

# Multi-omics single-cell data integration and regulatory inference with graph-linked embedding


---


Presenter: Soomin Song


23 Nov 2023

# Outline

 Background and objectives

 Material and methods

 Results

 Discussion and conclusion

 Q/A

# Outline

 Background and objectives

 Material and methods

 Results

 Discussion and conclusion

 Q/A



# Background

- **Rapid Advancements in Single-Cell Technologies:**
  - “Recent years have seen significant developments in experimental methods allowing for the measurement of multiple omics modalities in single cells”
- **Understanding Modalities:**
  - A ‘modality’ refers to a specific type of biological data, such as genomic, transcriptomic, proteomic, or epigenomic information. Each modality provides unique insights into the cellular processes.
- **Integration Challenge:**
  - Despite these advancements, most single-cell datasets include only one modality. A key obstacle in integrating multi-omics data is the distinct feature spaces of different omics layers.

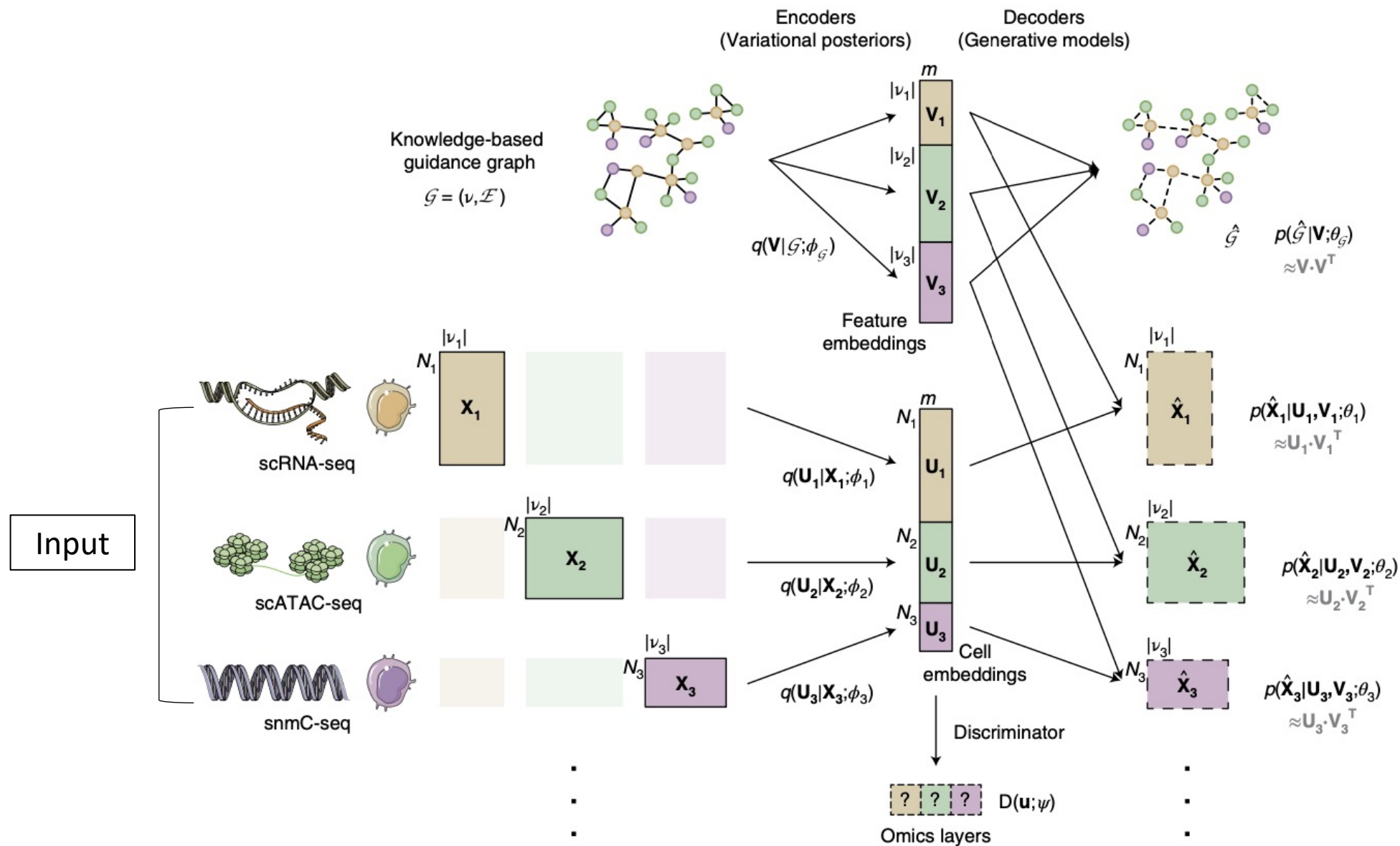


# Objectives

- **Introducing GLUE (Graph-Linked Unified Embedding):**
  - “GLUE” is a computational framework devised to overcome these integration challenges, enabling the combination of diverse omics data by modeling regulatory interactions across various modalities.
- **Aims of GLUE**
  - Integration Across Distinct Omics Modalities: GLUE is designed to unify different omics modalities, addressing the challenge of their distinct feature spaces and biological insights.
  - Enhanced Accuracy and Robustness: Through systematic benchmarking, GLUE has shown superior performance in accuracy, robustness, and scalability compared to existing methods.
  - Flexible and Scalable Design: Featuring a modular design, GLUE allows for flexible extensions and enhancements, suitable for various analysis tasks.




# Architecture of the GLUE framework



# Outline

 Background and objectives

 Material and methods

 Results

 Discussion and conclusion

 Q/A



# Materials and Methods

- The GLUE framework
  1. Input data
  2. Cell embedding
  3. Integration graph construction
  4. Alignment of cell type
  5. Regulatory network inference
  6. Network analysis





# Materials and Methods

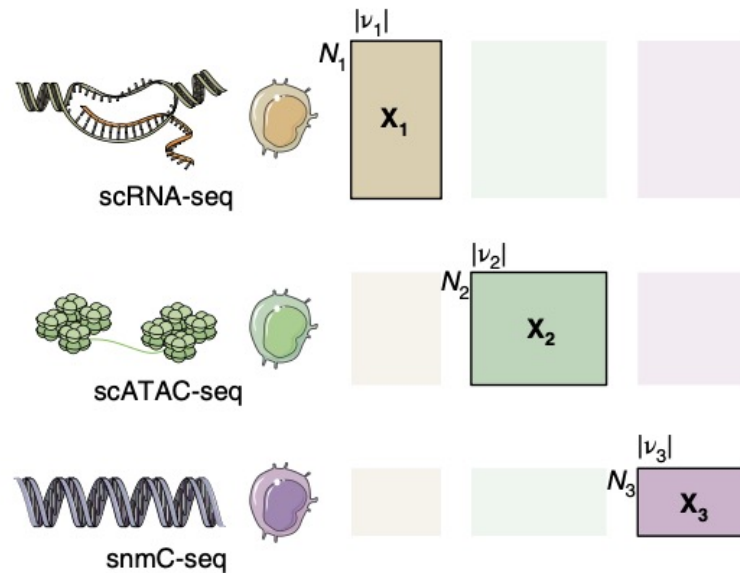
- Input data

GLUE starts with unpaired data from three different omics layers: scRNA-seq, scATAC-seq, and snmC-seq.

