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

Tianyu Liu
Yale University
tianyu.liu@yale.edu

Yuge Wang
Yale University
yuge.wang@yale.edu

Rex Ying
Yale University
rex.ying@yale.edu

Hongyu Zhao*
Yale University
hongyu.zhao@yale.edu
*: Corresponding author

Abstract

Discovering genes with similar functions across diverse biomedical contexts poses a significant challenge in gene representation learning due to data heterogeneity. In this study, we resolve this problem by introducing a novel model called Multimodal Similarity Learning Graph Neural Network, which combines Multimodal Machine Learning and Deep Graph Neural Networks to learn gene representations from single-cell sequencing and spatial transcriptomic data. Leveraging 82 training datasets from 10 tissues, three sequencing techniques, and three species, we create informative graph structures for model training and gene representations generation, while incorporating regularization with weighted similarity learning and contrastive learning to learn cross-data gene-gene relationships. This novel design ensures that we can offer gene representations containing functional similarity across different contexts in a joint space. Comprehensive benchmarking analysis shows our model’s capacity to effectively capture gene function similarity across multiple modalities, outperforming state-of-the-art methods in gene representation learning by up to 97.5%. Moreover, we employ bioinformatics tools in conjunction with gene representations to uncover pathway enrichment, regulation causal networks, and functions of disease-associated or dosage-sensitive genes. Therefore, our model efficiently produces unified gene representations for the analysis of gene functions, tissue functions, diseases, and species evolution.

1 Introduction

Progress in biological technology has broadened the variety of biological data, facilitating the examination of intricate biological systems. A prime example of such technology is single-cell sequencing, which allows for the comprehensive characterization of genetic information within individual cells [38, 78]. This technology provides access to the full range of a cell’s transcriptomics, epigenomics and proteomics information, including gene expression (scRNA-seq) [46, 114], chromatin accessibility (scATAC-seq) [20, 17], methylation [60, 68] and anti-bodies [85]. By sequencing cells from the same tissue at various time points, we can gain insight into patterns of cellular activity over time [30]. Moreover, spatial information of cellular activity represents an equally vital, additional dimension. Such data are defined as spatial data. All of these data are known as multi-omics data, and integrating multi-omics data for combined analysis poses a significant challenge.

However, the traditional idea of multi-omics data integration known as using cells as anchors [56, 87, 39, 12, 21, 76] is only partially suitable because of the following two challenges. 1. Different omics

data pose their own challenges. For example, the unit of observation of the spatial transcriptomic data is different from other single-cell data, as a single spatial location contains mixed information (see the left part of Figure 1 a) from various cells [23] and it is not appropriate to generate spatial clusters based on gene expression similarity [16]. Current research [86] also indicates that chromosome accessibility feature is not a powerful predictor for gene expression at the cell level. 2. The vast data volume in atlas-level studies challenges high-performance computing, risking out-of-memory or time-out errors [58, 101]. With nearly 37.2 trillion cells in the human body [77], comprehensive analysis is computationally infeasible. 3. Batch effects may adversely impact analysis results [54] by introducing noise. Consequently, an efficient and powerful method focusing on multi-omics and multi-tissue data (referred to as multimodal biological data) analysis is urgently needed to address these challenges.

Acknowledging the difficulties arising from the cell-oriented viewpoint, prior work shifted focus to the gene perspective. Using gene sets as a summary of expression profiles, based on natural selection during the species evolution, may provide a more robust anchor [8]. Protein-coding genes are also thought to interact with drugs [19], which are more relevant to diseases and drug discovery. Gene2vec [24] is a method inspired by Word2vec [18], which learns gene representations by generating skip-gram pairs from the co-expression network. Recently, the Gene-based data Integration and ANalysis Technique (GIANT) [16] has been developed, based on Node2vec [34] and OhmNet [119], to learn gene representations from both single-cell and spatial datasets. However, as shown in Figure 1 (b), significant functional clustering for genes from different datasets but the same tissue was not observed based on the gene embeddings from these two models, because these methods do not infer the similarity of associated genes from different multimodal data. Additionally, they did not offer metrics to quantitatively evaluate the performance of gene embeddings compared to baseline models.

Here we introduce a novel model called **Multimodal Similarity Learning Graph Neural Network (MuSe-GNN)**¹ for multimodal biological data integration from a gene-centric perspective. The overall workflow of MuSe-GNN is depicted in Figure 1 (a). Figure 1 (b) shows MuSe-GNN's superior ability to learn functional similarity among genes across datasets by suitable model structure and novel loss function, comparing to GIANT and Gene2vec. MuSe-GNN utilizes weight-sharing Graph Neural Networks (GNNs) to encode genes from different modalities into a shared space regularized by the similarity learning strategy and the contrastive learning strategy. At the single-graph level, the design of graph neural networks ensures that MuSe-GNN can learn the neighbor information in each co-expression network, thus preserving the gene function similarity. At the cross-data level, the similarity learning strategy ensures that MuSe-GNN can integrate genes with similar functions into a common group, while the contrastive learning strategy helps distinguish genes with different functions. Furthermore, MuSe-GNN utilizes more robust co-expression networks for training and applies dimensionality reduction [89] to high-dimensional data [73].

To the best of our knowledge, this is the first paper in gene representation learning that combines the Multimodal Machine Learning (MMML) concept [9, 27] with deep GNNs [49, 81] design. Both approaches are prevalent in state-of-the-art (SOTA) machine learning research, inspiring us to apply them to the joint analysis of large-scale multimodal biological datasets². As application examples, we first used the gene embeddings generated by MuSe-GNN to investigate crucial biological processes and causal networks for gene regulation in the human body. We also applied our model to analyze COVID and cancer datasets, aiming to unveil potential disease resistance mechanisms or complications based on specific differentially co-expressed genes. Lastly, we used gene embeddings from MuSe-GNN to improve the prediction accuracy for gene functions. Here genes are co-expressed means genes are connected with the same edge.

Given the lack of explicit metrics to evaluate gene embeddings, we proposed six metrics inspired by the batch effect correction problem [54] in single-cell data analysis. We evaluated our model using real-world biological datasets from one technique, and the benchmarking results demonstrated MuSe-GNN's significant improvement from **20.1%** to **97.5%** in comprehensive assessment. To summarize the advantages of our model, MuSe-GNN addresses the outlined problem about cross-data gene similarity learning and offers four major contributions: 1. Providing an effective representation learning approach for multi-structure biological data. 2. Integrating genes from different omics and tissue data into a joint space while preserving biological information. 3. Identifying co-located

¹Codes of MuSe-GNN: <https://github.com>HelloWorldLY/MuSe-GNN>

²Download links: Appendix M.

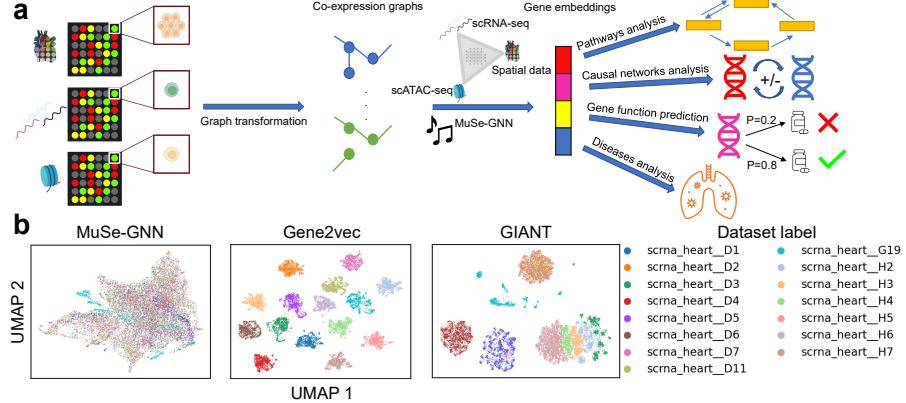


Figure 1: The workflow of MuSe-GNN, the visualization of gene embeddings, and the problems of two existing methods, GIANT and Gene2vec. **(a)** The process of learning gene embeddings by MuSe-GNN. Here we highlight the difference between single-cell data and spatial data, and the major applications of gene embeddings. Each spot from single-cell data represents one cell, while each spot from spatial data represents a mixture of cells. **(b)** UMAPs [63] for the gene embeddings of MuSe-GNN, Gene2vec and GIANT. They are colored by datasets. Based on UMAPs we can conclude that both Gene2vec and GIANT failed to learn the gene similarity based on the datasets from the same tissue, while MuSe-GNN produces meaningful embeddings agnostic to datasets.

genes with similar functions. 4. Inferring specialized causal networks of genes and the relationships between genes and biological pathways or between genes and diseases.

2 Related work

Co-expression Network Analysis. While direct attempts at joint analysis of gene functions across modalities are limited, there is relevant research in identifying correlation networks based on a single dataset. WGCNA [52] is a representative method that employs hierarchical clustering to identify gene modules with shared functions. However, as an early tool, WGCNA has restricted functionality. Its inference of co-expression networks necessitates more rigorous methods, and it cannot be directly applied to the analysis of multimodal biological data.

Network Based Biological Representation Learning. Apart from directly generating gene co-expression networks, learning quantitative gene representations in a lower dimensional space may better describe gene-gene relationships and facilitate downstream analyses. Gene2vec generates gene embeddings based on the co-expression network from a given database. However, it disregards the expression profile information and is based on the old Gene Expression Omnibus (GEO) up until 2017 [24]. GIANT leverages Node2vec [34] and OhmNet [119] to learn gene representations from both single-cell and spatial datasets by constructing graphs and hypergraphs for training. However, this approach still over-compresses multimodal biological datasets by removing expression profiles. Moreover, their co-expression networks created by Pearson correlation have a high false positive rate [82]. Additionally, some of the datasets used by GIANT are of low quality (see Appendix A). There are also methods sharing a similar objective of learning embeddings from other datasets. GSS [71] aims to learn a set of general representations for all genes from bulk-seq datasets [62] using Principal Component Analysis (PCA) and clustering analysis. However, it is for bulk-seq data and cannot be directly applied to single-cell datasets from different tissues. Gemini [103] focuses on integrating different protein function networks, with graph nodes representing proteins rather than genes.

Graph Transformer. In the deep learning domain, Transformer [95, 100] is one of the most successful models, leveraging seq2seq structure design, multi-head self-attention design, and position encoding design. Many researchers have sought to incorporate the advantages of Transformer to graph structure learning. TransformConv [81] introduces a multi-head attention mechanism to the graph version of supervised classification tasks and achieved significant improvements. MAGNA [98] considers higher-level hop relationships in graph data to enhance node classification capabilities. Graphomer [110] demonstrates the positive impact of the Transformer structure on various tasks

using data from the Open Graph Benchmark Large-Scale Challenge [44], which is further extended by GraphomerGD [111]. Recently, GPS [75] proposes a general Graph Transformer (GT) model by considering additional encoding types. Transformer architecture also contributes solutions to several biological questions [113]. scBERT [108] generates gene and cell embeddings using pre-training to improve the accuracy of cell type annotations. The substantial impact of these efforts highlights the crucial contribution of the Transformer architecture to graph data learning.

3 Methods

In the following sections of introducing MuSe-GNN, we will elaborate on our distinct approaches for graph construction that utilize multimodal biological data, followed by an explanation of our weight-sharing network architecture and the elements of the final loss function.

3.1 Preliminaries

GNN. GNNs aim to learn the graph representation of nodes (features) for data with a graph structure. Modern GNNs iteratively update the representation of a node by aggregating the representations of its k -order neighbors ($k \geq 1$) and combining them with the current representation. As described in [110], considering a graph $\bar{G} = \{V, E\}$ with nodes $V = \{v_1, v_2, \dots, v_n\}$, the AGGREGATE-COMBINE update for node v_i is defined as:

$$a_i^{(l+1)} = \text{AGGREGATE}^{(l+1)} \left(\left\{ h_j^{(l)} : j \in \mathcal{N}(v_i) \right\} \right); h_i^{(l+1)} = \text{COMBINE}^{(l+1)} \left(h_i^{(l)}, a_i^{(l+1)} \right), \quad (1)$$

where $\mathcal{N}(v_i)$ represents the neighbors of node v_i in the given graph, and $h_i^{(l)}$ and $h_i^{(l+1)}$ represent the node representation before and after updating, respectively.

Problem Definition. We address the gene embeddings generation task by handling multimodal biological datasets, denoted as $\mathcal{D} = (\{V_i, E_i\})_{i=1}^T$. Our goal is to construct a model $\mathcal{M}(\cdot, \theta)$, designed to yield gene embeddings set $\mathcal{E} = \{e_1, \dots, e_T\} = \mathcal{M}(\mathcal{D}, \theta)$. In this context, \mathcal{D} represents the input, θ represents the parameters, and \mathcal{E} represents the output. In other words, we aim to harmonize gene information from diverse modalities within a unified projection space, thereby generating consolidated gene representations.

3.2 Graph construction

Before constructing gene graphs, our first contribution involves the selection of highly variable genes (HVGs) for each dataset. These HVGs constitute a group of genes with high variance that can represent the biological functions of given expression profiles. Moreover, considering co-expression networks is important for gene representation learning because it allows us to characterize gene-gene relationships. As sequencing depth, or the total counts of each cell, often serves as a confounding factor in the co-expression networks inference [11], we employ two unique methodologies, scTransform [35] and CS-CORE [88], to process scRNA-seq and scATAC-seq data, thus creating gene expression profiles and co-expression networks unaffected by sequencing depth. For spatial transcriptomic data, our focus is on genes displaying spatial expression patterns. We use SPARK-X [116] to identify such genes and then apply scTransform and CS-CORE. For detailed algorithmic information regarding these methods, please see Appendix B. Additionally, we demonstrate the immunity of CS-CORE to batch effects when estimating gene-gene correlations in Appendix C. In all our generated graphs (equivalent to co-expression datasets), nodes represent **genes** and edges represent **co-expression** relation of genes.

3.3 Cross-Graph Transformer

To capitalize on the strengths of the Transformer model during our training process, we integrate a graph neural network featuring a multi-head self-attention design [81], called TransformerConv, to incorporate co-expression information and generate gene embeddings. Details of TransformerConv can be found in Appendix B.4. Incorporating multimodal information can estimate more accurate gene embeddings, supported by Appendix D. The cross-graph transformer can efficiently learn gene embeddings containing gene functions across different graphs, advocated by the comparison of different network structure choices in Appendix E.1.

Weight sharing. Given the variability among multimodal biological datasets, we employ a weight-sharing mechanism to ensure that our model learns shared information across different graphs, representing a novel approach for learning cross-graph relationships. We also highlight the importance of weight-sharing design in Appendix E.1.

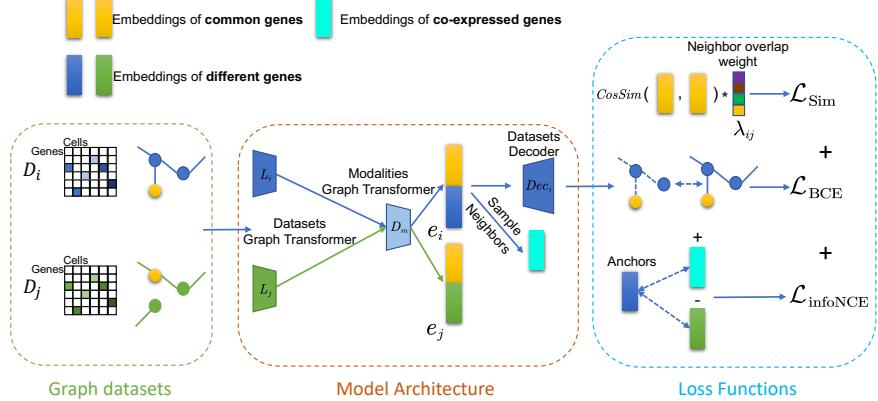


Figure 2: The overall model architecture and the design of loss functions for MuSe-GNN. The color of nodes in the green block represents common/different genes across two datasets. The brown block represents the network architecture of MuSe-GNN, and the blue block represents different loss function components of MuSe-GNN. The color gradients of the left two matrices represent different gene expression levels.

Datasets & Modalities Graph Transformer. Drawing inspiration from the hard weight-sharing procedure [15, 115], we not only employ dataset-specific Graph Transformer (GT) layers L_1, L_2, \dots, L_n for each graph (G_1, G_2, \dots, G_n) from the same modality m , but also connect all these dataset-specific layers to a set of shared GT layers, denoted as D_m . This design showcases our novel approach to incorporating weight-sharing into the GT framework. The forward process of MuSe-GNN, given dataset i with network parameter θ_* , is defined as follows:

$$X'_i = D_m(L_i(G_i; \theta_{L_i}); \theta_{D_m}). \quad (2)$$

Datasets Decoder. Here we propose a dataset-specific decoder structure based on Multi-layer Perceptrons (MLP). This decoder model is crucial in reconstructing the co-expression relationships among different genes, showcasing our inventive use of MLP for this purpose. Given a graph G_i and its corresponding gene embedding e_i , the decoding process of MuSe-GNN, with network parameter $\theta_{dec,i}$, is defined as follows:

$$E_{rec} = \text{DECODERMLP}(e_i e_i^T; \theta_{dec,i}), \quad (3)$$

where E_{rec} represents the reconstructed co-expression network.

3.4 Graph Reconstruction Loss (\mathcal{L}_{BCE})

Within a single graph, we implement a loss function inspired by the Graph Auto-encoder (GAE) [48]. This function is designed to preserve two key aspects: 1. the similarity among genes sharing common functions, and 2. the distinctions among genes with differing functions. This innovative use of a GAE-inspired loss function constitutes a significant contribution to the methodological design. For a graph $G_i = \{V_i, E_i\}$, the loss function for edge reconstruction is defined as:

$$\begin{aligned} e_i &= \text{ENCODERGNN}(\{V_i, E_i\}; \theta_{enc}) = D_m(L_i(G_i; \theta_{L_i}); \theta_{D_m}), \\ E_{rec} &= \text{DECODERMLP}(e_i e_i^T; \theta_{dec,i}); E'_{rec} = \text{FLATTEN}(E_{rec}); E'_i = \text{FLATTEN}(E_i), \\ \mathcal{L}_{\text{BCE}} &= -\frac{1}{|E'_{rec}|} \sum_{t=1}^{|E'_{rec}|} [E'_i[t] \cdot \log E'_{rec}[t] + (1 - E'_i[t]) \cdot \log (1 - E'_{rec}[t])], \end{aligned} \quad (4)$$

where ENCODERGNN and DECODERMLP represent the encoder and decoder parts of our model, respectively. \mathcal{L}_{BCE} denotes the computation of binary cross entropy loss (BCELoss) for the given input data. E_{rec} is the reconstructed adjacency matrix and $|E'_{rec}|$ is the length of its flatten version. Further justification of the model design can be found in Appendix F.

