

Systems biology

Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations

Daria Doncevic^{1,*} and Carl Herrmann  ^{1,*}

¹Health Data Science Unit and BioQuant, Medical Faculty Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

*Corresponding authors. Health Data Science Unit - Medical Faculty Heidelberg and BioQuant, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany.

E-mails: daria.doncevic@bioquant.uni-heidelberg.de (D.D.) and carl.herrmann@bioquant.uni-heidelberg.de (C.H.)

Associate Editor: Pier Luigi Martelli

BIBS seminar

2024. 4. 4.

Presenter: Jun Sik Kim

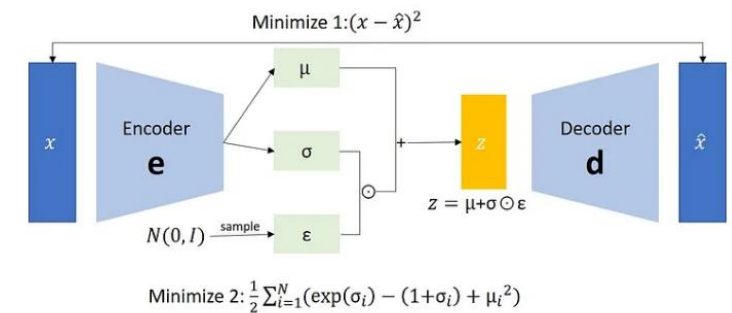
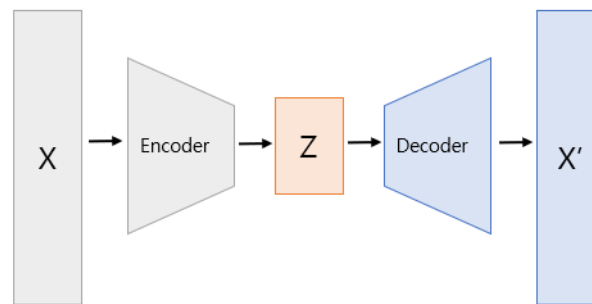
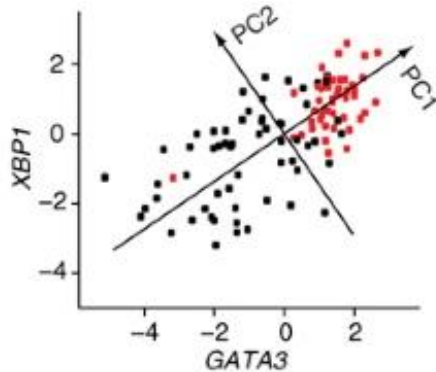
Table of Contents

- 1 Introduction
- 2 Materials & Methods
- 3 Results
- 4 Discussion

Introduction

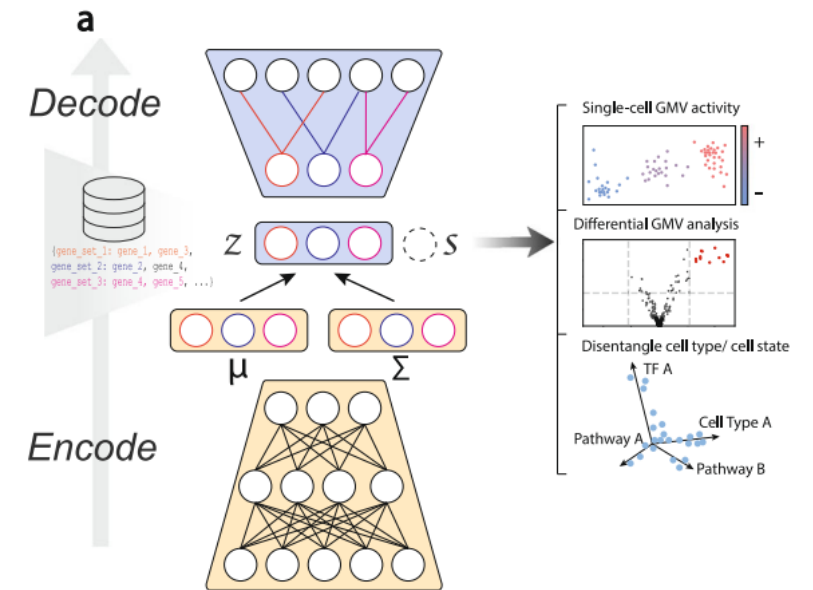
Introduction

- In recent years, deep learning (DL) has been widely used to analyze high-dimensional biological omics data, especially single-cell RNA sequencing.
- In contrast to linear methods, such as principal component analysis (PCA), non-linear models can capture more complex patterns in the data.
- One example of an unsupervised DL model that performs dimensionality reduction is the autoencoder (AE), which consists of 2 neural networks; an encoder and a decoder.
- A more recent variant of the AE is the variational autoencoder (VAE), which learns a probability distribution over the latent vectors of the data and thus belongs to the class of generative models.



- AE-based methods have successfully been applied in various ways:
 - cancer classification
 - data integration
 - data denoising
 - batch correction
 - cell clustering
 - multi-domain translation
 - prediction of drug treatment effects on single-cells
- However, in contrast to PCA, AE-based approaches **lack interpretability** as we cannot easily assign feature contributions to the latent vectors due to their non-linear nature.

- [Previous Methods 1] Different approaches have already been used to tackle the problem of limited interpretability by **modifying the neural network structure**.
- 1. **Tybal** tries to extract a biologically meaningful latent space by examining how different latent vectors separate covariates and then investigating the associated gene weights of the latent vectors of interest in the **one-layer decoder**.
- 2. In the **LDVAE** model, a **linear decoder** has been implemented to allow the assignment of feature weights to the different latent vectors.
- 3. In the **VEGA** model, the authors used a **one-layer, sparse decoder** that connects the latent variables to a set of annotated genes.
- One limitation of these models is the **simplicity of the structure**, which does not allow the incorporation of more complex, hierarchical biological information.



- [Previous Methods 2] Other methods have been aiming at **incorporating hierarchical biological networks** into a neural network.
 1. **DCell** is structured based on part of Gene Ontology (GO) to predict growth rates in yeast and the impact of double mutants on biological processes.
 2. **Gene Ontology Autoencoder (GOAE)** implements GO terms in one hidden layer of encoder and decoder by partial connectivity to the input and output layer.
 3. **Deep GONet** is a neural network classifier that imposes regularization on its weights to encourage the establishment of connections that mirror the GO-directed acyclic graph (DAG) and has been used to combine cancer classification with biological explanations.
- However, no attempts have yet been made to incorporate **full hierarchical biological networks** into a VAE to capture the different levels of description of biological processes in tasks that go beyond classification.

- **OntoVAE** (Ontology guided VAE) is a novel flexible VAE structure with a **multi-layer, sparse decoder** that allows for the incorporation of any kind of hierarchical biological information encoded as an ontology.
- OntoVAE provides **direct interpretability in its latent space and decoder**, as the activities of the neurons now correspond to activities of biological processes or phenotypes. This allows users to try different terms and monitor their activity changes, without the need to preselect specific processes.
- Importantly, OntoVAE can be used for predictive modeling. By OntoVAE we can modulate the values of input features in silico and then monitor how these changes propagate through the network. This allows to simulate the effects of drug treatment or genetic alterations.
- The investigation of subsequent alterations in the activation of hidden nodes representing processes or phenotypes allows to resolve complex genotype-phenotype relationships.

