NEURAL INFORMATION PROCESSING SYSTEMS

MuSe-GNN: Learning Unified Gene Representation From Multimodal Biological Graph Data

Tianyu Liu, Yuge Wang, Rex Ying, Hongyu Zhao

Yale University

Zhao lab

Ying lab

GitHub page

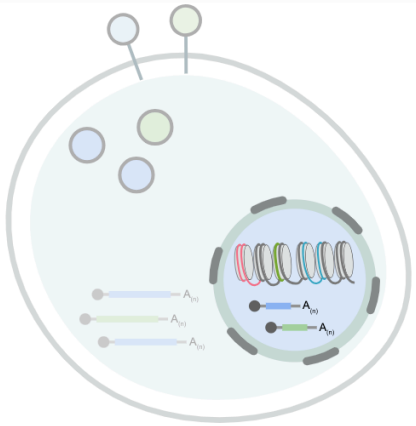
Success of current single-cell analysis

nature › technology features › article

TECHNOLOGY FEATURE | 19 July 2021 | Correction 21 July 2021 | Correction 24 August 2021

Single-cell analysis enters the multiomics age

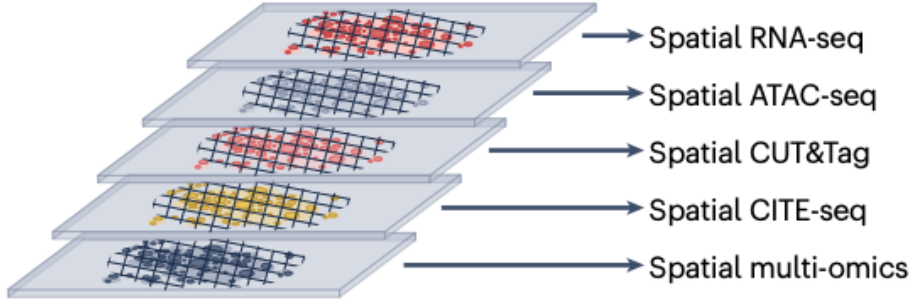
A rapidly growing collection of software tools is helping researchers to analyse multiple huge '-omics' data sets.