3.5 Weighted-similarity Learning (\mathcal{L}_{Sim})

To integrate shared biological information across graphs from disparate datasets, we fuse the reconstruction loss of the input graph structure with a cosine similarity learning loss. In this process, we treat common HVGs between each pair of datasets as anchors. Our aim is to maximize the cosine similarity ($\text{COSSIM}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \cdot \|\mathbf{b}\|_2}$) for each common HVG across two arbitrary datasets (represented by the yellow blocks in Figure 2), utilizing their gene embeddings. However, in practice, different common HVGs may have different levels of functional similarity in two datasets, which is hard to be directly quantified. Consequently, we employ the shared community score as an indirect measurement, which is incorporated as a weight for the cosine similarity of different common HVG pairs within the final loss function. Considering two graphs $G_i = \{V_i, E_i\}$ and $G_j = \{V_j, E_j\}$, and one of their shared genes g , we identify the co-expressed genes of the given gene g in both datasets, denoted as N_{ig} and N_{jg} . Thus, the weight λ_{ijg} for gene g can be expressed as follows:

$$\lambda_{ijg} = \frac{|N_{ig} \cap N_{jg}|}{|N_{ig} \cup N_{jg}|}. \quad (5)$$

We can iterate over all shared genes from 1 to n between these two graphs, ultimately yielding a vector as $\lambda_{ij} = [\lambda_{ij1}, \dots, \lambda_{ijn}]$. This vector encapsulates the community similarity between the two graphs across all common HVGs. We can then modify the cosine similarity of various gene pairs by first multiplying this vector with cosine similarities and then summing the resultant values across all genes. The negation of the outcome is our final weighted similarity loss, denoted as \mathcal{L}_{Sim} . The ablation test for weighed similarity loss can be found in Appendix E.1. More detailed explanations and evidence supporting this design in multimodal conditions can be found in Appendix G.

3.6 Self-supervised Graph Contrastive Learning ($\mathcal{L}_{\text{InfoNCE}}$)

Specifically, when integrating multimodal biological data, we employ the contrastive learning strategy [72] to ensure that functionally similar genes are clustered together as closely as possible, while functionally different genes are separated apart from each other. We utilize Information Noise Contrastive Estimation (InfoNCE) as a part of our loss function to maximize the mutual information between the anchor genes and genes with the same functions. This loss is applicable to different genes in two arbitrary graphs during the training process. In general, if we represent the embeddings of N genes as $\text{Gene}_N = \{e_1, \dots, e_N\}$, InfoNCE is designed to minimize:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E} \left[\log \frac{\exp(e \cdot k_+ / \tau)}{\sum_{i=0}^K \exp(e \cdot k_i / \tau)} \right], \quad (6)$$

where samples $\{k_0, k_1, k_2, \dots\}$ compose a set of gene embeddings known as keys of one dictionary and e is a query gene embedding. k_+ represents a positive sample of e and k_i denotes a negative sample of e . Equation 6 can be interpreted as a log loss of a $(K + 1)$ -way Softmax classifier, which attempts to classify e as k_+ . τ is a temperature parameter. τ is set to 0.07 referenced in MoCo [41].

3.7 Final Loss Function

In summary, the training objective of MuSe-GNN for graph i comprises three components:

$$\begin{aligned} & \min_{e_i, e_j} \mathcal{L}_{\text{BCE}}(\text{DECODERMLP}_i(e_i e_i^T), E_i) - \mathbb{E} [\text{COSSIM}(e_i[\text{Common}_{ij}], e_j[\text{Common}_{ij}]) \lambda_{ij}^T] \\ & + \lambda_c \mathcal{L}_{\text{InfoNCE}}(e_i[\text{Diff}_i] \oplus e_j[\text{Diff}_j], e_i[\text{Diff}_{\mathcal{N}(i)}] \oplus e_j[\text{Diff}_{\mathcal{N}(j)}]), \end{aligned} \quad (7)$$

where Common_{ij} denotes the index set for common HVGs, Diff_i represents the index set for different HVGs in graph i , and $\text{Diff}_{\mathcal{N}(i)}$ indicates the index set for neighbors of Diff_i in graph i . To expedite the training process and conserve memory usage, we sample graph j for each graph i during model training. We also employ multi-thread programming to accelerate the index set extraction process. λ_c is the weight for InfoNCE loss. All of the components in MuSe-GNN are supported by ablation experiments in Appendix E.1. Details of hyper-parameter tuning can be found in Appendix E.2.

4 Experiments

Datasets & Embeddings generation. Information on the different datasets used for different experiments is included at the beginning of each paragraph. The training algorithm of our model

Algorithm 1 Multimodal Similarity Learning Graph Neural Network (MuSe-GNN)

Input: Model ENCODERGNN, DECODERMLP; Dataset $\mathcal{D} = (\{V_i, E_i\})_{i=1}^T$; Number of epochs K ; Neighbor overlap list λ ; Weight coefficient for contrastive learning λ_c ;

Helper Functions: A function to find the genes with the same name FINDCOMMONGENES; A function to find the genes with different name FINDDIFFGENES; A function to calculate cosine similarity COSSIM; A function to sample data from a graph collection SAMPLE; A optimizer function to update weights ADAM.

Output: Model ENCODERGNN

- 1: INIT: initialize all parameters.
- 2: **for** s in K steps **do**
- 3: **for** id, $\{V, E\}$ in enumerate(\mathcal{D}) **do**
- 4: $e \leftarrow \text{ENCODERGNN}(\{V, E\}; \theta_{enc})$; $E_{rec} \leftarrow \text{DECODERMLP}(ee^T; \theta_{dec})$
- 5: $\mathcal{L} \leftarrow \mathcal{L}_{\text{BCE}}(E_{rec}, E)$
- 6: $\text{id}_{\text{new}}, \{V_{\text{new}}, E_{\text{new}}\} \leftarrow \text{SAMPLE}(G \setminus \{V, E\})$
- 7: $e_{\text{new}} \leftarrow \text{ENCODERGNN}(\{V_{\text{new}}, E_{\text{new}}\}; \theta_{enc})$
- 8: $e_{\text{Common}}, e'_{\text{Common}} \leftarrow \text{FINDCOMMONGENES}(e, e_{\text{new}})$
- 9: $\mathcal{L} \leftarrow \mathcal{L} - \text{COSSIM}(e_{\text{Common}}, e'_{\text{Common}}) \lambda_{\text{id}, \text{id}_{\text{new}}}^T$
- 10: $e_{\text{Diff}}, e'_{\text{Diff}}, e_{\text{Diff}_N}, e'_{\text{Diff}_N} \leftarrow \text{FINDDIFFGENES}(e, e_{\text{new}})$
- 11: $\mathcal{L} \leftarrow \mathcal{L} + \lambda_c \mathcal{L}_{\text{InfoNCE}}(e_{\text{Diff}} \oplus e_{\text{Diff}_N}, e'_{\text{Diff}} \oplus e'_{\text{Diff}_N})$
- 12: $\theta_{enc} \leftarrow \text{ADAM}(\mathcal{L}, \theta_{enc})$; $\theta_{dec} \leftarrow \text{ADAM}(\mathcal{L}, \theta_{dec})$
- 13: **Return** ENCODERGNN

is outlined in Algorithm 1. We stored the gene embeddings in the AnnData structure provided by Scanpy [104]. To identify groups of genes with similar functions, we applied the Leiden clustering algorithm [91] to the obtained gene embeddings. Details can be found in Appendix E.4.

Evaluation metrics. For the benchmarking process, we used six metrics: edge AUC (AUC) [74], common genes ASW (ASW) [59], common genes graph connectivity (GC) [59], common genes iLISI (iLISI) [59], common genes ratio (CGR), and neighbors overlap (NO) to provide a comprehensive comparison. Detailed descriptions of these metrics can be found in Appendix H. We computed the metrics for the different methods and calculated the average rank (Avg Rank $\in [1, 9]$). Moreover, to evaluate the model performance improvement, we applied min-max scaling to every metric across different models and computed the average score (Avg Score $\in [0, 1]$).

Baselines. We selected eight models as competitors for MuSe-GNN. The first group of methods stems from previous work on learning embeddings for biomedical data, including Principal Component Analysis (PCA) [74] used by GSS, Gene2vec, GIANT (the SOTA model for gene representation learning), Weight-sharing Multi-layer Auto-encoder (WSMAE) used by [109] and scBERT (the SOTA model with pre-training for cell type annotation). The second group of methods comprises common unsupervised learning baseline models, including GAE, VGAE [48] and Masked Auto-encoder (MAE) (the SOTA model for self-supervised learning) [42]. MuSe-GNN, GIANT, GAE, VGAE, WSMAE and MAE have training parameter sizes between 52.5 and 349 M, and all models are tuned to their best performance. Details are shown in Appendix E.2.

Biological applications. For the pathway analysis, we used Gene Ontology Enrichment Analysis (GOEA) [1] to identify specific biological pathways enriched in distinct gene clusters with common functions. Moreover, we used Ingenuity Pathway Analysis (IPA) [2] to extract biological information from the genes within various clusters, including causal networks [50] and sets of diseases and biological functions. Biological pathways refer to processes identified based on the co-occurrence of genes within a particular cluster. The causal network depicts the relationships between regulatory genes and their target genes. Disease and biological function sets facilitate the discovery of key processes and complications associated with specific diseases. Using gene embeddings also improves the performance of models for gene function prediction. To visualize gene embeddings in a low-dimensional space, we utilized Uniform Manifold Approximation and Projection (UMAP) [63].

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

4.1 Benchmarking Analysis

We executed each method 10 times by using the same setting of seeds to show the statistical significance based on datasets across different tissues. The performance comparison of nine gene embedding methods is presented in Tables 1 and 14. Based on these two tables, MuSe-GNN outperformed its competitors in terms of both average ranks and average scores across all the tissues. Based on Table 1, for major tissues, such as heart and lung, MuSe-GNN’s performance was 20.1% higher than the second-best method and 97.5% higher than the second-best method in heart and lung tissue, respectively. According to Appendix E.3, MuSe-GNN’s stability was also demonstrated through various metrics by comparing standard deviations, including AUC, GC, and NO. In contrast, methods such as Gene2vec, GAE, VGAE, MAE and scBERT exhibited significant instability in their evaluation results for kidney or thymus. Consequently, we concluded that MuSe-GNN is the best performing model for learning gene representation based on datasets from different tissues, making it applicable to learn gene embeddings from diverse multimodal biological data.

4.2 Analysis of Gene Embeddings from Multimodal Biological Data

In Figure 3, we displayed the integration results for multimodal biological data from Humans. Figure 3 (a) and (b) demonstrated that MuSe-GNN could successfully integrate genes from different modalities into a co-embedded space, allowing us to identify functional groups using the Leiden algorithm shown in Figure 3 (c). Furthermore, Figure 3 (d) revealed that most of the clusters in (c) were shared across different modalities. We also identified three significant functional groups in Figure 3 (a): the nervous system (predominantly composed of the cerebrum and cerebellum [26]), the cardiovascular system (mainly composed of heart, lung, and kidney [29]), and the immunology system (primarily consisting of spleen, liver, and peripheral blood mononuclear cells (PBMC) [70]). All systems are important in regulating the life activities of the body. We also uncovered a pre-epigenetics group (mainly consisting of scATAC-seq data without imprinting, modification, or editing), emphasizing the biological gap existing in multi-omics and the importance of post-transcriptional regulation.

Using GOEA, we could identify significant pathways enriched by different co-embedded gene clusters. For instance, Figure 3 (d) displayed the top 5 pathways in an immunology system cluster. The rank was calculated based on the negative logarithm of the false discovery rate. Since all top pathways were related to immunological defense and response, it further supported the accuracy of our embeddings in representing gene functions. For our analysis of shared transcription factors and major pathways across all the tissues, please refer to Appendix I. For our analysis of multi-species gene embeddings, please refer to Appendix J.

4.3 Analysis of Gene Embeddings for Diseases

We generated gene embeddings for human pancreas cells from samples with and without SARS-CoV-2 infection, as depicted in Figure 4 (a) and (b). We identified specific genes from COVID samples that did not align with control samples, which piqued our interest. These genes, highlighted by a red circle in Figure 4 (c), could be interpreted as differentially functional genes in diseased cells.

We conducted GOEA for the genes of interest and discovered a close relationship among these gene enrichment results and the top 5 pathways associated with immune activity. These results are displayed in Figure 4 (d). For the genes within our target cluster, we utilized IPA to identify the Entrez name of these genes, and 90.3% (122/135) genes in our cluster are related to immunoglobulin.

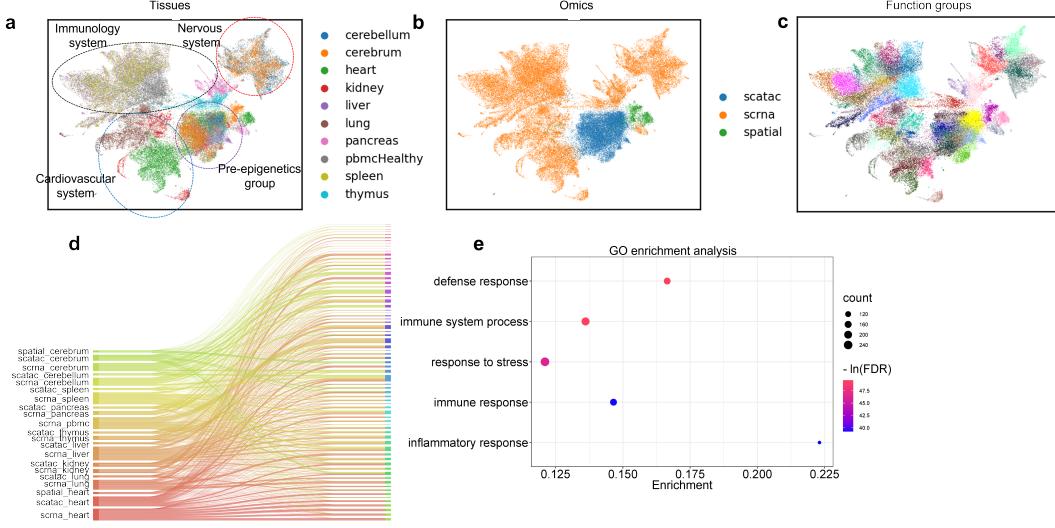


Figure 3: Gene representation learning results for multimodal biological data. (a) represents the UMAPs of gene embeddings colored by tissue type and highlighted by biological system. (b) represents the UMAPs of gene embeddings colored by omics type. (c) represents the UMAPs of gene embeddings colored by common function groups. (d) is a Sankey plot [80] to show the overlap of different modalities in the same clusters. (e) shows the top5 pathways related to the genes in the special cluster discovered by GOEA. The bubble plots in this paper were created based on ggplot2 [102].

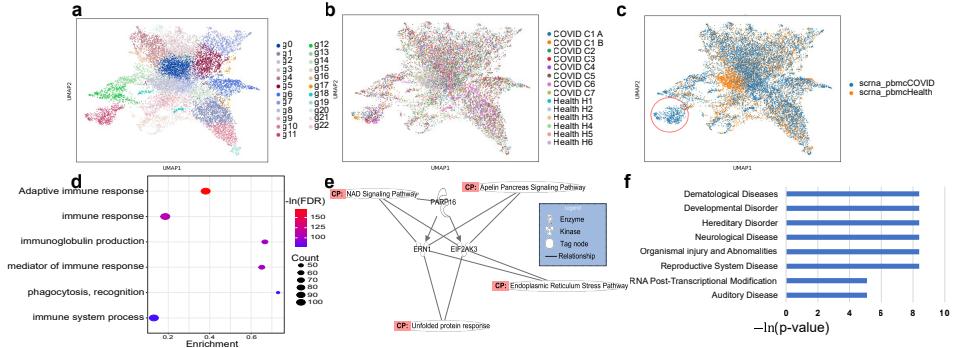


Figure 4: Gene embeddings from COVID samples and healthy samples. (a) represents the UMAPs of gene embeddings colored by functional groups. (b) represents the UMAPs of gene embeddings colored by datasets. (c) represents the gene embeddings colored by the conditions, and the red circle reflects the differential co-expression genes. (d) shows the top6 pathways related to the genes in the special cluster discovered by GOEA. (e) represents the causal network existing in the special cluster discovered by IPA. (f) represents the top diseases & biological functions discovered by IPA.

We could also infer the causal relationship existing in the gene regulatory activity of the immune system. For example, Figure 4 (e) showed a causal network inferred by IPA based on our genes cluster. PARP16, as an enzyme, can regulate ERN1 and EIF2AK3, and certain pathways are also related to this causal network. Moreover, we also showed the relation between the set of genes and Disease & Bio functions in Figure 4 (f). We identified top related Diseases & Bio functions ranked by negative logarithm of p-value, and all of these diseases could be interpreted as complications that may arise from new coronavirus infection [6, 64, 79, 69, 33, 7, 83, 84]. Our extra analyses for lung cancer data can be found in Appendix K.

4.4 Analysis of Gene Embeddings for Gene Function Prediction.

Here we intend to predict the dosage-sensitivity of genes related to genetic diagnosis (as dosage-sensitive or not) [90]. We used MuSe-GNN to generate gene embeddings for different datasets based on an unsupervised learning framework and utilized the gene embeddings as training dataset to predict the function of genes based on k-NN classifier. k-NN classifier is a very naive model and can reflect the contribution of gene embeddings in the prediction task.

In this task, we evaluated the performance of MuSe-GNN based on the dataset used in Geneformer [90, 31], comparing it to the prediction results based on raw data or Geneformer (total supervised learning). As shown in Table 2, the prediction accuracy based on gene emebddings from MuSe-GNN is the highest one. Moreover, the performance of gene embeddings from MuSe-GNN is better than Geneformer, which is a totally supervised learning model. Such finding proves the advantages of MuSe-GNN in the application of gene function prediction task. Further application analysis can be found in Appendix L.

