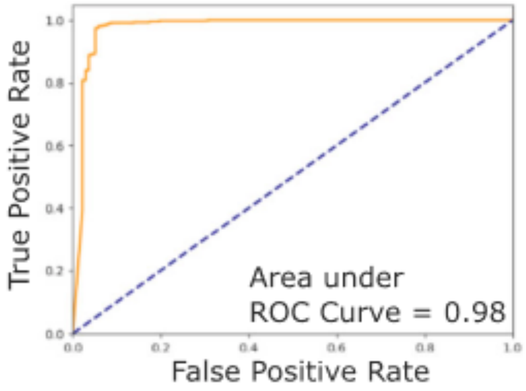


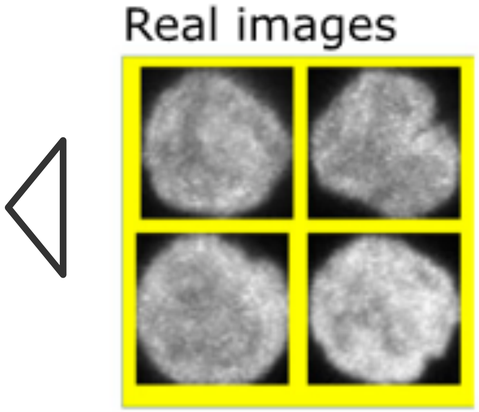


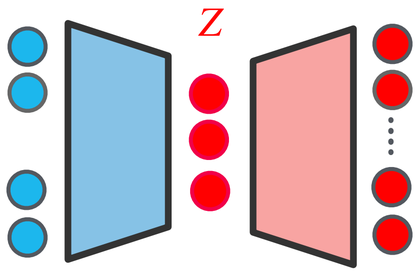


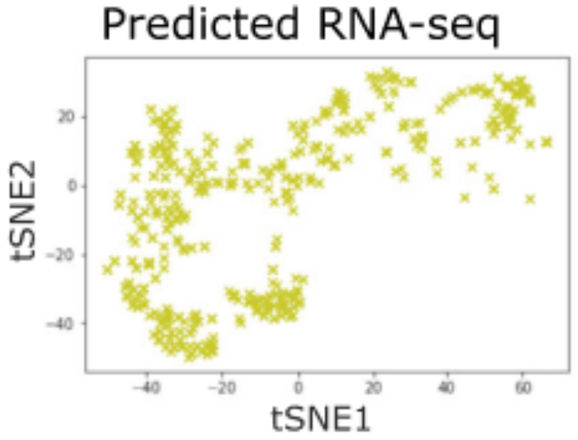


Classification peripheral/central



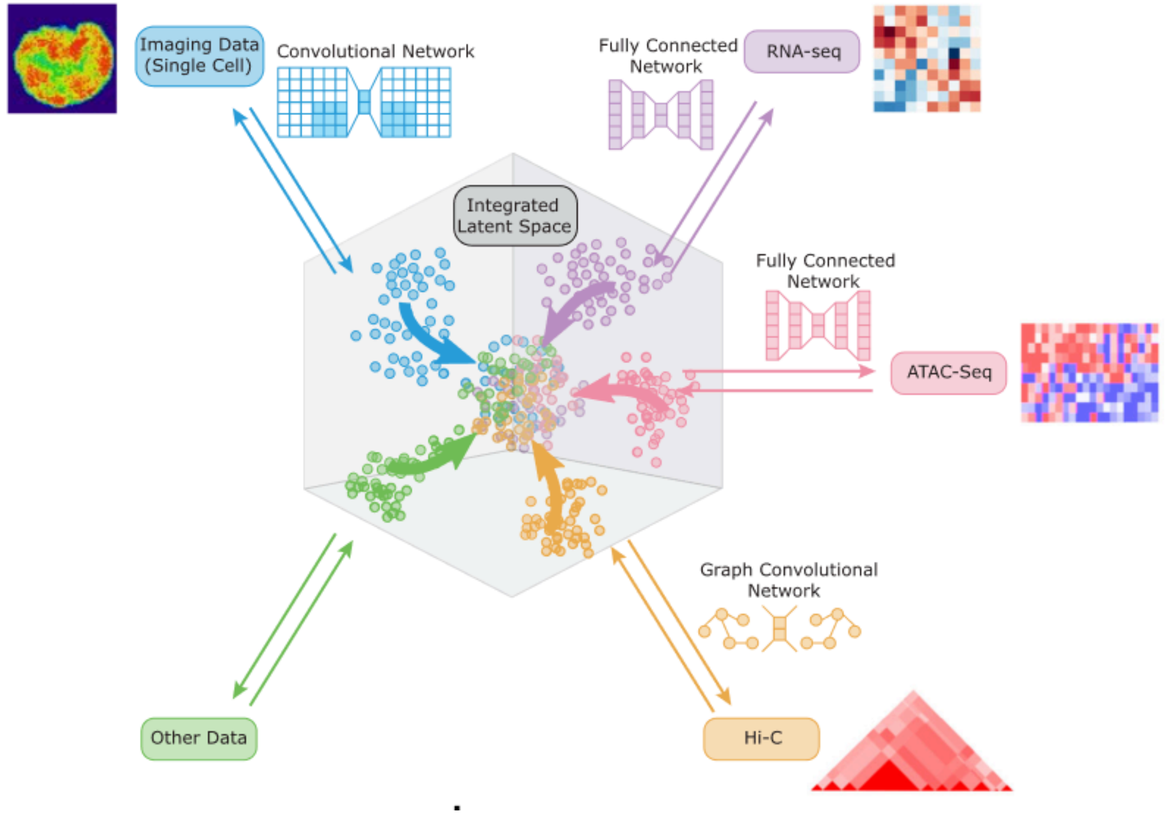






Limitations

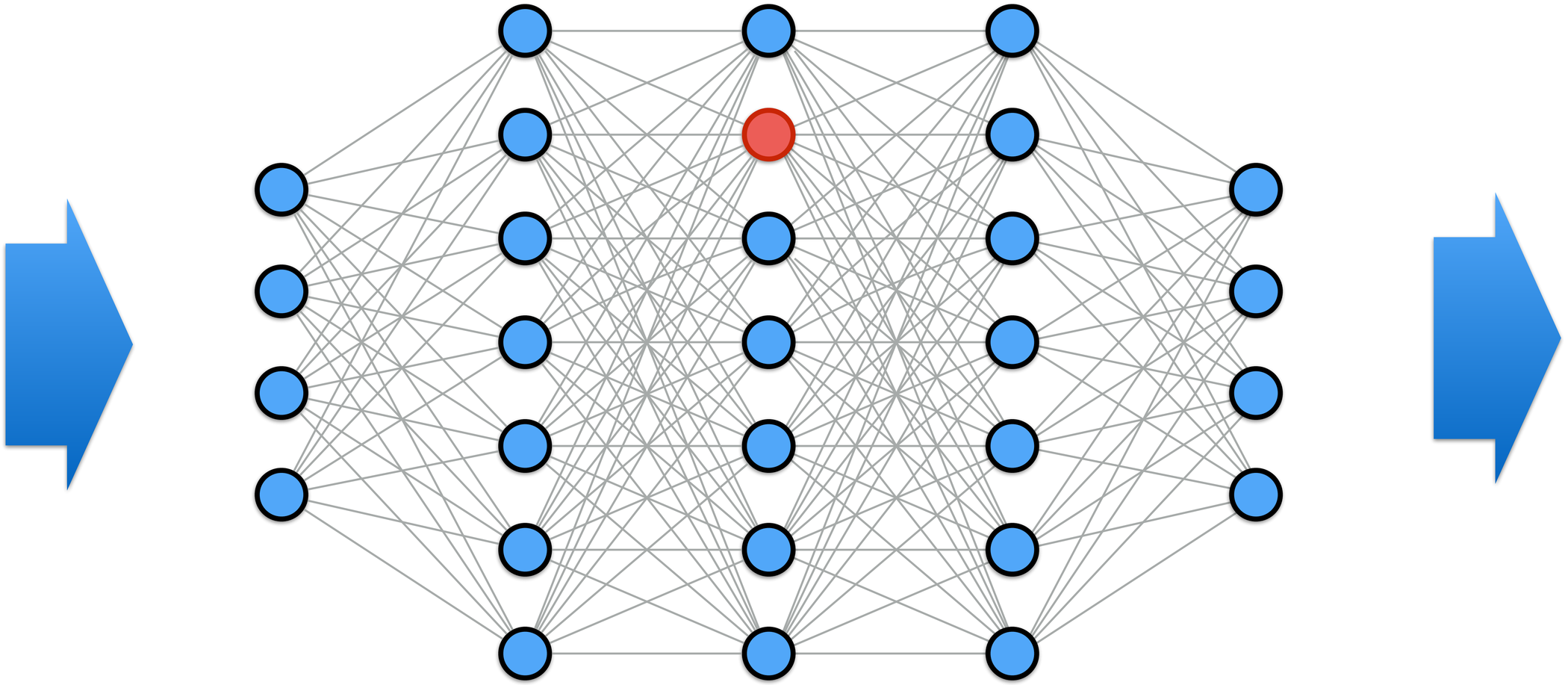
- Dimensions of the latent space cannot be biologically interpreted!
- "Latent space" engineering is not interpretable