Protein

RNA

Chromatin



HUMAN CELL ATLAS

HuBMAP
Human BioMolecular Atlas Program

SenNet

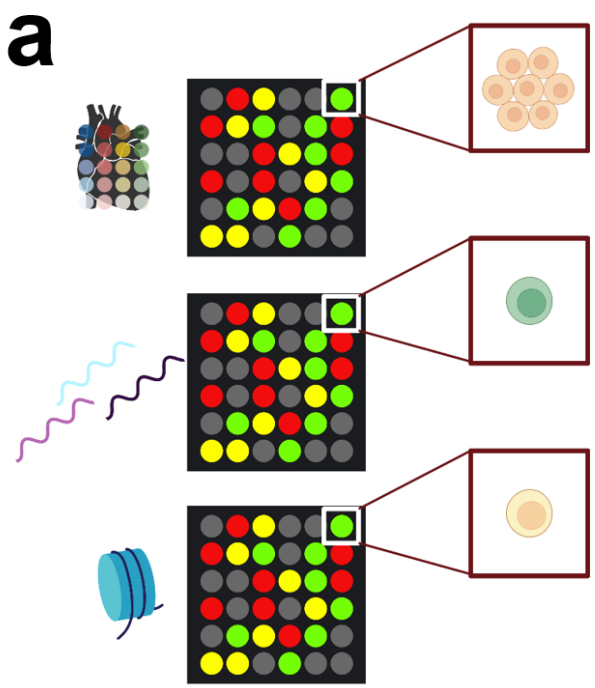
More data are coming…

Problem of multi-omics integration with cells

1. The vast data volume in atlas-level studies challenges high-performance computing.

2. Data from different omics pose their own challenges (mixture of cells, or problematic matching relation).

3. Batch effects may adversely impact analysis results by introducing noise.



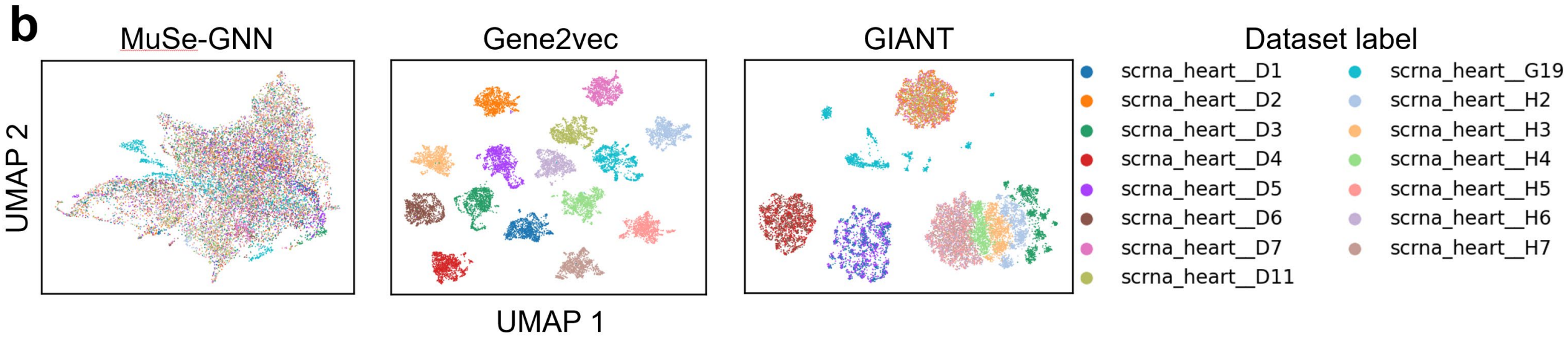
Each spots contains different number of cells.

Batch effect/data quality may affect the downstream analysis.

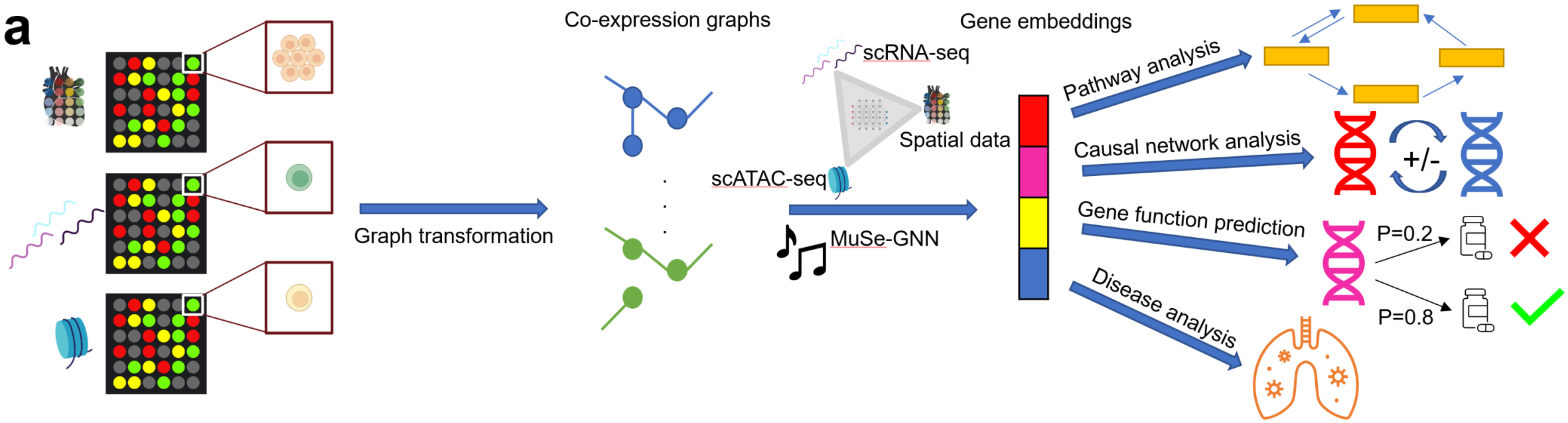
Peaks and genes are not perfectly matched.

Definition of gene embeddings

For multimodal biological datasets $\boldsymbol{D} = (\{\boldsymbol{V_i}, \boldsymbol{E_i}\})_{\boldsymbol{i=1}}^{\boldsymbol{T}}$, our goal is to construct a model $\boldsymbol{M}(\cdot, \boldsymbol{\theta})$, designed to yield gene embeddings set $\boldsymbol{\epsilon} = \{\boldsymbol{e_1}, \ldots, \boldsymbol{e_T}\} = \boldsymbol{M}(\boldsymbol{D}, \boldsymbol{\theta})$. We intend to harmonize gene information from diverse modalities.