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

5 Conclusion

In this paper, we introduce MuSe-GNN, a model based on Multimodal Machine Learning and Deep Graph Neural Networks, for learning gene embeddings from multi-context sequencing profiles. Through experiments on various multimodal biological datasets, we demonstrate that MuSe-GNN outperforms current gene embedding learning models across different metrics and can effectively learn the functional similarity of genes across tissues and techniques. Moreover, we performed various biological analyses using the learned gene embeddings, leveraging the capabilities of GOEA and IPA, such as identifying significant pathways, detecting diseases, and inferring causal networks. Our model can also contribute to the study of the pathogenic mechanisms of diseases like COVID and lung cancer, and improve the prediction performance for gene functions. Overall, the gene representations learned by MuSe-GNN are highly versatile and can be applied to different analysis frameworks.

At present, MuSe-GNN does not accept graphs with nodes other than genes as input. In the future, we plan to explore more efficient approaches for training large models related to Multimodal Machine Learning and extend MuSe-GNN to a more general version capable of handling a broader range of multimodal biological data.

6 Acknowledgements

This research was supported in part by NIH R01 GM134005, R56 AG074015, and NSF grant DMS 1902903 to H.Z. We appreciate the comments, feedback, and suggestions from Chang Su, Zichun Xu, Xinning Shan, Yuhang Xie, Mingze Dong, and Maria Brbic.

References

- [1] Gene ontology enrichment analysis. URL <http://geneontology.org>.
- [2] Ingenuity pathway analysis. URL <https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis>.
- [3] 10x Genomics. 10x genomics acquires spatial transcriptomics. *Science*, 2018. URL <https://www.10xgenomics.com/news/10x-genomics-acquires-spatial-transcriptomics>.
- [4] Adrian M Altenhoff, Clément-Marie Train, Kimberly J Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, Hale-Seda Radoykova, Victor Rossier, Alex

- Warwick Vesztrocy, et al. Oma orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic acids research*, 49(D1):D373–D379, 2021.
- [5] Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7, 2018.
 - [6] Khashayar Aram, Anant Patil, Mohamad Goldust, and Fateme Rajabi. Covid-19 and exacerbation of dermatological diseases: a review of the available literature. *Dermatologic Therapy*, 34(6):e15113, 2021.
 - [7] Arash Ardestani Zadeh and Davood Arab. Covid-19 and male reproductive system: pathogenic features and possible mechanisms. *Journal of molecular histology*, 52:869–878, 2021.
 - [8] Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nature biotechnology*, 39(10):1202–1215, 2021.
 - [9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.
 - [10] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.
 - [11] A Sina Booeshaghi, Ingileif B Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data. *bioRxiv*, pages 2022–05, 2022.
 - [12] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexis, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.
 - [13] Tianle Cai, Shengjie Luo, Keyulu Xu, Di He, Tie-yan Liu, and Liwei Wang. Graphnorm: A principled approach to accelerating graph neural network training. In *International Conference on Machine Learning*, pages 1204–1215. PMLR, 2021.
 - [14] Junyue Cao, Diana R O’Day, Hannah A Pliner, Paul D Kingsley, Mei Deng, Riza M Daza, Michael A Zager, Kimberly A Aldinger, Ronnie Blecher-Gonen, Fan Zhang, et al. A human cell atlas of fetal gene expression. *Science*, 370(6518):eaba7721, 2020.
 - [15] R Caruana. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Citeseer, 1993.
 - [16] Hao Chen, Nam D Nguyen, Matthew Ruffalo, and Ziv Bar-Joseph. A unified analysis of atlas single cell data. *bioRxiv*, 2022.
 - [17] Xi Chen, Ricardo J Miragaia, Kedar Nath Natarajan, and Sarah A Teichmann. A rapid and robust method for single cell chromatin accessibility profiling. *Nature communications*, 9(1):1–9, 2018.
 - [18] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.
 - [19] Kelsy C Cotto, Alex H Wagner, Yang-Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, and Malachi Griffith. Dgidb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic acids research*, 46(D1):D1068–D1073, 2018.
 - [20] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.
 - [21] C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabI5197, 2022.

- [22] Mingze Dong and Yuval Kluger. Towards understanding and reducing graph structural noise for gnns. 2023.
- [23] Rui Dong and Guo-Cheng Yuan. Spatialdwls: accurate deconvolution of spatial transcriptomic data. *Genome biology*, 22(1):145, 2021.
- [24] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(1):7–15, 2019.
- [25] Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. 2021. URL <https://openreview.net/forum?id=pW2Q2xLwIMD>.
- [26] Ramazan Durmaz, Metin Ant Atasoy, Gü̈l Durmaz, Baki Adapinar, Ali Arslantaş, Aydin Aydinli, and Esref Tel. Multiple nocardial abscesses of cerebrum, cerebellum and spinal cord, causing quadriplegia. *Clinical neurology and neurosurgery*, 103(1):59–62, 2001.
- [27] Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, pages 1–11, 2023.
- [28] HuBMAP Consortium et al. The human body at cellular resolution: the nih human biomolecular atlas program. *Nature*, 574(7777):187–192, 2019.
- [29] HE Fessler. Heart-lung interactions: applications in the critically ill. *European Respiratory Journal*, 10(1):226–237, 1997.
- [30] David S Fischer, Anna K Fiedler, Eric M Kernfeld, Ryan MJ Genga, Aimée Bastidas-Ponce, Mostafa Bakhti, Heiko Lickert, Jan Hasenauer, Rene Maehr, and Fabian J Theis. Inferring population dynamics from single-cell rna-sequencing time series data. *Nature biotechnology*, 37(4):461–468, 2019.
- [31] Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019:baz046, 2019.
- [32] Toni Gabaldón and Eugene V Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366, 2013.
- [33] Suhita Gayen Nee’Betal, Pedro Urday, Huda B Al-Kouatly, Kolawole Solarin, Joanna SY Chan, Sankar Addya, Rupsa C Boelig, and Zubair H Aghai. Covid-19 infection during pregnancy induces differential gene expression in human cord blood cells from term neonates. *Frontiers in Pediatrics*, 10:547, 2022.
- [34] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
- [35] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.
- [36] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [37] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.
- [38] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, et al. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

- [39] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.
- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [43] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022.
- [44] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.
- [45] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.
- [46] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.
- [47] Grigory Khromov and Sidak Pal Singh. Some fundamental aspects about lipschitz continuity of neural network functions. *arXiv preprint arXiv:2302.10886*, 2023.
- [48] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, 2016.
- [49] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=SJU4ayYg1>.
- [50] Andreas Krämer, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.
- [51] M Krassowski. Complexupset: Create complex upset plots using ggplot2 components. *R package version 0.5*, 18, 2020.
- [52] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.
- [53] Patty J Lee, Philip Blood, Katy Börner, Judith Campisi, Feng Chen, Heike Daldrup-Link, Phil De Jager, Li Ding, Francesca E Duncan, Oliver Eickelberg, et al. Nih sennet consortium: Mapping senescent cells in the human body to understand health and disease, 2022.
- [54] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.
- [55] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.
- [56] Xiang Lin, Tian Tian, Zhi Wei, and Hakon Hakonarson. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature communications*, 13(1):7705, 2022.

- [57] Yinyin Liu, Janusz A Starzyk, and Zhen Zhu. Optimized approximation algorithm in neural networks without overfitting. *IEEE transactions on neural networks*, 19(6):983–995, 2008.
- [58] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- [59] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.
- [60] Chongyuan Luo, Christopher L Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R Nery, Justin P Sandoval, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, 2017.
- [61] Kaili Ma, Haochen Yang, Han Yang, Tatiana Jin, Pengfei Chen, Yongqiang Chen, Barakeel Fanseu Kamhoua, and James Cheng. Improving graph representation learning by contrastive regularization. *arXiv preprint arXiv:2101.11525*, 2021.
- [62] Samuel Marguerat and Jürg Bähler. Rna-seq: from technology to biology. *Cellular and molecular life sciences*, 67:569–579, 2010.
- [63] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL <https://doi.org/10.21105/joss.00861>.
- [64] Bekir Fatih Meral. Parental views of families of children with autism spectrum disorder and developmental disorders during the covid-19 pandemic. *Journal of Autism and Developmental Disorders*, 52(4):1712–1724, 2022.
- [65] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.
- [66] Yujie Mo, Liang Peng, Jie Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Simple unsupervised graph representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7797–7805, Jun. 2022. doi: 10.1609/aaai.v36i7.20748. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20748>.
- [67] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [68] Ryan M Mulqueen, Dmitry Pokholok, Steven J Norberg, Kristof A Torkenczy, Andrew J Fields, Duanchen Sun, John R Sinnamon, Jay Shendure, Cole Trapnell, Brian J O’Roak, et al. Highly scalable generation of dna methylation profiles in single cells. *Nature biotechnology*, 36(5):428–431, 2018.
- [69] Edward J Needham, Sherry H-Y Chou, Alasdair J Coles, and David K Menon. Neurological implications of covid-19 infections. *Neurocritical care*, 32:667–671, 2020.
- [70] Rana Nikzad, Laura S Angelo, Kevin Aviles-Padilla, Duy T Le, Vipul K Singh, Lynn Bimler, Milica Vukmanovic-Stejic, Elena Vendrame, Thanmayi Ranganath, Laura Simpson, et al. Human natural killer cells mediate adaptive immunity to viral antigens. *Science immunology*, 4(35):eaat8116, 2019.
- [71] Sehyun Oh, Ludwig Geistlinger, Marcel Ramos, Daniel Blankenberg, Marius van den Beek, Jaclyn N Taroni, Vincent J Carey, Casey S Greene, Levi Waldron, and Sean Davis. Genomicsupersignature facilitates interpretation of rna-seq experiments through robust, efficient comparison to public databases. *Nature Communications*, 13(1):3695, 2022.
- [72] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

- [73] Subarna Palit, Christoph Heuser, Gustavo P De Almeida, Fabian J Theis, and Christina E Zielinski. Meeting the challenges of high-dimensional single-cell data analysis in immunology. *Frontiers in immunology*, 10:1515, 2019.
- [74] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [75] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.
- [76] Yanay Rosen, Maria Brbić, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec. Towards universal cell embeddings: Integrating single-cell rna-seq datasets across species with saturn. *Biorxiv: the Preprint Server for Biology*, 2023.
- [77] Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.
- [78] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.
- [79] Amanda Salih, Aaron Chin, Manisha Gandhi, Amir Shamshirsaz, Hennie Lombaard, and Joud Hajjar. Hereditary angioedema and covid-19 during pregnancy: Two case reports. *The Journal of Allergy and Clinical Immunology: In Practice*, 11(3):961–962, 2023.
- [80] Pierre Sassoulas, Anazalea, and Marcomanz. pySankey, 2018. URL <https://github.com/anazalea/pySankey>.
- [81] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1548–1554. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/214. URL <https://doi.org/10.24963/ijcai.2021/214>. Main Track.
- [82] Michael A Skinnider, Jordan W Squair, and Leonard J Foster. Evaluating measures of association for single-cell transcriptomics. *Nature methods*, 16(5):381–386, 2019.
- [83] Rajneesh Srivastava, Swapna Vidhur Daulatabad, Mansi Srivastava, and Sarath Chandra Janga. Role of sars-cov-2 in altering the rna-binding protein and mirna-directed post-transcriptional regulatory networks in humans. *International journal of molecular sciences*, 21(19):7090, 2020.
- [84] Won Sriwijitalai and Viroj Wiwanitkit. Hearing loss and covid-19: a note. *American journal of otolaryngology*, 2020.
- [85] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.
- [86] BJ Strober, Reem Elorbany, K Rhodes, Nirmal Krishnan, Karl Tayeb, Alexis Battle, and Yoav Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019.
- [87] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexis, William M Mauck, Yuhao Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
- [88] Chang Su, Zichun Xu, Xinning Shan, Biao Cai, Hongyu Zhao, and Jingfei Zhang. Cell-type-specific co-expression inference from single cell rna-sequencing data. *Nature Communications*, 14(1):4846, 2023.

- [89] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of machine learning research*, 8(5), 2007.
- [90] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, pages 1–9, 2023.
- [91] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.
- [92] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.
- [93] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.
- [94] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.
- [95] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [96] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=rJXMpikCZ>.
- [97] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.
- [98] Guangtao Wang, Rex Ying, Jing Huang, and Jure Leskovec. Multi-hop attention graph neural networks. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3089–3096. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/425. URL <https://doi.org/10.24963/ijcai.2021/425>. Main Track.
- [99] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1–11, 2021.
- [100] Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. VD-BERT: A Unified Vision and Dialog Transformer with BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3325–3338, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.269. URL <https://aclanthology.org/2020.emnlp-main.269>.
- [101] Yuge Wang, Tianyu Liu, and Hongyu Zhao. Respan: a powerful batch correction model for scRNA-seq data through residual adversarial networks. *Bioinformatics*, 38(16):3942–3949, 2022.
- [102] Hadley Wickham. ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 3(2):180–185, 2011.
- [103] Addie Woicik, Mingxin Zhang, Hanwen Xu, Sara Mostafavi, and Sheng Wang. Gemini: memory-efficient integration of hundreds of gene networks with high-order pooling. *Bioinformatics*, 39, 06 2023. doi: 10.1093/bioinformatics/btad247.

- [104] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.
- [105] Renjie Xie, Wei Xu, Yanzhi Chen, Jiabao Yu, Aiqun Hu, Derrick Wing Kwan Ng, and A. Lee Swindlehurst. A generalizable model-and-data driven approach for open-set rff authentication. *IEEE Transactions on Information Forensics and Security*, 16:4435–4450, 2021. doi: 10.1109/TIFS.2021.3106166.
- [106] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=ryGs6iA5Km>.
- [107] Mingcong Xu, Xuefeng Bai, Bo Ai, Guorui Zhang, Chao Song, Jun Zhao, Yuezhu Wang, Ling Wei, Fengcui Qian, Yanyu Li, et al. Tf-marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. *Nucleic acids research*, 50(D1):D402–D412, 2022.
- [108] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.
- [109] Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalam, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, GV Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature communications*, 12(1):1–10, 2021.
- [110] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.
- [111] Bohang Zhang, Shengjie Luo, Liwei Wang, and Di He. Rethinking the expressive power of GNNs via graph biconnectivity. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=r9hNv76KoT3>.
- [112] Di Zhang, Yanxiang Deng, Petra Kukanja, Eneritz Agirre, Marek Bartosovic, Mingze Dong, Cong Ma, Sai Ma, Graham Su, Shuzhen Bao, et al. Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature*, pages 1–10, 2023.
- [113] Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, and Wanwen Zeng. Applications of transformer-based language models in bioinformatics: A survey. *Bioinformatics Advances*, 2023.
- [114] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.
- [115] Kaixiong Zhou, Xiao Huang, Qingquan Song, Rui Chen, and Xia Hu. Auto-gnn: Neural architecture search of graph neural networks. *Frontiers in Big Data*, 5, 2022. ISSN 2624-909X. doi: 10.3389/fdata.2022.1029307. URL <https://www.frontiersin.org/articles/10.3389/fdata.2022.1029307>.
- [116] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):1–25, 2021.
- [117] Jiong Zhu, Ryan A. Rossi, Anup Rao, Tung Mai, Nedim Lipka, Nesreen K. Ahmed, and Danai Koutra. Graph neural networks with heterophily. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11168–11176, May 2021. doi: 10.1609/aaai.v35i12.17332. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17332>.
- [118] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.
- [119] Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017.

Appendix

Table of Contents

A GIANT’s selection of datasets	18
B Constructing co-expression network and training datasets	19
B.1 CS-CORE	20
B.2 scTransform	20
B.3 SPARK-X	20
B.4 Transformconv	21
C Differential co-expression analysis	22
D Theory analysis for multimodal machine learning	23
E Extra experiments, ablation test and model selection	24
E.1 Ablation tests	24
E.2 Hyper-parameter tuning	26
E.3 Details about benchmarking experiments	27
E.4 Details about function cluster identification	28
F Extra Discussion about Model Design	28
G Selecting anchors of common functional genes	33
H Metrics details	34
I Shared transcription factors and pathways analysis	35
J Multi-species gene embeddings	37
K Lung cancer analysis	38
L Gene Function Analysis	40
M Datasets information	41

A GIANT’s selection of datasets

The quality of datasets, as depicted in Figure 5, impacts the conclusions derived from gene embedding learning methods. GIANT employed low-quality spatial transcriptomic and scATAC-seq data, which ideally should be circumvented at the research inception. The spatial data, HBM477.BPMF.855 and HBM983.ZLKS.328, exhibit considerable barcode expression measurement loss. Under normal circumstances, gene reads distribution in sectioned samples should form complete circles instead of truncated patterns. The scATAC-seq datasets, HBM737.XZCH.767 and HBM793.TLNJ.249, present challenges in discerning differences in gene expression assays across distinct cells, as evidenced by the UMAP visualization results. In an ideal scenario, cells marked with varying colors should occupy different positions in the low-dimensional representation. HBM628.XTDP.423 and HBM292.XZZJ.243 serve as examples of high-quality datasets.

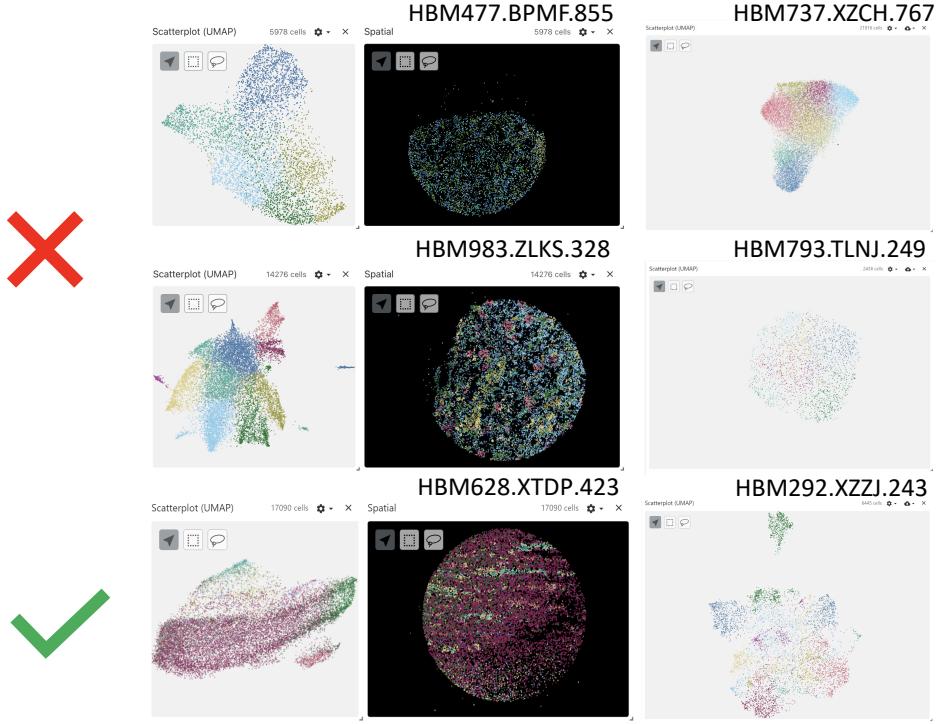


Figure 5: Problematic datasets selected by GIANT.

