Multi-omics Integration via Graph Learning

Computational Biology

Wenda Chu, Minsi Lu, Botian Wang, Sihang Zeng, Shiyu Zhao

Content

1. Introduction
2. Improvement: Graph Construction
3. Improvement: Principal Neighbor Aggregation
4. Experiment result & Case study

Introduction

Multi-omics Integration

Embed single-cell data from different sequencing methods (different omics) into a same latent space. In this project, we use scRNA-seq and scATAC-seq data.

Data

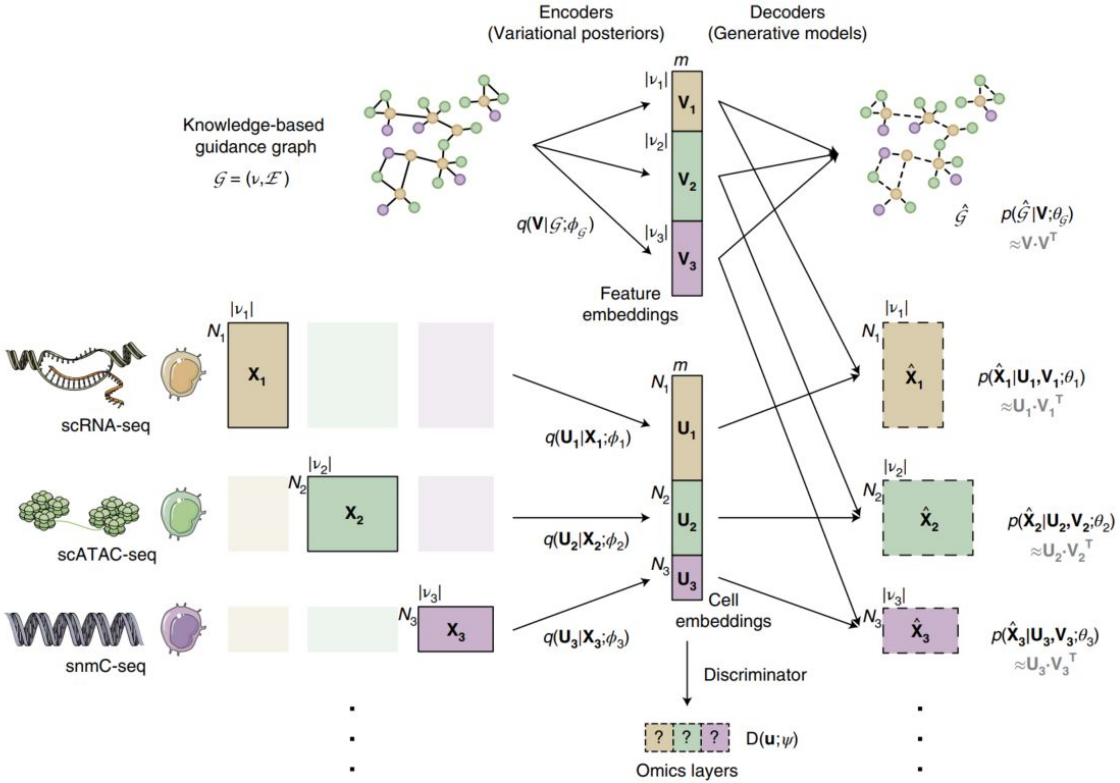
- Dataset: Chen-2019
- Sequencing method: SNARE-seq
- Cell: mouse cerebral cortex
- Data format:
  - scRNA-seq: (9190, 28930)
  - scATAC-seq: (9190, 241757)

genes	chrom	chromStart	chromEnd
0610005C13Rik	chr7	45567793	45575327
0610009B22Rik	chr11	51685385	51688874
0610009E02Rik	chr2	26445695	26459390
0610009L18Rik	chr11	120348677	120351190
0610010F05Rik	chr11	23564960	23633639

peaks	chrom	chromStart	chromEnd
chr1:3005833-3005982	chr1	3005833	3005982
chr1:3094772-3095489	chr1	3094772	3095489
chr1:3119556-3120739	chr1	3119556	3120739
chr1:3121334-3121696	chr1	3121334	3121696
chr1:3134637-3135032	chr1	3134637	3135032