- $X_1, X_2, X_3$  : different type of measurement

- $N_1, N_2, N_3$  : the number of cells

- $|V_1|, |V_2|, |V_3|$  : the number of features (like genes or accessible regions) in each layer.



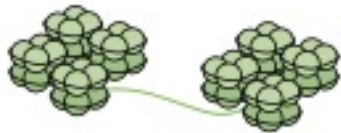


# Materials and Methods

- Input data



scRNA-seq



scATAC-seq



snmC-seq

## 1. Single-cell RNA Sequencing

- **What is it?** A technique to examine the gene expression profiles of individual cells.
- **Purpose:** Helps us understand what genes are active (being 'read' to make proteins) in each cell.
- **Importance:** Useful for identifying different cell types and states based on their unique gene expression patterns.

## 2. Single-cell Assay for Transposase-Accessible Chromatin using sequencing

- **What is it?:** A method to identify open regions in the chromatin structure of each cell.
- **Purpose:** Reveals parts of the DNA that are accessible and likely to be involved in controlling gene activity.
- **Importance:** Helps us understand how the organization of DNA impacts gene regulation at the single-cell level.

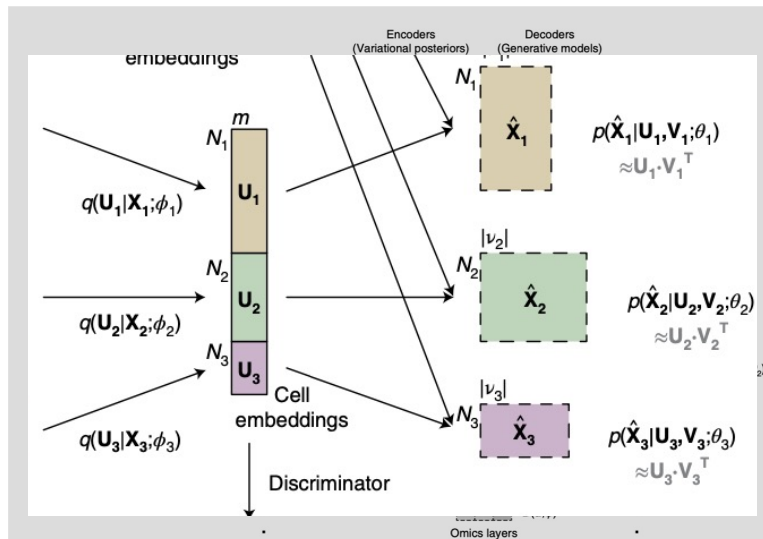
## 3. Single-nucleus Methylcytosine Sequencing

- **What is it?:** A method to identify open regions in the chromatin structure of each cell.
- **Purpose:** Reveals parts of the DNA that are accessible and likely to be involved in controlling gene activity.
- **Importance:** Helps us understand how the organization of DNA impacts gene regulation at the single-cell level.



# Materials and Methods

- Cell embedding: The preprocessed data from each omics layer is used to generate a low-dimensional representation of each cell called a cell embedding.
- Omics-specific variational autoencoders are utilized to learn low-dimensional cell embeddings  $|U_1|, |U_2|, |U_3|$  from each omics layer. These embeddings reduce the dimensionality of the data while preserving its essential characteristics.



encoders

$$q(\mathbf{u}|\mathbf{x}_k; \phi_k)$$

decoders

$$p(\mathbf{x}_k; \theta_k) = \int p(\mathbf{x}_k|\mathbf{u}; \theta_k) p(\mathbf{u}) d\mathbf{u}$$

$x_k$  = cells in the kth layer  
 $u$  = cell latent variable  
 $\theta_k$  = parameters in the decoder  
 $\phi_k$  = parameters in the encoders

Different autoencoders are independently parameterized and trained on separate data.

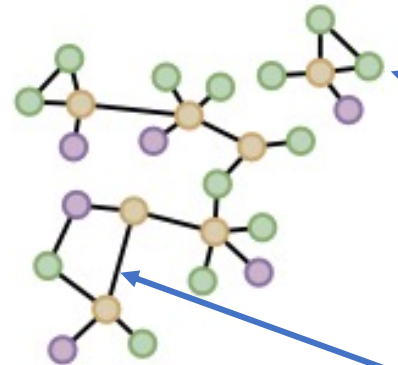


# Materials and Methods

- **Integration Graph Construction(Guidance Graph):** A knowledge-based guidance graph  $G=(V,E)$  is constructed, where vertices  $V$  represent the union of omics features across all layers. This graph encodes prior knowledge about regulatory interactions among these features. Guidance graph is treated as observed variable and model it as generated by low-dimensional feature latent variables (that is, feature embeddings).

Knowledge-based  
guidance graph

$$\mathcal{G} = (V, \mathcal{E})$$



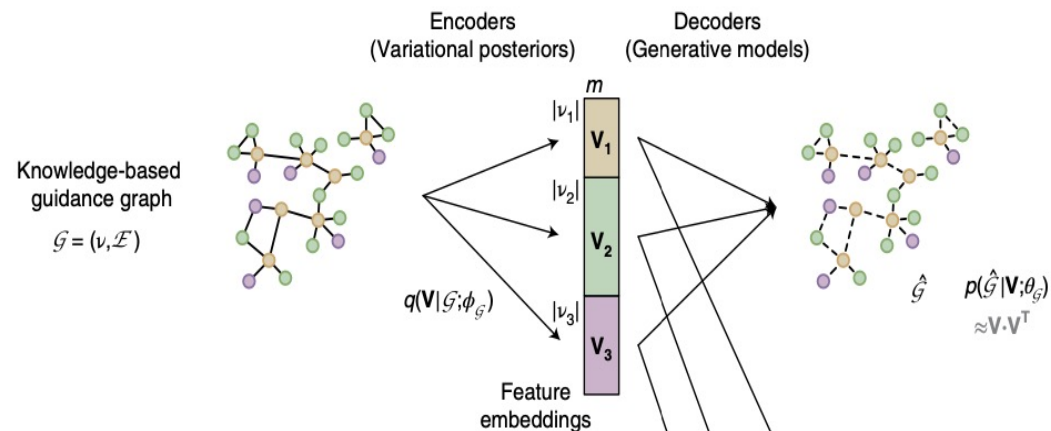
Vertices: features of different  
omics layers (ex. Genes,  
chromatin regions)

