Single-cell biological network inference using a heterogeneous graph transformer

Anjun Ma[1,2], Xiaoying Wang[3], Cankun Wang[1], Yuzhou Chang[1,2], Gang Xin[2], Zihai Li[2], Bingqiang Liu[3,*], Dong Xu[4,*] & Qin Ma[1,2,*]

1 Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA; **2** Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA; **3** School of Mathematics, Shandong University, Jinan, Shandong, China; **4** Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA

Background and significance

Background:

Single-cell multi-omics (**scMulti-omics**) allows the quantification of multiple modalities simultaneously to fully capture the intricacy of complex molecular mechanisms and cellular heterogeneity. Such analyses advance various biological studies when paired with robust computational analysis methods.

Challenges:

- Traditional statistical models and bioinformatic methods cannot reflect the influence of neighbor cell influences.
- How to find joint embeddings of cells and their features to build real connections without bias?
- However, most existing methods do not explicitly consider the topological information sharing among cells and modalities.

Highlights:

- We present **DeepMAPS** (Deep learning-based Multi-omics Analysis Platform for Single-cell data) for biological network inference from scMulti-omics.
- It models scMulti-omics in a **heterogeneous graph** and learns joint embedding of cells and genes considering both local and global contexts using a hypothesis-free **heterogeneous graph transformer** (HGT).
- It can predict and construct robust **cell-type-specific gene regulatory networks** (GRNs) from scMulti-omics.

The DeepMAPS workflow

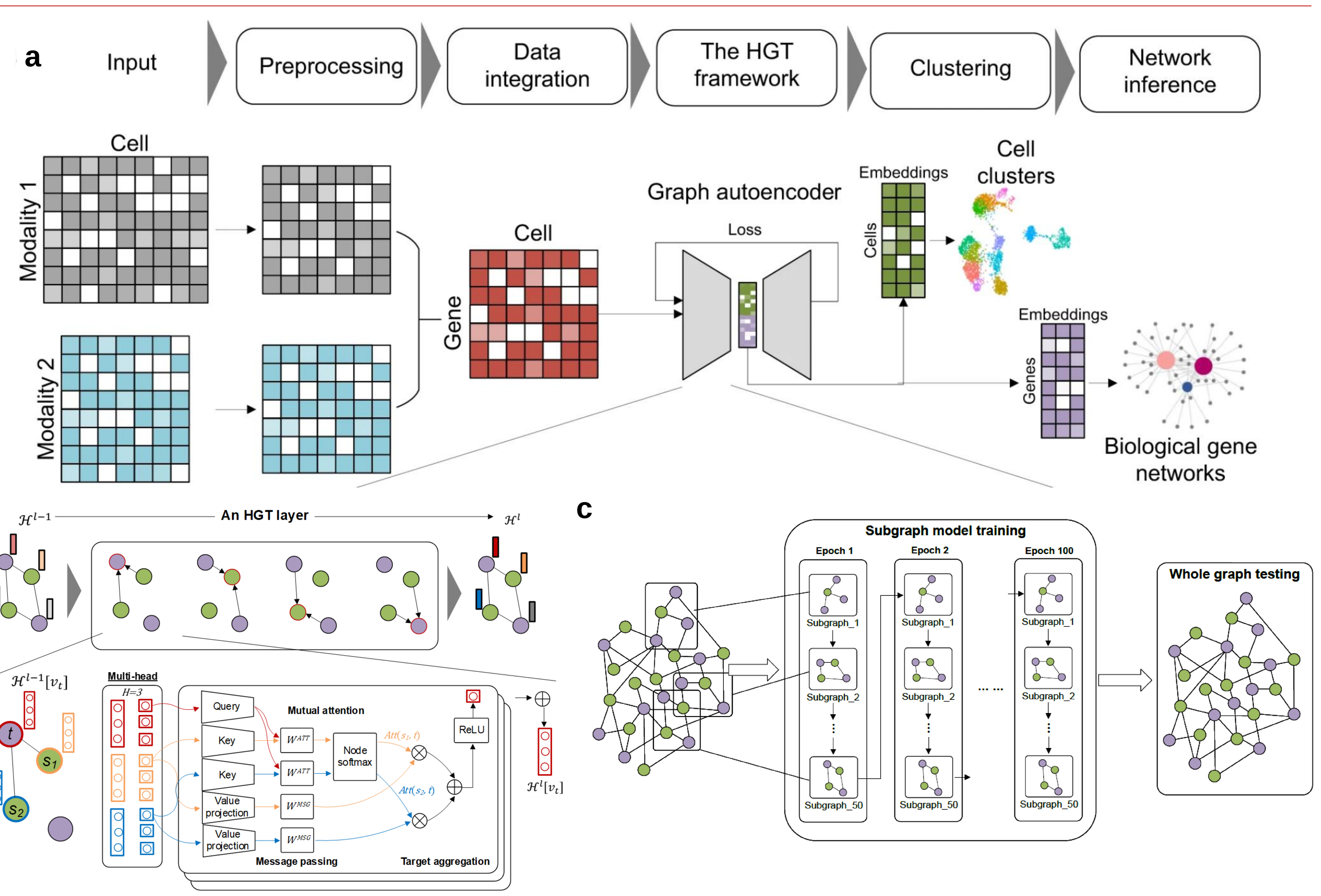
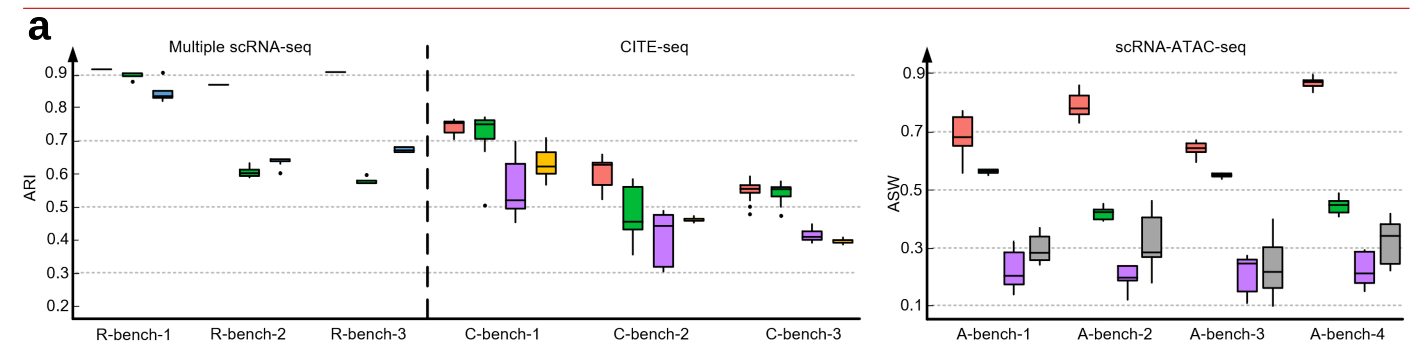