Materials & Methods

Variational Autoencoders

- A Variational Autoencoder(VAE) is a deep generative model which can learn meaningful data representations from high-dimensional input data.
- VAE can encode a particular distribution. After the encoding phase, there is a sampling phase in which we sample points from the distribution $q_{\phi}(z|x)$.
 - Encoding function q : $q_{\phi}(z|x)$ – variational distribution(encoding distribution)
 - Decoding function p : $p_{\theta}(x|z)$ – posterior
- Traditionally, the distributions in the VAE architecture are supposed to be Gaussian: the encoder function will learn the two-parameter vectors μ, σ that are used to generate samples in $q_{\phi}(z|x)$ using the reparameterization trick:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Variational Autoencoders

- The loss function for this architecture can be written as the sum of 2 distinct losses:
 1. Reconstruction loss

$$\mathcal{L}_{recon} = \mathbb{E}_{q_{\phi}(z|x)}[p_{\theta}(x|z)]$$

→ Interpretation: conditional entropy of x given z , quantifies the uncertainty one has over the joint distribution (x, z) , knowing z

2. Regularization loss: Kullback–Leibler(KL) divergence

$$\mathcal{L}_{reg} = D_{KL}(q_{\phi}(z|x) || p_{\theta}(z)) = \mathbb{E}_{q_{\phi}} \left[\log \frac{p_{\theta}(z)}{q_{\phi}(z|x)} \right]$$

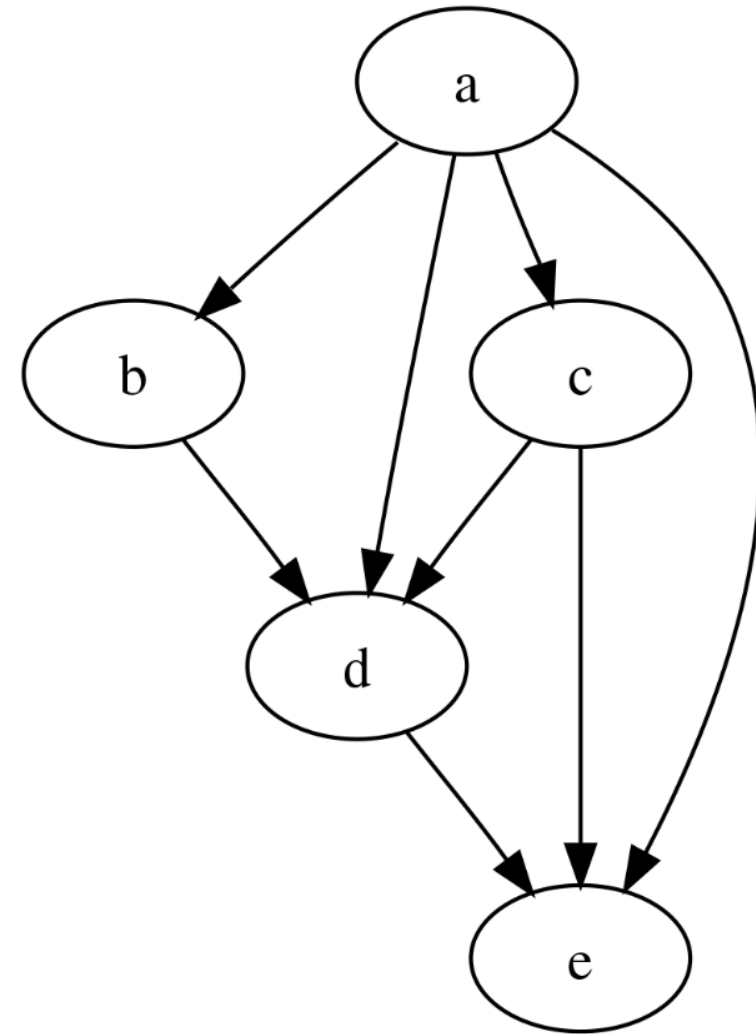
→ Interpretation: measurement of difference between the variational distribution and the prior distribution

- Total loss is defined with a hyperparameter β as:

$$\mathcal{L} = \mathcal{L}_{recon} + \beta \mathcal{L}_{reg}$$

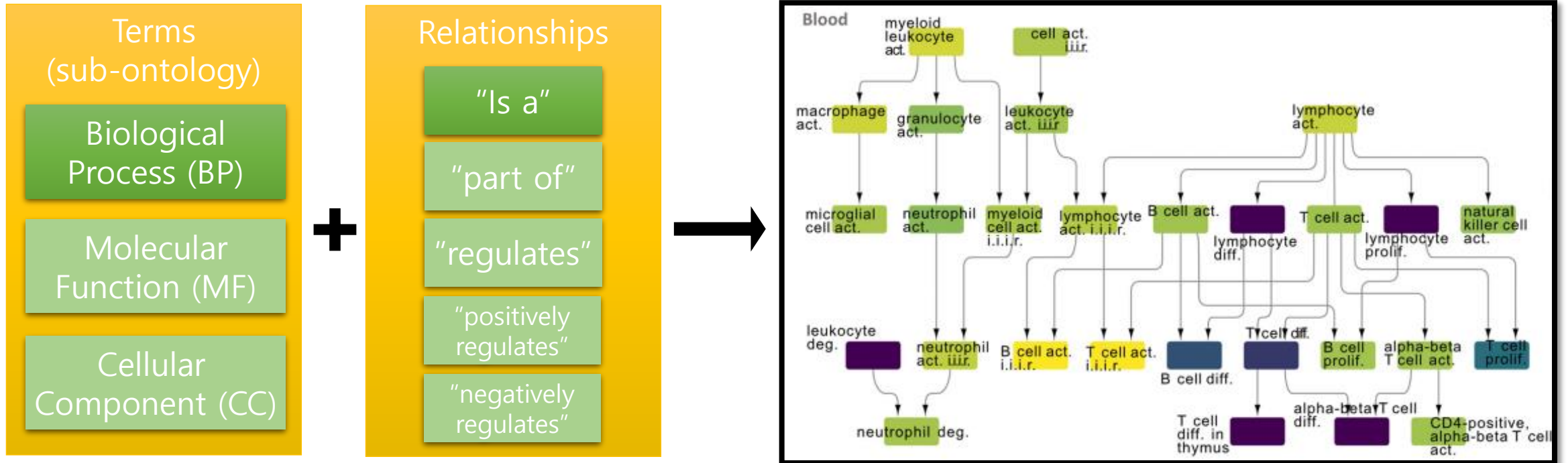
Directed Acyclic Graph (DAG)

- Graph with directions, but no directed cycles.
- If a graph is (1) directed, (2) topologically ordered, it is considered a DAG.
- DAG is composed of:
 1. Vertices (Nodes)
 2. Edges (Connections)
- Some examples of DAGs:
 - Family tree
 - Gene Ontology (GO)
 - Human Phenotype Ontology (HPO)



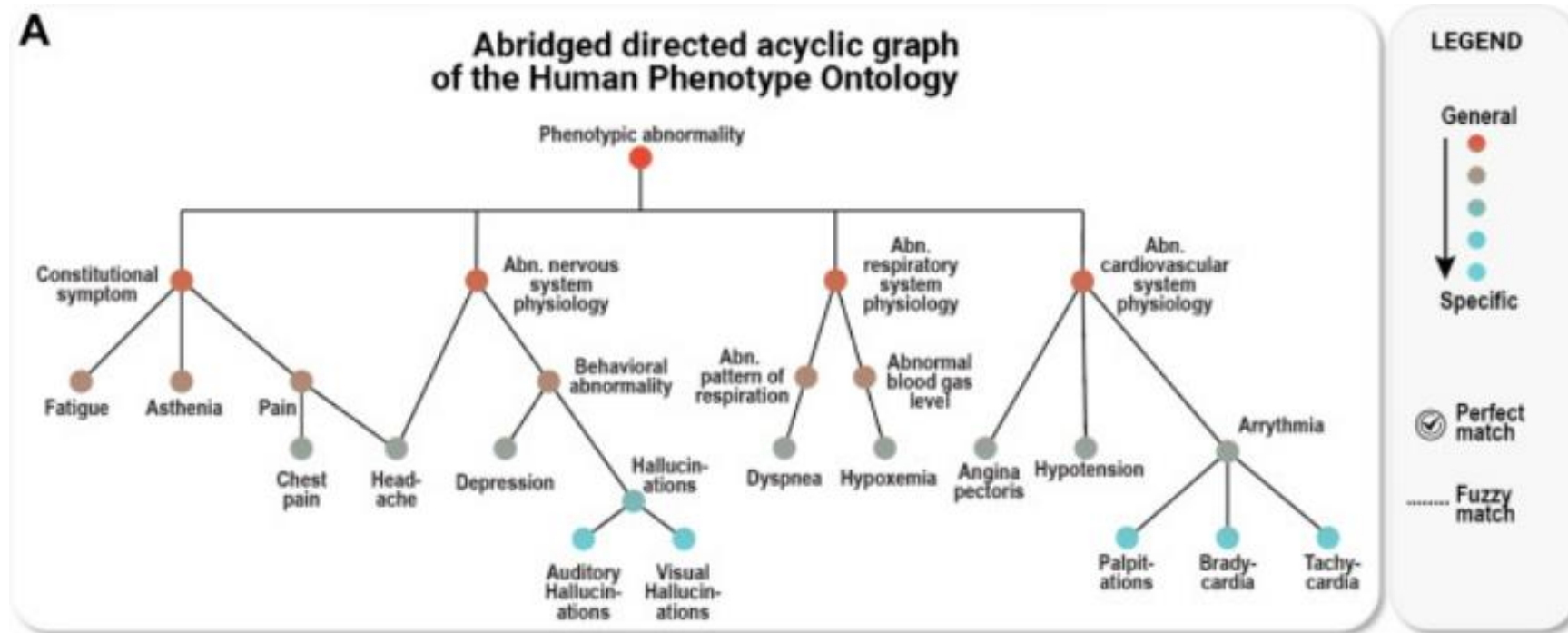
Directed Acyclic Graph (DAG) – Gene Ontology (GO)