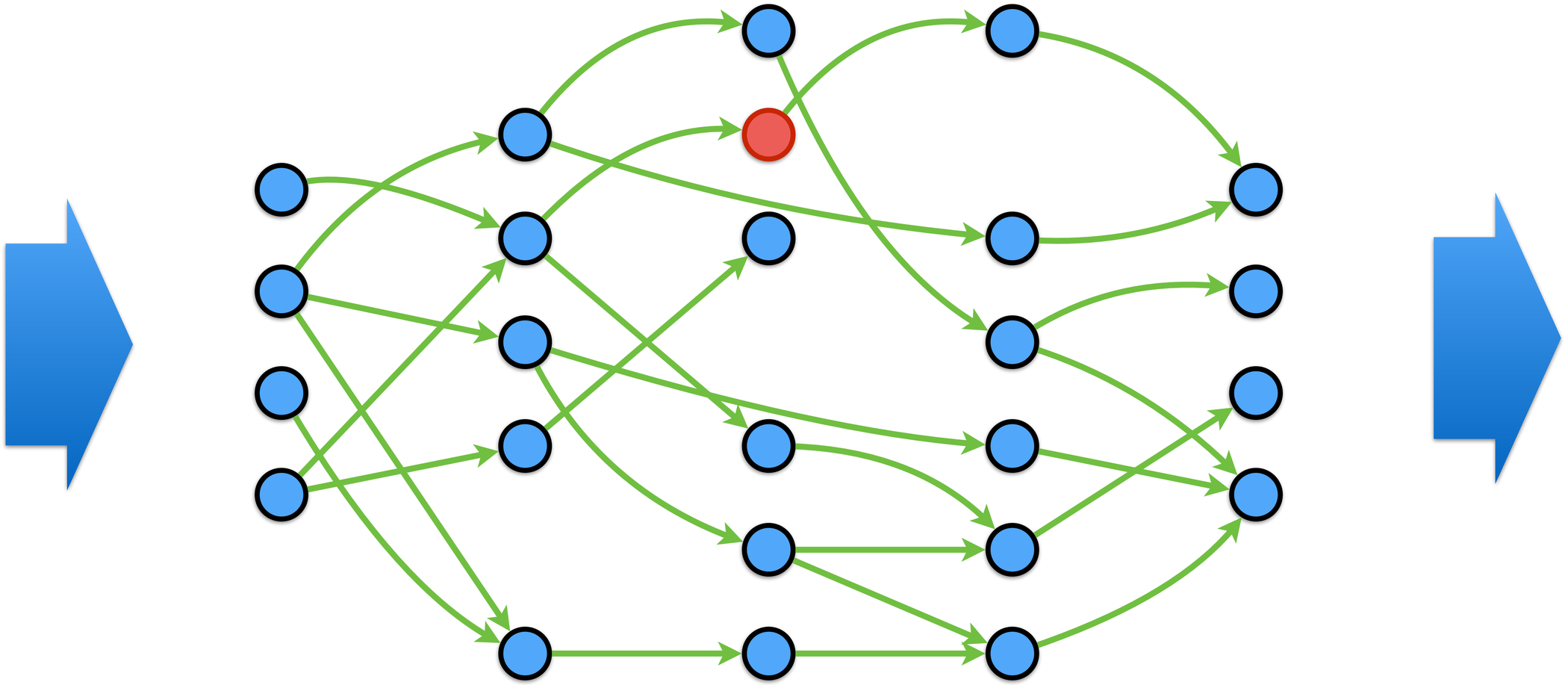
3 - Interpretable VAEs

Interpretable deep neural networks

DeepLearning inLifeSciences



Interpretable deep neural networks

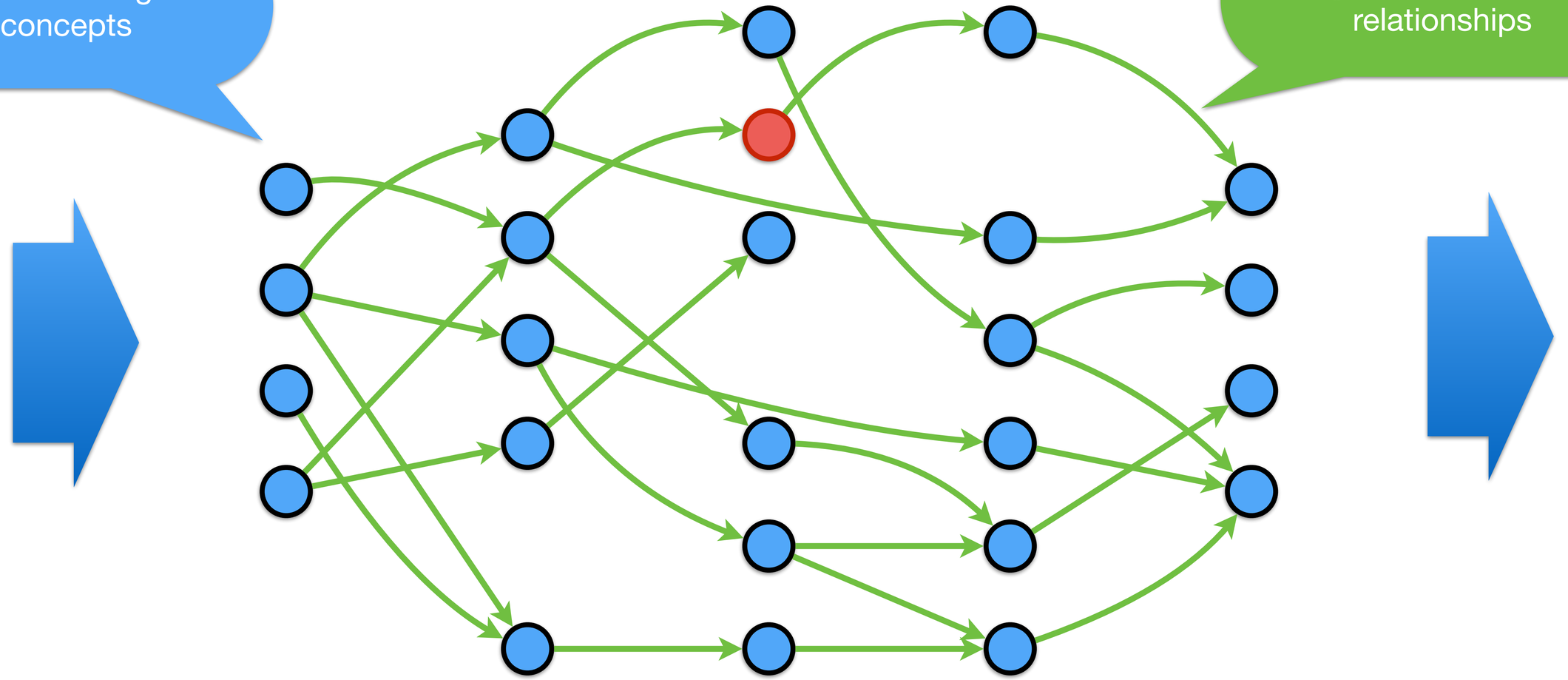


Interpretable deep neural networks

DeepLearning inLifeSciences

Nodes = biological concepts

Edges = interpretable relationships



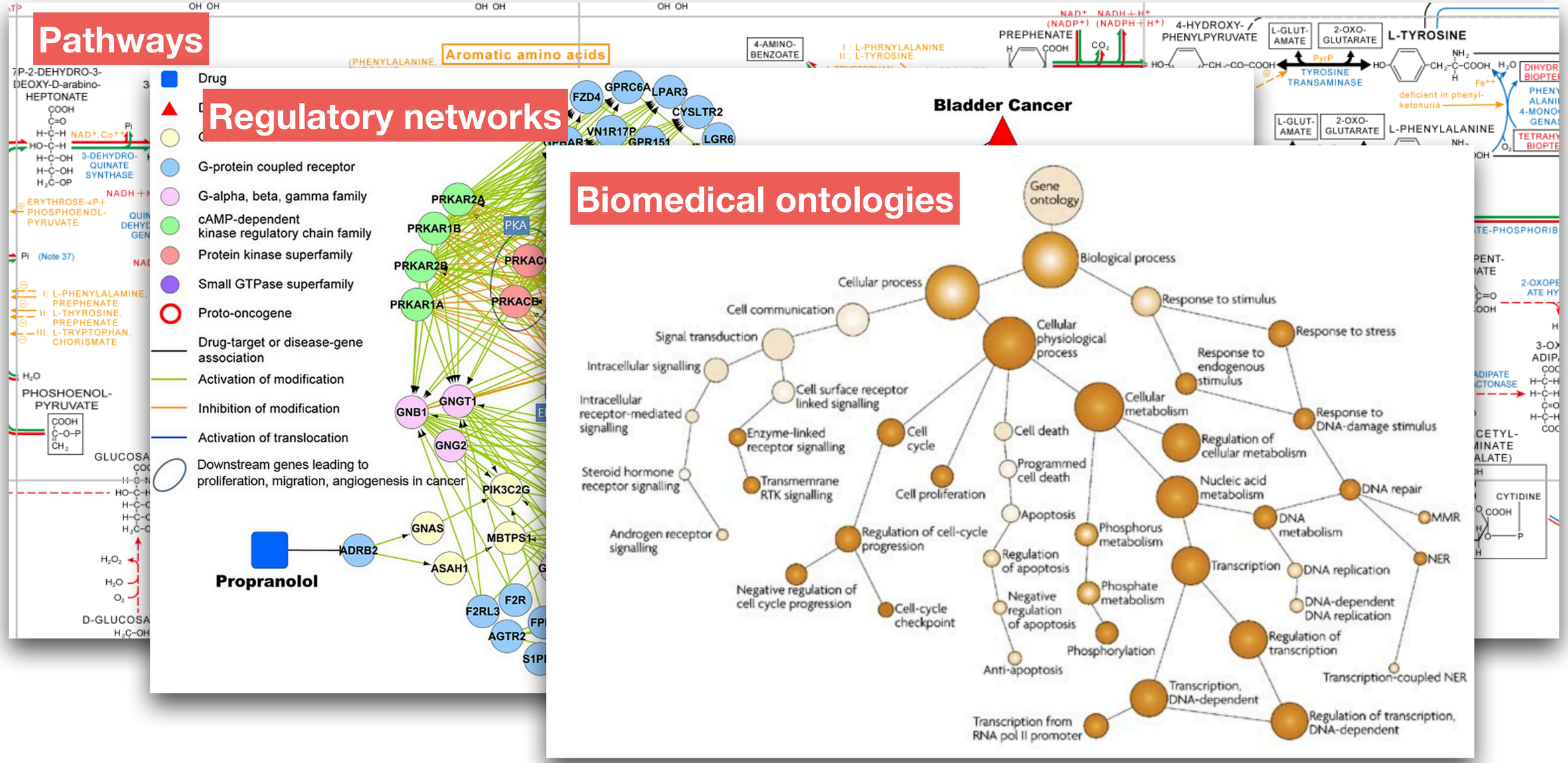
Biological networks

DeepLearning inLifeSciences