Overview of MuSe-GNN

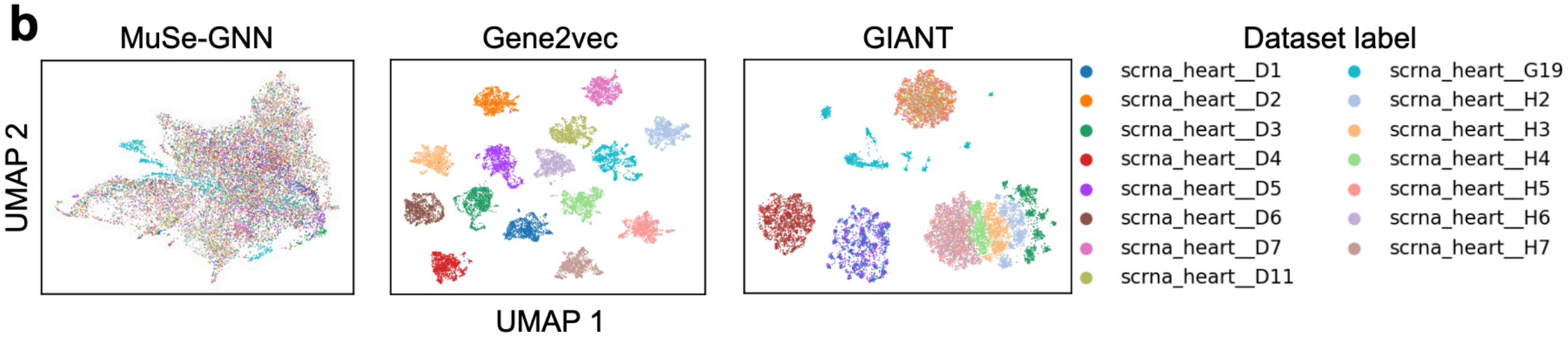


We introduce **Mu**ltimodal **S**imilarity L**e**arning **G**raph **N**eural **N**etwork (**MuSe-GNN**), to learn gene representations across different modalities/biomedical contexts.

MuSe-GNN = **Cross-graph Transformer** + **Weighted Similarity Learning** + Contrastive Learning.

Our model efficiently produces unified gene representations for the analysis of **gene functions, tissue functions, diseases, and species evolution**.