- 3,083 GO terms & 19,469 annotated genes (HGNC symbols)
3,245 GO terms & 19,387 annotated genes (Ensembl IDs)
- Filtered 'Biological Process' terms and 'is a' relationships only



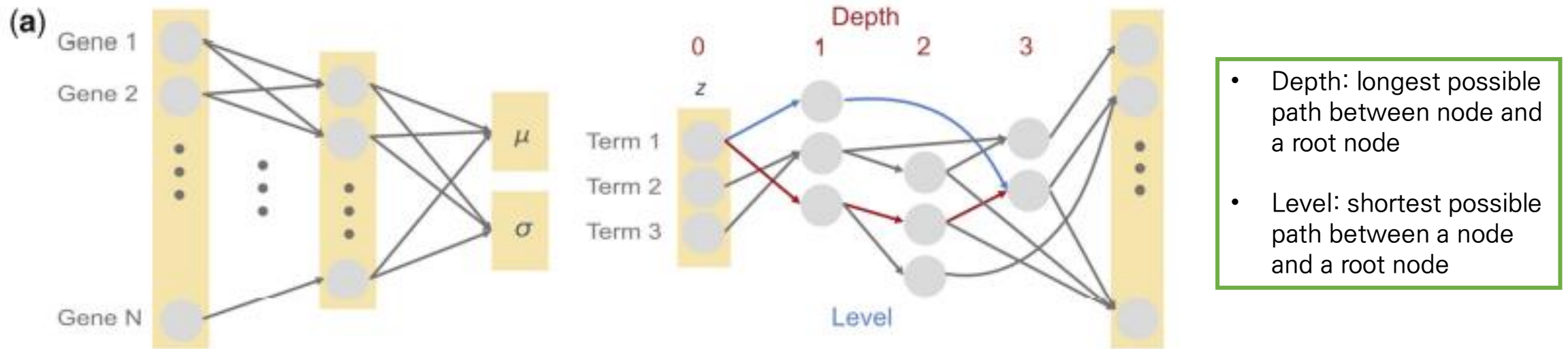
Directed Acyclic Graph (DAG) – Human Phenotype Ontology (HPO)

- Ontology that specifies disease-related terms, with more general disease in parent nodes
- 4,525 HPO terms & 4,774 annotated genes (HGNC symbols)
- Example)



OntoVAE architecture

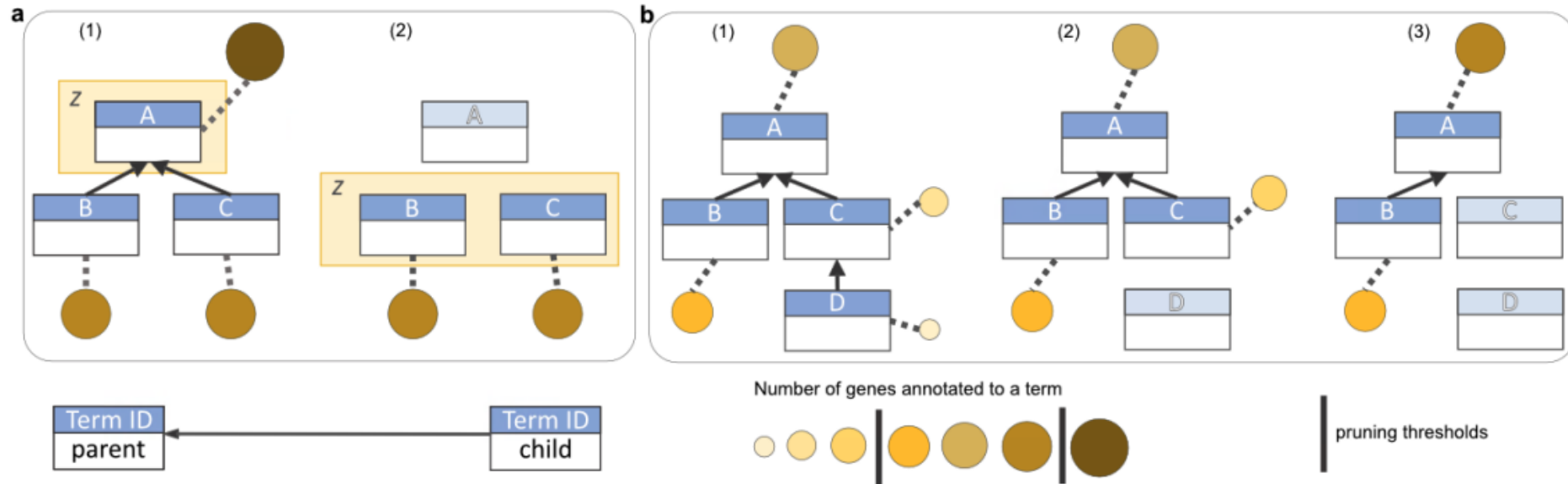
- OntoVAE is a modified VAE with a modified structure, so that it can **incorporate DAG**.
 - A non-linear encoder is coupled to a masked, multi-layer linear decoder representing biological ontology.
- Features of the model
 - Latent space is the root layer.
 - Each layer of the decoder corresponds to 1 depth of the ontology. (terms with same depth = same layer)
 - Decoder is linear, multi-layered, sparse, and have connections between non-neighboring layers(DenseNet).
 - This model has a trimming process to make a meaningful latent space.



```
def trim_dag(self, top_thresh=1000, bottom_thresh=30):
    """
    DAG is trimmed based on user-defined thresholds.
    Trimmed version is saved in the graph, annot and genes slots.
```

OntoVAE architecture – Trimming process

- Motivation: to make a meaningful latent space and avoid a 1D latent space
- Overall idea: Only terms with a specific #. of annotated genes are incorporated into the decoder nodes, so that generic terms and specific terms are removed.



(a) Top pruning

Root term is too general
(too much genes annotated)
→ use its children terms B,C

(b) Bottom pruning

Term D is too specific
(small #. of annotated genes)
→ transfer its genes to parent node C → A

Retrieval and comparison of pathway activities

- Activities are retrieved at each node in OntoVAE models.
 - attach pytorch hooks to each layer

```
# get activities from decoder
activation = {}
def get_activation(index):
    def hook(model, input, output):
        activation[index] = output.to('cpu').detach()
    return hook
hooks = {}
```

```
for i in range(len(self.decoder.decoder)-1):
    key = str(i)
    value = self.decoder.decoder[i][0].register_forward_hook(get_activation(i))
    hooks[key] = value
```

- Wilcoxon tests were performed on each GO term between two tissues.
 - Two-tailed Wilcoxon tests: compare pathway activities between groups of samples
 - One-tailed Wilcoxon tests: compare up/down-regulated pathway activities between groups of samples