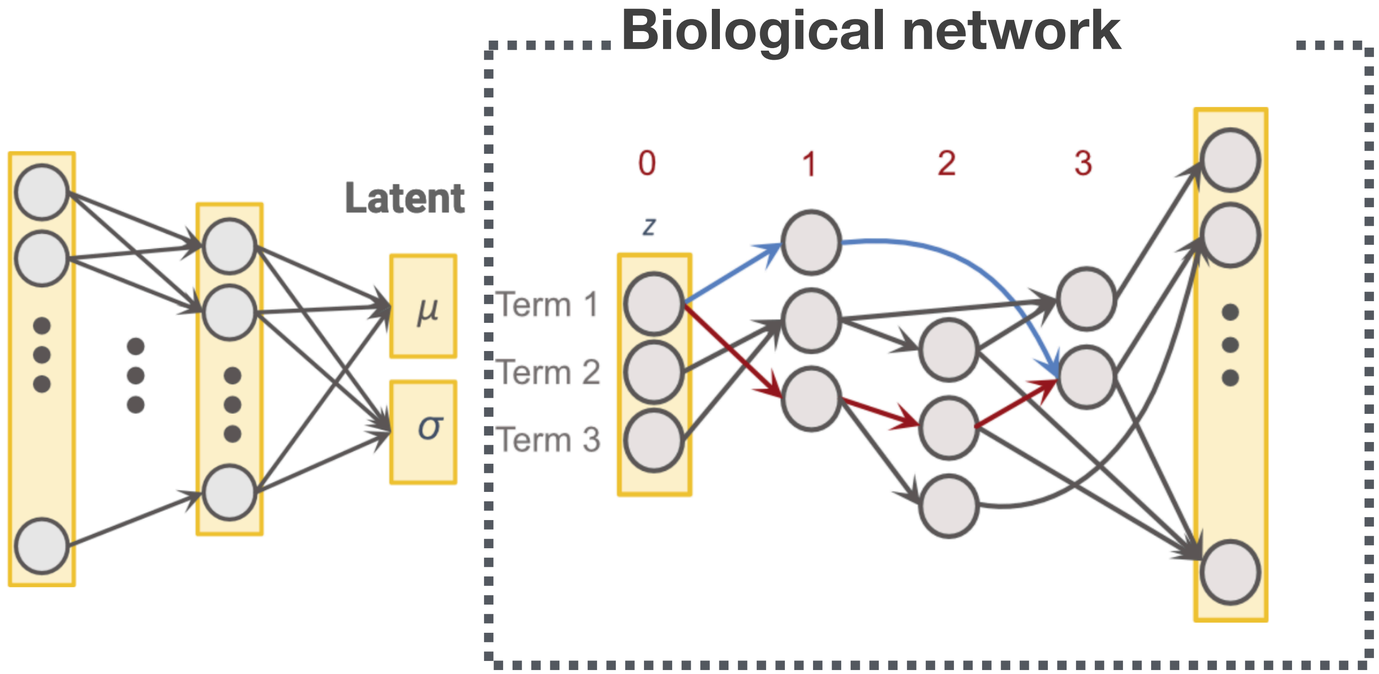
Pathways

Regulatory networks

Biomedical ontologies



Interpretable VAE

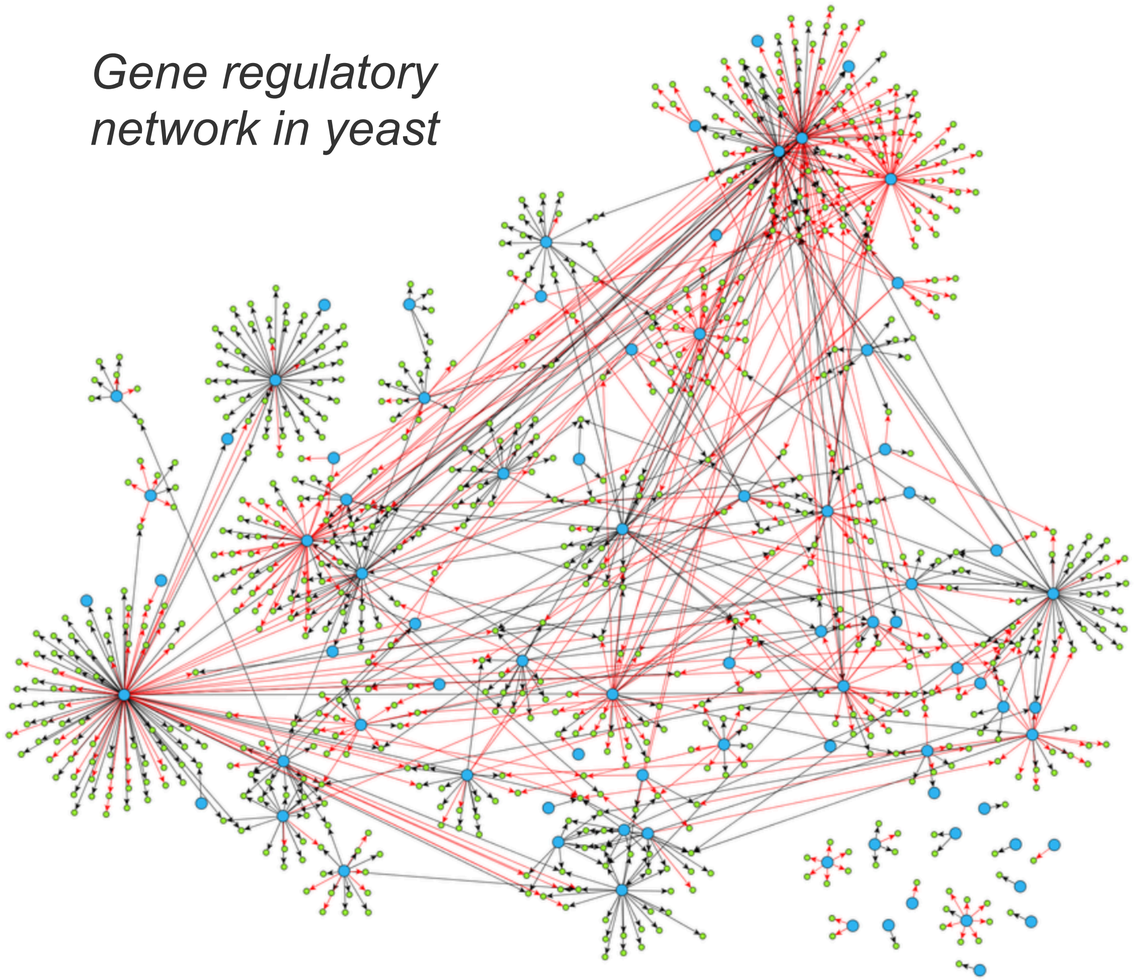


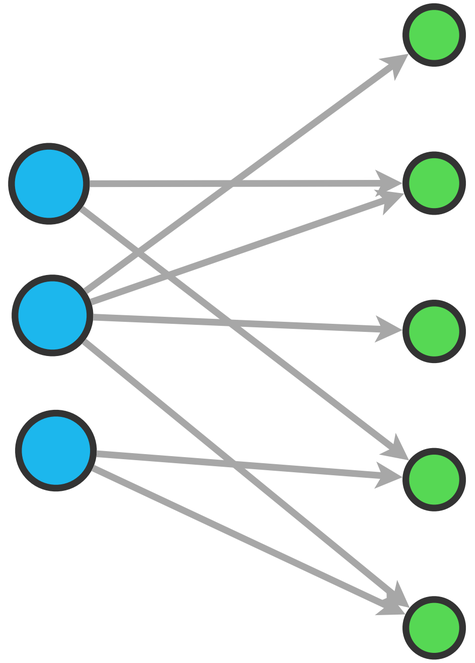
- Replace the decoder by a **biologically informed network**
- **Node activations** can be intepreted biologically as each node corresponds to a biological entity!
- **Edge weights** represent the strength of the influence between nodes