Chen, S., Lake, B.B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 37, 1452–1457 (2019). https://doi.org/10.1038/s41587-019-0290-0

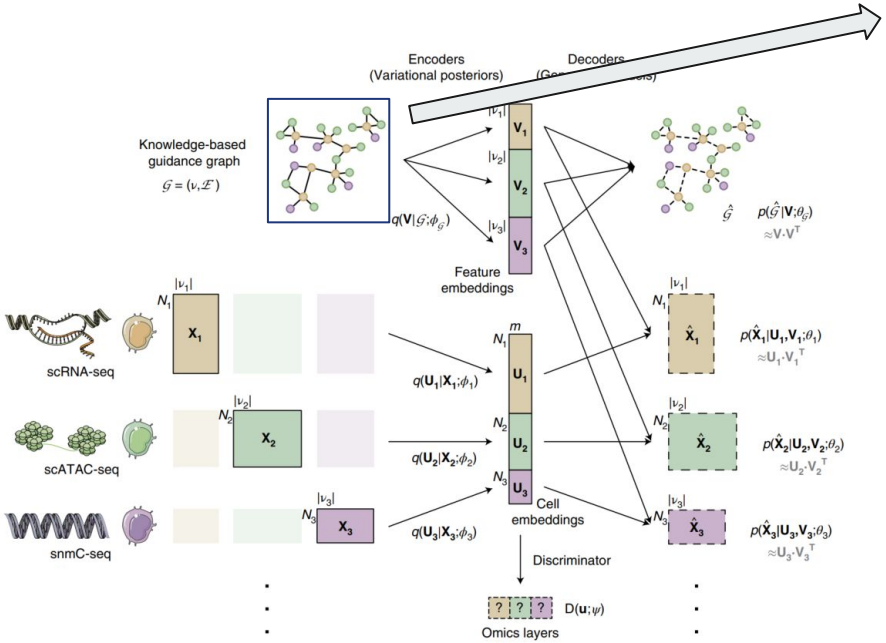
GLUE

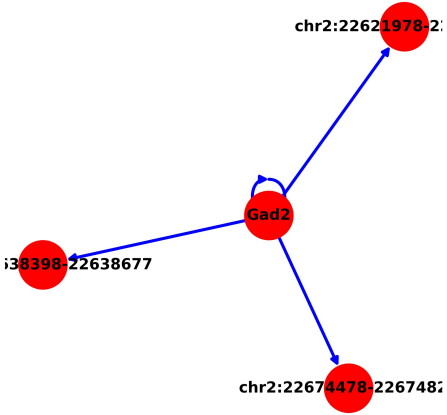


Cao, ZJ., Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 40, 1458–1466 (2022). https://doi.org/10.1038/s41587-022-01284-4

Improvement: Graph Construction

Graph structure in GLUE

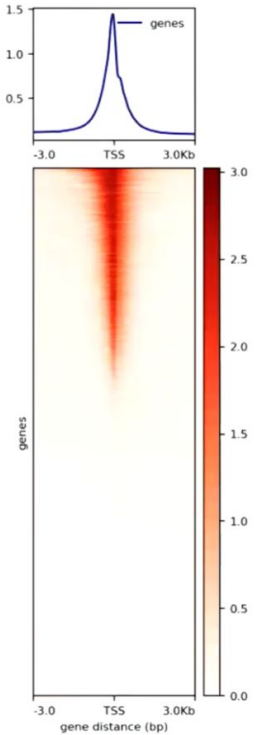




chr2:22621978-22622731 weight: 1.0
chr2:22638398-22638677 weight: 1.0
chr2:22674478-22674825 weight: 1.0
Gad2 weight: 1.0

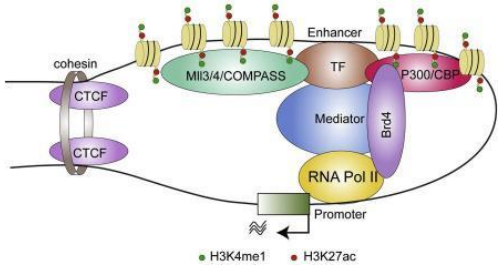
Simply using genomic overlap, regulatory relationships limited to promoters and gene regions