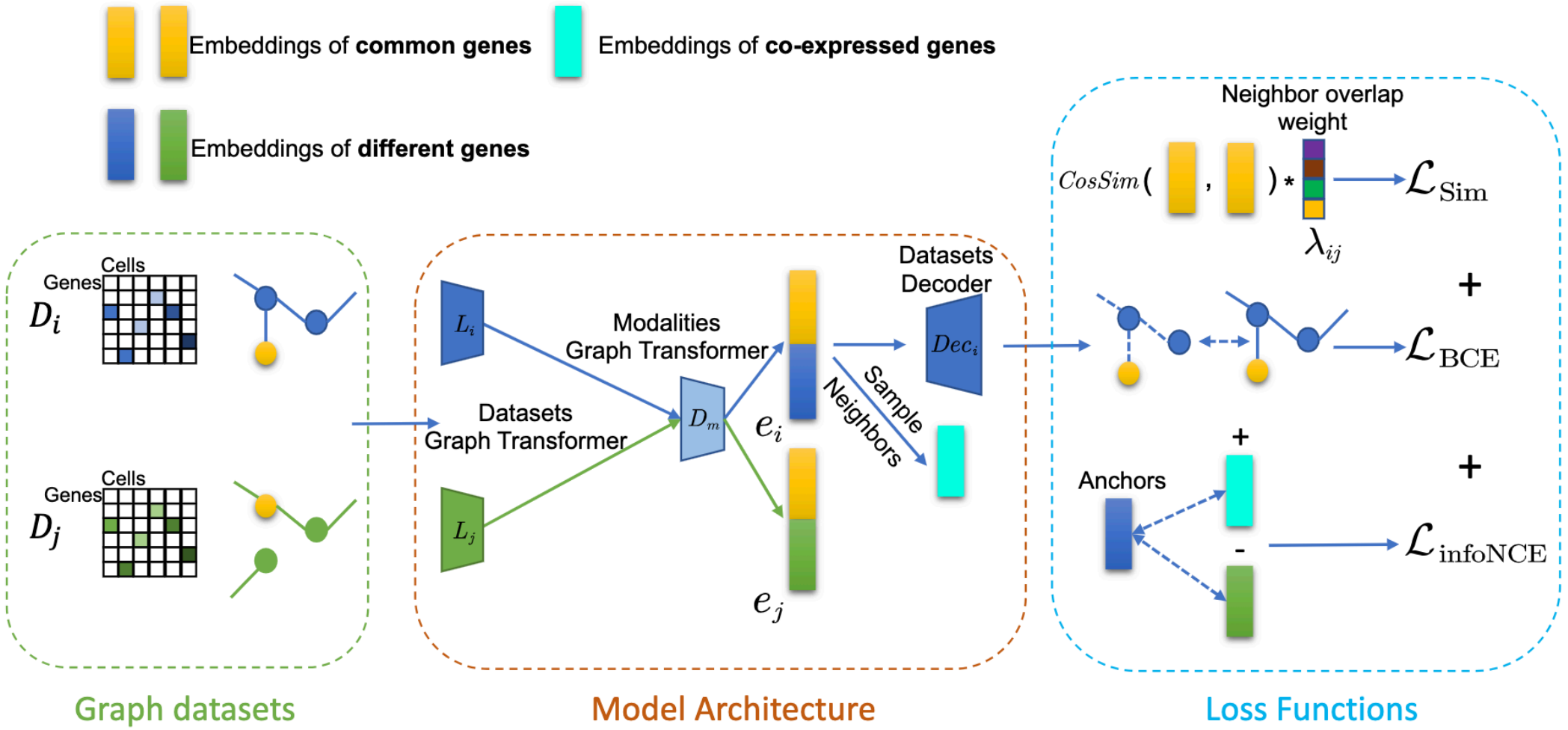
Overview of MuSe-GNN



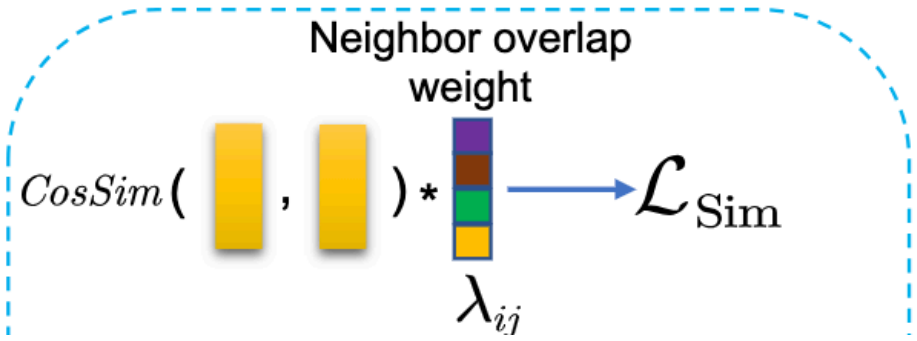
We leveraged **82** training datasets from **10** tissues, offering gene representations containing functional similarity across different contexts in a joint space.

MuSe-GNN outperforms SOTA methods in gene representation learning by up to **97.5%**.

Overview of MuSe-GNN



Weighted-similarity learning

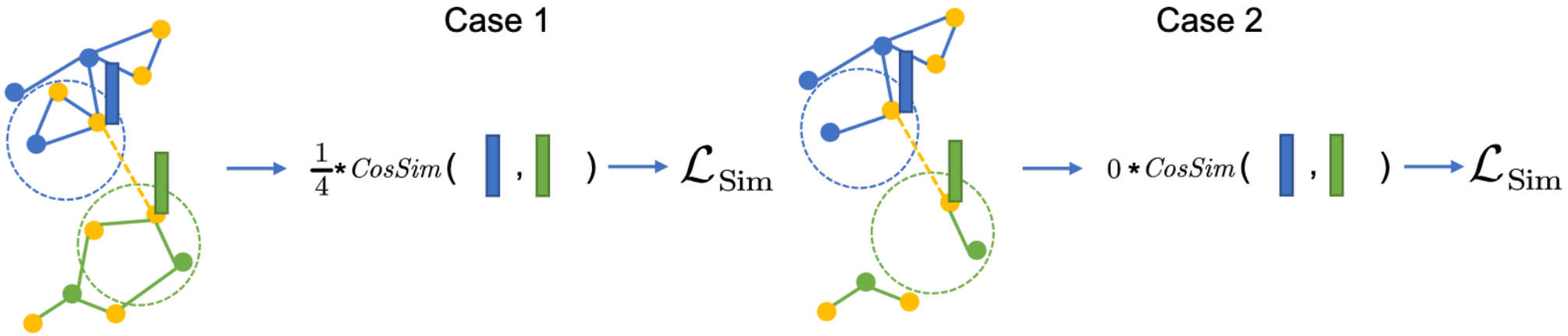


Embeddings of **common genes**

$$\lambda_{ijg} = \frac{|N_{ig} \cap N_{jg}|}{|N_{ig} \cup N_{jg}|}.$$

$N_{i(j)g}$: Gene neighbors of gene g of dataset $i(j)$.