Dataset

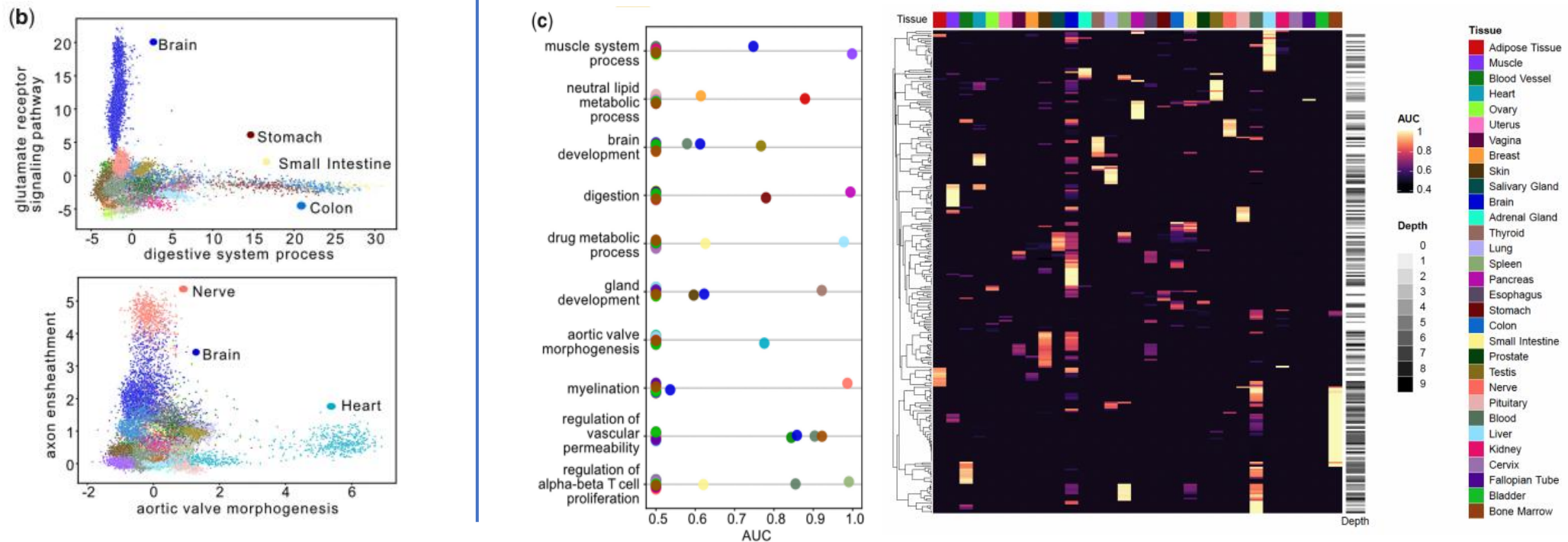
- Ontology: GO, HPO
- Bulk RNA-seq data
 - Genotype Tissue Expression(GTEx) dataset
 - Limb-girdle muscular dystrophy (LGMD; 지대형 근육영양장애) dataset (from Gene Expression Omnibus; GEO)
 - Preprocessed single-cell RNA-seq dataset of interferon(IFN)- β from Peripheral blood mononuclear cells(PBMCs) dataset

	Figure 1	Figure 2	Figure 3	Figure 4
Ontology	GO	GO	HPO	GO
Dataset	GTEx	GTEx muscle	GTEx muscle	PBMC
Task	Make GO-based decoder and capture key terms for each tissue	Simulate <i>in silico</i> gene knockouts	Apply <i>in silico</i> knockouts to find influential gene in a disease	Predict drug response by performing <i>in silico</i> stimulation

Results

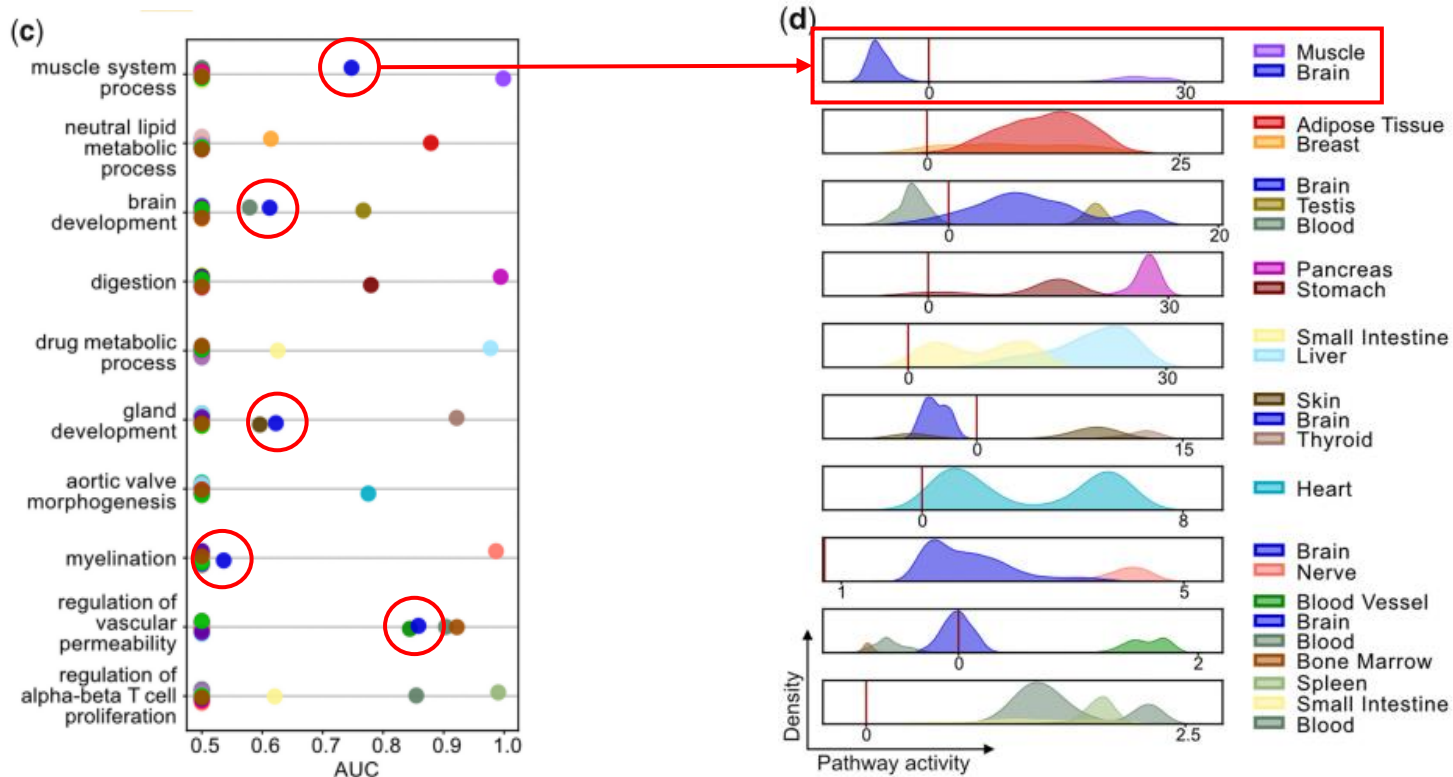
[Figure 1] OntoVAE provides biological interpretability in latent space and decoder

- Data: GTEx bulk RNA-seq data & GO
- (b) Retrieved the activities of all terms in the latent space and decoder for each sample and looked at some example terms to see if they were more active in the expected tissues.
- (c) median AUC from CV measured for each GO term(input) in classifying correct tissue(output)
 - Naïve Bayes classifier, 10-fold CV, 1-vs-all setting for each tissue