Fig. 1. (a) The overall framework of DeepMAPS. **(b)** An illustration of embedding update process of the target node in a single HGT layer. The attention mechanism in this HGT model enables the estimation of the importance of genes to specific cells, which can be used to discriminate gene contributions and enhances biological interpretability. **(c)** Subgraph strategy for HGT model training in DeepMAPS.

Benchmarking results



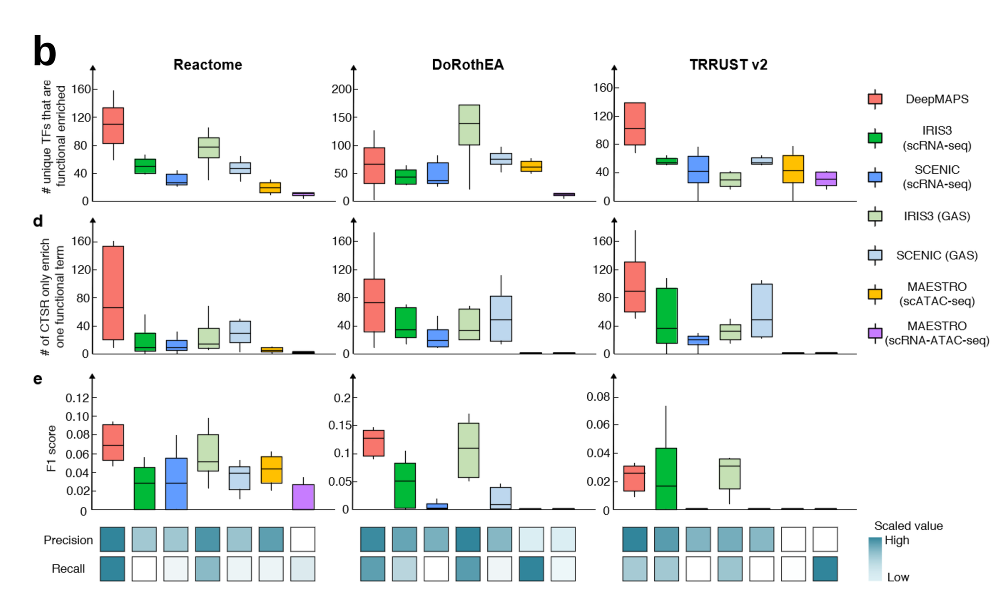
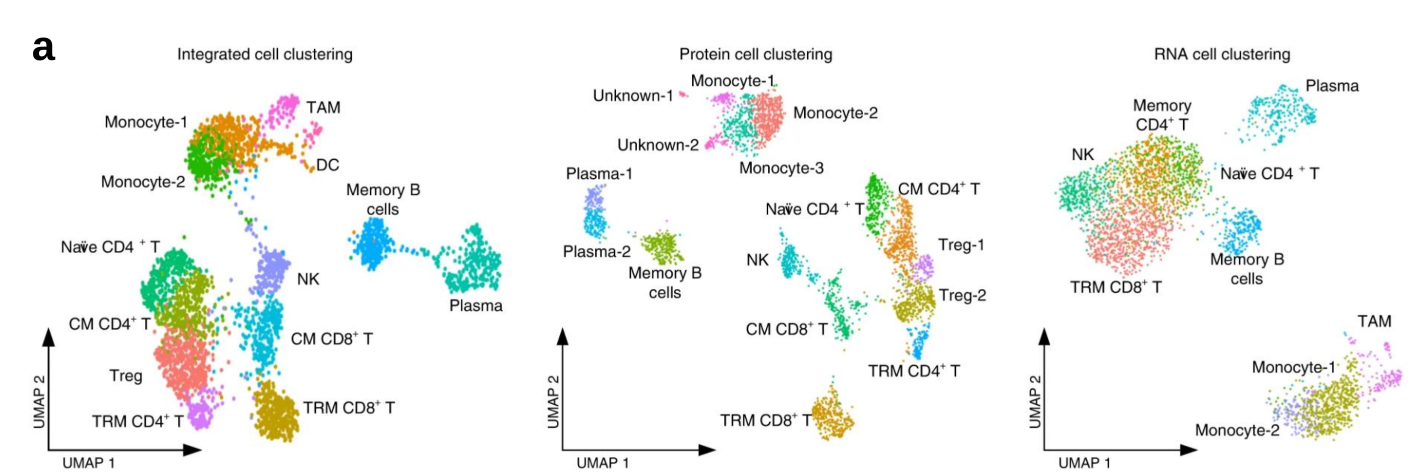


Fig. 2. (a) a Benchmark cell clustering results of ten datasets in ARI for the three multiple scRNA-seq data and the three CITE-seq data with benchmark labels, and ASW for the four scRNA-ATAC-seq data without labels. **(b)** Benchmark TF-regulons in specific cell types based on Reactome, DoRothEA, and TRRUST v2.

- DeepMAPS achieves superior performances in cell clustering and biological network inference from scMulti-omics data.
- DeepMAPS can infer statistically significant and biologically meaningful gene association networks from scMulti-omics data.