Our graph construction

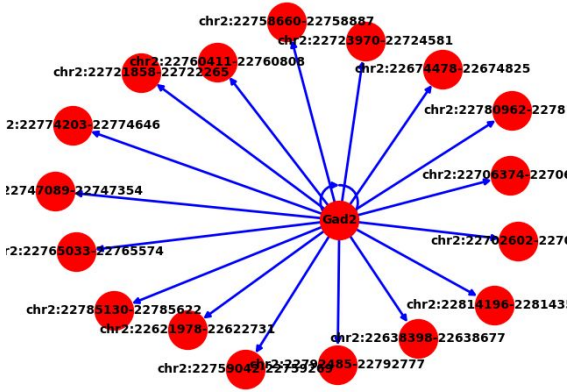


Connect atac peaks to RNA in the extended region and use a power-law function to decay the weights

More information, like cis-regulatory interactions :
CREs are often but not always upstream of the transcription site. CREs are non-coding sequences, but are often co-accessible with genes. We can infer cellular TF-gene interactions by TF-motif localization



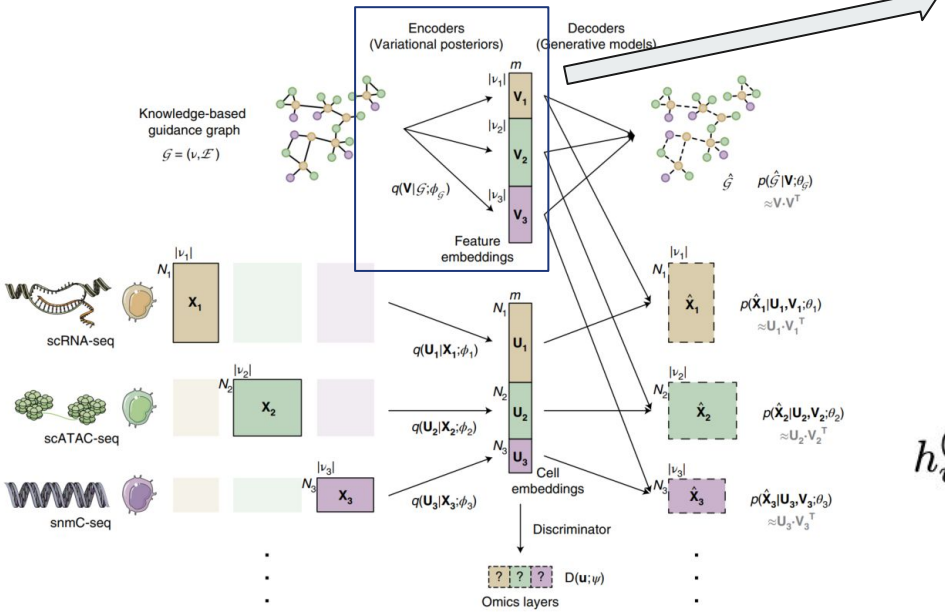
Simulated chromatin contact probability;
Genes are generally in an open state near the transcription start site (TSS).