Common genes
Different genes
Paired relation
Neighbors' extraction
Node features



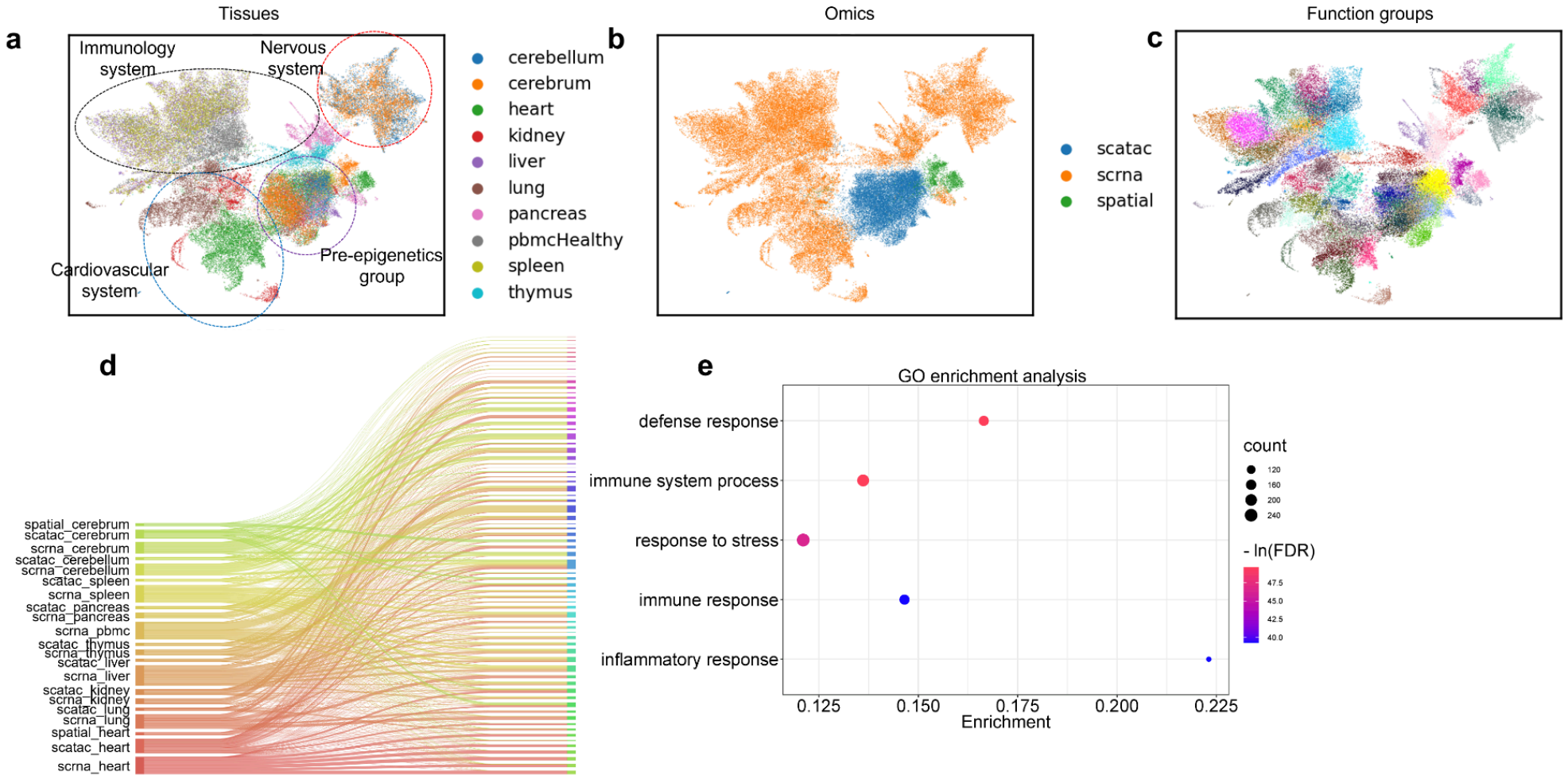
Results: Benchmarking analysis

We evaluate the gene embeddings for different tissues based on six metrics defined by ourselves.