Edges: Signed regulatory  
interactions (+ edge between an  
accessible region and its putative  
downstream gene)



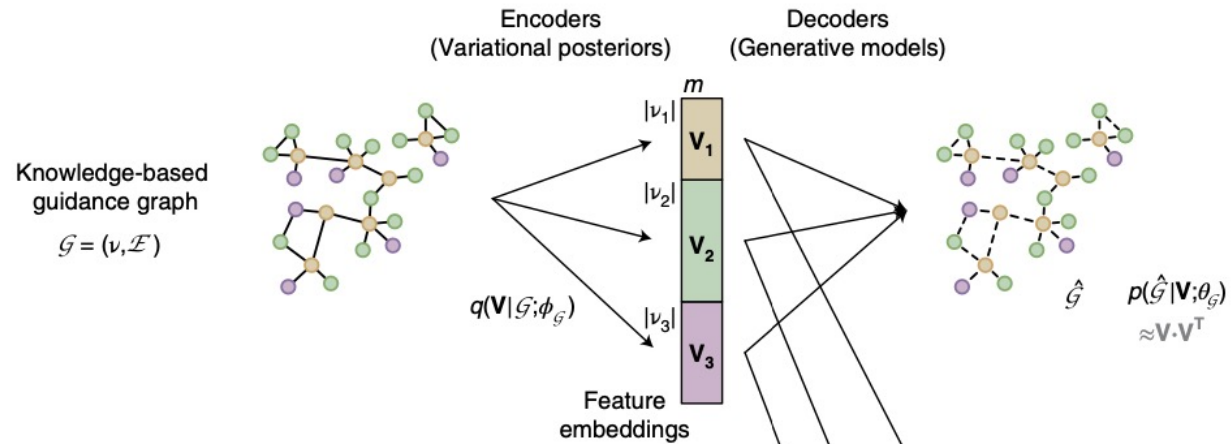
# Materials and Methods

- Alignment of Cell Type (Feature Embeddings and Discriminator):
  1. Feature embeddings  $V$  are learned from the guidance graph using a graph variational autoencoder. These embeddings are then used in data decoders to reconstruct omics data through an inner product with cell embeddings, linking the omics-specific data spaces.
  2. An omics discriminator  $D$  aligns the cell embeddings from different omics layers using adversarial learning, ensuring consistent orientation across embeddings.





# Materials and Methods



Treating the guidance graph as observed variable, Modeling with omics layers by the combination of feature latent variables, cell latent variable

$\mathbf{V}$  = all feature embedding into a single matrix

$$p(\mathbf{x}_k, \mathcal{G}; \theta_k, \theta_{\mathcal{G}}) = \int \underbrace{p(\mathbf{x}_k | \mathbf{u}, \mathbf{V}; \theta_k)}_{\text{Data decoder}} \underbrace{p(\mathcal{G} | \mathbf{V}; \theta_{\mathcal{G}})}_{\text{Guidance graph decoder}} \underbrace{p(\mathbf{u})}_{\text{cell embedding}} \underbrace{p(\mathbf{V})}_{\text{Feature embedding}} d\mathbf{u} d\mathbf{V}$$



# Materials and Methods

- Regulatory network inference (Decoders): The decoders, with learnable parameters  $\theta_1, \theta_2, \theta_3, \theta_G$  use the learned cell and feature embeddings to reconstruct the original omics data, thereby inferring regulatory networks.

$p(\mathbf{x}_k | \mathbf{u}, \mathbf{V}; \theta_k)$  (that is, data decoders) are built on the inner product between the cell embedding  $\mathbf{u}$  and feature embeddings  $\mathbf{V}_k$



# Materials and Methods

- Network analysis: The final step involves analyzing the inferred networks to identify key regulatory interactions. The discriminator  $\psi$  (hyperparameter) fine-tunes the cell embeddings to ensure that they align well across different omics layers, refining the network analysis. A discriminator  $D$  with a  $K$ -dimensional softmax output predicts the omics layers of cells based on their embeddings  $\mathbf{u}$ .

$$\mathcal{L}_D(\phi, \psi) = -\frac{1}{K} \sum_{k=1}^K \mathbb{E}_{\mathbf{x}_k \sim p_{data}(\mathbf{x}_k)} \mathbb{E}_{\mathbf{u} \sim q(\mathbf{u}|\mathbf{x}_k; \phi_k)} \log D_k(\mathbf{u}; \psi)$$

(The discriminator  $D$  is trained by minimizing the multiclass classification cross entropy)





# Materials and Methods

Iteratively learning through input data

- 1) The discriminator is updated according to objective equation (The discriminator D is trained by minimizing the multiclass classification cross entropy)

$$\min_{\psi} \lambda_D \cdot \mathcal{L}_D(\phi, \psi)$$

- 2) The data and graph autoencoders are updated according to equation

$$\max_{\theta, \phi} \lambda_D \cdot \mathcal{L}_D(\phi, \psi) + \lambda_G K \cdot \mathcal{L}_G(\theta_G, \phi_G) + \sum_{k=1}^K \mathcal{L}_{\mathcal{X}_k}(\theta_k, \phi_k, \phi_G)$$