Case study 1: CITE-seq data



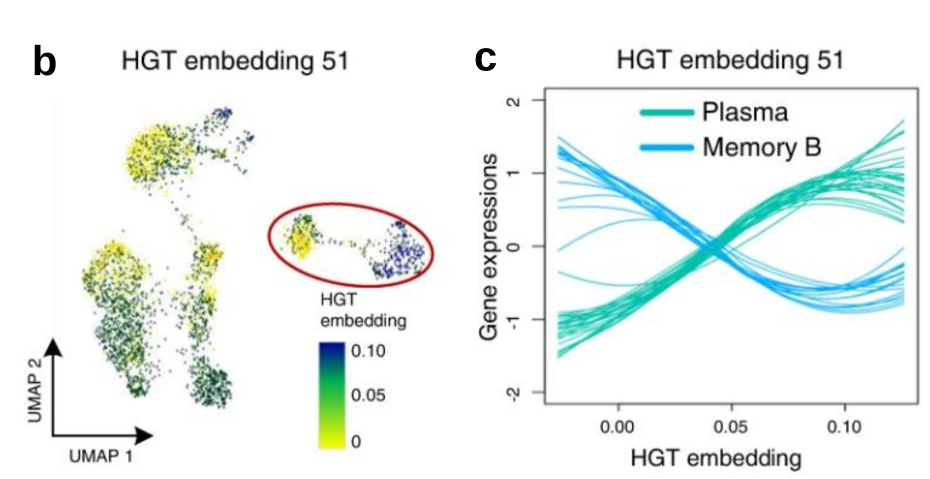
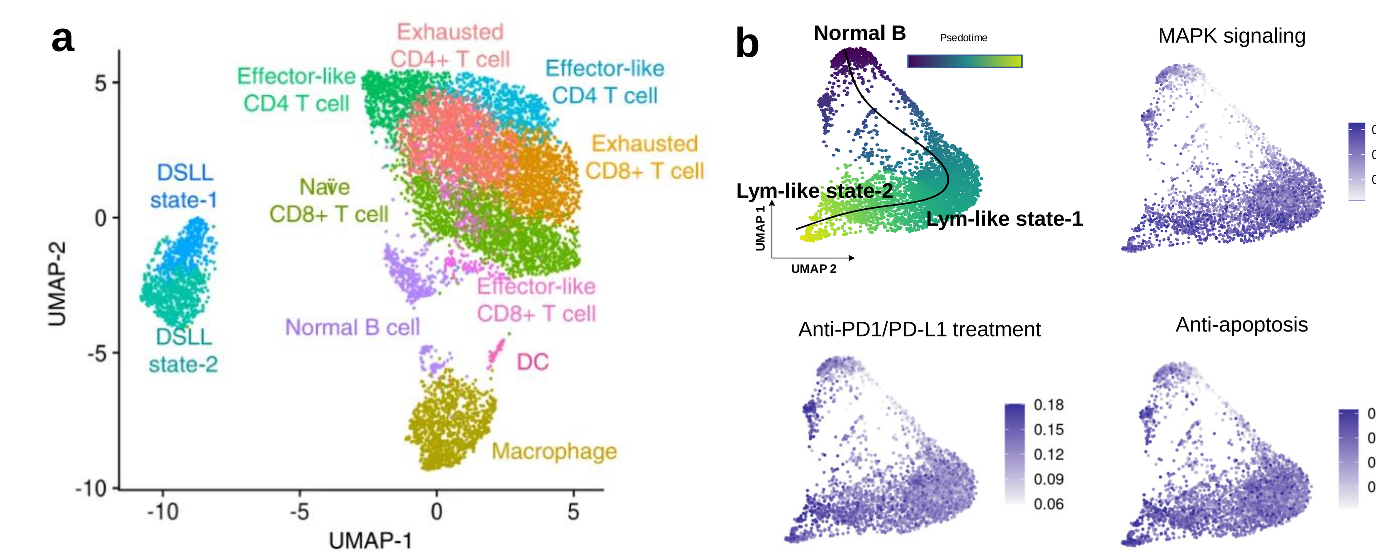


Fig. 3. (a) UMAPs for DeepMAPS cell clustering results from integrated RNA and protein data, protein data only, and RNA data only. **(b)** UMAP is colored by the 51st embedding, indicating distinct embedding representations in plasma cells and memory B cells. **(c)** Expression of top differentially expressed genes and proteins as a function of the 51st embedding to observe the pattern relations between plasma cells and memory B cells.

- DeepMAPS accurately identifies cell types and infers cell-cell communication in PBMC and lung tumor immune CITE-seq data.
- Each HGT embedding is interpretable that maintains gene/protein expression signals for the separation of specific cell groups.