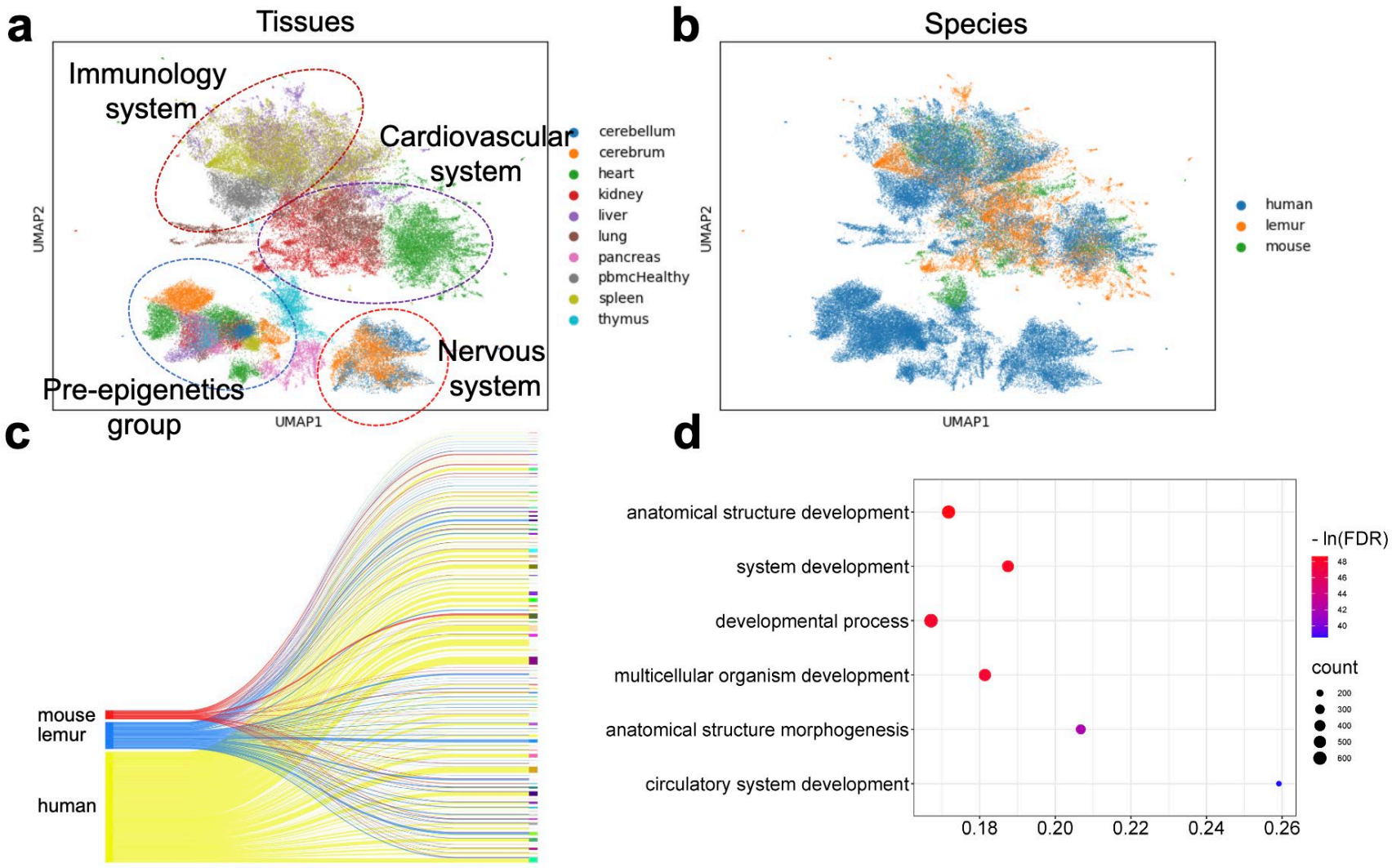
Table 1: Avg Score across different tissues. Standard deviations are reported in Appendix E.3.