Gene regulatory network

DeepLearning inLifeSciences

Gene regulatory network in yeast

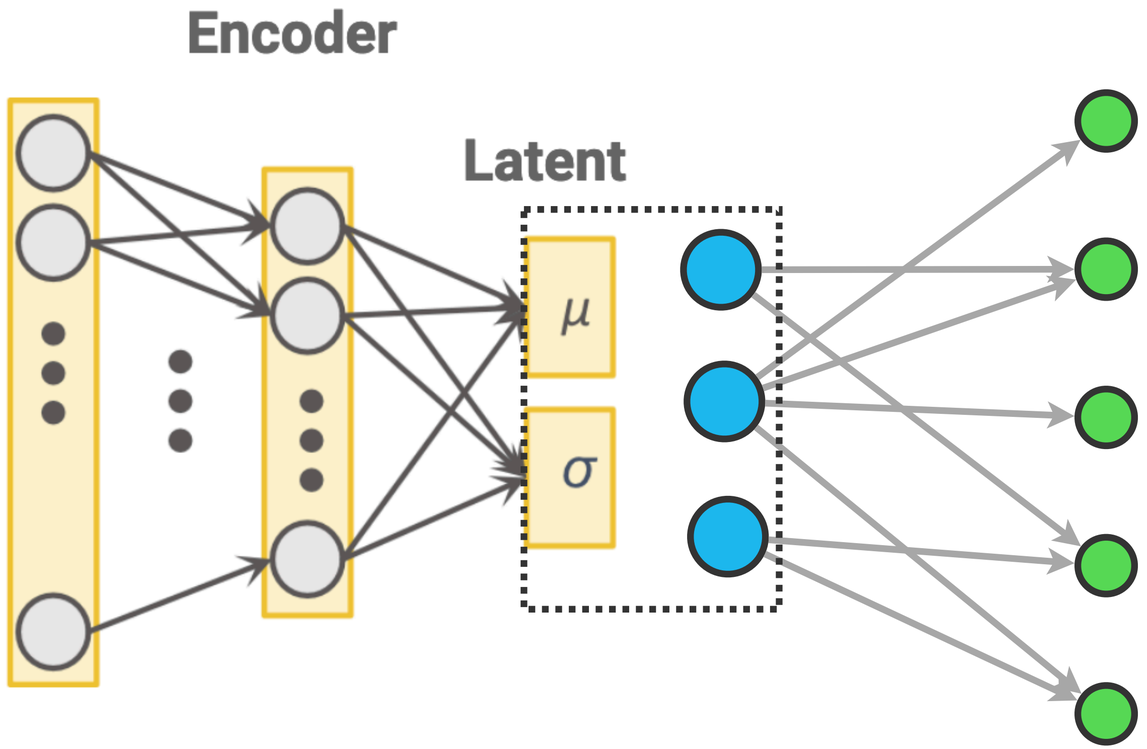




Transcription factors

Target genes

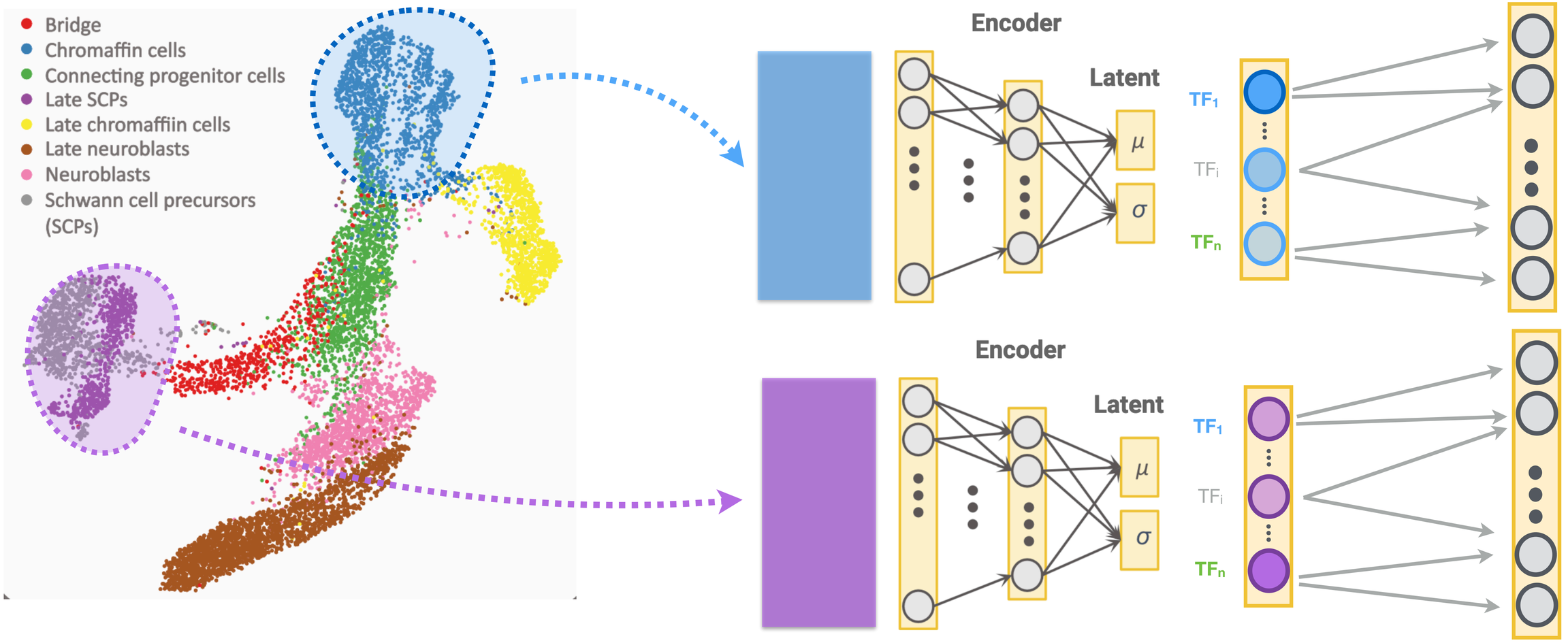
Interpretable deep-learning models: the VEGA model



- Latent space: **transcription factors**
- Decoder structure: **GRN** learned from prior knowledge or data
- **Activation values of latent nodes = biological activity of TFs**

[Seninge et al. (2021)]

