nature methods

Supplementary information

https://doi.org/10.1038/s41592-025-02627-0

scNET: learning context-specific gene and cell embeddings by integrating single-cell gene expression data with protein–protein interactions

In the format provided by the authors and unedited

Supplementary - scNET: Learning context-specific gene and cell embeddings by integrating single-cell gene expression data with protein-protein interaction information

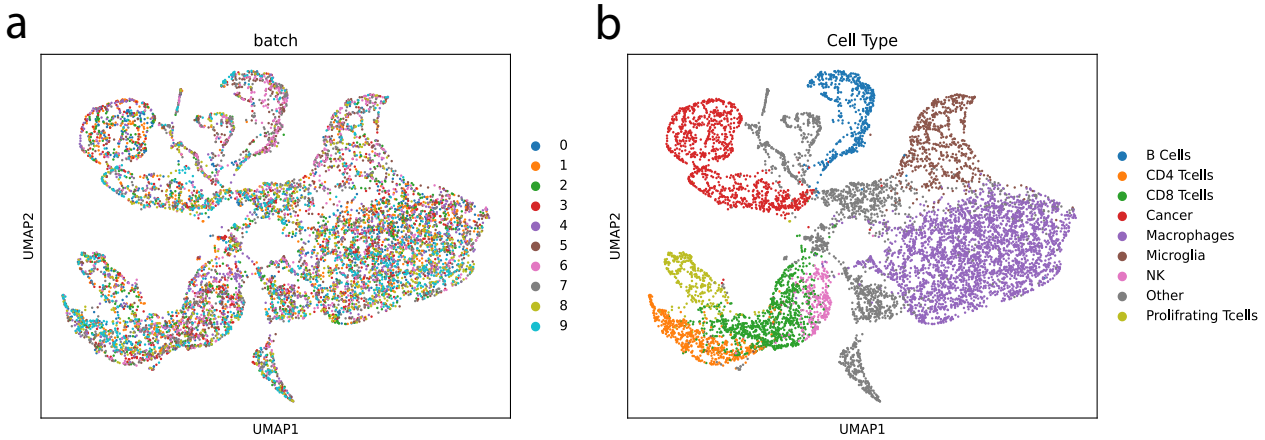
Ron Sheinin[1], Roded Sharan[1*†], Asaf Madi[2*†]

[1]Blavatnik School of Computer Science, Tel Aviv University, Israel.
[2]Department of Pathology, Faculty of Medicine, Tel Aviv University, Israel.

*Corresponding author(s). E-mail(s): roded@tauex.tau.ac.il ; asafmadi@tauex.tau.ac.il ;
Contributing authors: ronsheinin@mail.tau.ac.il;
†These authors contributed equally to this work.

1 scNET do not present significant mini batch effect

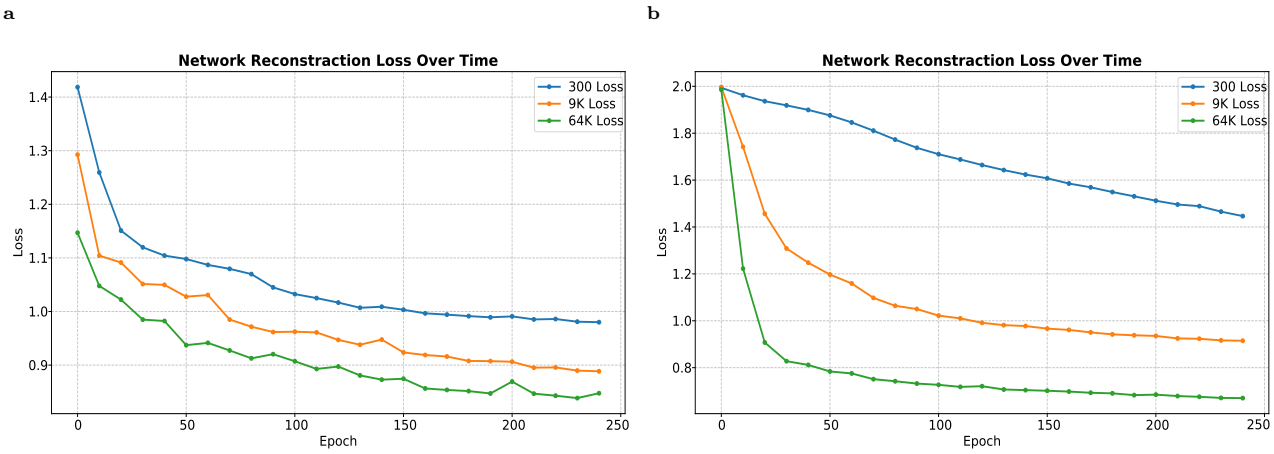
While each scNET mini-batch does not contain all the cells, it is large enough to provide a good representation of the overall data. Our experiments have shown that scNET successfully embeds cells into a joint latent space without significant batch effects. we tested the GBM dataset, which contains approximately 9,000 cells, by dividing it into mini-batches of 900 cells each. We then visualized the cells colored by batch number and by cell type.