B Constructing co-expression network and training datasets

In this section, we outline the algorithmic details of preprocessing steps and network construction using CS-CORE, scTransform, and SPARK-X. Numerous databases exist for atlas-scale transcriptomic and epigenomic data analysis, including the Human BioMolecular Atlas Program (HuBMAP) [28], the Human Cell Atlas (HCA) [77], the 20 Cellular Senescence Network (SenNet) [53], and more. We gather Single-cell RNA sequencing (scRNA-seq) from HCA datasets, [14] and [76], Single-cell sequencing Assay for Transposase-accessible Chromatin (scATAC-seq) from HuBMAP datasets and [14], and high-quality spatial data from [3, 112]. For distinct omics data, we implement one common and one specific process.

In the common process, we filter barcodes with gene expression counts below 200 and genes with expression counts below three in each barcode. We also filter Mitochondrial (MT) genes. For scRNA-seq datasets, we employ scTransform [35] to select HVGs and generate Pearson residuals, replacing raw expression with residuals. scTransform is the first model to incorporate sequencing depth as a covariate rather than directly applying the size factor to normalization. It eliminates confounding caused by sequencing depth in raw single-cell or spatial expression data, generating corrected gene expression profiles. These advantages make it a widely used normalization method [11].

To construct scRNA-seq dataset co-expression networks based on Unique Molecular Identifier (UMI), we use CS-CORE [88], a state-of-the-art tool for co-expression inference based on UMI count data. CS-CORE demonstrates increased robustness and a lower false positive rate compared to other tools. For scATAC-seq datasets, we use Seurat [39] to convert the original cells-peaks matrix into the cell-gene activity matrix, incorporating prior information. The cell-gene activity matrix can be processed similarly to the scRNA-seq data matrix, so subsequent preprocessing steps remain the same.

To construct spatial data co-expression networks, we consider spatial expression patterns (SE genes or spatially HVGs) and treat each barcode as a sample. We identify SE genes using SPARK-X [116], then generate corrected gene expression profiles based on scTransform and co-expression networks based on CS-CORE.

B.1 CS-CORE

Since our data are in UMI type, CS-CORE can be used to perform the co-expression network construction. Now considering we have n cells and for one cell i , its expression profile can be denoted as a vector (x_{i1}, \dots, x_{ip}) , where p is the number of genes. We can also use s_i to represent the sequencing depth of cell i . Given the underlying expression levels from cell i with p genes as (z_{i1}, \dots, z_{ip}) , the assumption of CS-CORE for the expression profile follows:

$$(z_{i1}, \dots, z_{ip}) \sim F_p(\mu, \Sigma), \quad x_{ij} | z_{ij} \sim \text{Poisson}(s_i z_{ij}) \quad (8)$$

where F_p is an unknown non-negative p-variate distribution with μ as mean and Σ as covariance matrix. Here the observed expression x_{ij} follows the Poisson measurement model depending on the underlying expression z_{ij} and the sequencing depth $s_i = \sum_{j=1}^n x_{ij}$. CS-CORE applies a moment-based iteratively reweighted least squares (IRLS) estimation procedure to estimate the covariance matrix. Once we have the covariance matrix $\Sigma_{p \times p} = [\sigma_{ij}]_{i=1 \dots p}^{j=1 \dots p}$, we can use the correlation $\rho_{jj'} = \frac{\sigma_{jj'}}{\sqrt{\sigma_{jj}\sigma_{j'j'}}}$ to estimate the co-expression relation between gene j and gene j' .

With such an assumption, CS-CORE can model the relationship between gene j and gene j' based on a statistical test:

$$T_{jj'} = \frac{\sum_i s_i^2 (x_{ij} - s_i \mu_j)(x_{ij'} - s_i \mu_{j'}) g_{ijj'}}{\sqrt{\sum_i s_i^4 (s_i \mu_j + s_i^2 \sigma_{jj})(s_i \mu_{j'} + s_i^2 \sigma_{j'j'}) g_{ijj'}^2}} \quad (9)$$

The statistic $T_{jj'}$ under the null hypothesis assume gene j and gene j' are independent, which means there is no edge between these two genes. We chose to use p-value ($p < 0.005$) to construct the edges in the co-expression network.

B.2 scTransform

To remove the confounding effect of sequencing depth from the expression level, we first process the single-cell data using scTransform to get the Pearson residuals and then use the Pearson residuals as the initial embeddings of different genes. By assuming that the UMI count data follow the negative binomial distribution, for a given gene g in cell c , we have:

$$\begin{aligned} x_{gc} &\sim \text{NB}(\mu_{gc}, \theta_g) \\ \ln \mu_{gc} &= \beta_{g0} + \ln s_c, \end{aligned} \quad (10)$$

scTransform regularizes θ as a function of gene means μ , by utilizing the Generalized Linear Model (GLM) with a log link function provided by the above equation. Furthermore, we can estimate the unknown parameters and calculate the Pearson residuals Z_{gc} based on:

$$\begin{aligned} Z_{gc} &= \frac{x_{gc} - \mu_{gc}}{\sigma_{gc}} \\ \mu_{gc} &= \exp(\beta_{g0} + \ln s_c) \\ \sigma_{gc} &= \sqrt{\mu_{gc} + \frac{\mu_{gc}^2}{\theta_{gc}}} \end{aligned} \quad (11)$$

We can finally replace the original expression matrix with the residual matrix Z generated by GLM, and store the expression matrix and the corresponding graph in Scanpy files.

B.3 SPARK-X

The primary distinctions between single-cell transcriptomic data and spatial transcriptomic data involve two aspects: 1. In most spatially resolved data, each barcode represents a mixture of different cells. 2. The additional spatial information introduces spatial gene expression patterns (SE genes) for spatially resolved data. To identify SE genes in spatial transcriptomic data, SPARK-X employs a statistical test comparing the distance covariance matrix, which is constructed using barcode positions, and the expression covariance matrix, which is built from the gene expression profiles.

More specifically, for a spatial transcriptomic gene expression matrix with size $n \times d$, we can denote the matrix for coordinates of samples as $S = (s_1^T, \dots, s_n^T)$, $s_i = (s_{i1}, s_{i2})$. Therefore,

the whole expression matrix can be represented as : $y = (y_1(s_1), \dots, y_n(s_n))^T$. Our target is to test whether y is independent from S , so we construct the expression covariance matrix based on $E = y(y^T y)^{-1} y^T$, and the distance covariance matrix based on $\Sigma = S(S^T S)^{-1} S^T$. For the distance covariance matrix S , SPARK-X considers different kernels to describe different spatial expression patterns, including 1. Gaussian kernel $(s'_{i1}, s'_{i2}) = (\exp(-\frac{s_{i1}^2}{2\sigma_1^2}), \exp(-\frac{s_{i2}^2}{2\sigma_2^2}))$ and 2. Cosine kernel $(s'_{i1}, s'_{i2}) = (\cos(\frac{2\pi s_{i1}}{\Phi_1}), \cos(\frac{2\pi s_{i2}}{\Phi_2}))$. After centering these two matrices, we have E_C and Σ_C . Therefore, the test statistic is:

$$T = \frac{\text{trace}(E_C \Sigma_C)}{n} \quad (12)$$

This statistic follows a χ_1^2 distribution and we record the p-value for each gene. The null hypothesis is that the gene expression is irrelevant to the position of barcodes. After finding the p-value list, we rank the p-value in ascending and select the top 1000 genes to perform normalization and co-expression graph construction.

The whole process of the graph construction is shown in Figure 6.

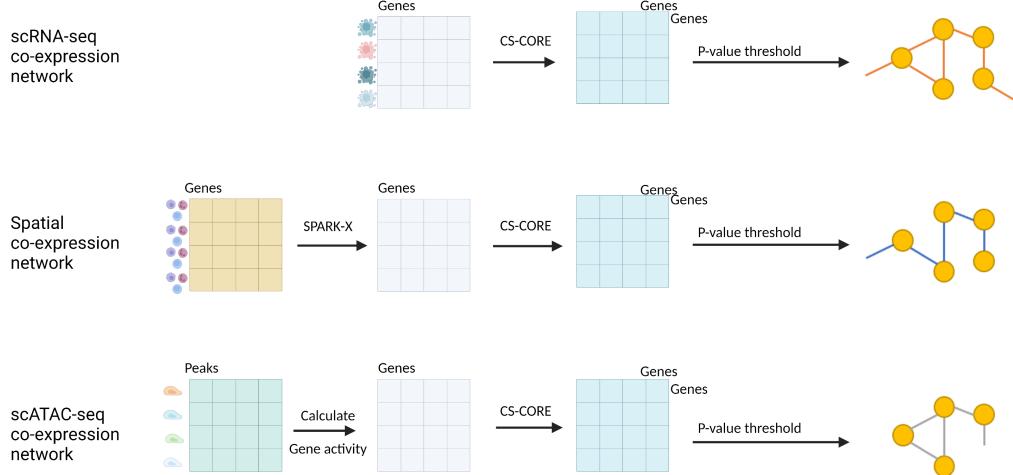


Figure 6: Schematic diagram of the graph data construction methods.

B.4 Transformconv

If we consider c as the index of attention heads, the multi-head attention for GNN from node j to node i can be expressed as:

$$\begin{aligned} q_{c,i}^{(l)} &= W_{c,q}^{(l)} h_i^{(l)} + b_{c,q}^{(l)}; k_{c,j}^{(l)} &= W_{c,k}^{(l)} h_j^{(l)} + b_{c,k}^{(l)}, \\ e_{c,ij} &= W_{c,e} e_{ij} + b_{c,e}; \alpha_{c,ij}^{(l)} &= \frac{\langle q_{c,i}^{(l)}, k_{c,j}^{(l)} + e_{c,ij} \rangle}{\sum_{u \in \mathcal{N}(i)} \langle q_{c,i}^{(l)}, k_{c,u}^{(l)} + e_{c,iu} \rangle}, \end{aligned} \quad (13)$$

where $\langle q, k \rangle = e^{\frac{q^T k}{\sqrt{d}}}$. \sqrt{d} is a scalar used to reduce gradient vanishing [95] as introduced in the original Transformer paper. The different vectors q, k correspond to the query vector and the key

vector, while e represents the edge features. The attention $\alpha_{c,ij}^{(l)}$ denotes the c_{th} attention value from node j to i for layer l . h represents the node embedding, $W_{c,q}^{(l)}, W_{c,k}^{(l)}, W_{c,e}$ are weight matrices, and $b_{c,q}^{(l)}, b_{c,k}^{(l)}, b_{c,e}$ are bias terms.

We define the embedding of node i in layer $l + 1$ as h_i^{l+1} and update the node embedding by:

$$\begin{aligned} v_{c,j}^{(l)} &= W_{c,v}^{(l)} h_j^{(l)} + b_{c,v}^{(l)}, \\ h_i^{(l+1)} &= \left\|_{c=1}^C \left[\sum_{j \in \mathcal{N}(i)} \alpha_{c,ij}^{(l)} (v_{c,j}^{(l)} + e_{c,ij}) \right] \right\|, \end{aligned} \quad (14)$$

where the vector v represents the value vector. The operation $\left\|_{c=1}^C$ denotes the concatenation of C total attention heads. $\mathcal{N}(i)$ represents the neighbors of node i . Additionally, we construct the residual link [40] based on h_i^{l-1} for h_i^{l+1} . The activation function employed to connect different layers is Mish ($\text{Mish}(x) = \tanh(\ln(1 + \exp(x)))$) [65]. We also incorporated a GraphNorm layer [13] to the connection path between hidden layers to enhance convergence.

C Differential co-expression analysis

In this section, we prove that using CS-CORE, we could obtain the co-expression relationships from multimodal biological data without the negative impact of confounding factors brought by the batch effect.

Theorem C.1 *Given the true observed gene expression level x and gene expression affected by batch effect x' , the construction of the co-expression network is the same based on the statistical test.*

Proof. We consider the point $x'_{ij} = ax_{ij} + b$, where a, b represent the effect caused by batch effect, and x_{ij} represents the true biological data. Moreover, based on the probabilistic model provided by CS-CORE, we have $\mathbb{E}(x_{ij}) = s_i\mu_j$, $\text{Var}(x_{ij}) = s_i\mu_j + s_i^2\sigma_{jj}$. Therefore, for the two points x_{ij} and x'_{ij} we have:

$$\begin{aligned} x_{ij} &= s_i\mu_j + \epsilon_{ij} \\ (x_{ij} - s_i\mu_j)^2 &= s_i\mu_j + s_i^2\sigma_{jj} + \eta_{ij} \\ (x_{ij} - s_i\mu_j)(x_{ij'} - s_i\mu_{j'}) &= s_i^2\sigma_{jj'} + \xi_{ijj'} \end{aligned} \quad (15)$$

and

$$\begin{aligned} x'_{ij} &= as_i\mu_j + b + a\epsilon_{ij} \\ (x'_{ij} - (as_i\mu_j + b))^2 &= a^2s_i\mu_i + s_i^2a^2\sigma_{jj} + \eta_{ij} \\ (x'_{ij} - (as_i\mu_j + b))(x'_{ij'} - (as_i\mu_{j'} + b)) &= s_i^2a^2\sigma_{jj'} + \xi_{ijj'} \end{aligned} \quad (16)$$

Based on this assumption, we can calculate the test statistic $T'_{jj'}$ with batch effect, which is:

$$\begin{aligned} T'_{jj'} &= \frac{\sum_i s_i^2 (x'_{ij} - as_i\mu_j - b) (x'_{ij'} - as_i\mu_{j'} - b) g_{ijj'}}{\sqrt{\sum_i s_i^4 (a^2s_i\mu_j + a^2s_i^2\sigma_{jj}) (a^2s_i\mu_{j'} + a^2s_i^2\sigma_{j'j'}) g_{ijj'}^2}} \\ &= \frac{a^2 \sum_i s_i^2 (x_{ij} - s_i\mu_j) (x_{ij'} - s_i\mu_{j'}) g_{ijj'}}{a^2 \sqrt{\sum_i s_i^4 (s_i\mu_j + s_i^2\sigma_{jj}) (s_i\mu_{j'} + s_i^2\sigma_{j'j'}) g_{ijj'}^2}} \\ &= \frac{\sum_i s_i^2 (x_{ij} - s_i\mu_j) (x_{ij'} - s_i\mu_{j'}) g_{ijj'}}{\sqrt{\sum_i s_i^4 (s_i\mu_j + s_i^2\sigma_{jj}) (s_i\mu_{j'} + s_i^2\sigma_{j'j'}) g_{ijj'}^2}} \\ &= T_{jj'} \end{aligned} \quad (17)$$

This statistic is irrelevant to a, b . Therefore, the batch effect cannot affect our co-expression calculation, and our correlation defined by CS-CORE can reflect the correlation for true biological content.

D Theory analysis for multimodal machine learning

Here we present a series of explanation for why using more modalities is more suitable in our task based on arguing that MMML method can generate more accurate estimate of the latent space representation for genes as the number of modalities increases.

Here we consider our datasets set \mathcal{D} and different types of modalities \mathcal{M} and \mathcal{N} , and their corresponding gene embeddings as $e_{\mathcal{M}}$ and $e_{\mathcal{N}}$, where $\text{CARD}(\mathcal{N}) < \text{CARD}(\mathcal{M})$. We also have the true representation e^* . Based on our model, we can estimate $\hat{e}_{\mathcal{M}}, \hat{e}_{\mathcal{N}}$. Here we intend to prove the theorem:

Theorem D.1 $\hat{e}_{\mathcal{M}}$ is a better estimation for e^* comparing to $\hat{e}_{\mathcal{N}}$ with $\text{CARD}(\mathcal{N}) < \text{CARD}(\mathcal{M})$.

Here we consider dataset $\mathcal{S} = \{(X_i, Y_i)\}_{i=1}^m$ as a general training dataset. In our specific cases, we can treat X_i carries information with graph structure and Y_i contains the edges information of X_i . Therefore, we have formalized the dataset we used in our unsupervised learning task. Our target is to minimize the empirical risk:

$$\begin{aligned} \min \quad & \hat{r}(h \circ g_{\mathcal{M}}) \triangleq \frac{1}{m} \sum_{i=1}^m \ell(h \circ g_{\mathcal{M}}(\mathbf{X}_i), Y_i), \\ \text{s.t.} \quad & h \in \mathcal{H}, g_{\mathcal{M}} \in \mathcal{G}_{\mathcal{M}} \end{aligned} \quad (18)$$

where $\ell(\cdot, \cdot)$ represents the loss function, $h \circ g_{\mathcal{M}}(\mathbf{X}_i) = h(g_{\mathcal{M}}(\mathbf{X}_i))$ represents the composite function of h and $g_{\mathcal{M}}$. Here function g maps data from X_i to the latent space, while function h maps data from latent space to the space of Y_i . We have \mathcal{H} as the function class of h and $\mathcal{G}_{\mathcal{M}}$ is the function class of $g_{\mathcal{M}}$. We define the true map functions as h^* and g^* . We denote $r(\cdot, \cdot)$ as risk.

To finish our proof, we need three assumptions.

Assumption D.1. The loss function is L -smooth with respect to the first ordinate, and is bounded by a constant C [67, 92, 93].

In our case, we have three components. BCELoss and Cosine Similarity follow this property [105, 47]. InfoNCE loss also follows this property proved by [61].

Assumption D.2. The true latent representation g^* is contained in \mathcal{G} , and the task mapping h^* is contained in \mathcal{H} [25, 92, 93].

Assumption D.3. For any $g' \in \mathcal{G}'$ and $\mathcal{M} \subset [K]$, $g' \circ p'_{\mathcal{M}} \in \mathcal{G}'$ [45].

Here the set $[K]$ represents modality set with K modalities. $p'_{\mathcal{M}}$ is a diagonal matrix using 1 for ii -th entry $\in \mathcal{M}$ and 0 otherwise.

We also need one metric to evaluate the quality of our generated embeddings, so here we introduce our definition of *latent representation quality*.