# Outline

 Background and objectives

 Material and methods

 Results

 Discussion and conclusion

 Q/A



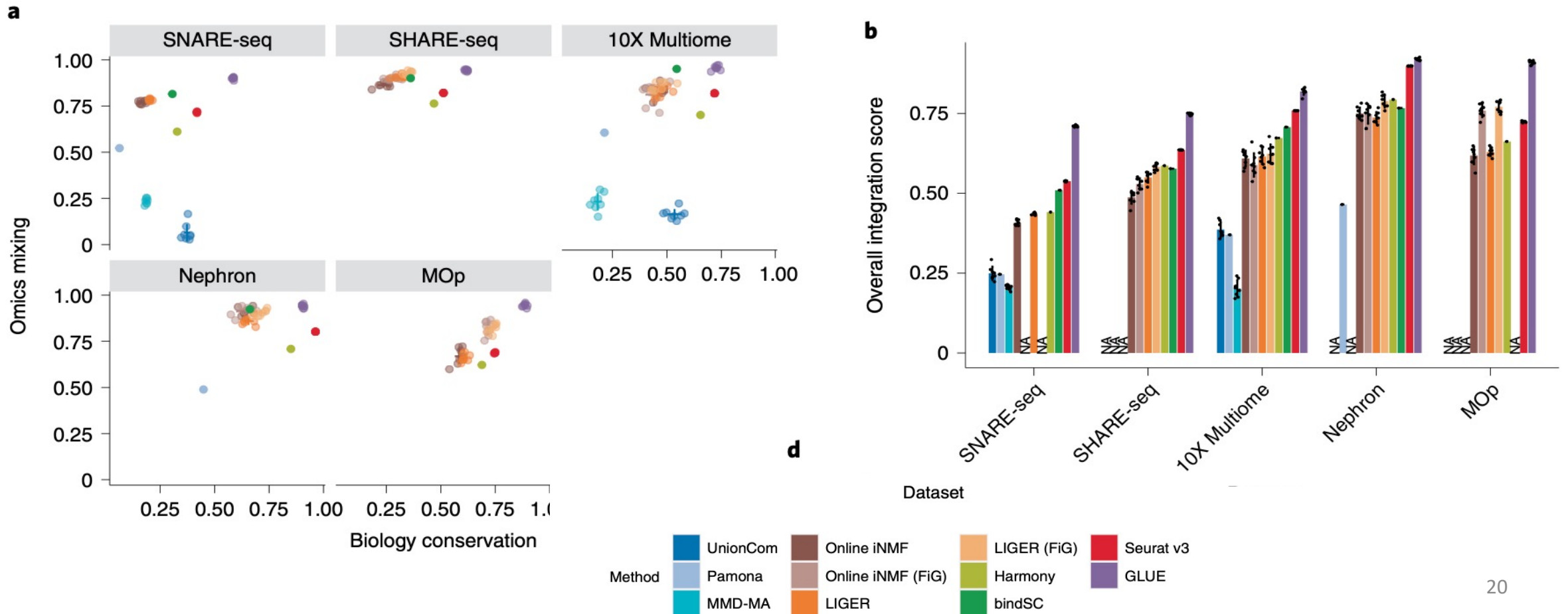
# Results

- Result 1: Integrating Performance of GLUE
  - GLUE was compared with multiple popular unpaired multi-omics integration methods.
- Result 2: Application of GLUE to triple-omics integration and regulatory inference
  - GLUE was applied to integrate gene expression, chromatic accessibility, and DNA methylation in neuronal cells of the adult mouse cortex.
- Result 3: Application of GLUE to regulatory inference
  - GLUE's incorporation of a regulatory graph enables a Bayesian-like approach for regulatory inference.
- Result 4: Application of GLUE to constructing a multi-omics human cell atlas
  - GLUE achieved the integration of gene expression and chromatic accessibility data into a unified multi-omics human cell atlas.



# Result 1: Integrating Performance of GLUE

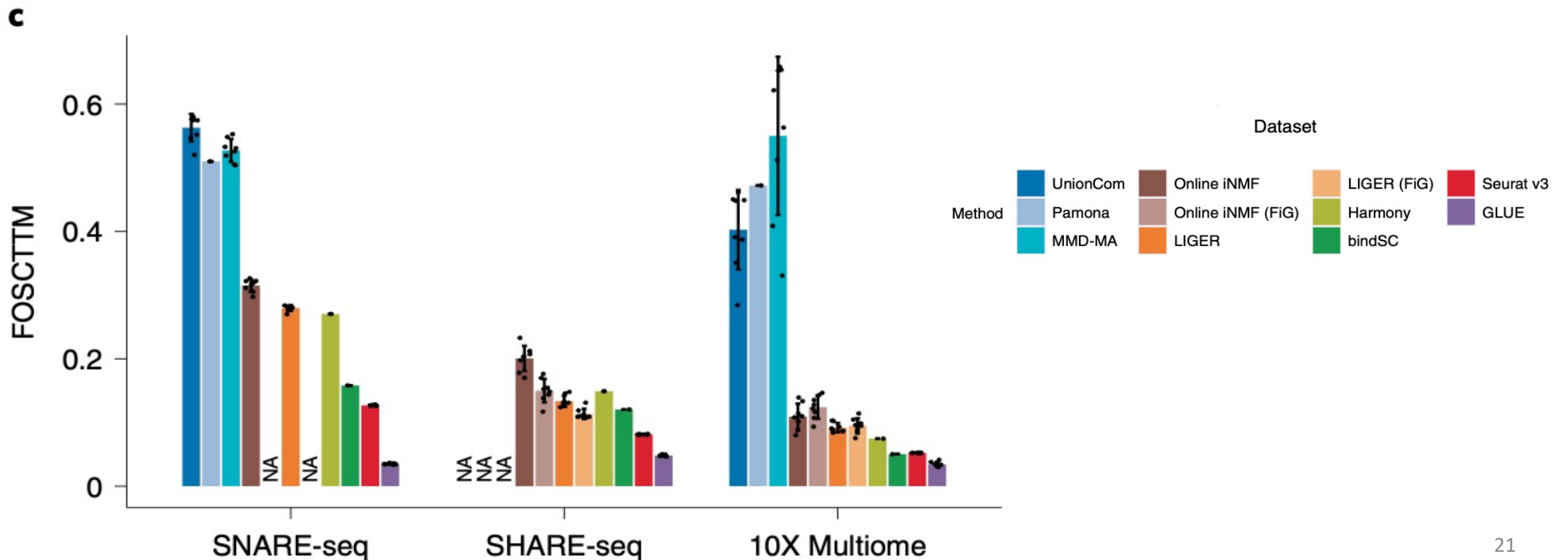
- GLUE was compared with multiple popular unpaired multi-omics integration methods
- GLUE achieved high level of biology conservation and omics mixing simultaneously





# Result 1: Integrating Performance of GLUE

- GLUE was compared with multiple popular unpaired multi-omics integration methods  
GLUE achieved the lowest FOSCTTM, decreasing the alignment error by large margins compared to the second-best method on each dataset (the decreases were 3.6-fold for SNARE-seq, 1.7-fold for SHARE-seq and 1.5-fold for 10X Multiome).