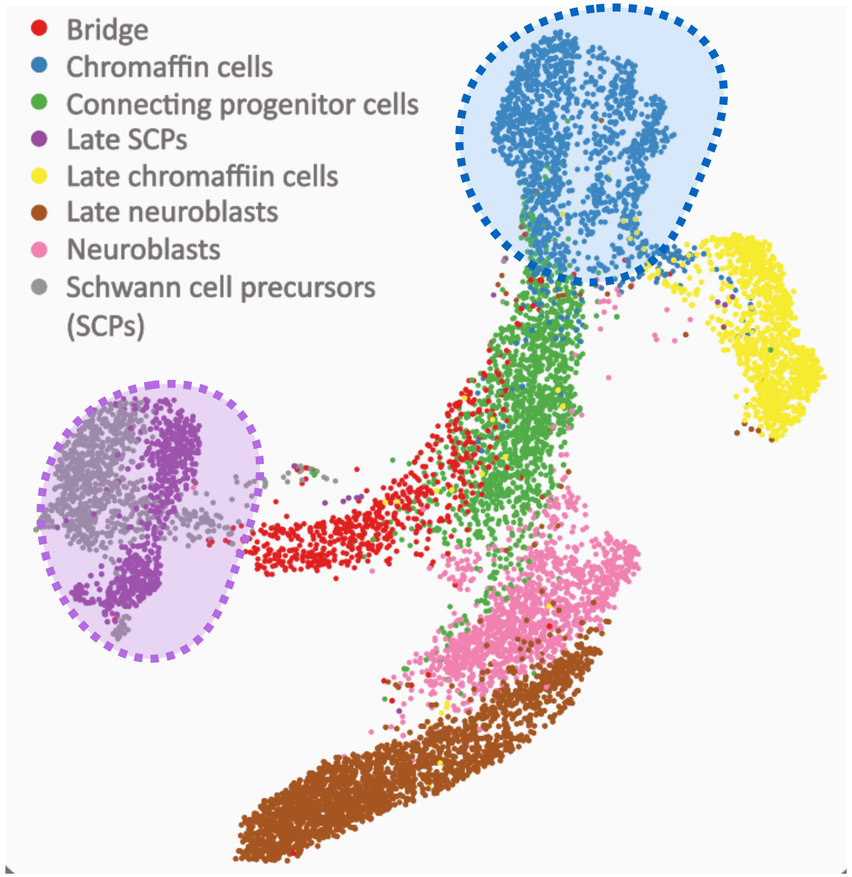
Interpretable deep-learning models

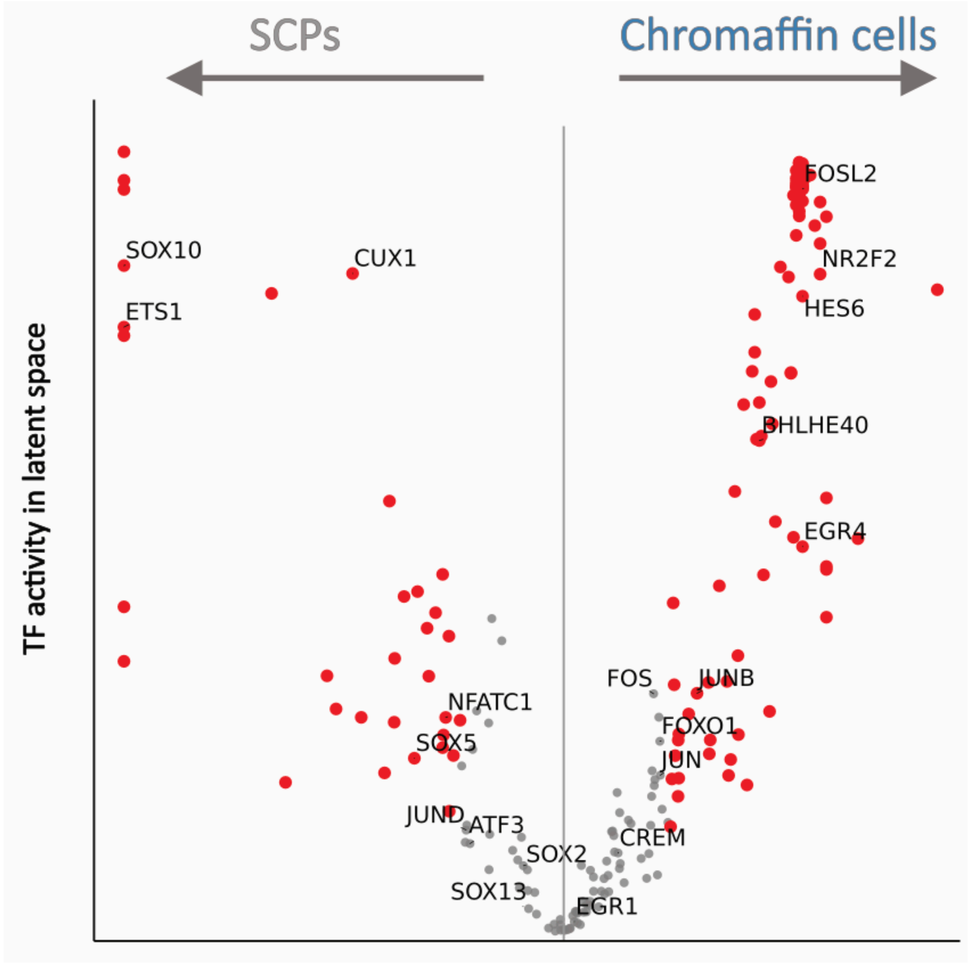


[Jansky et al. (2021)]

[Qian-Wu Liao; Anna von Bachmann]

Interpretable deep-learning models





[Jansky et al. (2021)]

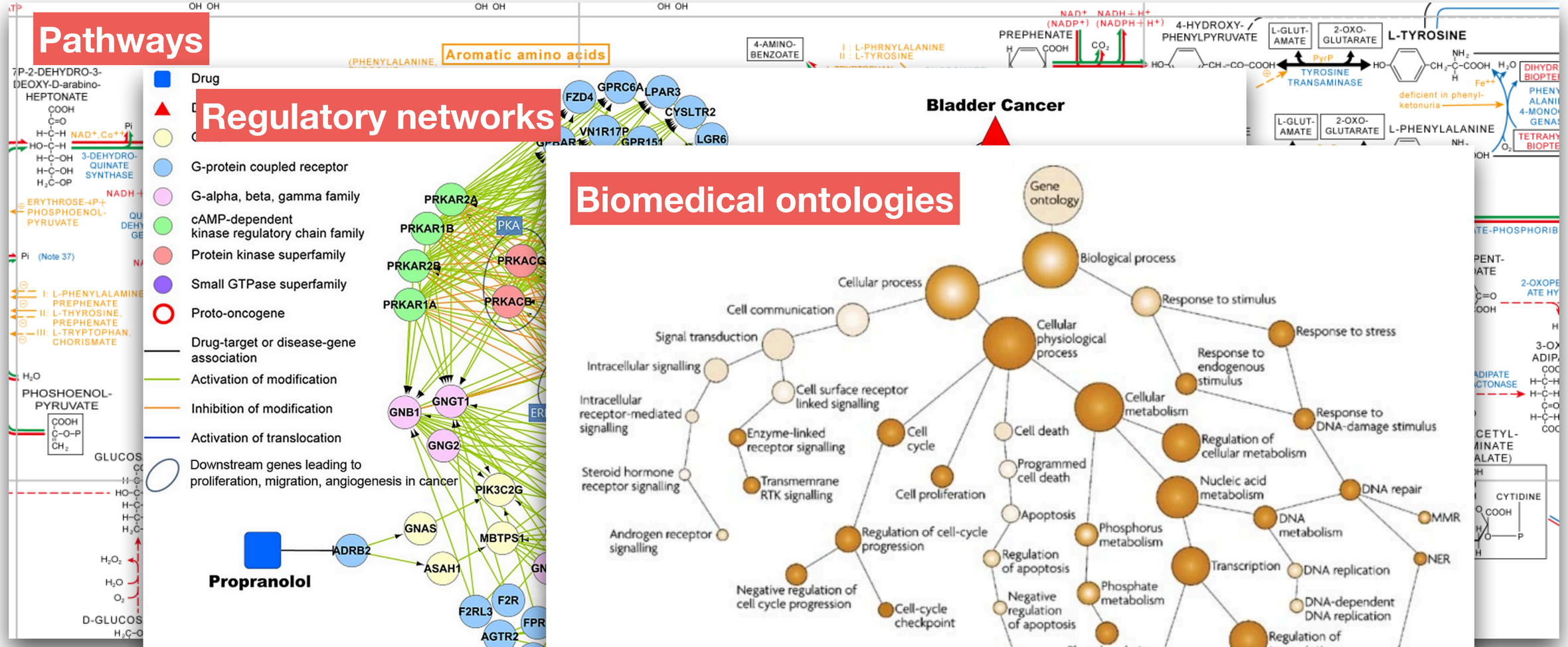
[Qian-Wu Liao; Anna von Bachmann]