Supplementary Figure 1. Batch Effect Analysis. UMAP visualization of the scNET cell embedding of the GBM dataset [1], with cells colored by training mini-batch (a) and by cell type (b). The dataset consists of approximately 9,000 cells, which were split into 10 training mini-batches of 900 cells each. The results demonstrate that scNET effectively embeds the cells into a joint learned space, minimizing batch effects.

2 scNET Running Time and Loss Convergence

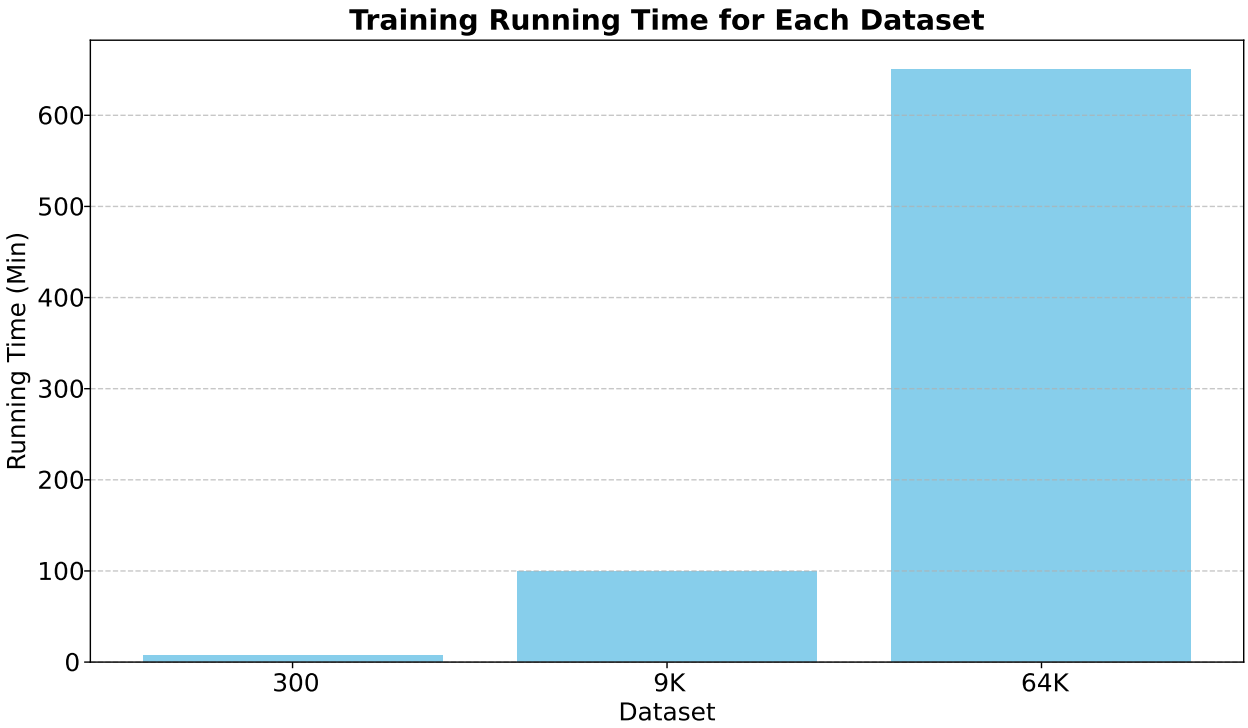
To examine the training process of scNET, we evaluated the loss convergence and running time of scNET on three different scRNA-seq datasets representing three different scales: Cell Cycle (300 cells)[2], Mouse GBM (9,000 cells)[1], and Mouse Brain Cortex (64,000 cells)[3]. It is important to emphasize that scNET is designed to assist in the analysis of specific scRNA-seq datasets rather than to handle atlas-scale data. Thus, 64,000 cells represent the upper limit for a dataset size that is appropriate for scNET.