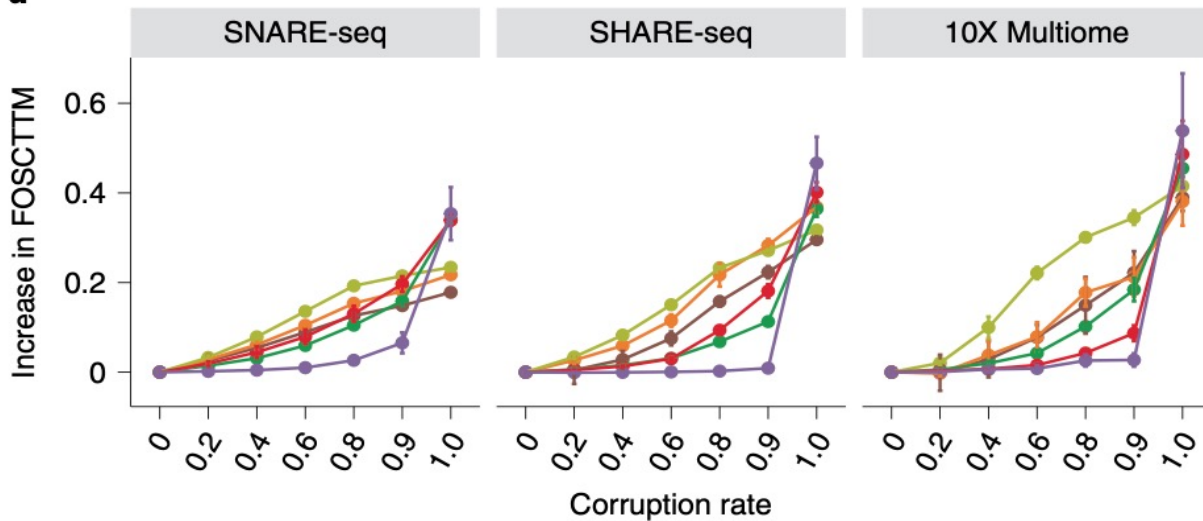




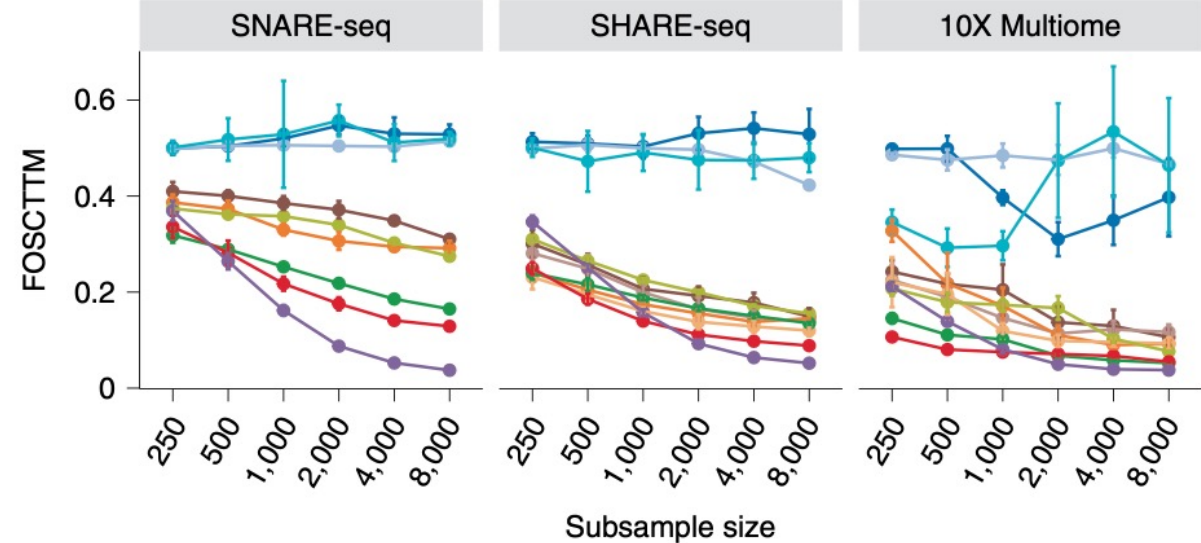
# Result 1: Integrating Performance of GLUE

- GLUE was compared with multiple popular unpaired multi-omics integration methods
- GLUE exhibited the smallest performance changes even at corruption rates as high as 90% suggesting its superior robustness. GLUE even remained the top-ranking method with as few as 2,000 cells, but the alignment error increased more steeply when the data volume decreased to less than 1,000 cells.