logFC Bayes factor $\log(\frac{\bullet}{\bullet})$

Biological networks

DeepLearning inLifeSciences

Pathways



Regulatory networks

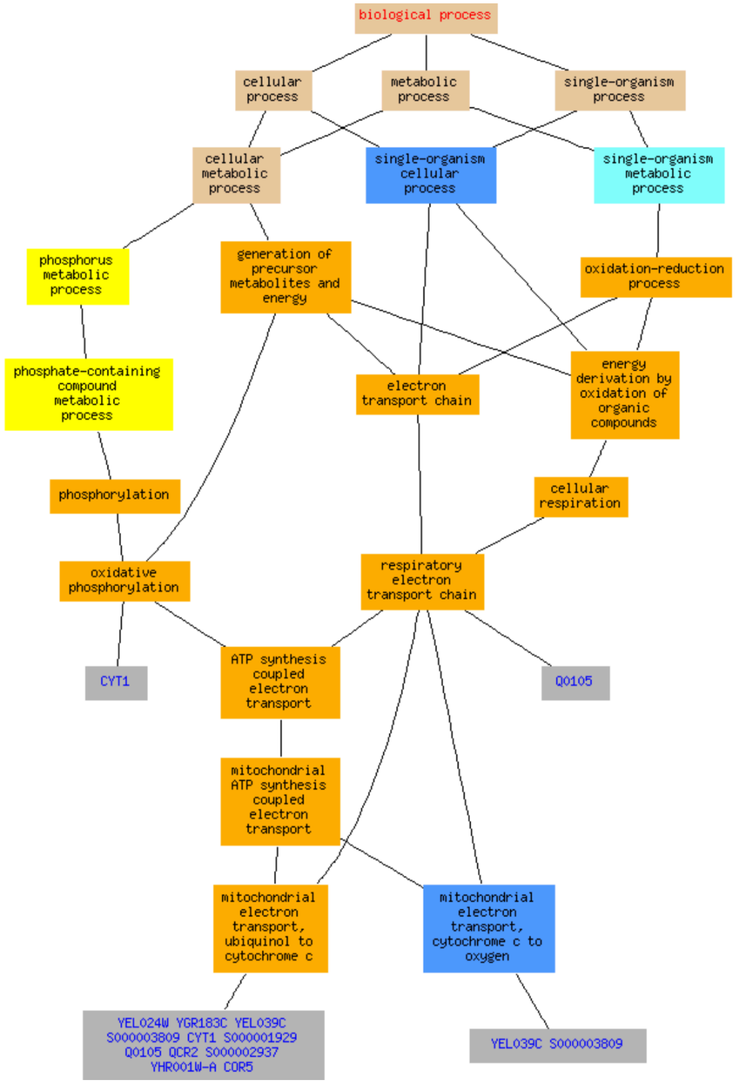
Biomedical ontologies

Biomedical ontologies

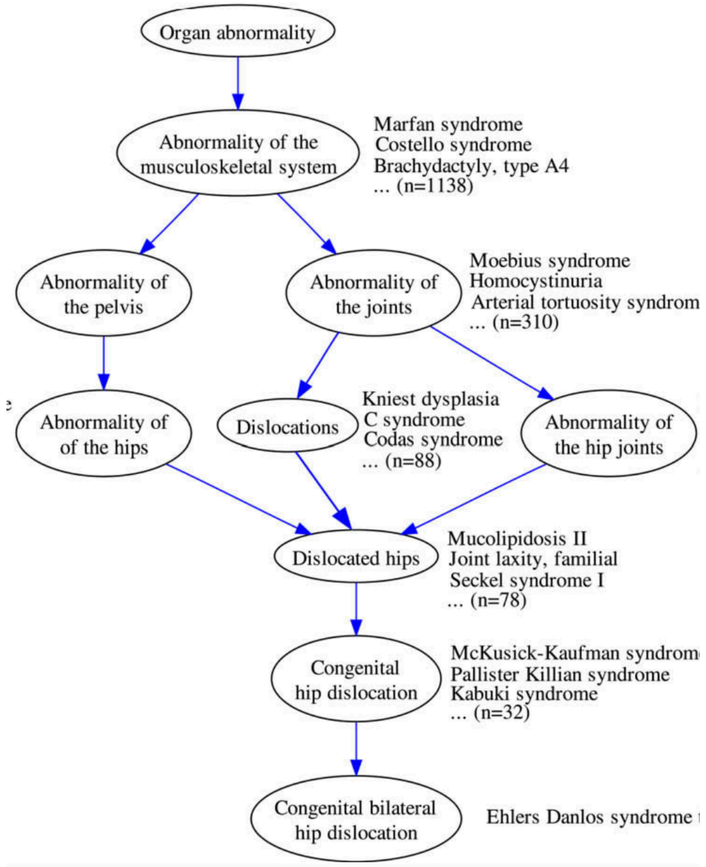
Ontologies are hierarchical vocabularies to describe biological processes