Definition D.1. For any latent space generated by g , the **latent representation quality** is defined as [45]

$$\eta(g) = \inf_{h \in \mathcal{H}} [r(h \circ g) - r(h^* \circ g^*)] \quad (19)$$

One approach we can use to prove **Theorem D.1** is to explore the order relationship between the upper bound of $\eta(g_{\mathcal{M}})$ and the upper bound of $\eta(g_{\mathcal{N}})$. That is, we intend to prove:

$$\sup_{g_{\mathcal{M}} \in \mathcal{G}_{\mathcal{M}}} \eta(g_{\mathcal{M}}) \leq \sup_{g_{\mathcal{N}} \in \mathcal{G}_{\mathcal{N}}} \eta(g_{\mathcal{N}})$$

To prove the above relation, we rely on a new theorem to compute the upper bound of $\eta(\cdot)$.

Theorem D.2. [45] Let $\mathcal{S} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ be a dataset of m examples drawn i.i.d. according to \mathcal{D} . Let \mathcal{M} be a subset of $[K]$. Assuming we have produced the empirical risk minimizers $(\hat{h}_{\mathcal{M}}, \hat{g}_{\mathcal{M}})$ training with the \mathcal{M} modalities. Then, for all $1 > \delta > 0$ with probability at least $1 - \delta$:

$$\eta(\hat{g}_{\mathcal{M}}) \leq 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_{\mathcal{M}}) + 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}) + 6C\sqrt{\frac{2 \ln(2/\delta)}{m}} + \hat{L}(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}, \mathcal{S}),$$

where $\hat{L}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}, \mathcal{S}) \triangleq \hat{r}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}) - \hat{r}(h^* \circ g^*)$ is the centered empirical loss and L represents the smoothness coefficient.

Remark. Now we consider sets $\mathcal{N} \subset \mathcal{M} \subset [K]$, and based on Assumption D.3, we have $\mathcal{G}_\mathcal{N} \subset \mathcal{G}_\mathcal{M} \subset \mathcal{G}$. We can interpret this relation as the size of parameter space PARAM follows the relation: $\text{PARAM}_\mathcal{N} \subset \text{PARAM}_\mathcal{M} \subset \text{PARAM}_\mathcal{G}$. Therefore, larger function class has a smaller empirical risk, so we have

$$\hat{L}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}, \mathcal{S}) \leq \hat{L}(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}, \mathcal{S}) \quad (20)$$

Based on Theorem D.2, we can derive the upper bound of $\eta(\hat{g}_\mathcal{N})$ as:

$$\eta(\hat{g}_\mathcal{N}) \leq 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_\mathcal{N}) + 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}) + 6C\sqrt{\frac{2\ln(2/\delta)}{m}} + \hat{L}(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}, \mathcal{S}),$$

where $\mathfrak{R}_m(\cdot)$ represents the Rademacher complexity [10]. Here we can use the Right-hand Side (RHS) to approximate term $4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_\mathcal{M}) \sim \sqrt{C(\mathcal{H} \circ \mathcal{G}_\mathcal{M})/m}$ and term $4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_\mathcal{N}) \sim \sqrt{C(\mathcal{H} \circ \mathcal{G}_\mathcal{N})/m}$. Based on the basic structural property of Rademacher complexity [10], we have

$$C(\mathcal{H} \circ \mathcal{G}_\mathcal{N}) \leq C(\mathcal{H} \circ \mathcal{G}_\mathcal{M})$$

Therefore, we can compute:

$$\begin{aligned} \eta(\hat{g}_\mathcal{M}) - \eta(\hat{g}_\mathcal{N}) &= 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_\mathcal{M}) + \hat{L}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}, \mathcal{S}) - 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_\mathcal{N}) - \hat{L}(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}, \mathcal{S}) \\ &= \sqrt{\frac{C(\mathcal{H} \circ \mathcal{G}_\mathcal{M})}{m}} - \sqrt{\frac{C(\mathcal{H} \circ \mathcal{G}_\mathcal{N})}{m}} + \hat{L}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}, \mathcal{S}) - \hat{L}(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}, \mathcal{S}) \\ &= \frac{\sqrt{C(\mathcal{H} \circ \mathcal{G}_\mathcal{M})} - \sqrt{C(\mathcal{H} \circ \mathcal{G}_\mathcal{N})}}{\sqrt{m}} + \hat{L}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}, \mathcal{S}) - \hat{L}(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}, \mathcal{S}) \\ &\rightarrow \hat{L}(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}, \mathcal{S}) - \hat{L}(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}, \mathcal{S})|_{m \rightarrow \infty} \\ &\leq 0 \end{aligned}$$

Remark. Therefore, as the number of modalities increases, the representation quality of more modalities will be better than the representation quality of fewer modalities. Hence we proved Theorem D.1 under large number of modalities cases. In the real biological application, genes can be active in many different scenarios, including different cells, different tissues and even different species. Therefore, our proof fits the context of MuSe-GNN's real application.

E Extra experiments, ablation test and model selection

E.1 Ablation tests

We further investigate the contributions of various loss functions and GNN models to MuSe-GNN's performance. Our final loss function comprises three components: the graph reconstruction loss $\mathcal{L}_{\text{reconst}}$, the similarity maximization loss $\mathcal{L}_{\text{CosSim}}$, and the contrastive learning loss $\mathcal{L}_{\text{InfoNCE}}$. The reconstruction loss serves as the basic loss, and we assess our model's performance with different loss function components across all scRNA-seq datasets. The results are presented in Table 3 and 4.

Table 3: The overall benchmarking average rank for ablation tests based on different tissues.

$\mathcal{L}_{\text{CosSim}}$	$\mathcal{L}_{\text{InfoNCE}}$	Heart	Lung	Liver	Kidney	Thymus	Spleen	Pancreas	Cerebrum	Cerebellum	PBMC	Avg Rank
✓		2.83	3.50	3.17	3.33	2.67	3.17	3.00	3.67	3.00	3.33	3.17
	✓	2.17	1.67	2.00	2.17	2.50	2.00	1.50	2.00	2.17	5.67	2.38
✓	✓	3.00	3.00	3.17	2.83	3.17	3.17	3.33	3.00	2.83	5.00	3.25
		2.00	1.83	1.67	1.67	1.67	1.67	2.17	1.33	2.00	3.83	1.98

Based on the results of this ablation test, we conclude that combining all three components to construct our loss function is the optimal choice, achieving a performance that is 224.0% higher than the version without any additional regularization. Furthermore, using only contrastive learning may

Table 4: The overall benchmarking average score for ablation tests based on different tissues.

$\mathcal{L}_{\text{CosSim}}$	$\mathcal{L}_{\text{InfoNCE}}$	Heart	Lung	Liver	Kidney	Thymus	Spleen	Pancreas	Cerebrum	Cerebellum	PBMC	Avg Score
✓		0.42	0.13	0.21	0.24	0.36	0.19	0.28	0.09	0.37	0.23	0.25
		0.65	0.81	0.77	0.75	0.57	0.70	0.83	0.71	0.68	0.66	0.71
	✓	0.34	0.27	0.25	0.23	0.23	0.20	0.18	0.26	0.35	0.23	0.25
✓	✓	0.78	0.81	0.94	0.80	0.70	0.96	0.53	0.91	0.74	0.90	0.81

reduce the performance of our model, implying that learning similarity is more crucial than learning differences for the gene embedding generation task. When both components are included, the final average rank is 14.1% higher than the version with only the similarity learning part. Thus, we need to incorporate both weighted similarity learning and contrastive learning in our final design to learn unified gene embeddings with improved performance.

Additionally, we compare the TransformConv framework with other GNN models, including GCN, Graph Attention Network (GAT) [96], SUGRL (GCN+MLP+Contrastive Learning) [66], GPS [75], GRACE (GCN+Contrastive Learning) [118], GraphSAGE [36], Graph Isomorphism Network (GIN) [106], GraphMAE [43] and Graphomer [110]. We set the number of epochs to 1000 for this comparison. The results are displayed in Table 5.

Table 5: The overall benchmarking score and rank for model selection.

Methods	ASW	AUC	iLISI	GC	CGR	NO	Avg Rank	Avg Score
GCN	0.74±0.02	0.82±0.01	0.32±0.02	0.44±0.02	0.32±0.04	0.31±0.01	3.5	0.51
GAT	0.74±0.09	0.83±0.01	0.14±0.03	0.42±0.03	0.00±0.00	0.15±0.03	5	0.17
SUGRL	0.88±0.01	0.62±0.00	0.45±0.01	0.43±0.03	0.38±0.02	0.29±0.00	3.5	0.56
GPS	0.77±0.02	0.63±0.00	0.50±0.01	0.72±0.03	0.57±0.02	0.30±0.01	2.67	0.68
GRACE	0.82±0.02	0.81±0.00	0.35±0.01	0.42±0.04	0.11±0.01	0.25±0.01	4.17	0.48
GraphSAGE	OOM	OOM	OOM	OOM	OOM	OOM	7.00	0
GIN	OOM	OOM	OOM	OOM	OOM	OOM	7.00	0
GraphMAE	OOM	OOM	OOM	OOM	OOM	OOM	7.00	0
Graphomer	OOM	OOM	OOM	OOM	OOM	OOM	7.00	0
Transformconv	0.75±0.01	0.78±0.02	0.51±0.02	0.68±0.04	0.61±0.02	0.31±0.00	2.17	0.80

In this table, OOM means out of memory. We can conclude that using TransformConv as the basic graph neural network framework is the best choice, which is 17.6% higher than the version based on GPS (the top2 model). Moreover, based on Figure 7, we can also discover that methods based on GCN or GAT failed to learn the gene function similarity across different datasets. Therefore, choosing TransformConv is reasonable, and the major contribution of Transformer model towards gene embedding learning task is the multi-head attention design.

We also conduct experiments to justify the need for weight-sharing design. In this comparison, we evaluate the performance and model statistics of two models (with weight-sharing design (ws) and without weight-sharing design (w/o ws)). As shown in Table 6, the combined high or equivalent rank of weight-sharing designs in comparison to w/o weight-sharing designs is nearly consistent across all tissue types. Furthermore, the weight-sharing design can also reduce the model’s training cost by decreasing the model size, the number of trainable parameters, and the total training time, as confirmed in Table 7.

Table 6: Comparison between with weight-sharing design and without weight-sharing design.

Methods	Heart	Lung	Liver	Kidney	Thymus	Spleen	Pancreas	Cerebrum	Cerebellum	PBMC	Avg Rank
ws	1.50	1.33	1.33	1.33	1.33	1.33	1.67	1.50	1.17	1.17	1.37
w/o ws	1.50	1.67	1.67	1.67	1.67	1.67	1.33	1.50	1.83	1.83	1.63

To demonstrate the necessity of both expression profiles and graph structure for learning gene embeddings with biological information, we design an experiment in which all nodes in different graphs are replaced with the same features while maintaining the original graph structure. Based on the results shown in Figure 8 and Table 8, we can conclude that without the information from expression profiles, it becomes difficult for our model to learn unified gene embeddings because of the decline in scores of different metrics.

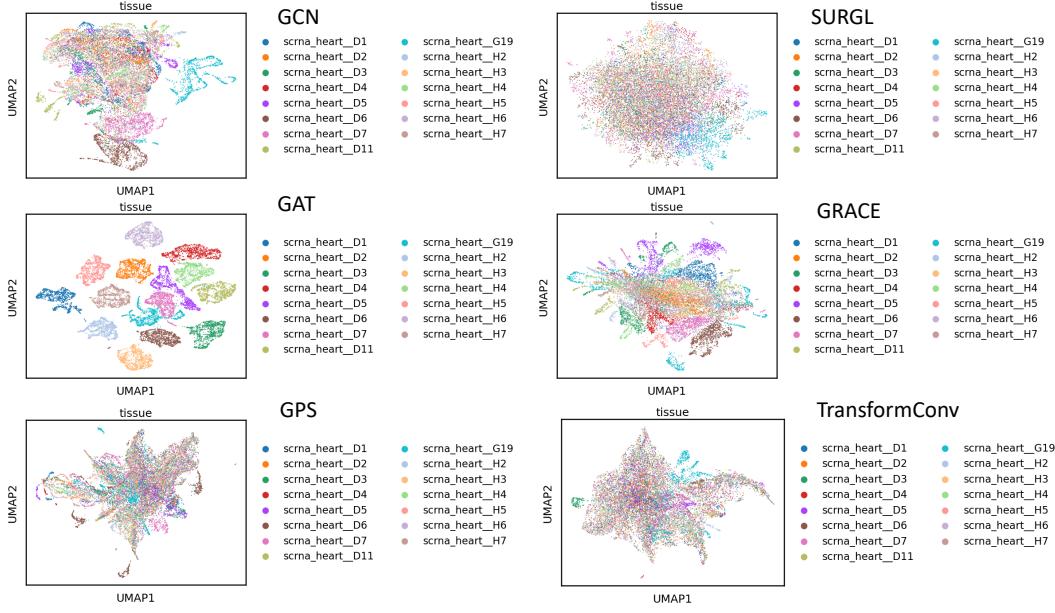


Figure 7: Gene embeddings based on different GNNs.

Table 7: Model statistics.

Methods	model size (MB)	# of parameters (M)	training time per epoch (s)
ws	1333	349	3.64
w/o ws	1346.8	353	4.85

To demonstrate the necessity of Weighted Similarity Learning (WSL) design, we design an experiment to analyze the performance of different model structures with WSL or w/o WSL. According to Table 9, we found that MuSe-GNN with WSL performed better than MuSe-GNN without WSL across all of the metrics in heart tissue. Therefore, we need to utilize WSL mechanism to effectively learn the gene-gene relationships across different datasets.

E.2 Hyper-parameter tuning

Hyper-parameter settings for various models can be divided into two groups: 1. Methods such as PCA, Gene2vec, GIANT, WSMAE, GAE, VGAE, and MAE, which are free of pre-training, are configured with the best hyper-parameters based on parameter searching to ensure fairness. Details are included in Table 10. 2. Method such as scBERT uses the pre-trained model. The optimal hyper-parameters for MuSe-GNN are provided in Table 11. Number of trainable parameters for different models is summarized in Table 12.

The average rank results for various hyper-parameters are presented in Table 13. We set the initial parameter settings as follows: Epoch = 2000, LR of encoder = 1e-2, LR of decoder = 1e-3, λ_c = 1e-2, Dim = 32, and Sample size (for contrastive learning) = 100. Notably, for the value choice of λ_c , we have two candidates with the same average rank result. To further compare the gene

Table 8: Comparison between with feature design and without feature design.

Methods	ASW	AUC	iLISI	GC	CGR	NO
with feature	0.75±0.01	0.78±0.02	0.53±0.01	0.73±0.04	0.65±0.02	0.31±0.00
w/o feature	0.75±0.05	0.64±0.02	0.24±0.04	0.43±0.03	0.02±0.01	0.10±0.03

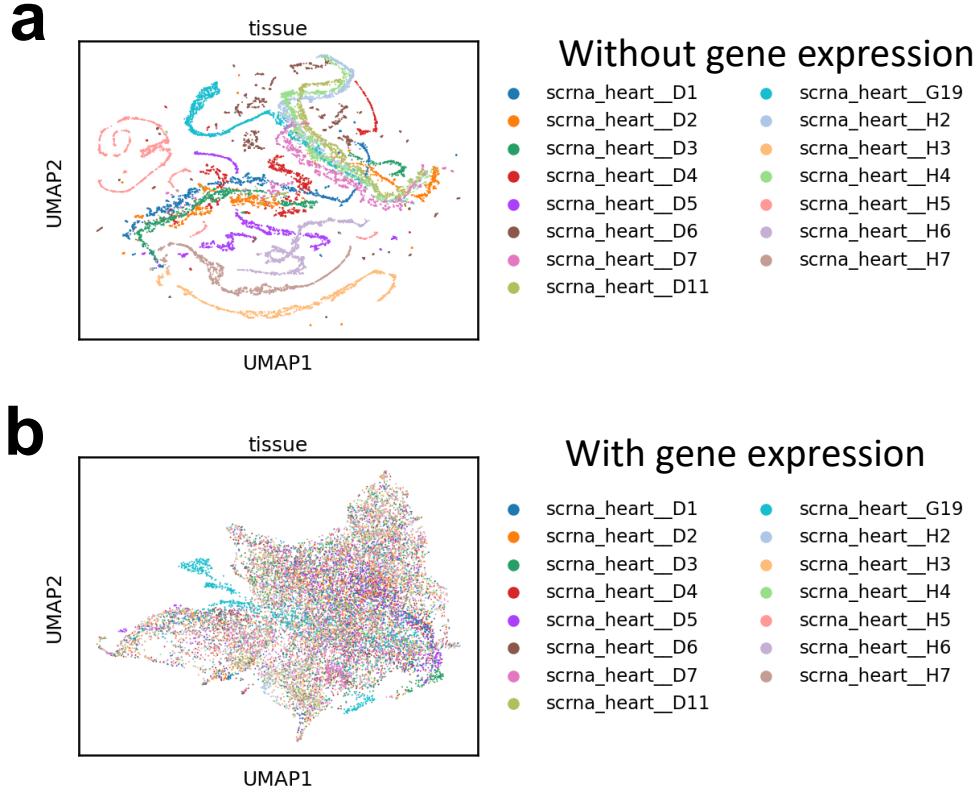


Figure 8: Gene embeddings based on different input information. **(a)** represents the UMAPs of gene embeddings without expression profiles and **(b)** represents the UMAPs of gene emebddings with expression profiles.

Table 9: Comparison between with weighted similarity learning design and without weighted similarity learning design.

Methods	ASW	AUC	iLISI	GC	CGR	NO
with WSL	0.75±0.01	0.78±0.02	0.53±0.01	0.73±0.04	0.65±0.02	0.31±0.00
w/o WSL	0.69±0.02	0.66±0.03	0.46±0.03	0.57±0.07	0.48±0.05	0.29±0.00

embeddings generated by models based on these two choices, we plot the UMAPs for these two models. According to Figure 9, we observe that choosing $\lambda_c = 0.1$ fails to integrate genes from dataset G19. Therefore, we ultimately set $\lambda_c = 1e-2$.

Additionally, the dimensions of gene embeddings should not be set to a very large value, not only due to the results in our table but also because of theoretical constraints and practical feasibility. Generally, an excessively large number of embedding settings may cause the number of hidden layer neurons to exceed the number of features in the input partial dataset, increasing the probability of overfitting by introducing more noise [57]. Furthermore, a larger number of dimensions will consume more computational resources, potentially leading to out-of-memory issues, such as when setting the Dim of embeddings = 256.

E.3 Details about benchmarking experiments

The average ranks result for all the scRNA-seq datasets is displayed in Table 14. The experimental results for all the scRNA-seq datasets are displayed in Tables 15 to 24 (unscaled). Based on the benchmark results for different tissues, we can conclude that MuSe-GNN consistently outperforms

Table 10: Hyper-parameter candidates for benchmarking methods.