Methods	Heart	Lung	Liver	Kidney	Thymus	Spleen	Pancreas	Cerebrum	Cerebellum	PBMC
PCA	0.52	0.48	0.56	0.47	0.56	0.60	0.51	0.62	0.53	0.51
Gene2vec	0.40	0.37	0.33	0.29	0.21	0.31	0.24	0.27	0.31	0.19
GIANT	0.50	0.40	0.33	0.38	0.58	0.33	0.56	0.29	0.28	0.28
WSMAE	0.50	0.47	0.54	0.46	0.57	0.53	0.52	0.55	0.59	0.50
GAE	0.61	0.45	0.58	0.40	0.56	0.58	0.52	0.56	0.60	0.54
VGAE	0.64	0.32	0.33	0.38	0.56	0.31	0.33	0.41	0.33	0.47
MAE	0.36	0.47	0.50	0.45	0.41	0.52	0.39	0.50	0.49	0.50
scBERT	0.41	0.49	0.55	0.62	0.17	0.58	0.46	0.60	0.61	0.58
MuSeGNN	**0.77**	**0.96**	**0.92**	**0.89**	**0.89**	**0.94**	**0.80**	**0.95**	**0.90**	**0.92**

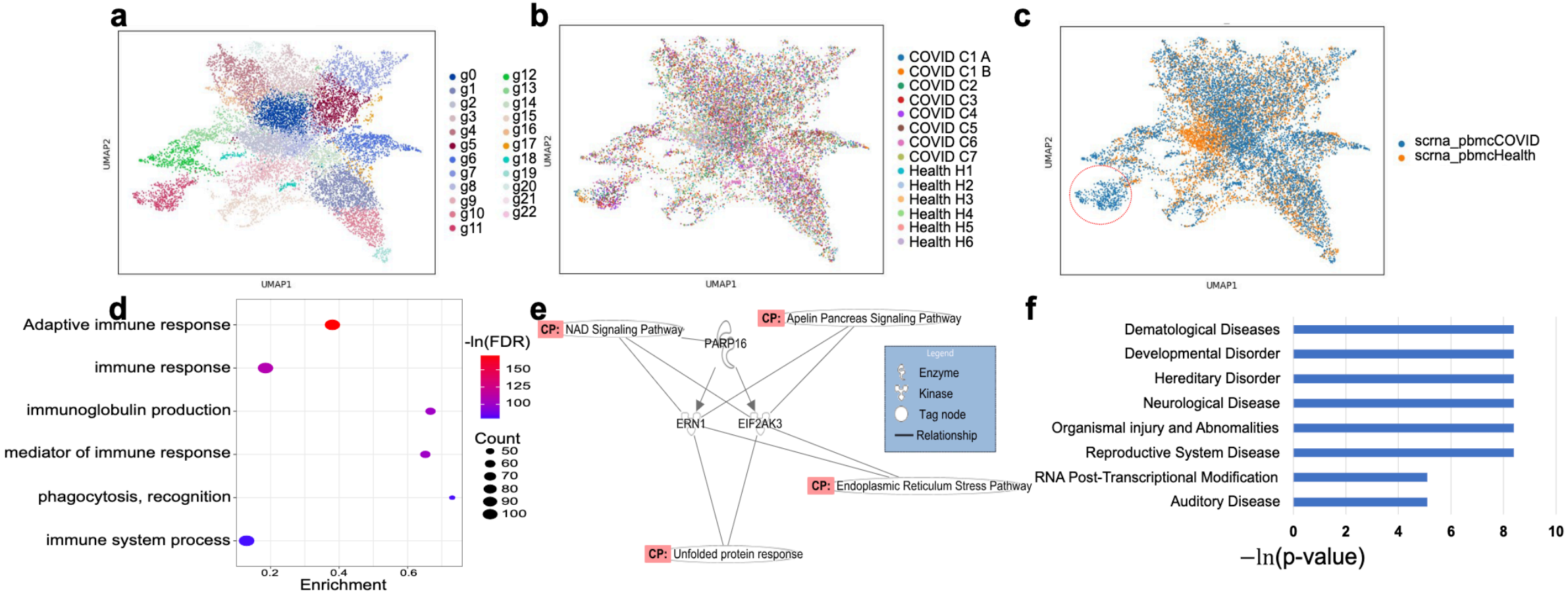
Results: Multi-omics gene embeddings



Results: Multi-species gene embeddings



Results: Gene embeddings for disease analysis



Results: Gene embeddings for gene function prediction

Table 2: Accuracy for dosage-sensitivity prediction

	MuSe-GNN (unsup)	Geneformer (sup)	Raw