Supplementary Figure 2. Evaluation of Network Reconstruction Loss and Gene Expression Loss for Different Dataset Sizes. This figure evaluates the network reconstruction loss and gene expression loss across different dataset sizes: Cell Cycle (300 cells), Mouse GBM (9,000 cells), and Mouse Brain Cortex (64,000 cells). The evaluation was performed over 250 epochs. (a) **Network Reconstruction Loss:** The analysis shows that larger datasets tend to converge to a lower loss. Additionally, the loss for all datasets nearly converges around epoch 150. (b) **Gene Expression Reconstruction Loss:** The results indicate that larger datasets not only achieve lower losses but also converge faster. For datasets with 9,000 or 64,000 cells, the reduction in loss becomes less significant after 100 epochs.

In Supplementary Figure 2, we show the convergence of the network and gene expression reconstruction loss for each dataset. We observed that for larger datasets, the convergence in both losses is to smaller values and occurs within fewer epochs.

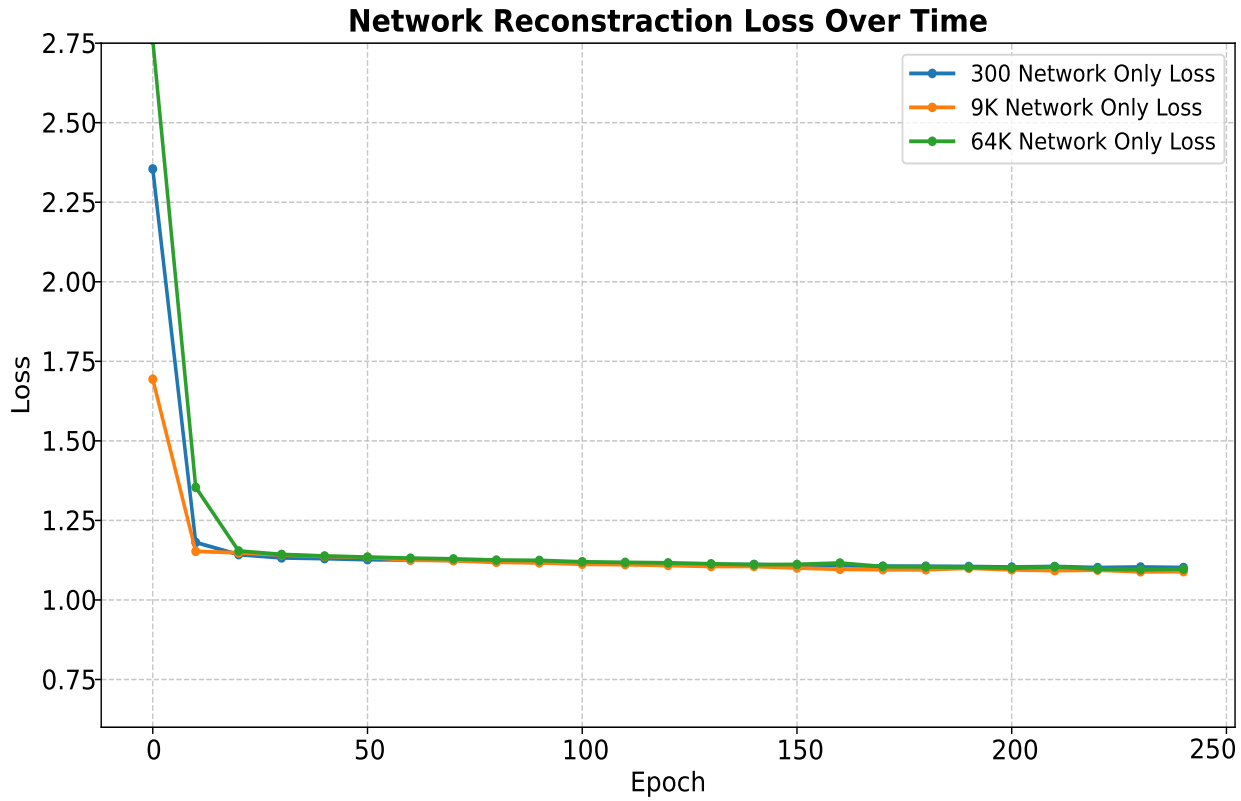
Next, we evaluated the running time for 250 epochs for each dataset, as shown in Supplementary Figure 3. Our analysis reveals that the running time of scNET scales approximately linearly with the size of the dataset. Specifically, for a dataset with 10,000 cells, the running time is around 100 minutes, while for the larger dataset with 64,000 cells, it requires approximately 650 minutes. These results indicate that scNET remains feasible to use even with larger datasets, particularly as larger datasets tend to converge with fewer epochs.