Models	Hyper-parameters
PCA	Dim:[32,128]
Gene2vec	Dim:[32,128]
GIANT	Dim:[32,128]
WSMAE	LR:[1e-4,1e-2]; Batch size:[1000,3000]; Dim:[32,128]
GAE	LR:[1e-4,1e-2]; Batch size:[1000,3000]; Dim:[32,128]
VGAE	LR:[1e-4,1e-2]; Batch size:[1000,3000]; Dim:[32,128]
MAE	LR:[1e-4,1e-2]; Batch size:[1000,3000]; Dim:[32,128]

Table 11: Hyper-parameters list for MuSe-GNN.

Parameter	Value
Epoch	2000
LR of encoder	1e-4
LR of decoder	1e-3
λ_c	1e-2
Dim of embeddings	32
Sample size	100

other methods according to specific metrics, such as iLISI and NO. Furthermore, the gene embeddings learned by MuSe-GNN are relatively more stable than those from other methods, as evidenced by the analysis of standard deviation. Considering the overall performance, MuSe-GNN, as the top-performing method, also surpasses the second-best methods from 20.1% to 97.5% and surpasses the third-best methods from 26.9% to 99.7%.

E.4 Details about function cluster identification

In this paper, we utilized Leiden cluster method [91] to identify genes with common functions based on the value of their embeddings. The genes with distance smaller than a community-discovering based threshold (known as co-embedded genes) will be identified into a group. Here we sample 100 genes in our final gene embedding result from heart tissue to illustrate the distribution of such distance. Based on Figure 10, we can conclude that our method has the ability to identify genes with similar embeddings, and further, it can be used to derive genes with similar functions.

All of the experiments of this paper are finished based on Intel Xeon Gold 6240 Processor and one NVIDIA A100 GPU. The RAM setting is 30 GB.

F Extra Discussion about Model Design

This section aims to demonstrate the reasoning behind our model design in greater details. As a novel graph neural network model, MuSe-GNN offers three significant advantages.

Firstly, our method’s graph construction process is robust and reliable. Graph construction is a critical step for GNN models [117], and we place considerable emphasis on both the graph construction method and data quality throughout our work. In single-cell data analysis research, some researchers

Table 12: Number of trainable parameters for different models

Models	# of parameters (M)
MuSe-GNN	349
GIANT	260
VGAE	210
WSMAE	105
GAE	52.5
MAE	52.5

Table 13: Hyper-parameter searching results.

LR for encoder	Avg Rank	InfoNCE penalty	Avg Rank
0.01	3.17	1	3.5
0.001	2.50	0.1	2.13
0.0005	2.33	0.01	2.13
0.0001	1.83	0.001	2.25
LR for decoder	Avg Rank	Dim	Avg Rank
0.01	3.43	32	1.86
0.001	1.86	64	2.00
0.0005	2.29	128	2.14
0.0001	2.43	256	4.00
batch size	Avg Rank	sample size	Avg Rank
1000	3	50	2.38
1500	3	100	1.75
2000	1.5	150	2.88
3000	2.5	200	2.5

Table 14: The average ranks for gene embeddings benchmarking across different tissues.

Methods	Heart	Lung	Liver	Kidney	Thymus	Spleen	Pancreas	Cerebrum	Cerebellum	PBMC
PCA	4.33	4.33	4.67	5.50	4.00	4.17	4.83	4.50	5.50	4.00
Gene2vec	5.67	6.67	6.50	6.67	7.33	7.17	6.83	6.67	6.00	7.33
GIANT	6.00	4.67	5.33	4.67	3.67	5.67	3.83	5.50	6.00	6.17
WSMAE	5.83	5.33	4.83	4.83	3.83	5.17	5.00	5.33	4.33	5.00
GAE	4.33	5.83	5.33	5.50	5.17	5.33	4.50	6.00	5.00	5.33
VGAE	3.50	6.33	6.33	6.67	4.83	6.50	6.17	6.67	6.83	5.83
MAE	7.00	5.17	5.17	5.17	5.50	5.17	5.67	5.33	5.33	5.33
scBERT	5.67	4.50	4.33	3.17	7.83	3.83	4.83	3.17	3.17	3.67
MuSe-GNN	2.67	2.17	2.50	2.83	2.67	2.00	2.83	1.67	2.67	2.33

Table 15: Benchmark score table for Heart.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.77 ± 0.01	0.64 ± 0.00	0.43 ± 0.00	0.49 ± 0.02	0.30 ± 0.01	0.27 ± 0.00
Gene2vec	0.82 ± 0.03	0.76 ± 0.00	0.09 ± 0.01	0.65 ± 0.03	0.00 ± 0.00	0.10 ± 0.02
GIANT	0.84 ± 0.01	0.55 ± 0.00	0.26 ± 0.02	0.48 ± 0.03	0.10 ± 0.01	0.01 ± 0.00
WSMAE	0.79 ± 0.01	0.72 ± 0.00	0.34 ± 0.01	0.43 ± 0.04	0.11 ± 0.02	0.26 ± 0.01
GAE	0.79 ± 0.01	0.98 ± 0.00	0.38 ± 0.01	0.45 ± 0.02	0.12 ± 0.02	0.28 ± 0.01
VGAE	0.80 ± 0.01	0.97 ± 0.00	0.39 ± 0.01	0.44 ± 0.02	0.18 ± 0.03	0.28 ± 0.00
MAE	0.82 ± 0.01	0.52 ± 0.00	0.30 ± 0.02	0.24 ± 0.05	0.03 ± 0.00	0.22 ± 0.01
scBERT	0.77 ± 0.01	0.47 ± 0.00	0.39 ± 0.01	0.37 ± 0.03	0.26 ± 0.01	0.28 ± 0.00
MuSe-GNN	0.75 ± 0.01	0.78 ± 0.02	0.53 ± 0.01	0.73 ± 0.04	0.65 ± 0.01	0.31 ± 0.00

Table 16: Benchmark score table for Lung.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.79 ± 0.00	0.69 ± 0.00	0.58 ± 0.00	0.70 ± 0.01	0.07 ± 0.01	0.10 ± 0.00
Gene2vec	0.71 ± 0.40	0.76 ± 0.00	0.00 ± 0.00	0.89 ± 0.01	0.00 ± 0.00	0.01 ± 0.00
GIANT	0.90 ± 0.01	0.49 ± 0.00	0.12 ± 0.03	0.88 ± 0.03	0.07 ± 0.01	0.01 ± 0.00
WSMAE	0.76 ± 0.01	0.69 ± 0.01	0.45 ± 0.03	0.77 ± 0.03	0.05 ± 0.01	0.09 ± 0.01
GAE	0.73 ± 0.03	0.86 ± 0.00	0.23 ± 0.04	0.78 ± 0.03	0.02 ± 0.00	0.07 ± 0.01
VGAE	0.51 ± 0.10	0.87 ± 0.01	0.00 ± 0.00	0.87 ± 0.03	0.00 ± 0.00	0.01 ± 0.01
MAE	0.85 ± 0.02	0.62 ± 0.01	0.35 ± 0.05	0.81 ± 0.03	0.02 ± 0.01	0.08 ± 0.01
scBERT	0.87 ± 0.01	0.48 ± 0.00	0.29 ± 0.02	0.86 ± 0.05	0.06 ± 0.01	0.10 ± 0.00
MuSe-GNN	0.84 ± 0.02	0.86 ± 0.01	0.64 ± 0.04	0.87 ± 0.03	0.37 ± 0.03	0.19 ± 0.00

Table 17: Benchmark score table for Liver.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.77±0.00	0.58±0.00	0.39±0.00	0.69±0.01	0.21±0.00	0.15±0.00
Gene2vec	0.63±0.34	0.69±0.00	0.00±0.00	0.88±0.01	0.00±0.00	0.03±0.01
GIANT	0.90±0.01	0.55±0.00	0.40±0.02	0.44±0.03	0.11±0.01	0.00±0.00
WSMAE	0.80±0.01	0.62±0.01	0.26±0.03	0.83±0.03	0.09±0.01	0.13±0.01
GAE	0.74±0.02	0.84±0.00	0.17±0.02	0.80±0.01	0.05±0.01	0.14±0.01
VGAE	0.49±0.05	0.87±0.00	0.00±0.00	0.87±0.02	0.01±0.00	0.01±0.00
MAE	0.83±0.01	0.58±0.01	0.25±0.04	0.88±0.04	0.04±0.01	0.09±0.01
scBERT	0.87±0.01	0.50±0.00	0.23±0.05	0.90±0.03	0.11±0.02	0.14±0.00
MuSe-GNN	0.86±0.01	0.78±0.03	0.65±0.02	0.83±0.04	0.46±0.02	0.18±0.00

Table 18: Benchmark score table for Kidney.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.69±0.01	0.55±0.00	0.39±0.01	0.98±0.00	0.00±0.00	0.21±0.00
Gene2vec	0.33±0.42	0.74±0.00	0.00±0.00	1.00±0.01	0.00±0.00	0.02±0.03
GIANT	0.92±0.00	0.55±0.00	0.89±0.01	0.78±0.03	0.02±0.00	0.01±0.00
WSMAE	0.74±0.05	0.68±0.01	0.08±0.04	0.99±0.00	0.01±0.00	0.15±0.04
GAE	0.52±0.10	0.86±0.00	0.01±0.01	0.98±0.01	0.00±0.00	0.05±0.02
VGAE	0.51±0.16	0.88±0.00	0.00±0.00	0.98±0.01	0.00±0.00	0.00±0.01
MAE	0.84±0.04	0.52±0.01	0.17±0.07	1.00±0.00	0.00±0.00	0.16±0.04
scBERT	0.89±0.02	0.47±0.00	0.59±0.18	1.00±0.00	0.05±0.01	0.30±0.00
MuSe-GNN	0.88±0.03	0.71±0.02	0.84±0.04	0.98±0.02	0.31±0.04	0.31±0.01

Table 19: Benchmark score table for Thymus.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.76±0.01	0.79±0.00	0.93±0.00	0.44±0.05	0.00±0.00	0.00±0.00
Gene2vec	0.00±0.00	0.75±0.00	0.00±0.00	0.58±0.03	0.00±0.00	0.00±0.00
GIANT	0.88±0.02	0.55±0.00	0.98±0.01	0.80±0.02	0.00±0.00	0.00±0.00
WSMAE	0.81±0.03	0.74±0.01	0.93±0.06	0.48±0.06	0.00±0.00	0.00±0.00
GAE	0.72±0.12	0.81±0.00	0.89±0.10	0.69±0.05	0.00±0.00	0.00±0.00
VGAE	0.77±0.05	0.84±0.01	0.89±0.05	0.44±0.10	0.00±0.00	0.00±0.00
MAE	0.77±0.28	0.59±0.01	0.94±0.05	0.33±0.05	0.00±0.00	0.00±0.00
scBERT	0.09±0.27	0.50±0.00	0.41±0.18	0.42±0.04	0.00±0.00	0.00±0.00
MuSe-GNN	0.71±0.05	0.80±0.02	0.98±0.01	0.65±0.07	0.01±0.00	0.00±0.00

Table 20: Benchmark score table for Spleen.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.81±0.01	0.56±0.00	0.47±0.00	0.73±0.01	0.23±0.01	0.17±0.00
Gene2vec	0.69±0.36	0.68±0.01	0.00±0.00	0.85±0.01	0.00±0.00	0.01±0.01
GIANT	0.88±0.01	0.55±0.00	0.42±0.02	0.45±0.02	0.10±0.00	0.01±0.00
WSMAE	0.79±0.01	0.61±0.01	0.26±0.03	0.86±0.02	0.08±0.01	0.15±0.01
GAE	0.75±0.02	0.84±0.00	0.18±0.02	0.81±0.01	0.05±0.01	0.16±0.01
VGAE	0.52±0.05	0.86±0.00	0.00±0.00	0.85±0.03	0.00±0.00	0.01±0.00
MAE	0.84±0.02	0.56±0.01	0.30±0.04	0.92±0.02	0.03±0.00	0.12±0.00
scBERT	0.87±0.01	0.50±0.00	0.34±0.06	0.94±0.02	0.11±0.01	0.16±0.00
MuSe-GNN	0.86±0.01	0.79±0.02	0.70±0.02	0.89±0.03	0.47±0.01	0.19±0.00

Table 21: Benchmark score table for Pancreas.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.61±0.01	0.62±0.00	0.64±0.00	0.57±0.01	0.00±0.00	0.01±0.00
Gene2vec	0.17±0.36	0.73±0.00	0.00±0.00	0.51±0.05	0.00±0.00	0.00±0.00
GIANT	0.91±0.01	0.55±0.00	0.97±0.03	0.65±0.05	0.00±0.00	0.00±0.00
WSMAE	0.67±0.09	0.70±0.01	0.66±0.10	0.49±0.06	0.00±0.00	0.00±0.00
GAE	0.09±0.27	0.82±0.00	0.90±0.07	0.65±0.03	0.00±0.00	0.00±0.00
VGAE	0.37±0.40	0.83±0.01	0.13±0.10	0.51±0.04	0.00±0.00	0.00±0.00
MAE	0.71±0.11	0.56±0.01	0.83±0.07	0.37±0.05	0.00±0.00	0.00±0.00
scBERT	0.76±0.05	0.47±0.00	0.54±0.15	0.39±0.04	0.00±0.00	0.01±0.00
MuSe-GNN	0.72±0.08	0.65±0.01	0.95±0.13	0.62±0.06	0.01±0.00	0.00±0.00

Table 22: Benchmark score table for Cerebrum.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.82±0.01	0.63±0.00	0.34±0.00	0.95±0.00	0.12±0.00	0.18±0.00
Gene2vec	0.07±0.23	0.69±0.00	0.00±0.00	0.99±0.00	0.00±0.00	0.01±0.01
GIANT	0.89±0.00	0.55±0.00	0.40±0.02	0.51±0.02	0.03±0.00	0.01±0.00
WSMAE	0.80±0.02	0.67±0.01	0.21±0.04	0.97±0.01	0.03±0.01	0.13±0.02
GAE	0.68±0.04	0.83±0.00	0.03±0.01	0.95±0.01	0.02±0.00	0.13±0.01
VGAE	0.47±0.09	0.84±0.01	0.00±0.00	0.98±0.01	0.00±0.00	0.01±0.00
MAE	0.83±0.02	0.53±0.01	0.20±0.05	0.99±0.01	0.02±0.01	0.14±0.01
scBERT	0.89±0.01	0.48±0.00	0.50±0.07	0.99±0.00	0.10±0.01	0.17±0.00
MuSe-GNN	0.90±0.01	0.73±0.02	0.79±0.03	0.99±0.01	0.54±0.01	0.21±0.00

Table 23: Benchmark score table for Cerebellum.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.82±0.00	0.63±0.00	0.34±0.00	0.95±0.00	0.12±0.01	0.18±0.00
Gene2vec	0.31±0.40	0.71±0.00	0.00±0.00	1.00±0.00	0.00±0.00	0.02±0.02
GIANT	0.89±0.00	0.55±0.00	0.34±0.03	0.73±0.02	0.03±0.00	0.01±0.00
WSMAE	0.83±0.02	0.66±0.01	0.26±0.04	1.00±0.00	0.06±0.01	0.29±0.02
GAE	0.74±0.02	0.83±0.00	0.11±0.08	0.99±0.01	0.07±0.04	0.25±0.04
VGAE	0.19±0.26	0.83±0.01	0.00±0.00	0.99±0.01	0.00±0.00	0.00±0.00
MAE	0.86±0.01	0.51±0.00	0.16±0.03	1.00±0.00	0.02±0.00	0.26±0.03
scBERT	0.90±0.01	0.47±0.00	0.54±0.06	1.00±0.00	0.15±0.02	0.31±0.01
MuSe-GNN	0.92±0.01	0.64±0.02	0.86±0.02	0.98±0.02	0.65±0.00	0.38±0.00

Table 24: Benchmark score table for PBMC.

Methods	ASW	AUC	iLISI	GC	CGR	NO
PCA	0.86±0.01	0.53±0.00	0.46±0.00	0.66±0.01	0.22±0.01	0.13±0.00
GENE2VEC	0.00±0.00	0.59±0.00	0.00±0.00	0.92±0.01	0.00±0.00	0.00±0.00
GIANT	0.85±0.01	0.55±0.00	0.38±0.02	0.55±0.03	0.04±0.00	0.00±0.00
WSMAE	0.81±0.01	0.58±0.01	0.29±0.02	0.81±0.05	0.10±0.02	0.12±0.01
GAE	0.74±0.01	0.82±0.00	0.15±0.03	0.79±0.03	0.06±0.02	0.12±0.01
VGAE	0.67±0.04	0.89±0.00	0.00±0.00	0.93±0.02	0.02±0.00	0.01±0.00
MAE	0.85±0.02	0.53±0.01	0.30±0.05	0.94±0.03	0.02±0.00	0.10±0.01
scBERT	0.88±0.01	0.52±0.00	0.48±0.05	0.93±0.03	0.15±0.02	0.11±0.00
MuSe-GNN	0.86±0.01	0.79±0.02	0.75±0.02	0.88±0.02	0.54±0.01	0.18±0.00

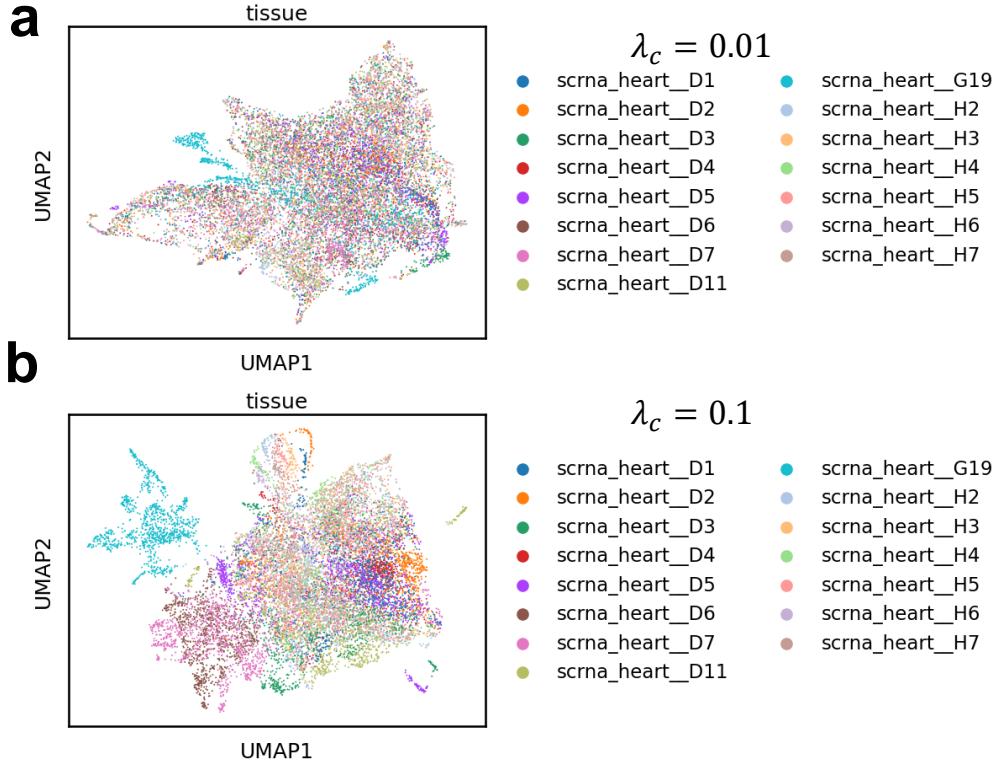


Figure 9: Gene embeddings on different λ_c .

model cell-cell similarity based on the nearest neighbor graph [99, 94]. This approach may be affected by batch effects resulting from different sequencing profiles, leading to incorrect conclusions. For instance, in scRNA-seq data imputation tasks, such a design may result in a high false positive rate [5]. A gene co-expression network derived from a robust statistical framework is more trustworthy, and biological experiments can validate the co-expression performance of various genes. Moreover, to understand the difference of co-expression networks from different tissues or omics data, we calculate the Edge Signal-Noise Ratio (ESNR) [22] for each network, and figure out that the co-expression network based on scRNA-seq has the best quality, while the co-expression networks from other omics have very small ESNR. Such conclusions are also consistent with the model performance.