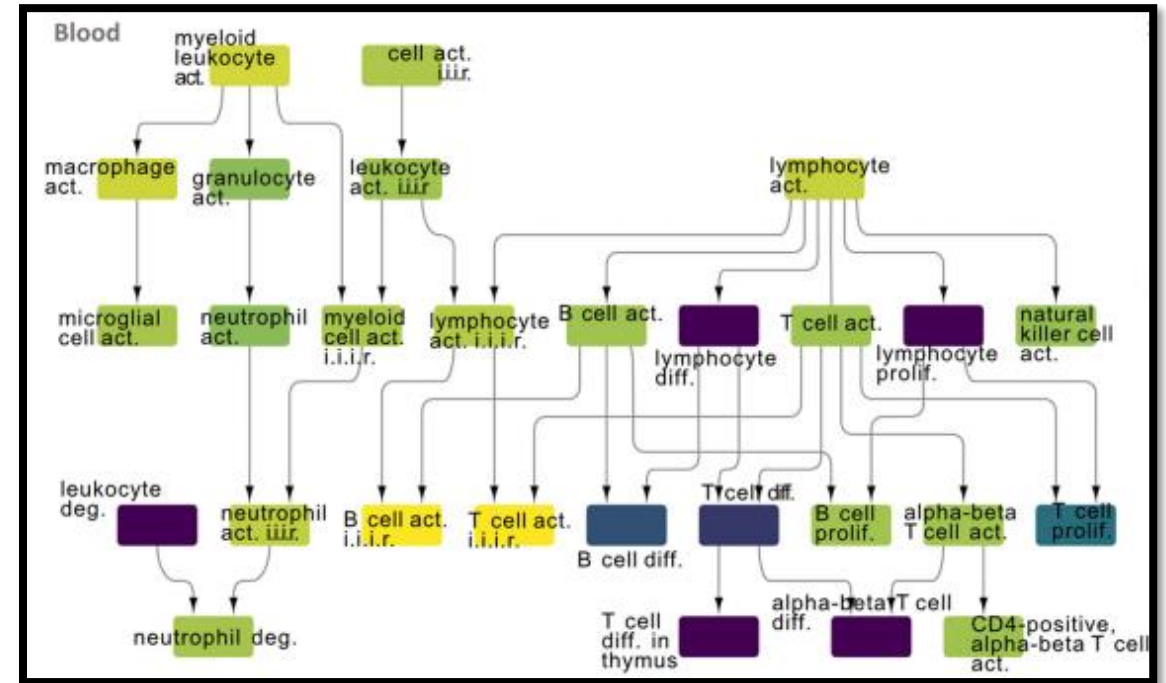
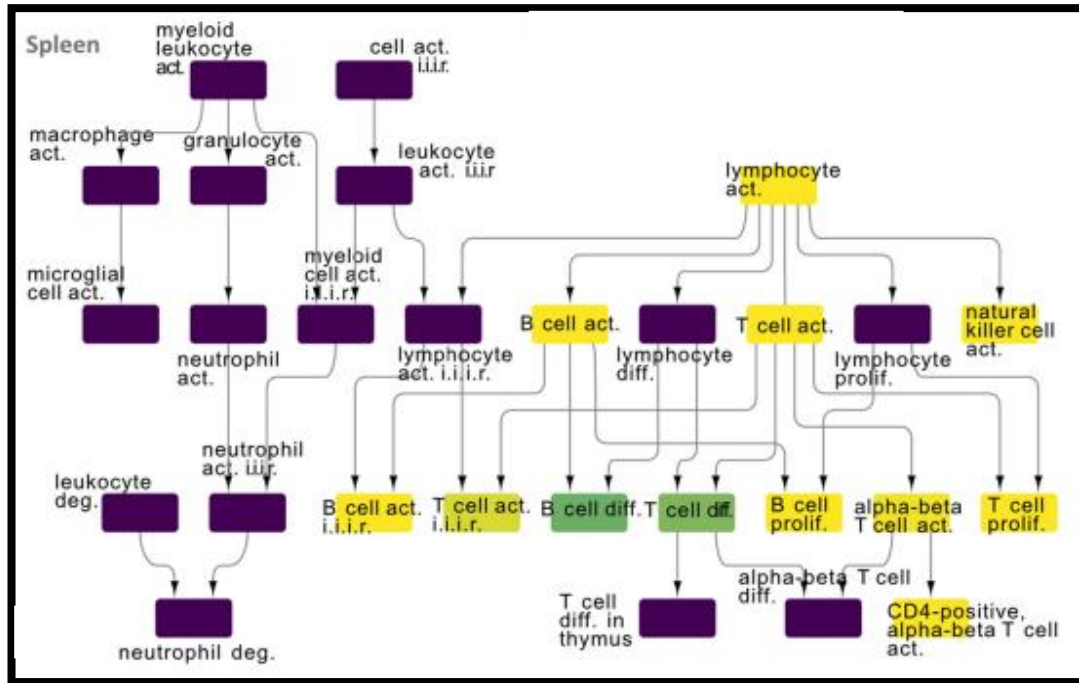
[Figure 1] OntoVAE provides biological interpretability in latent space and decoder

- Why does ‘muscle system process’ show high AUC in classifying ‘Brain tissue’?
- (d) Density of brain tissues over ‘muscle system process’ pathway is very low
 - Low expression of pathways, as well as high expression, affect the classification results
 - High AUC **does not** necessarily mean high pathway activity.



[Figure 1] OntoVAE provides biological interpretability in latent space and decoder

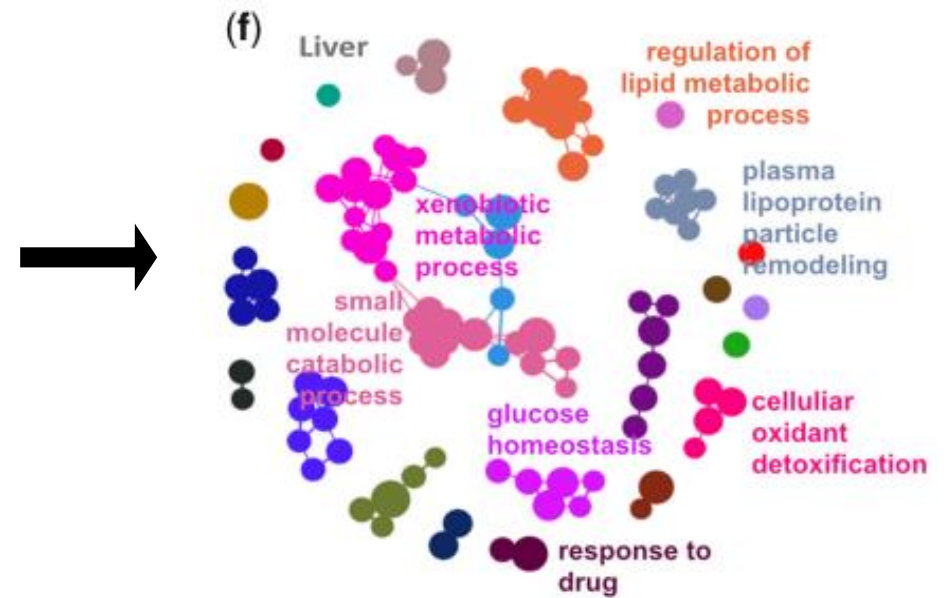
- (e) Subnetwork of the GO graph, colored by median AUC
 - 2 main branches are shown, and active nodes are shown with bright color (yellow, green)



[Figure 1] OntoVAE provides biological interpretability in latent space and decoder

- (f) Process of extracting the top GO terms for a given tissue
 - The authors retrieve the activities at each node when running samples through a pre-trained model. (Pytorch hooks)
 - One-sided Wilcoxon tests were performed on each GO term between two tissues.
 - Terms with high hits(#. of significant results from multiple tests) were selected to further group the terms for a given tissue into a network.

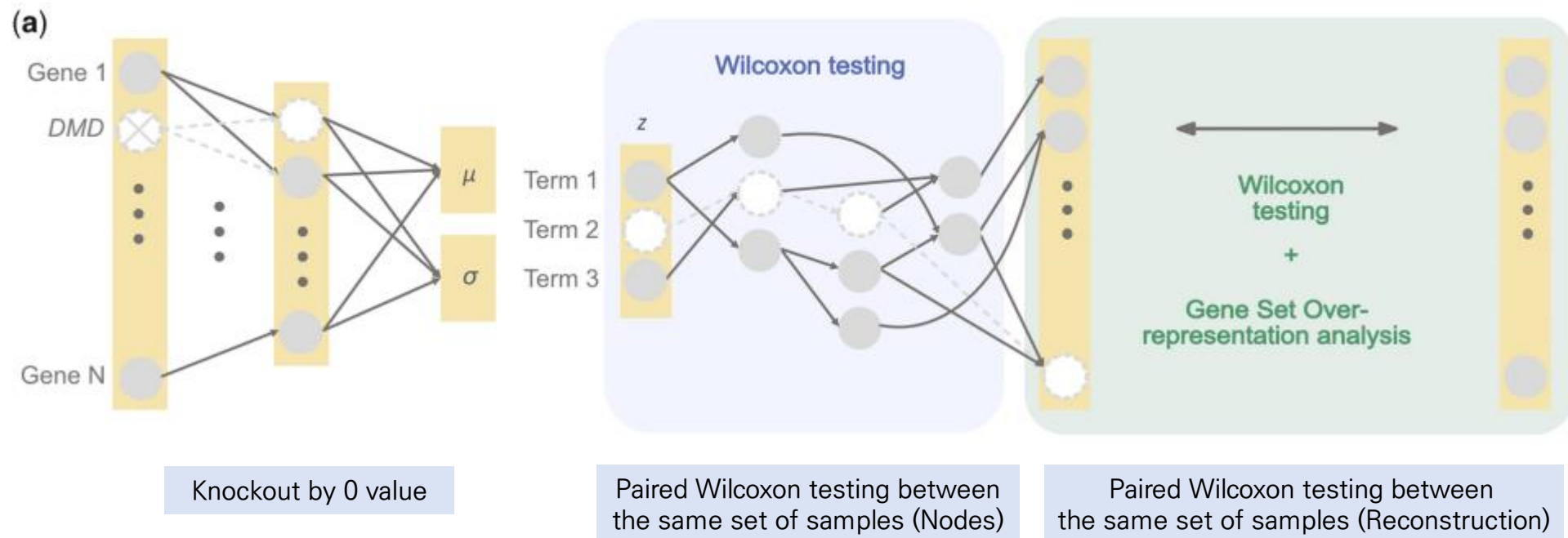
id	term	genes	tissue	hits	med_stat	rank
GO:0009164	nucleoside catabolic process	35	Liver	30	21.15347889	1
GO:0072329	monocarboxylic acid catabolic process	118	Liver	30	21.1525784	1
GO:0019320	hexose catabolic process	52	Liver	30	21.15256714	1
GO:0006007	glucose catabolic process	32	Liver	30	21.15256714	1
GO:0043574	peroxisomal transport	76	Liver	30	21.15255587	1
GO:0006625	protein targeting to peroxisome	72	Liver	30	21.1525446	1
GO:0034656	nucleobase-containing small molecule catabolic process	50	Liver	30	21.15214506	1
GO:0044282	small molecule catabolic process	421	Liver	30	21.15144435	1
GO:0016054	organic acid catabolic process	250	Liver	30	21.15144435	1