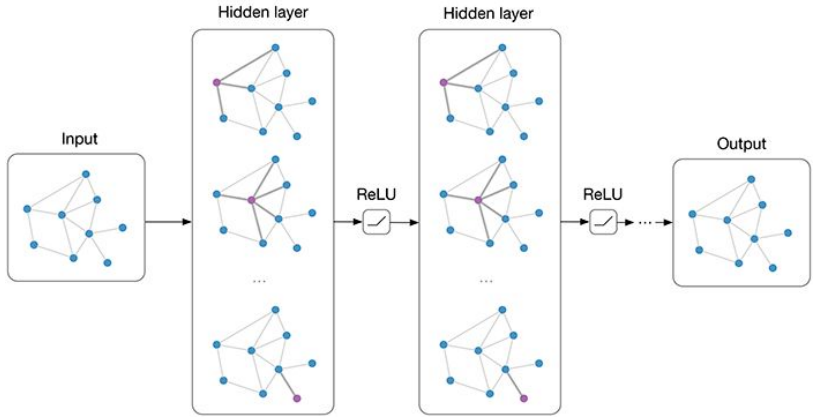


chr2:22621978-22622731 weight: 1.0
chr2:22638398-22638677 weight: 1.0
chr2:22674478-22674825 weight: 1.0
chr2:22702602-22702837 weight: 0.18153018383277017
chr2:22706374-22706918 weight: 0.1419794688239361
chr2:22721858-22722265 weight: 0.0800516008915812
chr2:22723970-22724581 weight: 0.0759383453268991
chr2:22747089-22747354 weight: 0.050050036153407765
chr2:22758660-22758887 weight: 0.043290732893392636
chr2:22759042-22759269 weight: 0.04310315612950408
chr2:22760411-22760808 weight: 0.042446203628414045
chr2:22765033-22765574 weight: 0.04039034907576054
chr2:22774203-22774646 weight: 0.036924270237414264
chr2:22780962-22781291 weight: 0.0347783312880787
chr2:22785130-22785622 weight: 0.03359313257336422
chr2:22792485-22792777 weight: 0.03171511241245008
chr2:22814196-22814358 weight: 0.027355373222849172
Gad2 weight: 1.0

Improvement:
Principal Neighbor Aggregator

Graph encoder in GLUE





Graph Convolutional Network

$$h_v^{(k)} = \phi\left(h_v^{(k-1)}, f\left(\left\{h_u^{(k-1)} : u \in \mathcal{N}(v)\right\}\right)\right)$$

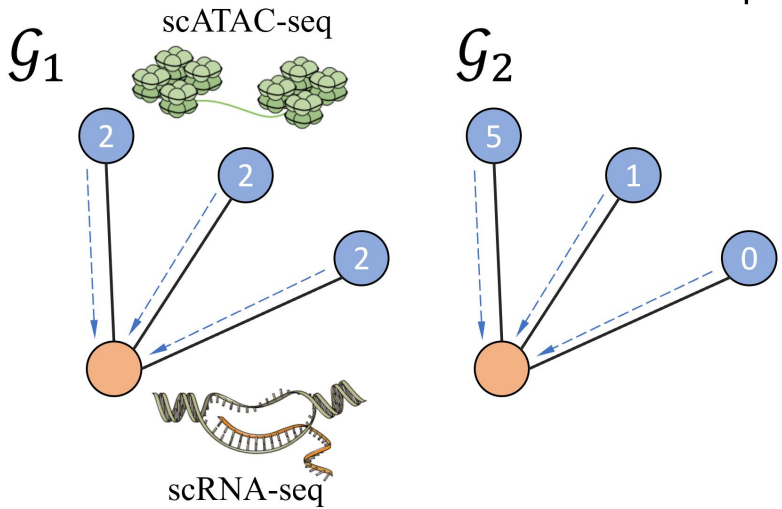
combinator aggregator

Is there any problem in MEAN aggregator? ⇦ Crucial in aggregating multi-omics info

Indistinguishability in single aggregator

A failure case for mean aggregator:

Mean Aggregator: $\mu_i(X) = \frac{1}{|N(i)|} \sum_{j \in N(i)} X_j$

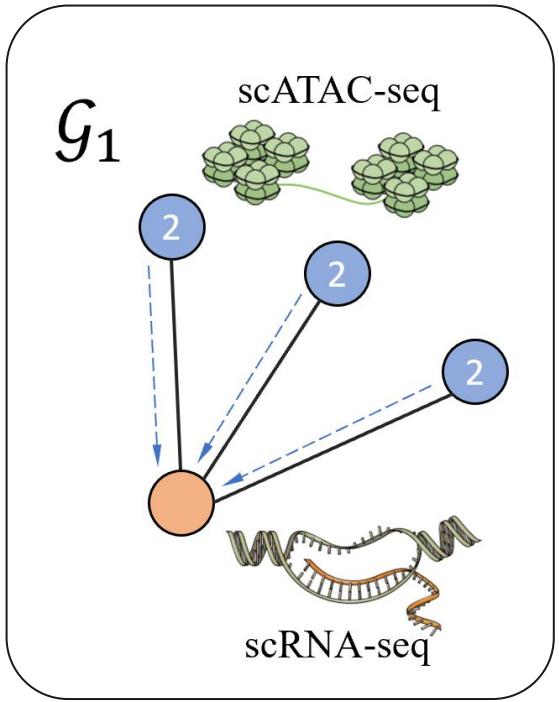


These two graphs are indistinguishable!