Secondly, our model utilizes a sophisticated and stable training strategy. By employing the cosine similarity penalty, we ensure that genes with the same functions from different tissues or techniques can be co-embedded in similar positions. Additionally, we use graph auto-encoder and contrastive learning strategies to retain the original biological information of various datasets. Each of our training strategies serves a specific purpose.

Finally, our method showcases the high effectiveness of MMML in addressing heterogeneous data learning challenges. MMML can successfully learn similar information across multimodal data and represent it in a unified space through model fusion [55]. This approach can be applied to more complex data studies in the future, such as integrating data from Magnetic Resonance Imaging (MRI) results, spatial resolution image information, and current expression profiles into a single space. Furthermore, GNN remains the primary model for handling tasks related to graph-structured data. By combining these methods and uncovering biological insights, our research contributes to a deeper understanding of efficient human genome representation.

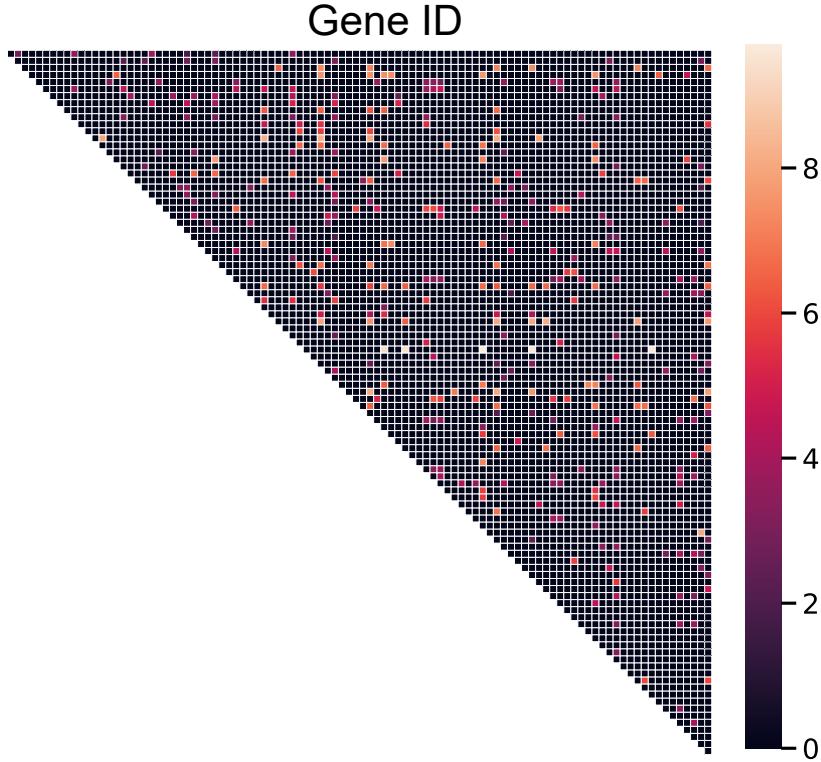


Figure 10: Heatmap of gene embedding distances. The heatmap color represents the distance between genes, with brighter colors indicating greater distances and more distinct functions.

G Selecting anchors of common functional genes

In this paper, we employ explicit common functional genes as anchors across different datasets to implement regularization in our training process. The common functional genes we select can potentially recover a portion of the functional overlap of the genes, while the remaining implicit genes can be recovered during the training process.

In this paper, we choose to use common highly variable genes (HVGs) combined from different datasets as anchors, and employ their co-expression network overlap as weights. An example is illustrated in Figure 12 case 1. In addition to the proof provided by the ablation test results, there are four reasons for this choice: 1. HVGs represent the active genes of a expression profile (dataset), which are correlated to gene functions. 2. For datasets from the same tissue, the overlap of HVGs and the overlap of neighbor genes (known as the co-expressed genes in the same graph) of HVGs pair are similar. This demonstrates a type of consistency in biological function inference. 3. For datasets from different modalities (tissues, techniques, etc.), only using the overlap of HVGs is not reliable. The overlap of neighbor genes of HVGs pair from different tissues shows functional similarity based on hierarchical clustering. We can observe the clear difference of HVGs overlap score and neighbors overlap score of HVGs by comparing Figure 11 (a) to (b). Furthermore, we can consider an extreme case, illustrated in Figure 12 case 2. If two HVGs from different tissues do not contain shared neighbors, the training process will be processed for such genes without weighted constraints. 4. Since it is quite challenging to select all the true genes with similar functions, we can perform a trade-off between learning the similarity of potential functionally similar genes and learning the difference of potential functionally distinct genes. The neighbor genes overlap in common HVGs is significantly different from the overlap in different HVGs for graphs from the same tissue, as shown in Figure 13. Such statistical results also support our choice of using common HVGs as anchors.

In conclusion, we aim to generate similar embeddings for functionally related genes, incorporating common HVGs and their neighbors in co-expression networks, by utilizing the regularization term

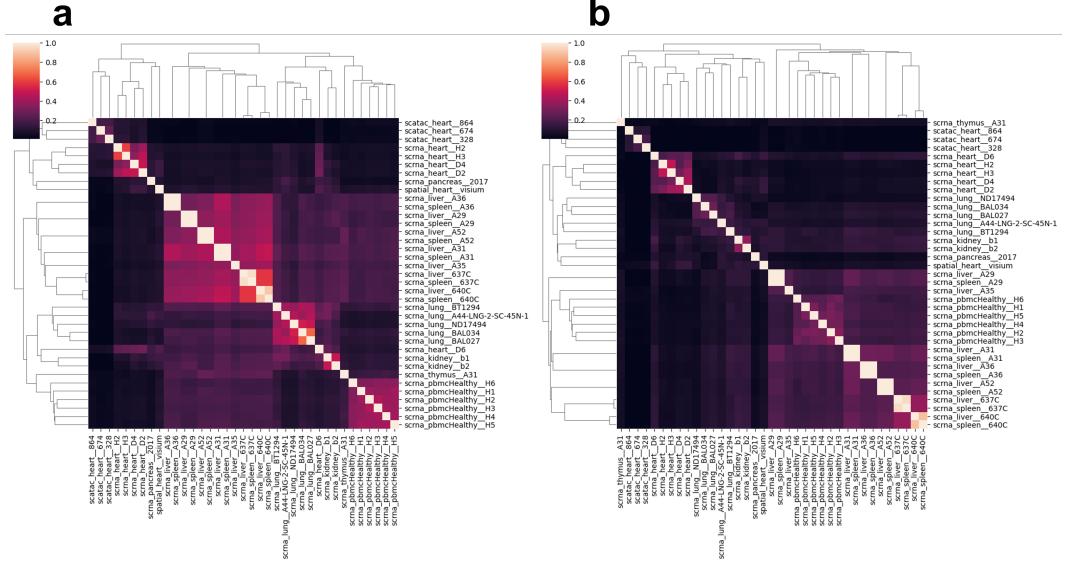


Figure 11: Cluster heatmaps of anchor genes for multimodal data. **(a)** represents the cluster heatmap based on HVGs overlap, **(b)** represents the cluster heatmap based on the neighbor genes overlap of HVGs.

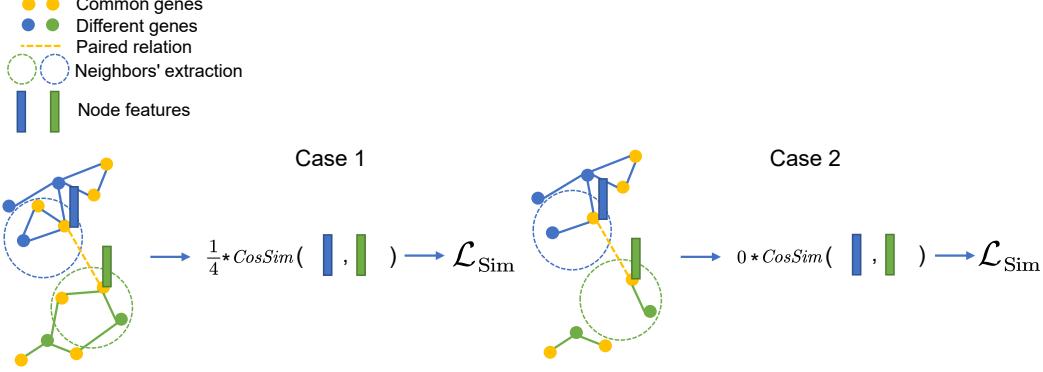


Figure 12: The need for weight similarity learning illustrated by two cases.

associated with common HVGs. Conversely, embeddings for functionally distinct genes should differ, enabling their identification through community detection-based clustering algorithms.

H Metrics details

Here we assume that HVGs from the same tissues with the same name have similar functions across different datasets, and we also consider evaluating the preservation of the known gene co-expression relation by using node similarity. These metrics include:

- edge AUC: Edge Area under the ROC Curve (AUC) is a metric widely used in the edge prediction task. We calculate the AUC score between the multiplication of normalized embeddings with Sigmoid threshold ($\text{Sigmoid}(zz^T)$) and the true edge relation. This metric can reflect the co-expression information we have in our embedding space.
- common genes ASW: Common genes ASW represents the ASW score we calculated based on common genes across different tissues. Since the overlap genes of different tissues are

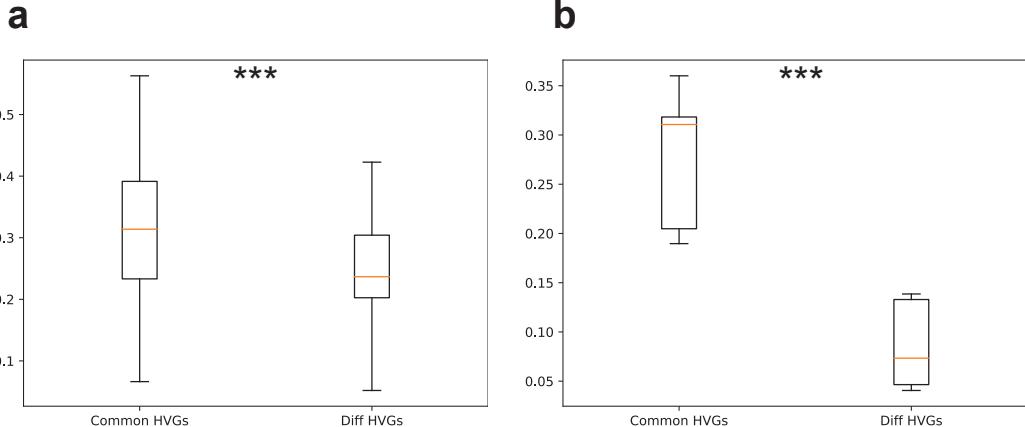


Figure 13: Boxplots for the neighbor overlap score of HVGs from different tissues. We used Kolmogorov–Smirnov test [97] to compare the difference between the two groups of genes. The significance level of group difference is shown by the stars (* * * : P-value ≤ 0.01). **(a)** represent the distribution of overlap scores for heart tissue, grouped by common HVGs and different HVGs. **(b)** represent the distribution of overlap score for cerebrum tissue, grouped by common HVGs and different HVGs.

not very large, based on the assumption that different highly variable genes have similar functions in different datasets, we can utilize common genes ASW as a metric to evaluate our integration performance.

- common genes Graph connectivity: Graph connectivity is a metric to evaluate the connectivity based on the ratio between the number of nodes in the largest connected component of the graph constructed by the gene cluster and the number of nodes belonging to the graph. We calculate the graph connectivity score for each gene cluster and take the average.
- common genes iLISI: The inverse Simpson’s index (iLISI) determines the number of common genes that can be drawn from a neighbor list before one dataset (or datasets from the same tissue) is observed twice. Therefore, this metric can be used to evaluate the integration level of the common genes across different datasets or tissues based on the choice of our index type.
- common genes ratio: The common genes ratio is defined based on the weighted average of one minus the ratio of unique genes from one cluster to the total genes in the cluster across all the clusters. We utilize the Leiden method to generate the clusters. This metric can be used to evaluate the level of integration of the same functional genes by a given method.
- neighbors overlap: The neighbor overlap value is defined based on the neighbors of one gene in the co-expression network. For each cluster, we calculate the similarity of genes’ neighbors among different datasets and perform a weighted average of the similarity. A larger value means that the method is able to integrate similar genes.

I Shared transcription factors and pathways analysis

Transcription factors (TFs) are crucial in transcriptional regulation [37]. By examining the overlap of TFs across diverse datasets, we can assess the similarities between distinct modal data in terms of transcriptional regulation. We employed the TRRUST v2 database [37] to identify regulators for each gene within the gene embedding space. Subsequently, for every cluster, we extracted the transcription factors associated with the genes and computed the transcription factor overlap for the various multimodal data linked to the genes. We iteratively repeated this process for each cluster and documented the number of shared transcription factors based on different combinations of biological multimodal data. The results are illustrated in Figure 14.

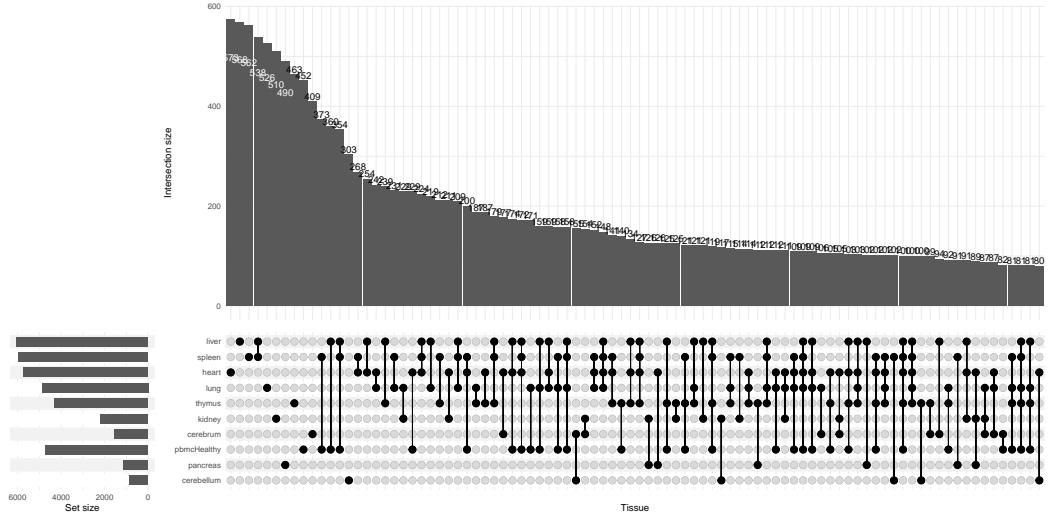


Figure 14: Results of common TFs discovery.

Figure 14 reveals that tissues with similar functions exhibit greater TF overlap, such as the heart and lung group, and the spleen and liver group. These findings demonstrate that our generated embeddings effectively capture the similarities between various data types. ComplexUpset [51] was utilized to create this figure.

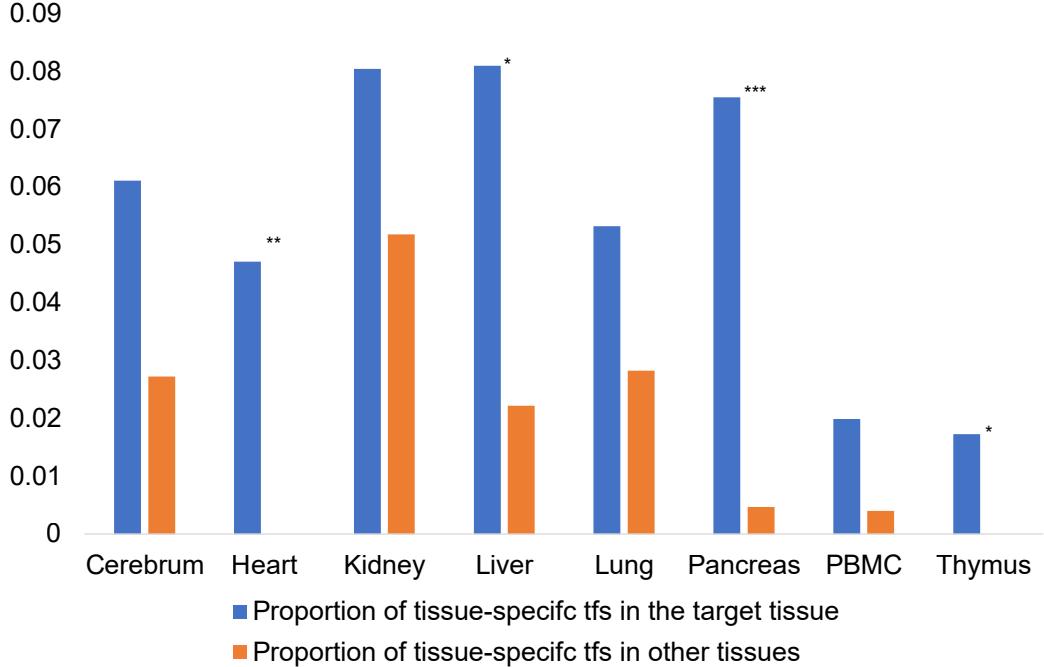


Figure 15: The proportion of tissue-specific TFs enrichment in the target tissue (blue) and in other tissues (orange). Here the significant enrichment in blue groups are marked by stars (* : P-value ≤ 0.1 ; ** : P-value ≤ 0.05 ; *** : P-value ≤ 0.01).