d



e



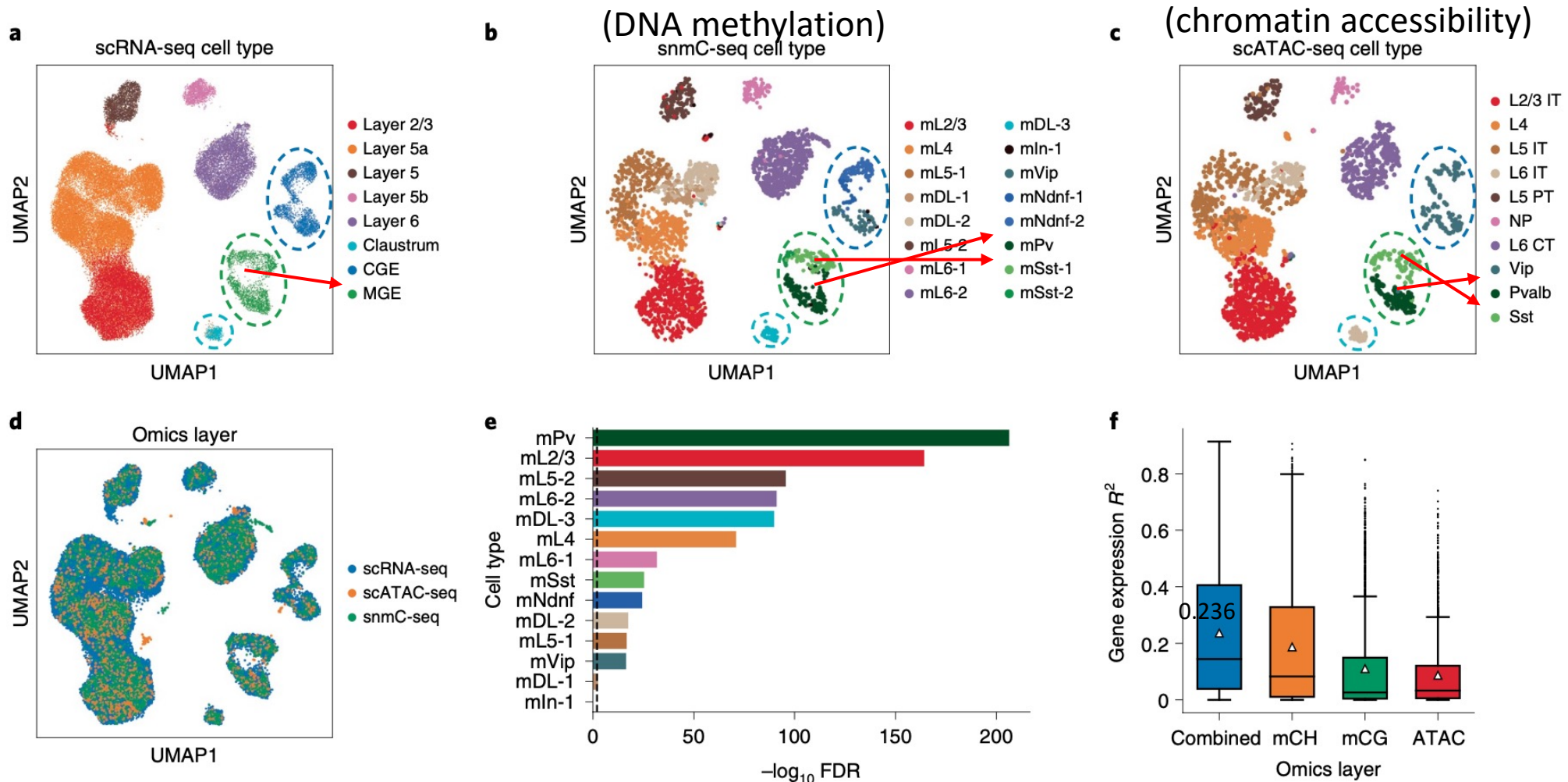
Dataset





# Result 2: Triple-omics integration

- As a case study, GLUE integrated three distinct omics layers (gene expression, chromatin accessibility, and DNA methylation) in neuronal cells of the adult mouse cortex.
- The GLUE alignment successfully revealed a shared manifold of cell states across the three omics layers





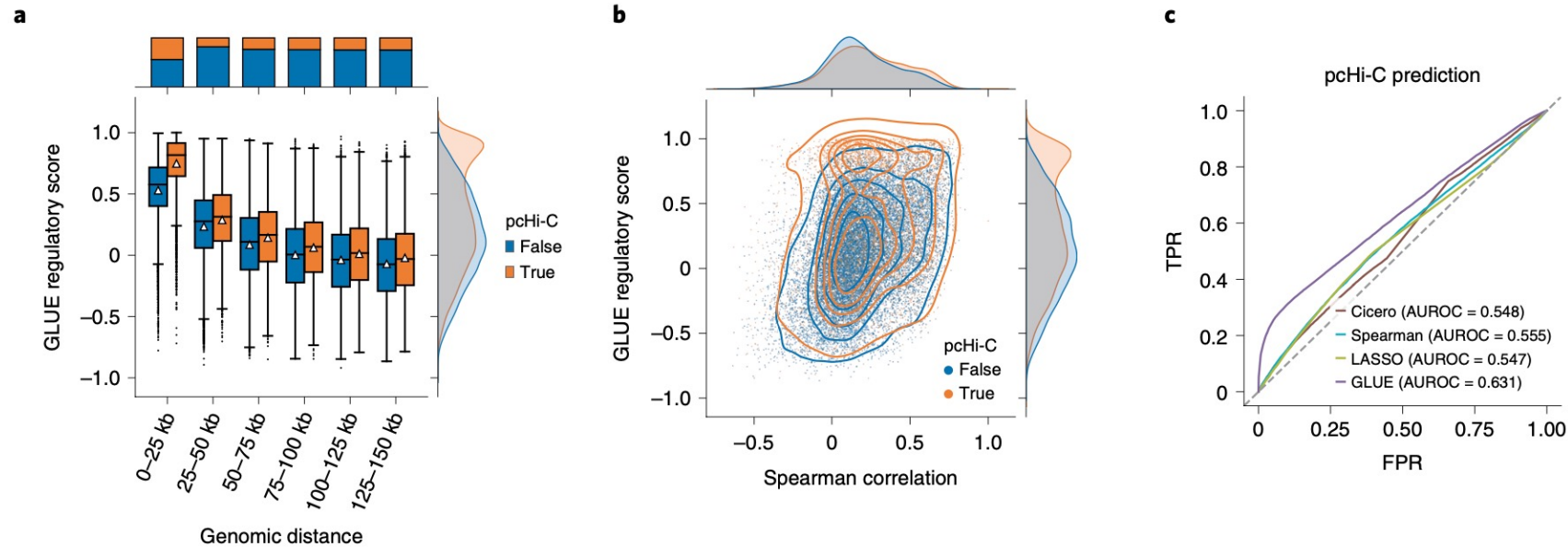
# Result 3: Regulatory inference capabilities

- The performance of the GLUE regulatory score in comparison with pHi-C data and other methods

**a(boxplot):** Close peak-gene pairs tend to have higher regulatory scores

**b(density plot):** GLUE's scores generally align well with the empirical data, particularly for those peak-gene pairs that pHi-C data supports.