Indistinguishability in single aggregator

Failure cases for all single aggregators:

Combined aggregators: Principal Neighborhood Aggregator



Mean

$$\mu_i(X) = \frac{1}{|N(i)|}\sum_{j \in N(i)} X_j$$

Max

$$MAX_i(X) = \max_{j \in N(i)} X_j$$

Min

$$MIN_i(X) = \min_{j \in N(i)} X_j$$

Standard deviation

$$\sigma_i(X) = \sqrt{\mu_i(X^2) - \mu_i(X)^2}$$

MLP

Enriched info

Stronger distinguishability

Stronger expressiveness

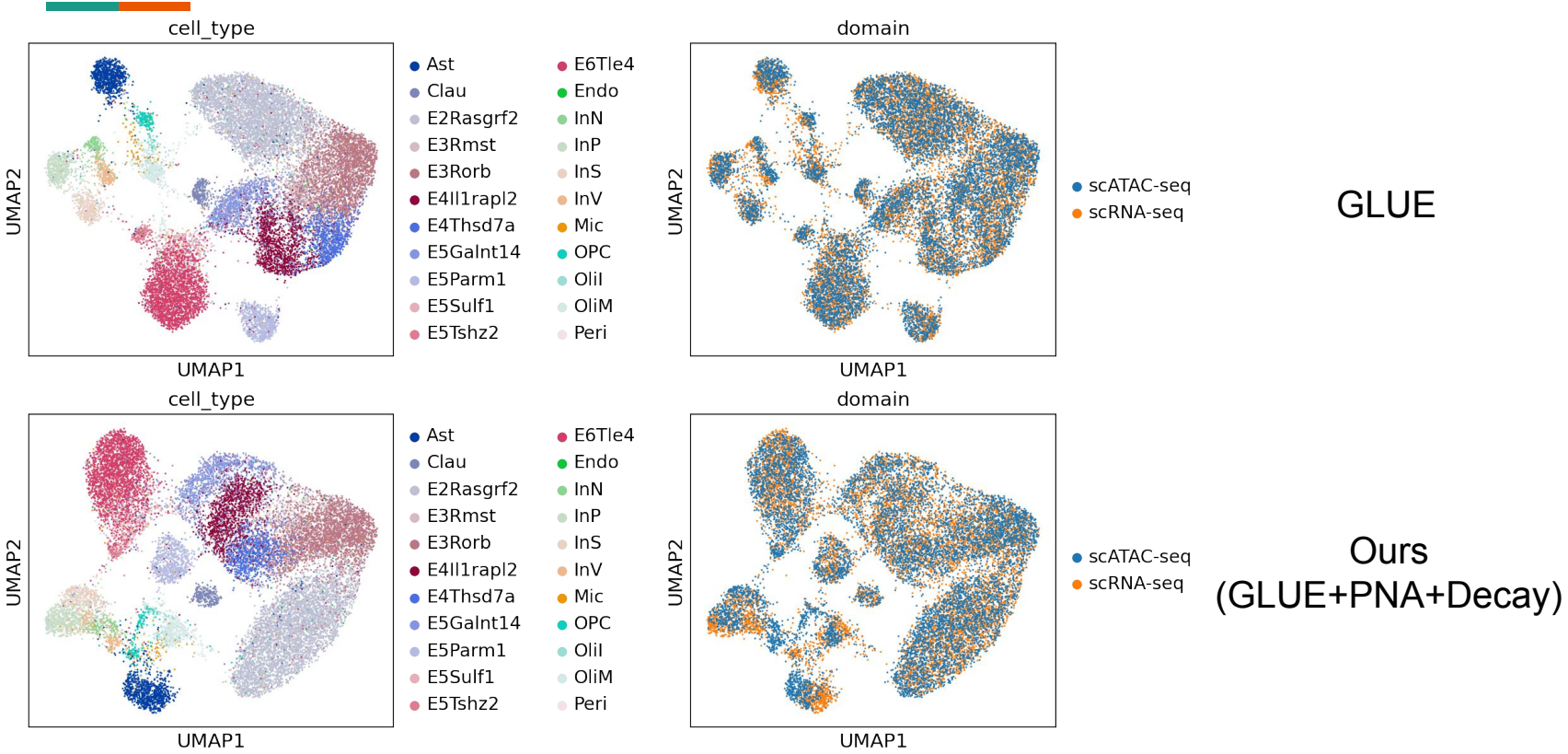
Better multi-omics integration

Result & Case Study

Results - Cell Integration Metrics

Metric Type	Mertics	GLUE	GLUE+PNA	GLUE+PNA+Decay
Biological Conservation Metrics	Average Silhouette Width	**0.5772**	0.5688	0.5707
	Mean Average Precision	**0.7809**	0.7463	0.7529
Omics Mixing Metrics	Graph Connectivity	**0.8502**	0.8377	0.8423
	Average Silhouette Width Batch	**0.9540**	0.9256	0.9141
	Seurat Alignment Score	0.9367	**0.9508**	0.9192
Cluster Performance	Normalized Mutual Information	**0.6997**	0.6823	0.6933

Results - Cell Integration Plots



Results - Regulatory Inference

Gene2peak graph statistics

Model / Graph Property	Number of nodes	Number of edges
GLUE	27266	25552
GLUE + PNA + Decay	27168	26645

Genes included in our graph but excluded in the original graph:

- Pcp2, Apom, Figla, Ctsw, Nmrk2, etc. (41 in total)

More relationships can be extracted using our graph!

Results - Regulatory Inference

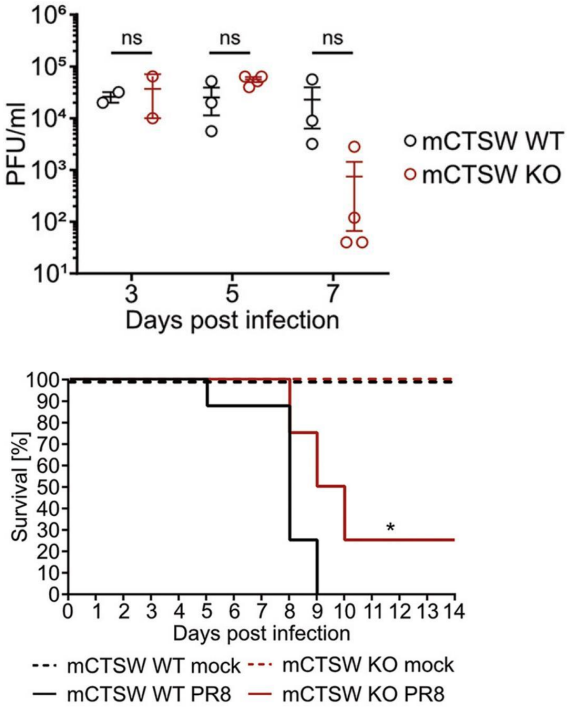
Case Study:

Cathepsin W is is a cysteine protease that is encoded by the ***CTSW*** gene.

It is required for IAV replication.

Reducing the levels of expression of CtsW reduces viral titers for IAV.

CTSW-deficient mice display a 25% increase in survival and a delay in mortality



Günther, S. C., Martínez-Romero, C., Sempere Borau, M., Pham, C., García-Sastre, A., & Stertz, S. (2022). Proteomic Identification of Potential Target Proteins of Cathepsin W for Its Development as a Drug Target for Influenza. *Microbiology spectrum*, *10*(4), e0092122.

Comments on the performance

- Cell integration metrics cannot fully measure the improvements of node features
- Our analysis on the gene2peak graph demonstrates the strengths of our approach
- Gene2TFMotif graph can also be generated using the Motif data

Conclusion

- We design a better graph construction method and graph encoder to integrate multi-omics data via graph learning
- Our method shows comparative results on several metrics and better results on regulatory inference.

Thank you for listening!