- 1 node in the network = 1 GO term
- Size of the node: #. of genes associated with the GO term

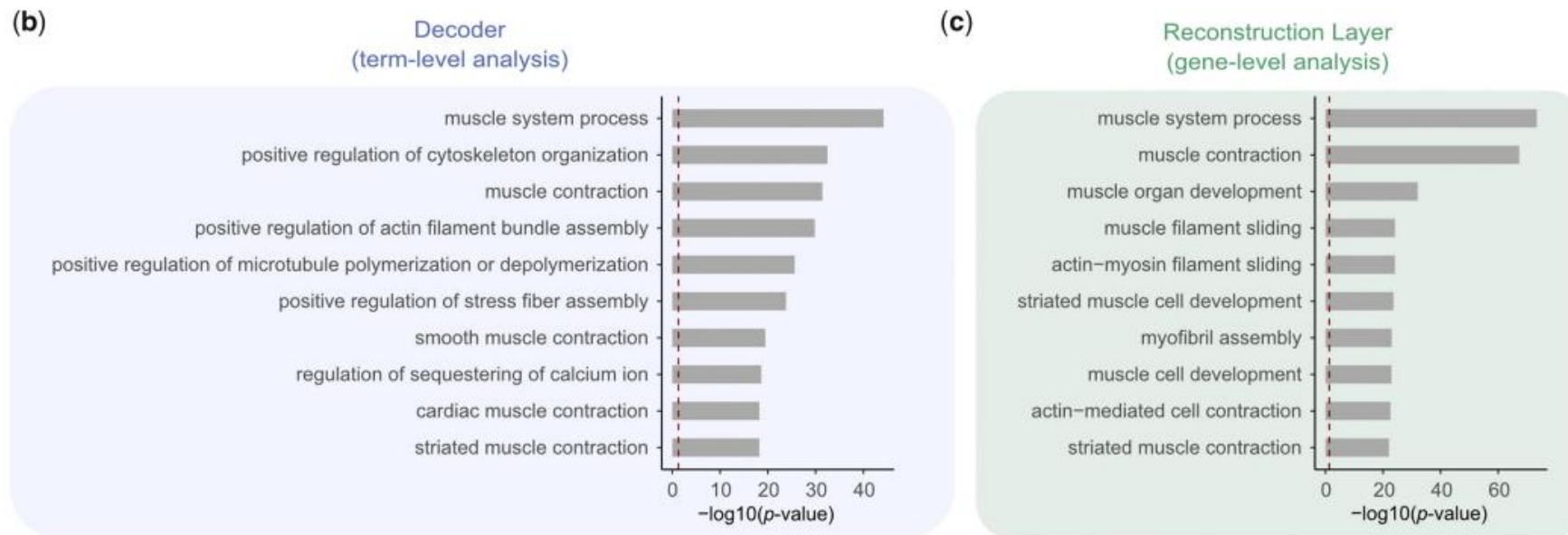
[Figure 2] OntoVAE can be used for *in silico* phenotype predictions of gene knockouts

- Data: GTEx muscle(881 samples) + GO
- Performed an *in silico* knockout of the DMD gene in the GTEx muscle samples
 - DMD: encodes protein dystrophin, located in the muscles, attaches the cytoskeleton to the extracellular matrix.
 - Depletion of functional dystrophin protein → cause muscle weakness & degradation
- (a) Schematic of how OntoVAE is used for *in silico* gene knockout experiments



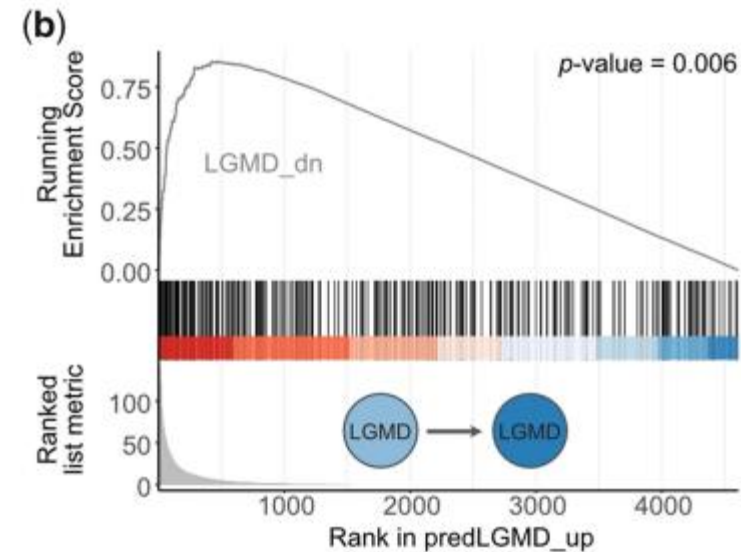
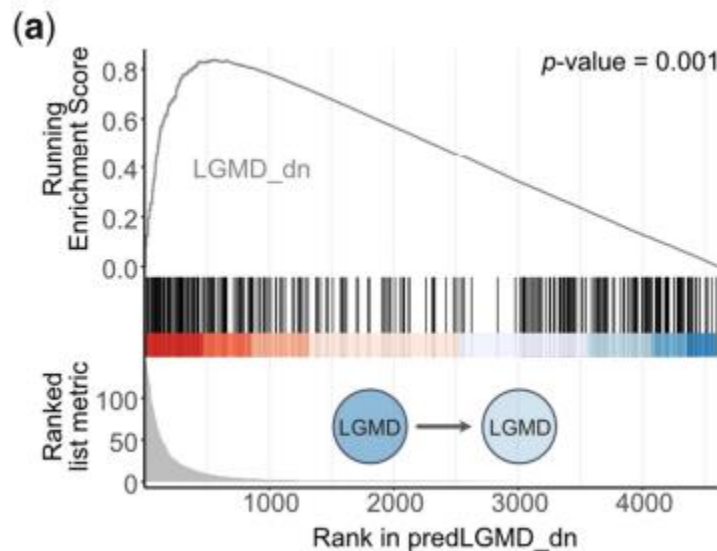
[Figure 2] OntoVAE can be used for *in silico* phenotype predictions of gene knockouts

- (b) Term-level analysis obtained terms highly related to DMD function.
- (c) Gene-level analysis obtained terms highly related to DMD function, from significant genes.
- These show that OntoVAE captures meaningful relationships between the genes, and therefore can be used to predict the consequences of a gene knockout, with direct interpretability.