Genes are annotated to specific processes ("nodes") of the ontology

Gene ontology



Phenotype ontology

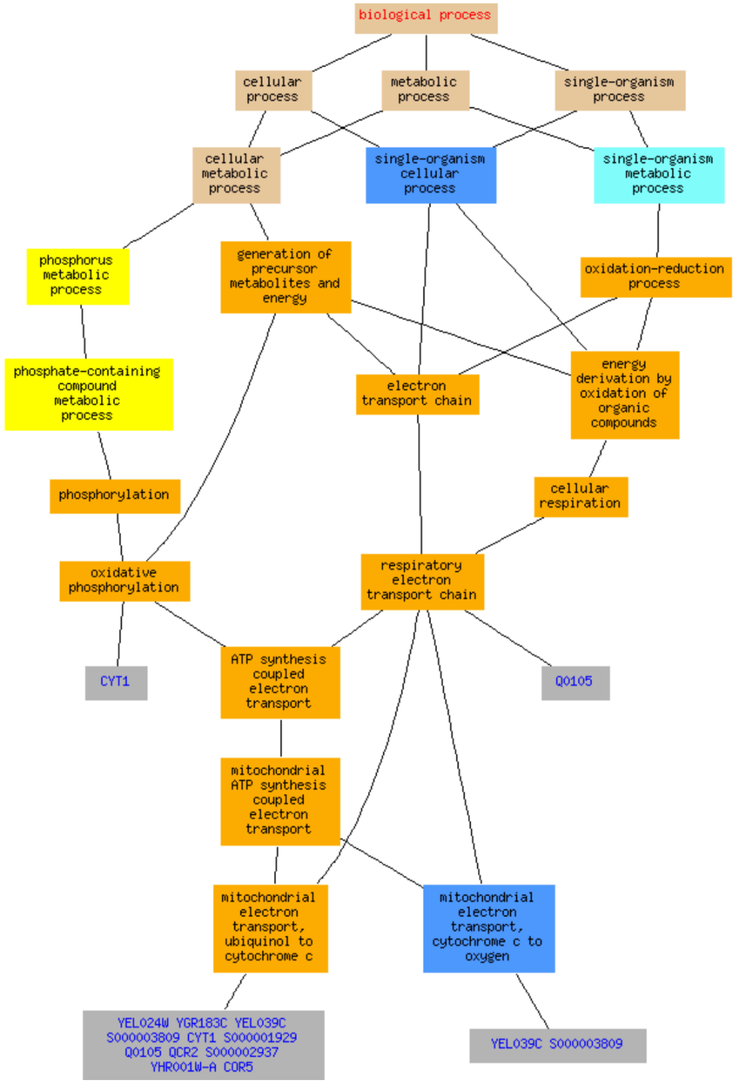


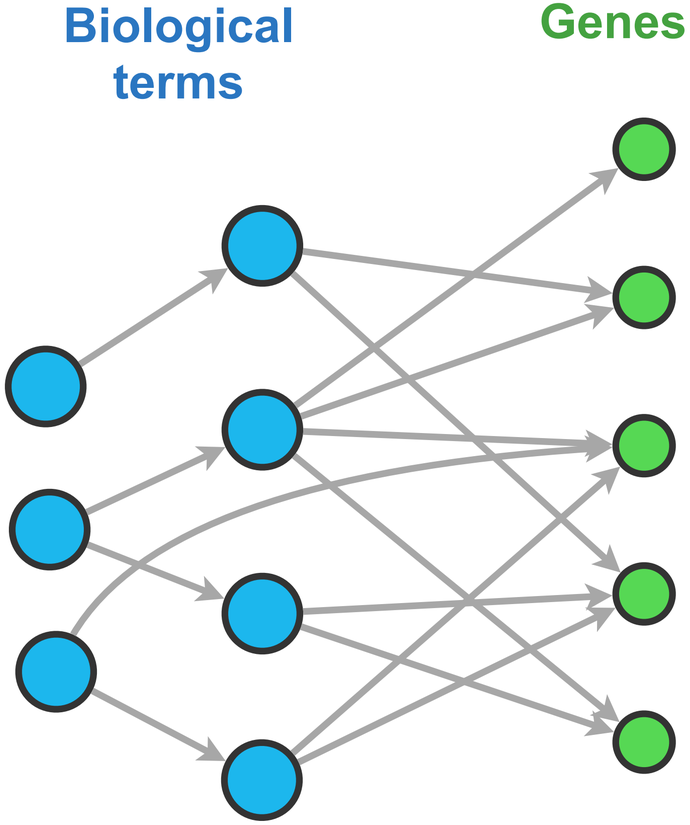
Biomedical ontologies

Ontologies are hierarchical vocabularies to describe biological processes

Genes are annotated to specific processes ("nodes") of the ontology

Gene ontology

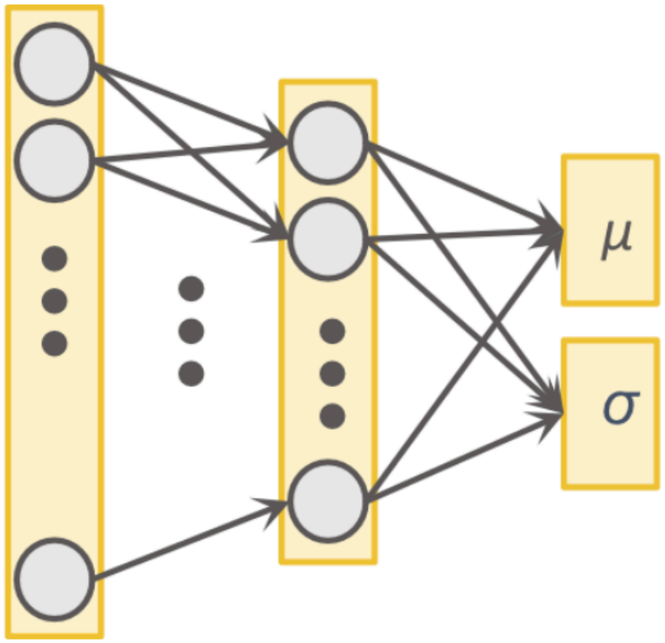


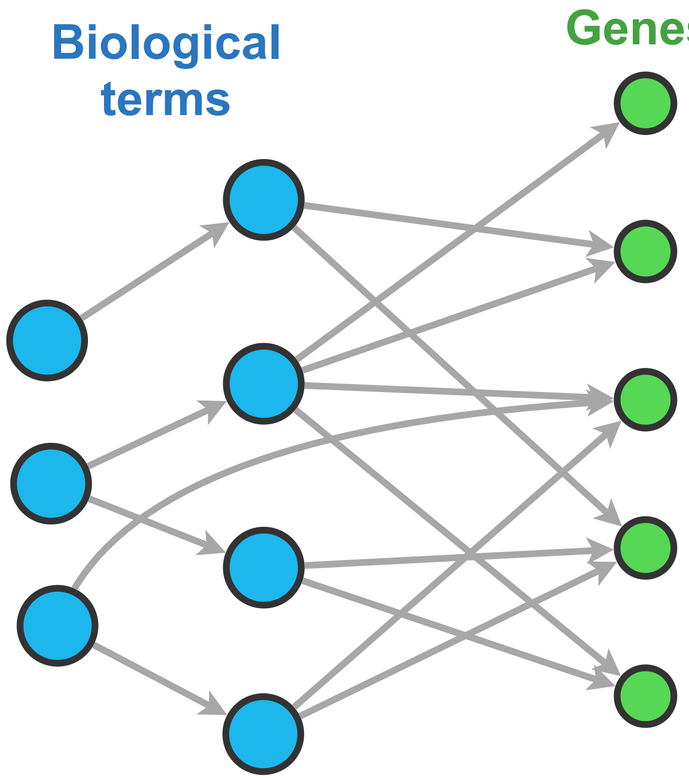


Biomedical ontologies

DeepLearning inLifeSciences

OntoVAE





GENEONTOLOGY Unifying Biology

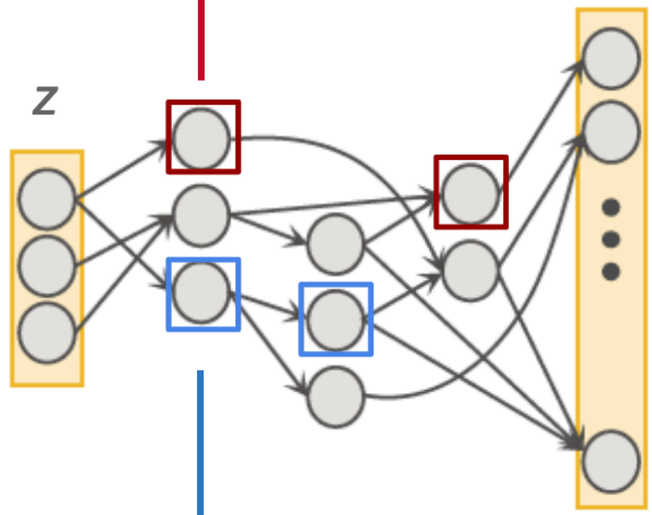
human phenotype ontology

- Each decoder node corresponds to a biological term of the ontology
- **Node activation = Biological activity**

[Doncevic & Herrmann, Bioinformatics (2023)]