Case-study 2: scRNA-seq and scATAC-seq



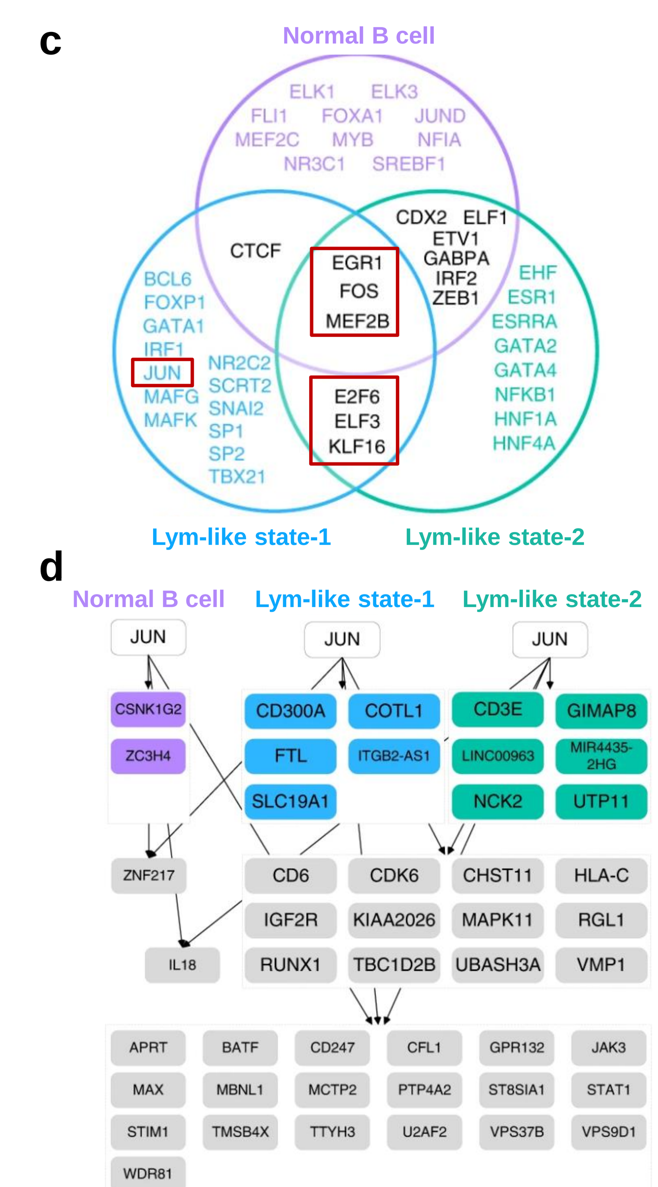


Fig. 4. (a) The UMAP shows the clustering results of DeepMAPS. Cell clusters were manually annotated based on curated marker genes. **(b)** Trajectory analysis and functional enrichment of B cells. **(c)** Selected 20 TF in each of the three clusters, representing the top 20 regulons with the highest centrality scores. **(d)** The downstream-regulated genes of JUN (the most differentially active regulon in DSLL state-1) in the three clusters. Gene expressions, peak accessibilities, TF-motif binding affinity, peak-to-gene distance, and TF expressions are considered in defining the activity of TF-gene regulations.

- DeepMAPS identifies cell-state-specific GRNs in diffuse small lymphocytic lymphoma scRNA-seq and scATAC-seq.
- DeepMAPS can construct GRNs and identify cell-type-specific regulatory patterns to offer a better understanding of cell states and developmental orders in subpopulations.

Conclusion and discussion

- DeepMAPS predicts both cell clusters and corresponding gene (regulatory) networks from single-cell multi-omics data, better than existing tools.
- The key framework, multi-head heterogeneous graph transformer, can exchange "message" among cells and features, making it anti-noise and increase discrepancies and similarity of cells
- The attention score calculated in the graph transformer explain the importance of a feature to a cell.

Reference

Anjun Ma, Xiaoying Wang, Jingxian Li, Cankun Wang, Tong Xiao, Yuntao Liu, Hao Cheng, Juexin Wang, Yang Li, Yuzhou Chang, Jinpu Li, Duolin Wang, Yuexu Jiang, Li Su, Gang Xin, Shaopeng Gu, Zihai Li, Bingqiang Liu, Dong Xu & Qin Ma. **Single-cell biological network inference using a heterogeneous graph transformer.** *Nat Commun* **14**, 964 (2023). https://doi.org/10.1038/s41467-023-36559-0

Acknowledgement

This work was supported by awards R01-GM131399 (Q.M.), R35-GM126985 (D.X.), and U54-AG075931 (Q.M.) from the National Institutes of Health.
The work was also supported by the award NSF1945971 (Q.M.) from the National Science Foundation.
This work was supported by the Pelotonia Institute of Immuno-Oncology (PIIO). The content is solely the responsibility of the authors and does not necessarily represent the official views of the PIIO.