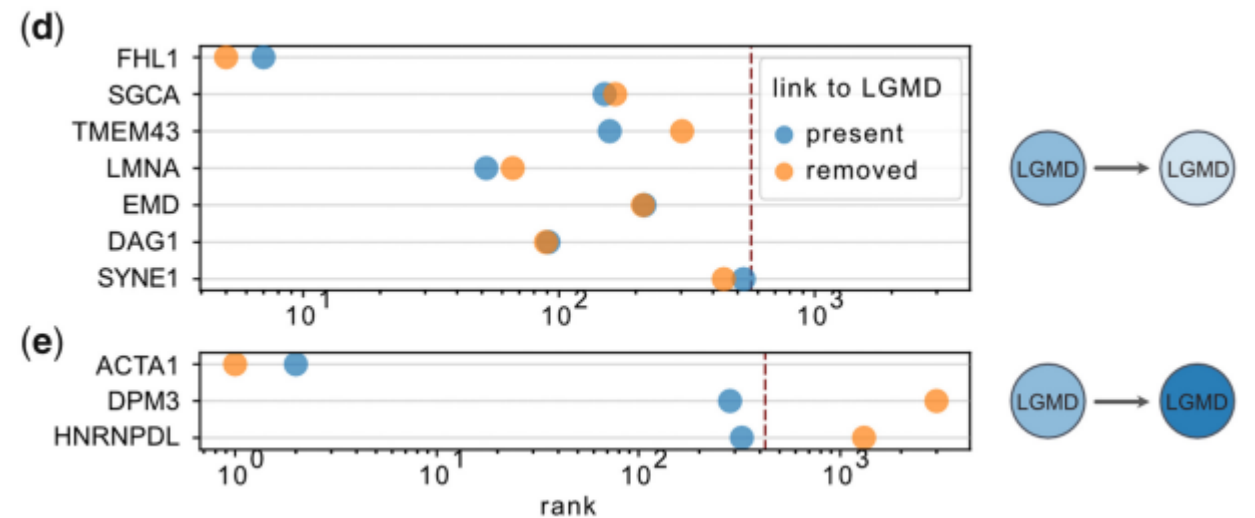
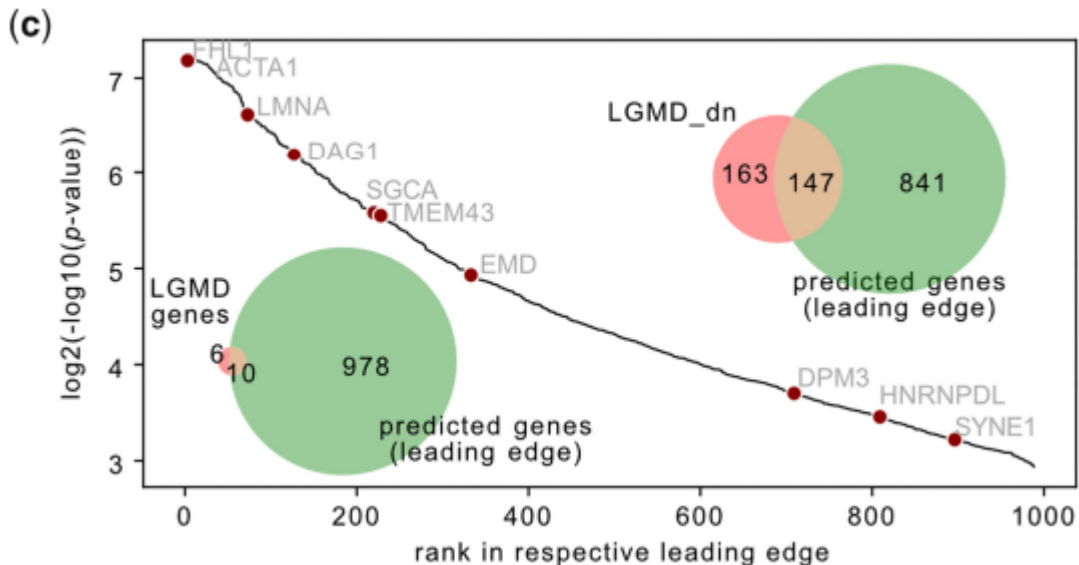
[Figure 3] OntoVAE can predict disease-specific gene expression changes

- Data: GEO LGMD(881 samples) / HPO(disease-related ontology)(4774 genes)
- Performed knockout for every gene (4609), followed by paired Wilcoxon test at the LGMD node
 - 2 separate one-sided paired Wilcoxon tests for each gene
- GSEA results using the ranking of genes that significantly (a) down-regulated, (b) up-regulated the LGMD node in the decoder



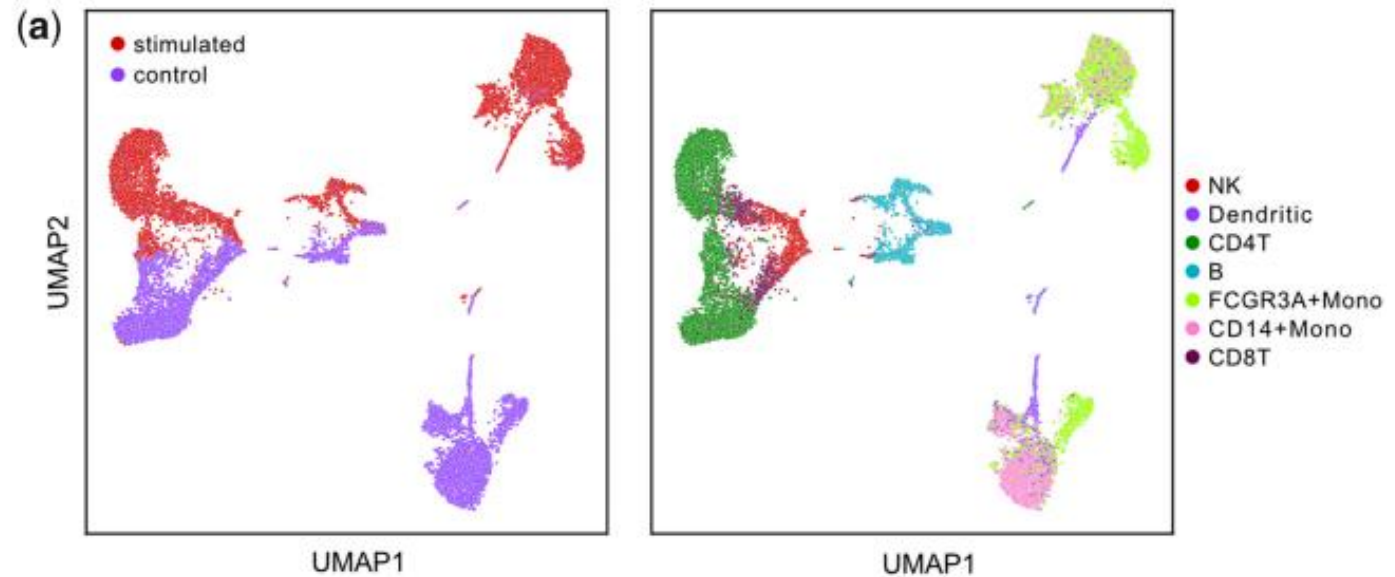
[Figure 3] OntoVAE can predict disease-specific gene expression changes

- Validity data: bulk RNA-seq data on **LGMD** patients ($N = 16$) and control ($N = 15$) (Depuydt *et al.* 2022)
 - Made a list of **Up- & Down-regulated genes** based on real data
- Predicted genes: **988 'leading edge' genes** from the **GEO LGMD** data
- (c) check **overlapping genes** from **GEO LGMD** and **validation LGMD** data: 147 / 10 genes
- (d), (e) check whether 10 genes that were directly annotated to LGMD in the HPO dataset would affect the model performance



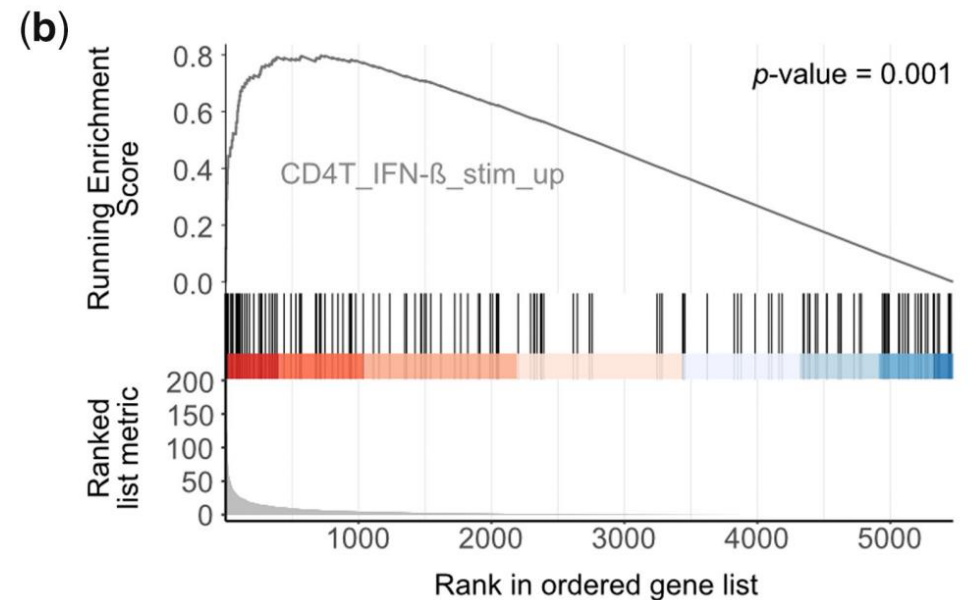
[Figure 4] OntoVAE can predict treatment effects *in silico*

- Data: PBMC dataset (Kang *et al.*, 2018) / GO (need terms for IFN response)
 - PBMCs from lupus patients were treated with IFN- β
 - Performed single-cell RNA-seq
- (a) UMAP representation of the dataset
 - Clustering based on treatment vs control (left)
 - Clustering based on each cell type (right)
- Investigated CD4T cells for OntoVAE: largest population from all cell types