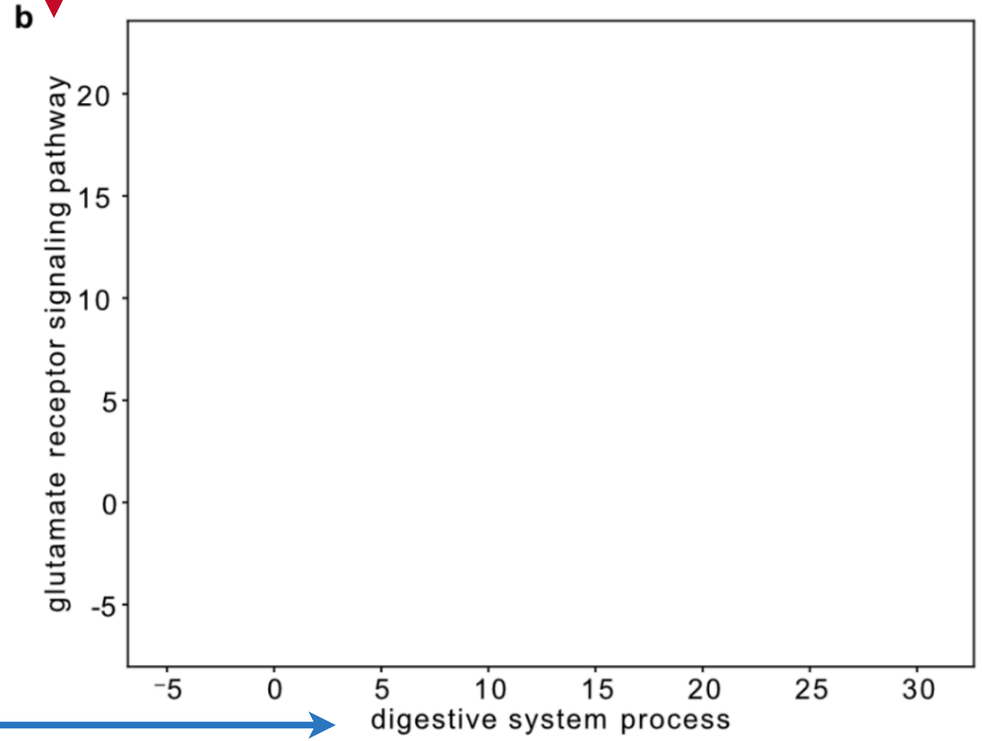
Tissue expression data

RNA-seq data of human tissues



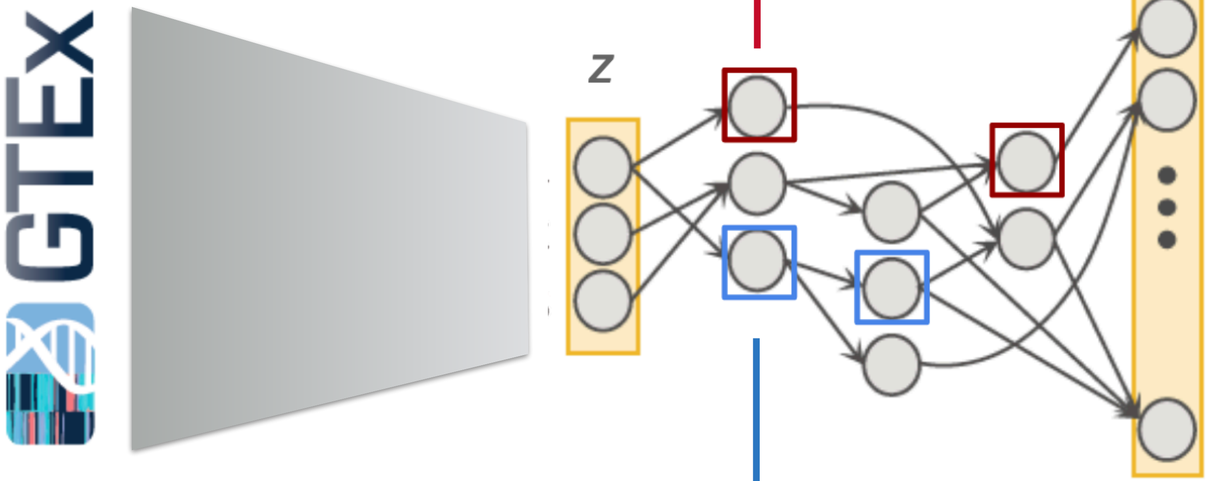
Glutamate receptor signaling pathway

Digestive system process



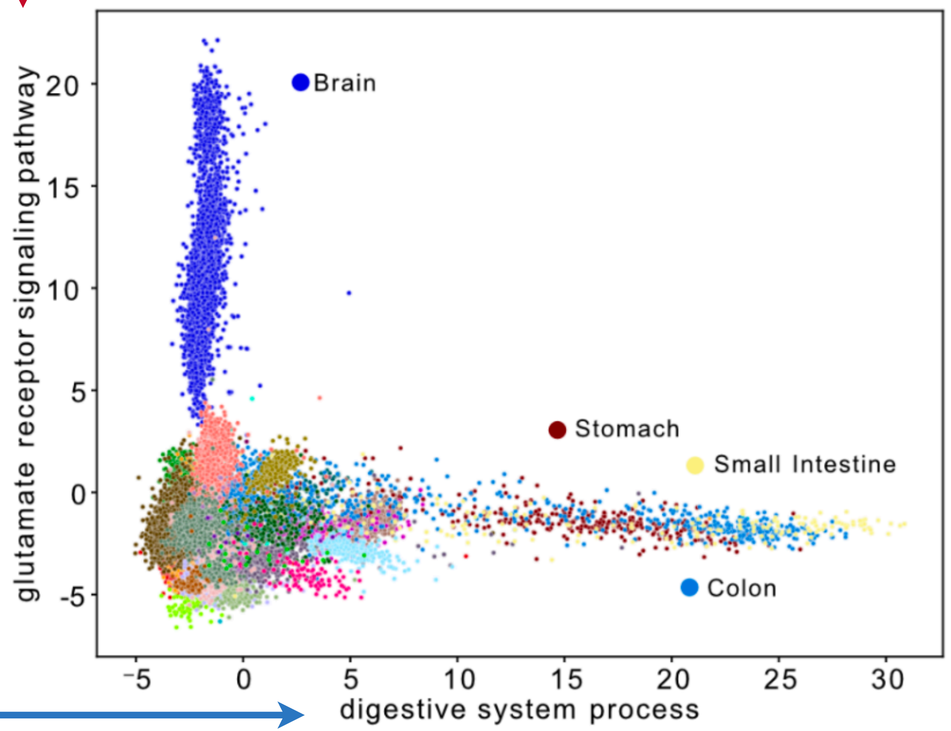
Tissue expression data

RNA-seq data of human tissues



Glutamate receptor signaling pathway

Digestive system process

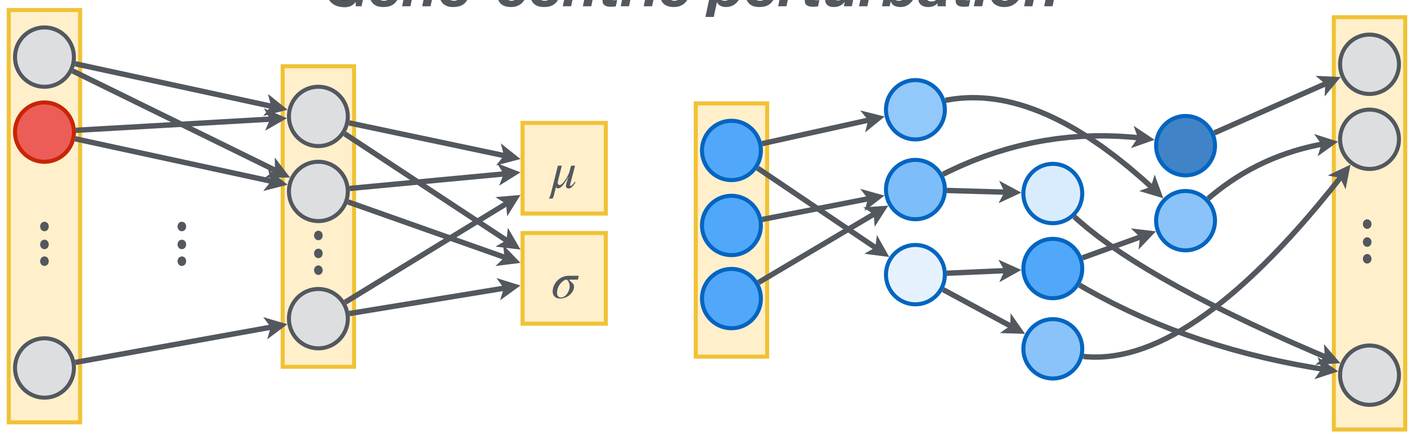


Modeling perturbations

DeepLearning inLifeSciences

Gene-centric perturbation

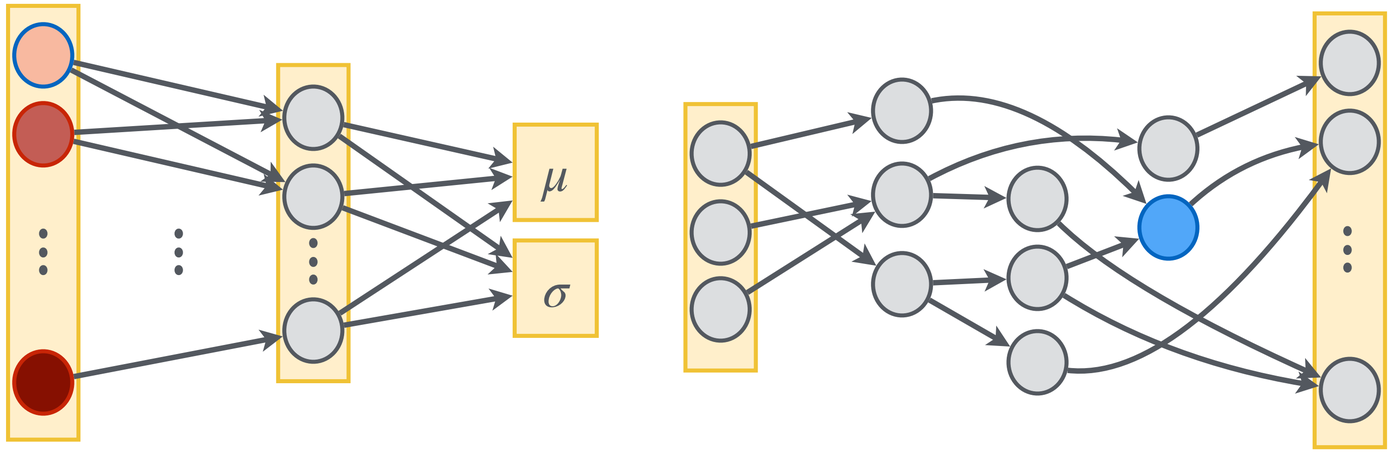
Perturb specific gene



which processes are most affected?

Process-centric perturbation

which genes have the greatest impact?



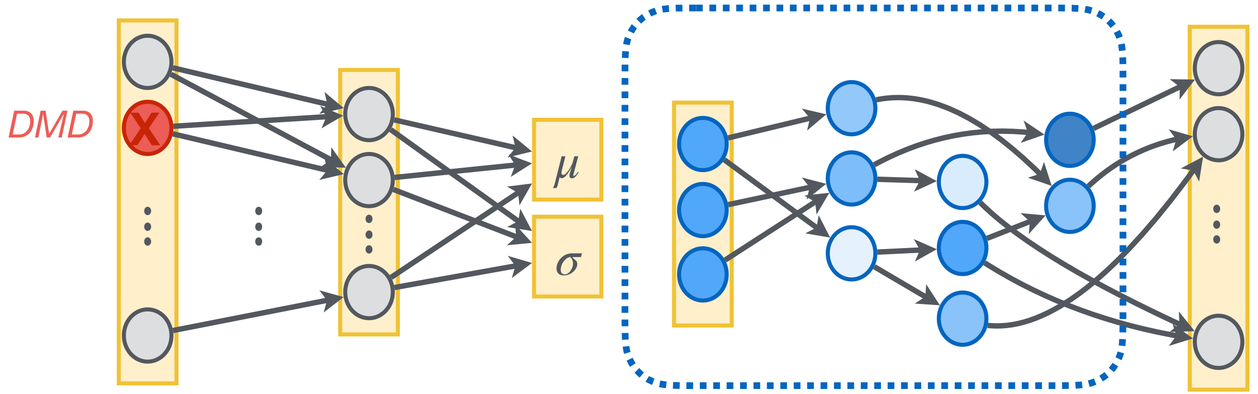
Gene-centric: genetic disease

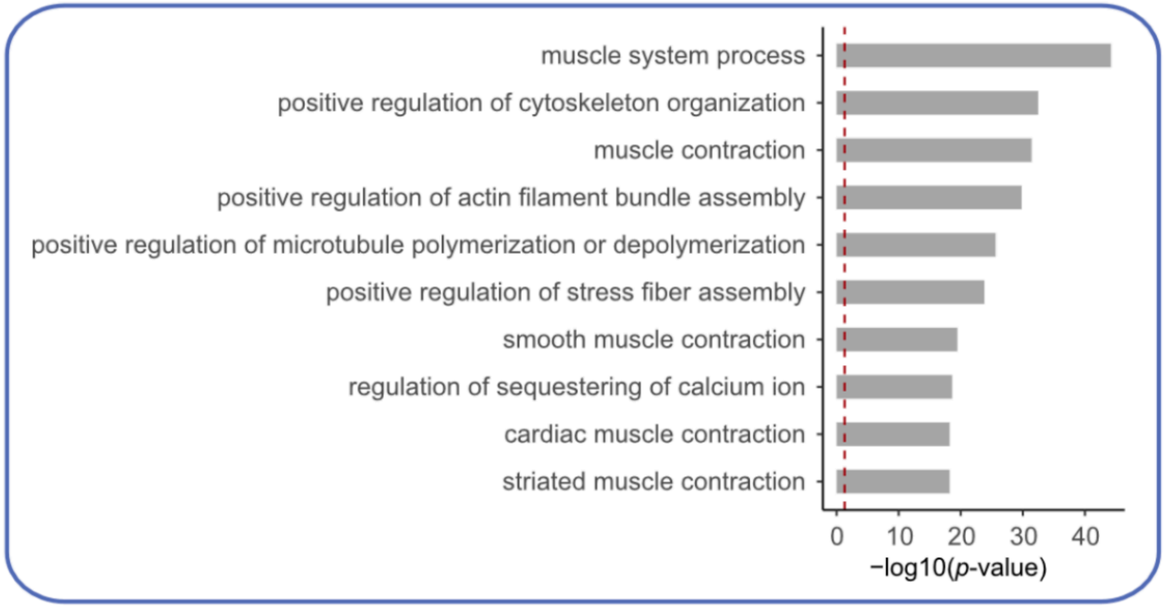
DeepLearning inLifeSciences

- **Duchenne muscular dystrophy** is is a genetic disorder leading to progressive muscle degeneration
- Monogenic disorder due to alterations in the dystrophin (*DMD*) gene
- **Simulate knockout** of *DMD* gene by comparing WT to in-silico knockout (in muscle samples)

  - *Model trained on all tissue samples*
  - *Knockout in muscle samples*
  - *Gene-ontology terms*

Examine differential activation of terms





Take-home messages

DeepLearning inLifeSciences

- VAEs are used in (single-cell) genomics to perform **various tasks** such as
  - dimensional reduction
  - data integration
  - obtaining interpretable/predictive models

- Requires to take into account the **specific distribution of the data** (espcially sparsity in single-cell genomics)

- Integrating **a-priori biological knowledge** (e.g. as prior network) can yield interesting insights into non-trivial connections between biological processes