**c(ROC curve):** GLUE's prediction is highly accurate

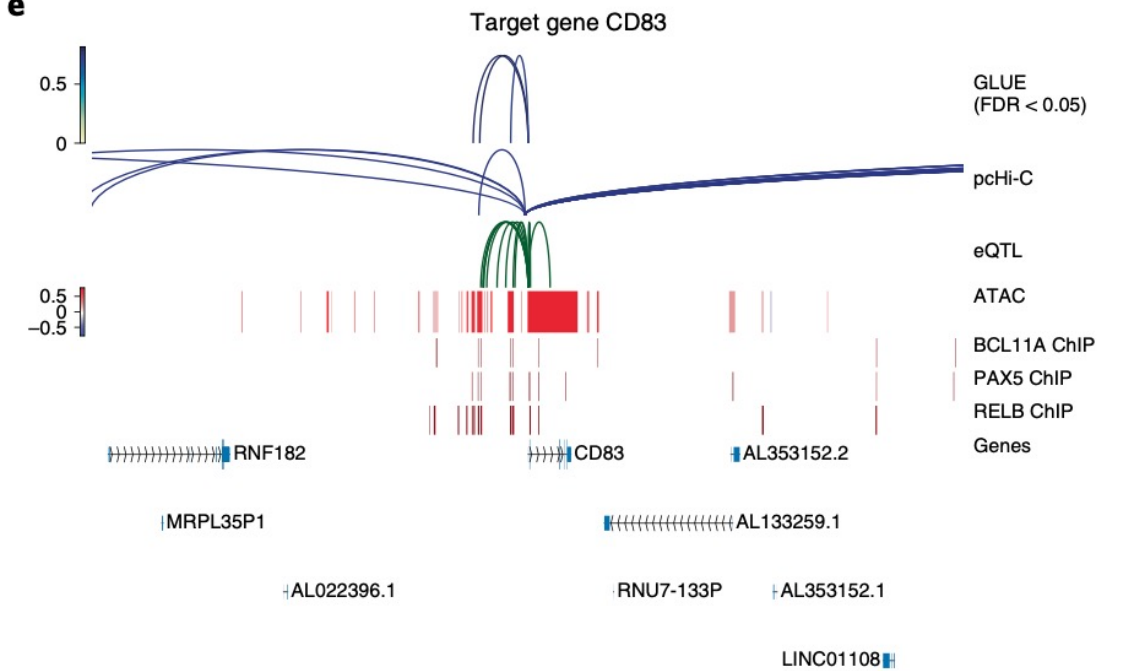
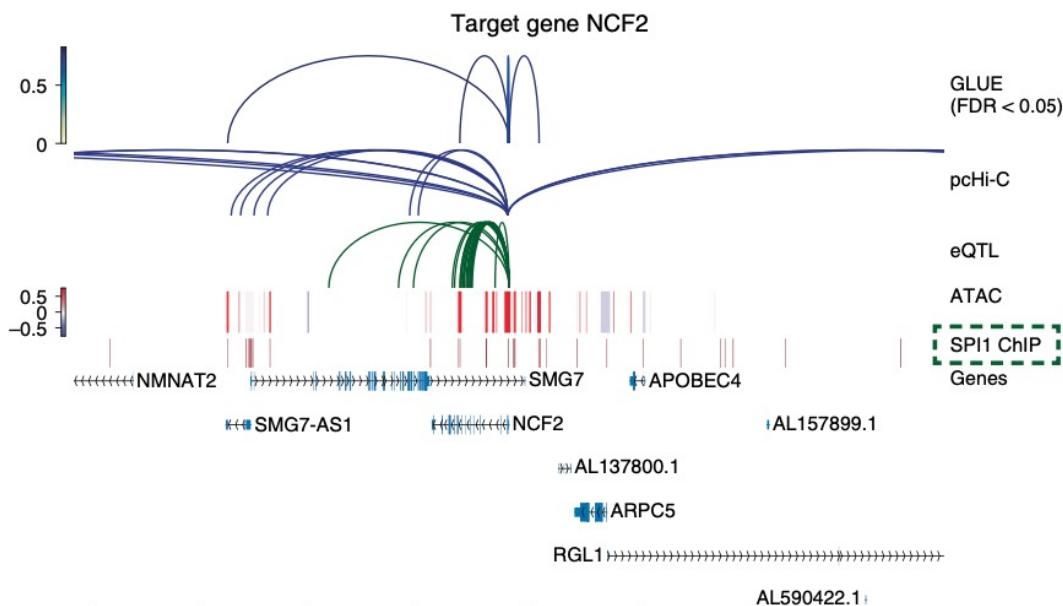






# Result 3: Regulatory inference capabilities

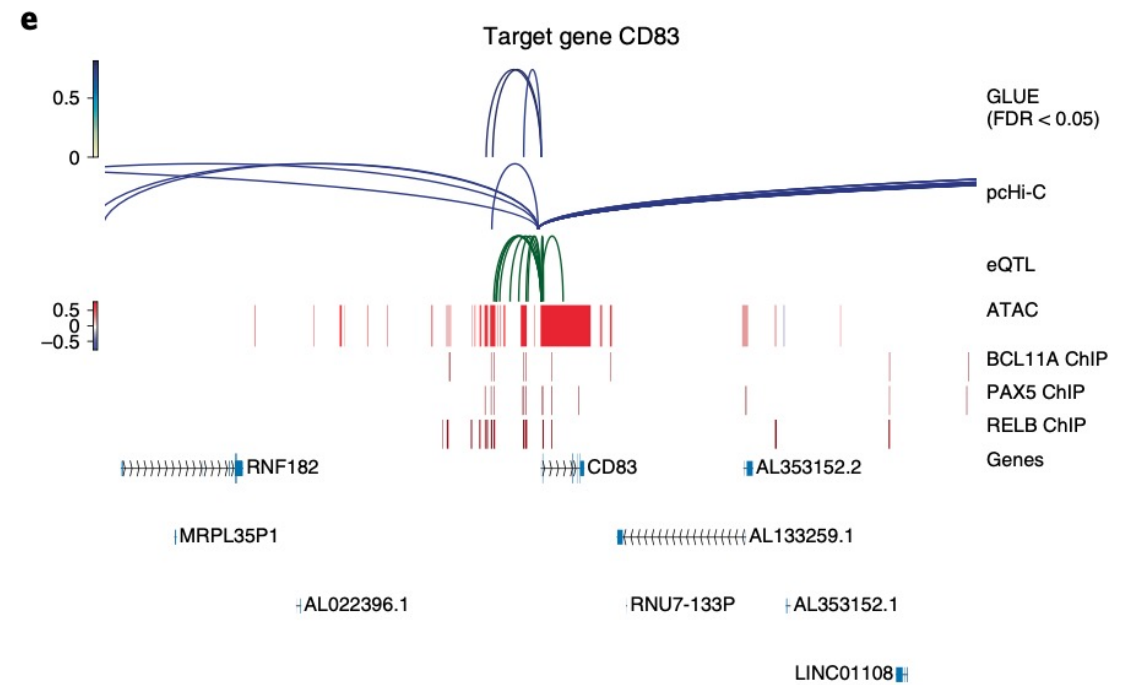
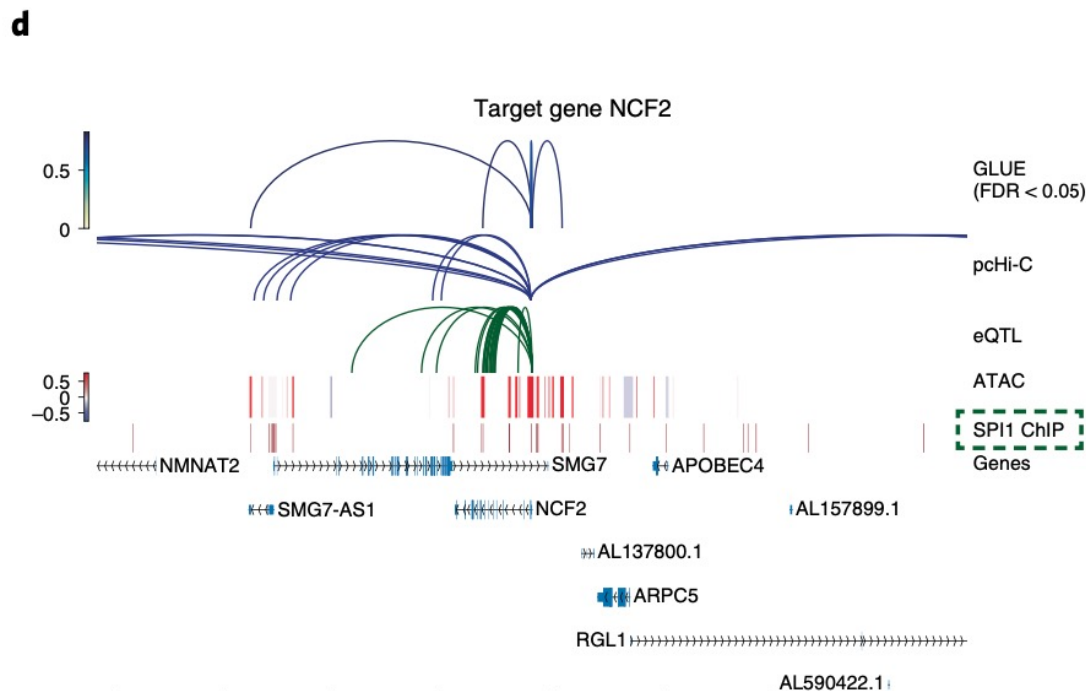
- Validation of Predictions; Visually represent how well GLUE's predictions align with other experimental
  - pHi-C (Physical Chromatin Interactions): The blue lines indicate interactions identified by pHi-C, supporting GLUE's predictions
  - eQTL (Genetic Variants Associated with Gene Expression): The green lines represent eQTL interactions, providing a genetic backing to the gene expression patterns seen. The arcs represent the potential regulatory interactions.
  - ATAC( open Chromatic Regions): The red lines show regions of open chromatic, suggesting areas that are active in controlling gene expression.
  - SPI1 ChIP (Protein-DNA Binding Data): The black line represents SPI1 protein binding, which is involve in gene regulation