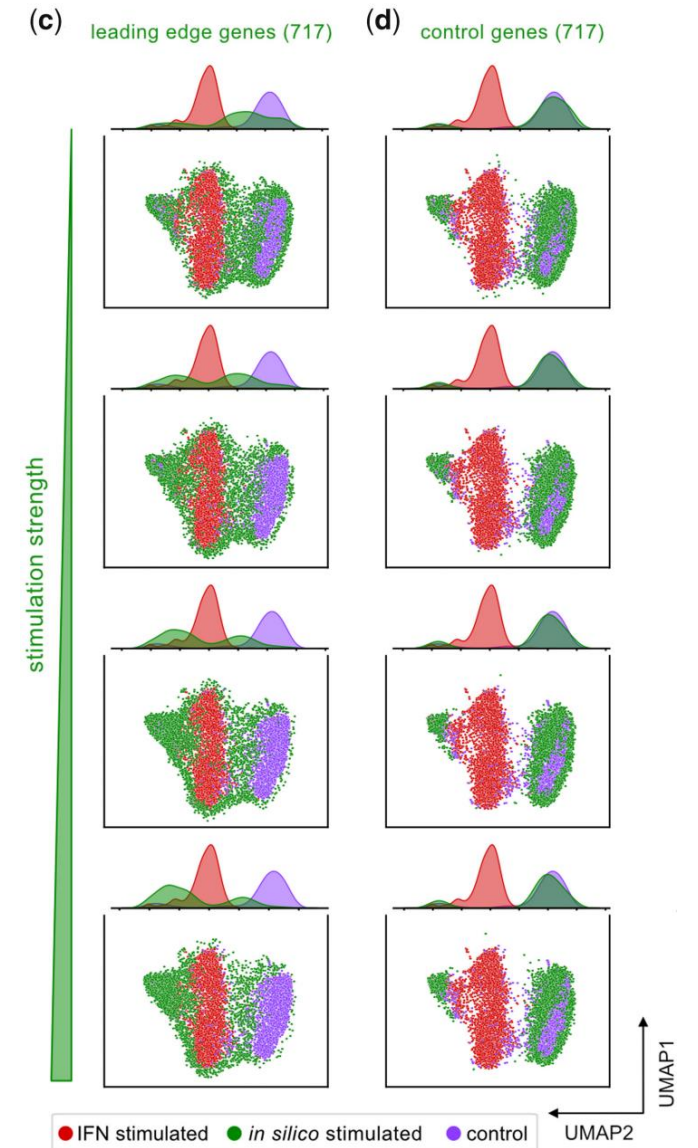
[Figure 4] OntoVAE can predict treatment effects *in silico*

- Extract “reference genes” (CD4T_IFN- β _stim_up) from CD4T cells by performing GSEA between stimulated and unstimulated CD4T cells → 146 genes significantly up-regulated
- Use GO-decoder OntoVAE to train model for unstimulated CD4T cells, then perform one-by-one *in silico* stimulation of all input genes
 - Set the expression value of each gene to a higher value: simulate a specific type of gene expression as enhanced
- (b) Paired Wilcoxon test on ‘type I IFN signaling pathway’ node in the GO-decoder, then performed GSEA with the reference genes



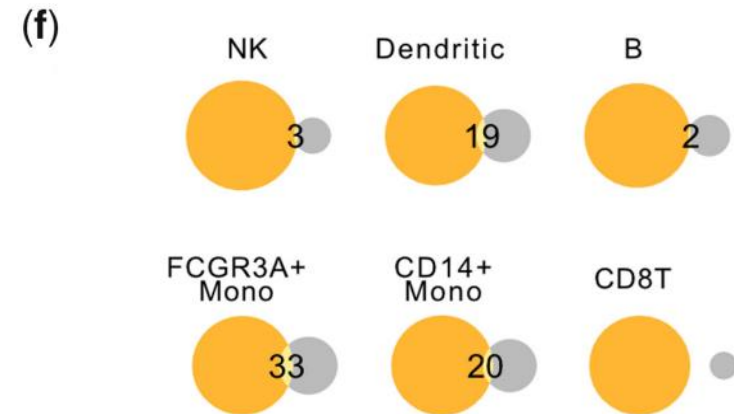
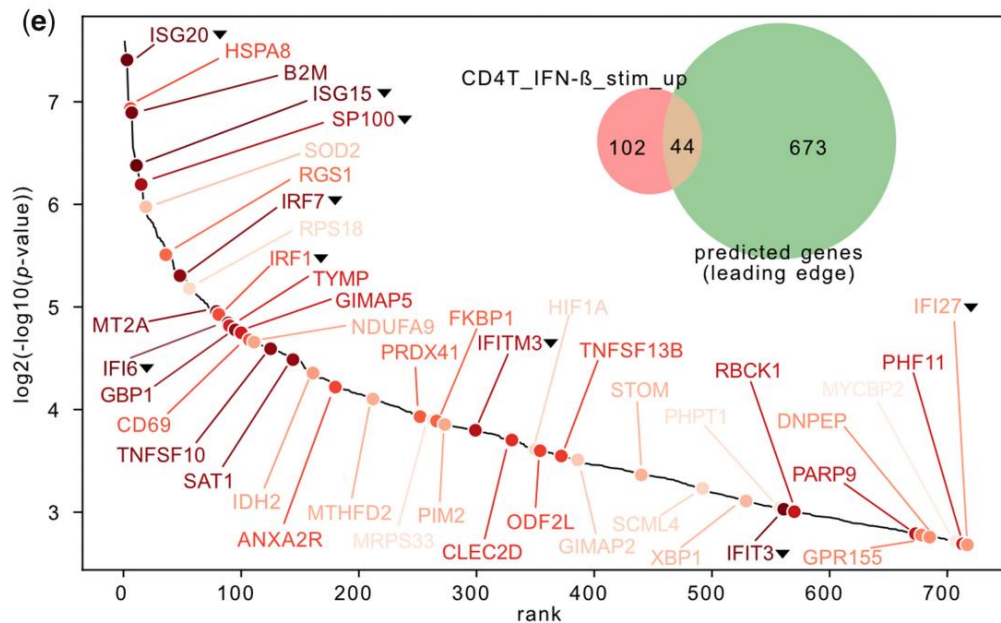
[Figure 4] OntoVAE can predict treatment effects *in silico*

- Find the similarity of the perturbed(modified) CD4T cells to IFN-stimulated CD4T cells
- Compared UMAP projection of **actual stimulated cells**, **CD4T control**, and ***in silico*-stimulated cells**
 - (c) *in silico*-stimulation performed for 717 leading edge genes
 - (d) Same #. of bottom genes were *in silico*-stimulated for comparison
- Results show that while stimulating significant genes from GSEA, the distribution of the cells get closer to the actual stimulated cells as the stimulation strength gets larger.
 - This is not seen from results stimulating control genes.



[Figure 4] OntoVAE can predict treatment effects *in silico*

- (e) 44 genes overlapped between the predicted genes and reference genes.
 - 35 out of 44 genes are not directly annotated to the target term or its children terms.
- (f) Some of the remaining 673 predicted genes overlap with DEGs in other cell types.



Discussion

- In this work, the authors designed OntoVAE, a novel VAE, where any hierarchical biological network that has the structure of a DAG can be incorporated in its latent space and decoder. OntoVAE can also be trimmed to have a reasonable number of root nodes that represent the latent space variables.
- The model was applied on different ontologies(GO, HPO) and datasets(GTEEx, PBMC).

	Figure 1	Figure 2	Figure 3	Figure 4
Ontology	GO	GO	HPO	GO
Dataset	GTEEx	GTEEx muscle	GTEEx muscle	PBMC
Task	Make GO-based decoder and capture key terms for each tissue	Simulate <i>in silico</i> gene knockouts	Apply <i>in silico</i> knockouts to find influential gene in a disease	Predict drug response by performing <i>in silico</i> stimulation

- One interesting and central finding is the fact that the model can predict influential genes on processes or phenotypes that go beyond the ones directly annotated to the term in question, indicating that the model is capable of learning more complex gene-term relationships in a data-driven way.
- OntoVAE has advantages in its model that it does not limit the number of biological terms under scrutiny or use a single-layer linear decoder. This makes the model capable of encoding thousands of terms without the need for preselection and maintaining the hierarchical information contained in the ontology.
- In summary, OntoVAE can be adapted to any ontology and dataset. It is used to compute pathway activities and predict disease or treatment-induced changes in gene expression. This model fully exploits the conceptual complexity of the hierarchical structure of the ontology and can highlight differences between samples at different levels of the ontology.

Thank You!