Accuracy	0.77±0.01	0.74±0.06	0.75±0.01

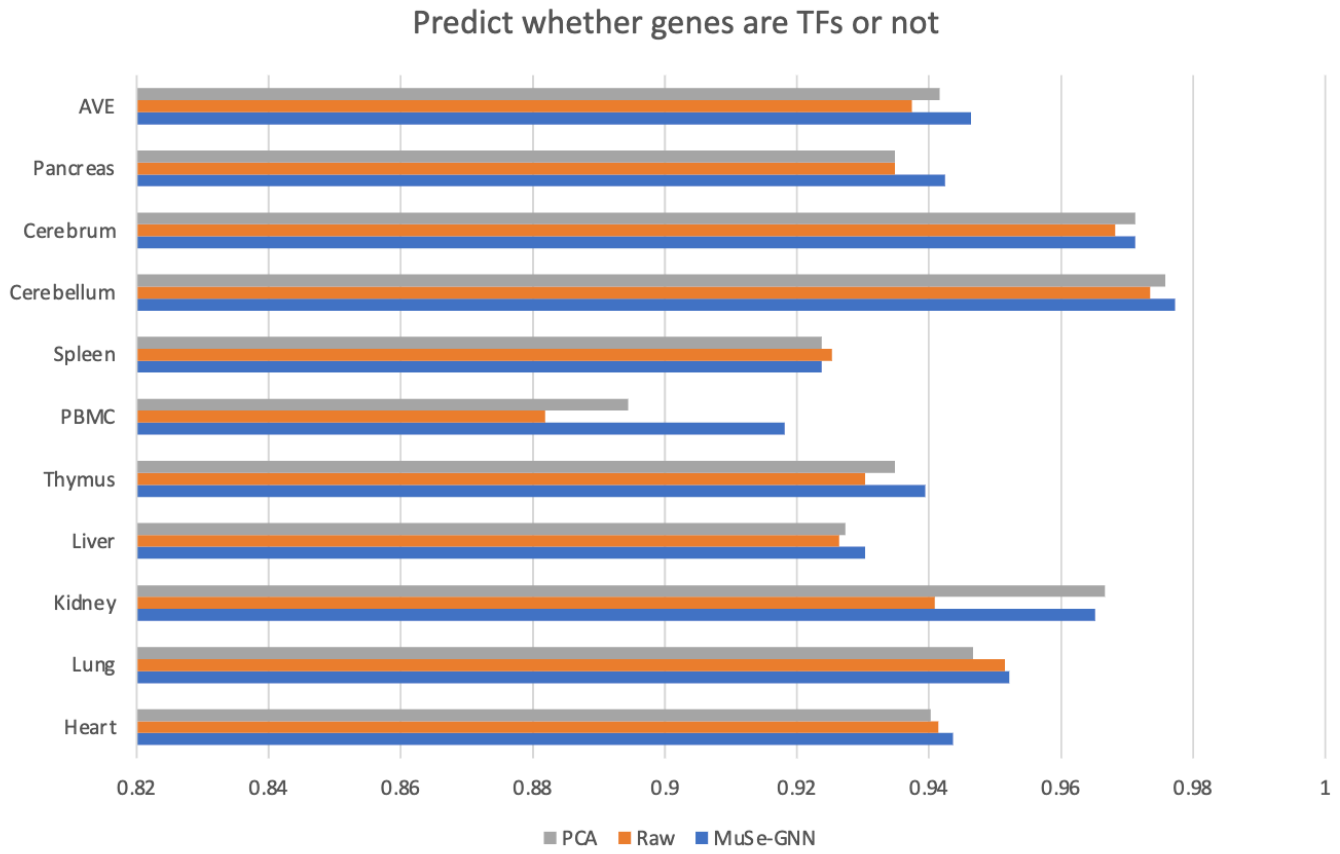


Figure 20: Accuracy for Gene-TF prediction across different tissues.

Discussion & Conclusion

1. GNN + MMML = MuSe-GNN

2. MuSe-GNN outperforms current gene embedding learning models across different metrics and can effectively learn the functional similarity of genes across tissues and techniques.

3. The gene representations learned by MuSe-GNN are highly versatile and can be applied to different analysis frameworks.

4. In the future, we plan to explore more efficient approaches for training and extend MuSe-GNN to handle a broader range of multimodal biological data.

Acknowledgement

We appreciate the comments, feedback, and suggestions from Chang Su, Zichun Xu, Xinning Shan, Yuhan Xie, Mingze Dong, and Maria Brbic.

Yale Center for Research Computing

NIH National Institutes of Health
Turning Discovery Into Health

NSF National Science Foundation

Manuscript

Poster