Additionally, we can emphasize the enrichment status of tissue-specific TFs, as depicted in Figure 15. We utilized the TF information from [107] to compute the enrichment proportion for tissue-specific

TFs enriched in the given tissue and other tissues. A one-tailed Fisher’s exact test was also conducted to assess the significance of enrichment differences. The figure suggests that our gene embeddings consistently enrich tissue-specific TFs across all evaluated tissues. Notably, for some tissues, the enrichment is significant, which aligns with the conclusions from [16]. However, in other tissues, the difference is not significant. A potential explanation is that such tissues may play a critical role in the interaction between different tissues for specific biological processes.

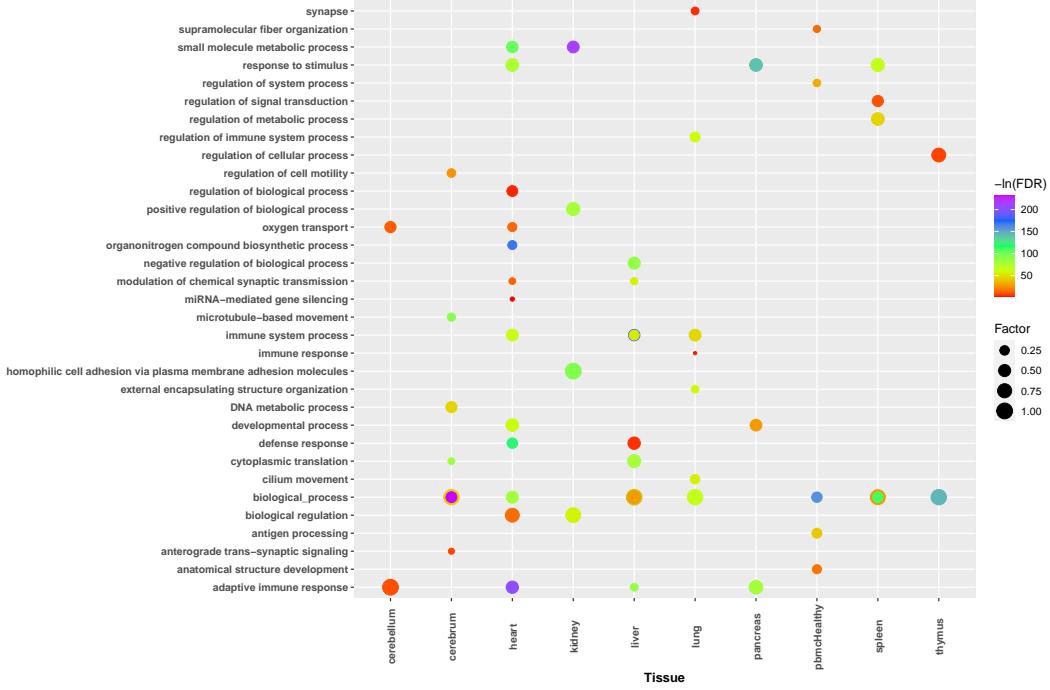


Figure 16: Top1 enriched pathways across different gene function clusters grouped by major tissue in each cluster. Here the color of bubbles represents the $-\ln(FDR)$ value and the size of bubbles represents the value of rich factor for different pathways.

In Figure 16, we determined the most enriched pathway for each cluster and classified it into the most representative tissue within that cluster by GOEA. The representative tissue was chosen based on the highest proportion of the source tissue for genes. This figure allows us to conclude that the heart tissue is the most versatile in the human body, followed by liver and lung tissues, which are also crucial tissues in humans.

J Multi-species gene embeddings

Integrating datasets from different species based on cells is challenging because we always need extra information, either orthology genes (genes from different species whose evaluation is only related to speciation events) information or protein embeddings information [76]. However, using genes as anchors simplifies this task. We examined gene embeddings generated by MuSe-GNN across three mammalian species, including humans, lemurs, and mice. As shown in Figure 17 (a) and (b), our method effectively integrated information from different species into a shared space, preserving gene similarity across various techniques, tissues, and species. We also observed the co-embedding of genes from different species but the same tissue, supporting the similarity of evaluation direction. Since the enrichment of orthology genes can also be used as proof to evaluate whether genes with similar functions are grouped [32], we collected genes from different clusters with three species pairs and computed the proportion of these genes in the orthology genes database [4]. Furthermore, we compared the results based on gene embeddings to random selection as a null approach. From

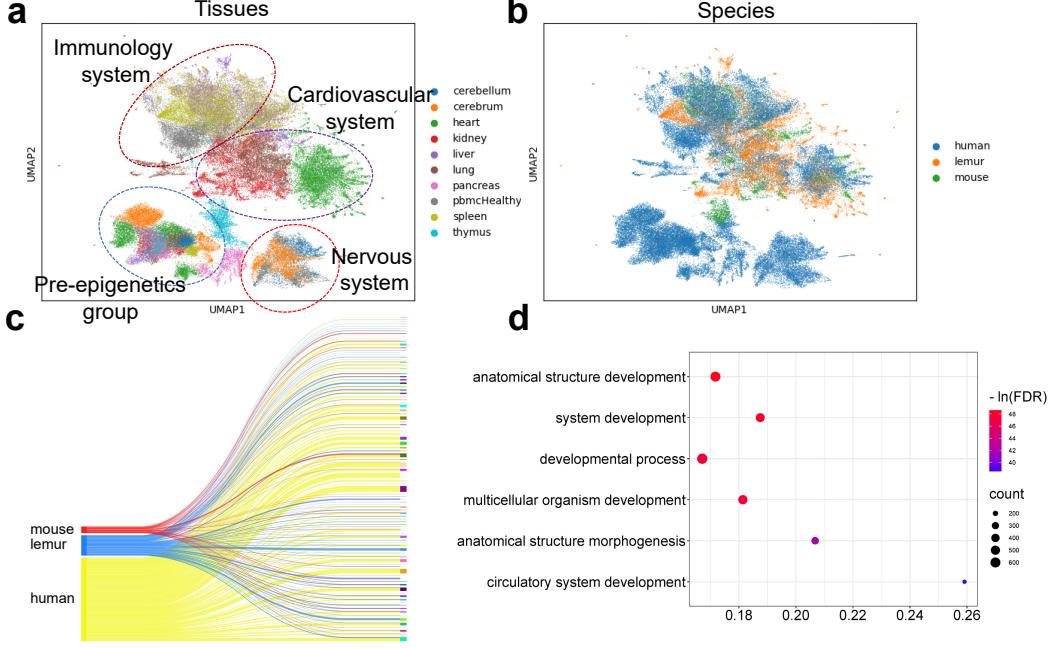


Figure 17: Gene representation learning results for multimodal biological data with different species. (a) represents the UMAPs of gene embeddings colored by tissue type. (b) represents the UMAPs of gene embeddings colored by common function groups. (c) is a Sankey plot [80] to show the genes overlap of different species in the same clusters. (d) shows the top5 pathways related to the genes in the special cluster discovered by GOEA. The bubble plots in this paper were created based on ggplot2 [102].

Figure 18, we could conclude that gene embeddings from our model can effectively enrich genes from different species but with similar functions.

Furthermore, Figure 17 (c) reveals that genes from different species are distributed across multiple common function clusters. To explore shared gene functions across species, we selected one common function cluster containing genes from all three species for GOEA analysis. Since most genes in this cluster originated from heart tissue, we anticipated pathway enrichment related to multicellular systems and circulatory systems, as displayed in Figure 3 (d). Thus, we demonstrated that MuSe-GNN can learn gene embeddings from multiple species and identify genes with similar functions across various species.

K Lung cancer analysis

In this section, we applied MuSe-GNN to analyze differentially co-expressed genes in two types of datasets, comprising lung cancer samples and healthy samples. We converted the scRNA-seq results into distinct graphs according to our graph construction strategy and trained MuSe-GNN to generate gene embeddings. We then employed the Leiden algorithm to identify common functional gene groups. After collecting essential genes with cancer-related special functions, we used GOEA and IPA to uncover more biological information.

As shown in Figure 19 (a)-(c), we observed that for lung tissue, the co-expression relationships of genes in lung tumor cells and normal cells differ significantly. Figure 19 (d) displays the top pathways identified by GOEA for genes recognized as specifically co-expressed in cancer cells, with most pathways related to tissue development and cell migration. These pathways align with the fact that cancer cells often undergo multiple divisions and frequently relocate. Furthermore, using IPA, we analyzed causal networks and disease functions. Figure 19 (e) illustrates an example of a causal network discovered in a specific gene group, primarily regulated by ethyl protocatechuate, indicating

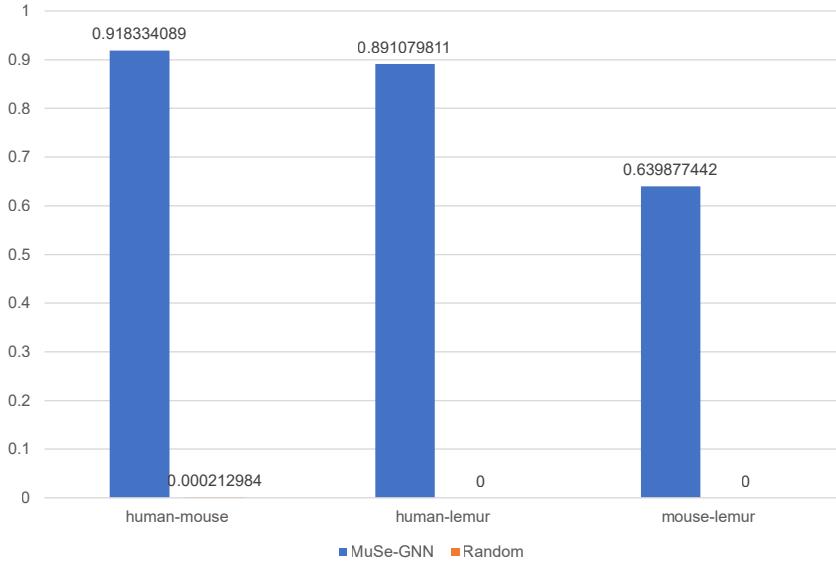


Figure 18: Compare the proportion of orthology genes discovered by MuSe-GNN and random selection.

a chemical-dominated regulatory network. Thus, with the genes identified by MuSe-GNN, we can uncover various specific regulatory networks present in cancer and investigate strategies to inhibit cancer cell growth and metastasis. Lastly, Figure K (f) reveals that functions with low FDR rates include tissue development and cancer disease.

Our analysis demonstrates that MuSe-GNN can be utilized to discover functional gene clusters exhibiting specificity in cancer, assisting researchers in gaining a deeper understanding of gene expression and regulatory relationships in cancer.

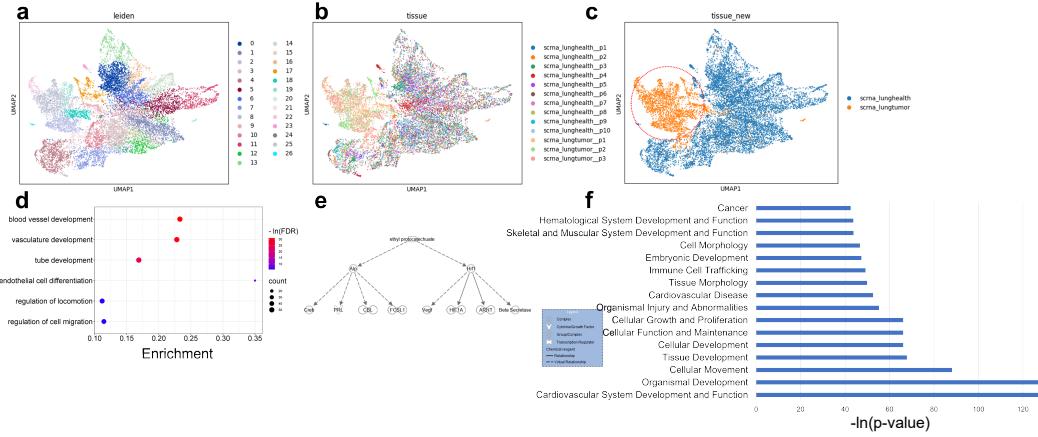


Figure 19: Gene representation learning results for samples with cancer and healthy samples. (a) represents the UMAPs of gene embeddings colored by functional groups. (b) represents the UMAPs of gene embeddings colored by datasets. (c) represents the gene embeddings colored by the conditions, and the red circle reflects the differential co-expression genes. (d) shows the top6 pathways related to the genes in the special cluster discovered by GOEA. (e) represents the causal network existing in the special cluster discovered by IPA. (f) represents the top diseases & biological functions discovered by IPA.

L Gene Function Analysis

Here we continued discussing the application of gene embeddings in the gene function prediction task. We further intended to predict the characters of genes in the regulation process (as transcript factor or not) [37]. We used MuSe-GNN to generate gene embeddings for different datasets based on an unsupervised learning framework and utilized the gene embeddings as training dataset to predict the function of genes based on k-NN classifier.

In this task, we evaluated the performance of MuSe-GNN based on scRNA-seq datasets from different tissues, comparing it to the prediction results based on raw data or PCA. On average, the gene embeddings from MuSe-GNN are the most powerful embeddings, yielding highest accuracy in the validation set. Moreover, the accuracy based on gene embeddings from MuSe-GNN also surpass the score based on raw data or output of PCA in different tissues: Heart, Lung, Liver, Thymus, PBMC, Cerebellum and Pancreas. Such result is shown in Figure 20.

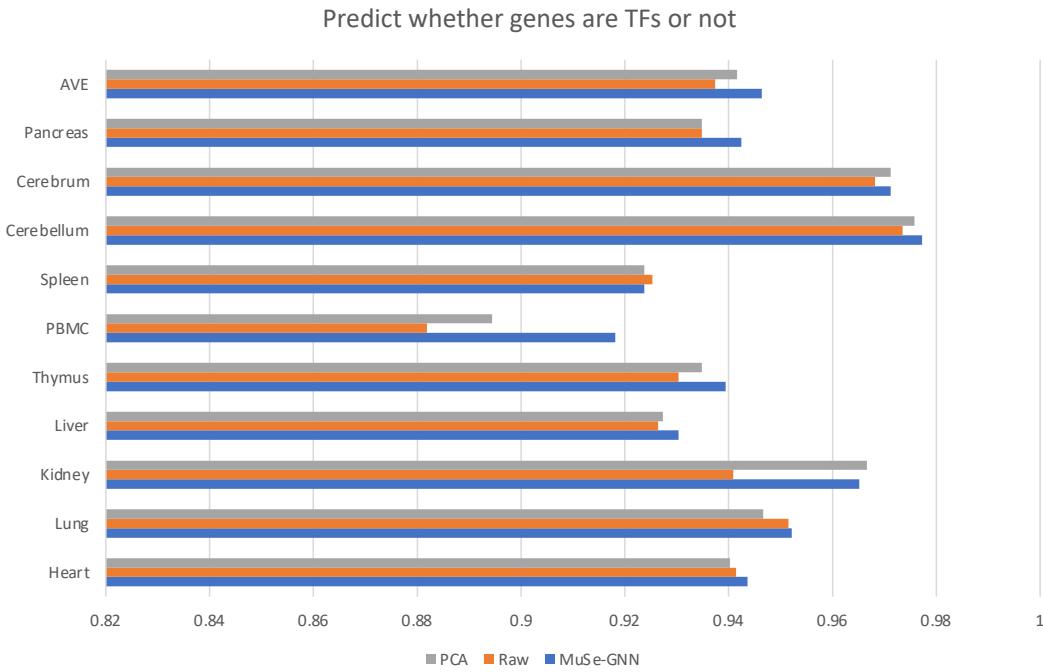


Figure 20: Accuracy for Gene-TF prediction across different tissues.

M Datasets information

Tissue	Datatype	Link
Heart	scRNA-seq	https://www.heartcellatlas.org/ https://data.mendeley.com/datasets/mbvhhf8m62/2
Lung	scRNA-seq	https://beta.fastgenomics.org/datasets/detail-dataset-427f1eee6dd44f50bae1ab13f0f3c6a9
Kidney	scRNA-seq	https://www.kidneycellatlas.org/
Liver, Thymus, Spleen	scRNA-seq	https://singlecell.broadinstitute.org/single_cell/study/SCP1845/cross-tissue-immune-cell-analysis-reveals-tissue-specific-features-in-humans?scpbr=human-cell-atlas-main-collection#study-download https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/
Cerebrum, Cerebellum	scRNA-seq	https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development
PBMC (with COVID)	scRNA-seq	https://www.covid19cellatlas.org/index.patient.html PBMCs
Lung (with Cancer)	scRNA-seq	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE196303
Pancreas	scRNA-seq	https://azimuth.hubmapconsortium.org/references/#Human%20-%20Pancreas
Heart	scATAC-seq	https://portal.hubmapconsortium.org/browse/dataset/ba41e71358136f6a202114681a217a95 https://portal.hubmapconsortium.org/browse/dataset/d180976b42a9ed8b5fb12a508f79a238 https://portal.hubmapconsortium.org/browse/dataset/d39a1b1f40bbab8990078d343b332cdc https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/
Lung, Kidney, Thymus, Spleen, Cerebrum, Cerebellum, Pancreas	scATAC-seq	https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/
Heart	Spatial Transcriptomics	https://support.10xgenomics.com/spatial-gene-expression/datasets/1.1.0/V1_Human_Heart
Cerebrum	Spatial Transcriptomics	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205055

Figure 21: Information about multimodal biological datasets from Human.

Tissue	Datatype	Link
Heart, Lung, Kidney, Liver, Thymus, Spleen	scRNA-seq	https://figshare.com/articles/dataset/Tabula_Microcebus_v1_0/14468196?file=31777475 https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_microfluidic_emulsion_v2_/5968960/2

Figure 22: Information about multimodal biological datasets from Lemur and Mouse.

The statistics of different graphs can be found in Supplementary File 1.