# Result 3: Regulatory inference capabilities

- Validation of Predictions; Visually represent how well GLUE's predictions align with other experimental
  - GLUE's predictions align well with the data from pHi-C and eQTL methods, indicating it was not only accurate in identifying known interactions but also effective in predicting new ones that other methods validated.





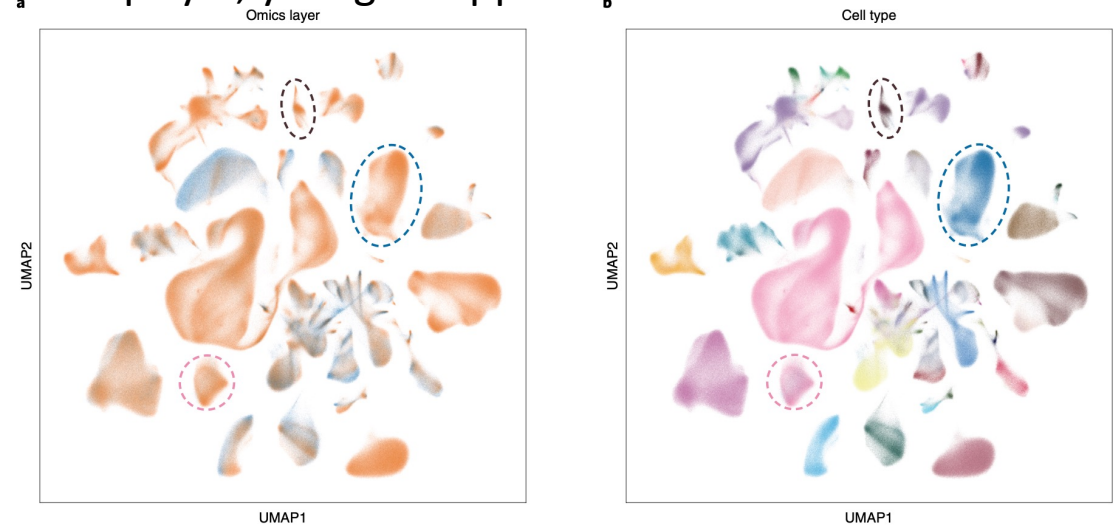
## Result 4: Atlas-scale Integration

- UMAP visualizations representing the results of cell data integration using the GLUE framework.

Figures demonstrate how GLUE can be used to reconcile differences in cell type identification across different omics layers, indicating the framework's ability to integrate complex datasets.

- **Pink Circles:** These highlight cells that were identified as 'excitatory neurons' when looking at RNA-seq data but were labeled as 'Astrocytes' in the ATAC-seq data.
- **Blue Circles:** labeled as 'Astrocytes' in the RNA-seq data but as a mixed group of 'Astrocytes/oligodendrocytes' when analyzed with ATAC-seq.
- **Brown Circles:** identified as 'Oligodendrocytes' in the RNA-seq layer, yet again appear as 'Astrocytes/oligodendrocytes' in the ATAC-seq layer.

The highlighted circles draw attention to discrepancies or ambiguities in cell type identification between RNA and ATAC sequencing data. These discrepancies are key findings that GLUE aims to address.




# Outline

 Background and objectives

 Material and methods

 Results

 Discussion and conclusion

 Q/A



# Discussion and Conclusion

- Innovative Framework Design
  - GLUE uniquely integrates omics-specific autoencoders with graph-based coupling and adversarial alignment.
  - This innovative design ensures superior accuracy and robustness for unpaired single-cell multi-omics data integration
- Regulatory Interactions Modeling
  - A standout feature of GLUE is its explicit modeling of regulatory interactions across different omics layers
  - This enables GLUE to support integrative regulatory inference, setting it apart from other methods
- Transformative Impact in Genomics
  - GLUE marks a significant advancement in genomics, revolutionizing the integration of multi-omics data at the single-cell level.
  - Its effectiveness across various applications underscores its potential to transform multi-omics data analysis
- Future Research Prospects
  - GLUE creates opportunity toward effectively understanding gene regulatory maps via large-scale multi-omics integration at single-cell resolution.

The whole package of GLUE, along with tutorials and demo cases, is available online at <https://doi.org/10.1038/s41587-022-01284-4>.

# Outline

 Background and objectives

 Material and methods

 Results

 Conclusion

 Q/A



Q/A



Q/A