Supplementary Figure 3. Evaluation of scNET Running Time. Evaluation of running time of scNET across three datasets: Cell Cycle (300 cells), Mouse GBM (9,000 cells), and Mouse Brain Cortex (64,000 cells). Each dataset was run for 250 epochs. The analysis reveals that the running time of scNET scales approximately linearly with the size of the dataset. Specifically, for a dataset with 10,000 cells, the running time is around 100 minutes, while for the larger dataset with 64,000 cells, it requires approximately 650 minutes. These results indicate that scNET remains feasible for use with larger datasets, especially as larger datasets tend to converge with fewer epochs.

Finally, we evaluated the advantage of our joint learning task (both network and gene expression) relative to separate learning tasks. To this end, we built a GCN for reconstructing the PPI network using the gene expression matrix. We applied this model to three of our datasets, as shown in Supplementary Figure 5.

The results show that for all three datasets, the loss values converged around 1.1, while in scNET, the network convergence was between 1.0 and 0.8. Notably, we observed that for all datasets, the losses of the separate GCN models converged to similar values, indicating a tendency to overfit the network topology without adequately capturing the dynamic nature of the gene expression data. This suggests that separating the GCN models results in a loss of performance, as the gene-specific dynamics driven by the expression data are not effectively integrated into the embeddings. In contrast, the dual-view GCN encoder in scNET benefits from jointly optimizing both gene and cell embeddings, allowing it to capture both the network topology and the underlying biological signals.



Supplementary Figure 4. GCN Autoencoder with Only Network Reconstruction Loss Shows Overfitting to Network Topology. To examine the advantage of our joint model, we built a vanilla GCN and GAN autoencoder model with the same number of layers as scNET, but using only the network reconstruction loss.

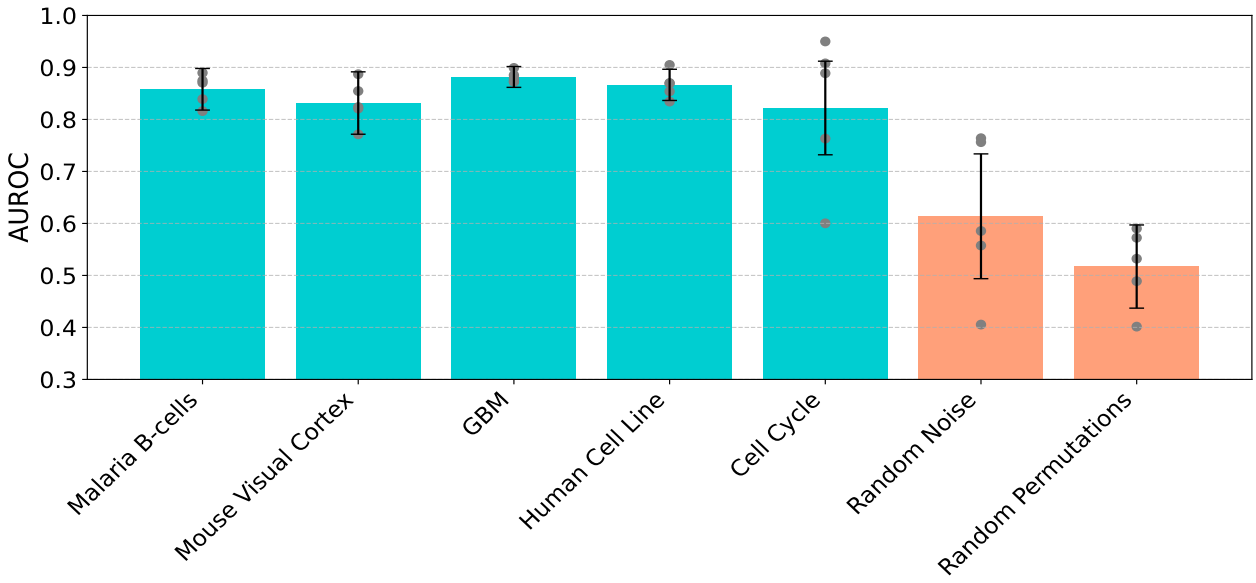
2.1 scNET Captures PPI Network Information Contingent Upon Gene Expression Data

Our initial aim was to validate that scNET effectively captures PPI network information within its gene embeddings. As a preliminary assessment, we evaluated scNET's ability to predict PPIs in a 5-fold cross-validation setting. For each fold, we calculated the AUROC by selecting a random set of negative edges (non-edges) equal in size to the test set. The results, summarized in Supplementary Figure5 (left), demonstrate that scNET effectively captures the network topology within its gene embeddings, achieving a robust AUROC score of approximately 0.85 with a relatively small standard deviation across folds (except for the Cell Cycle dataset, which exhibited a higher standard deviation of 0.09).

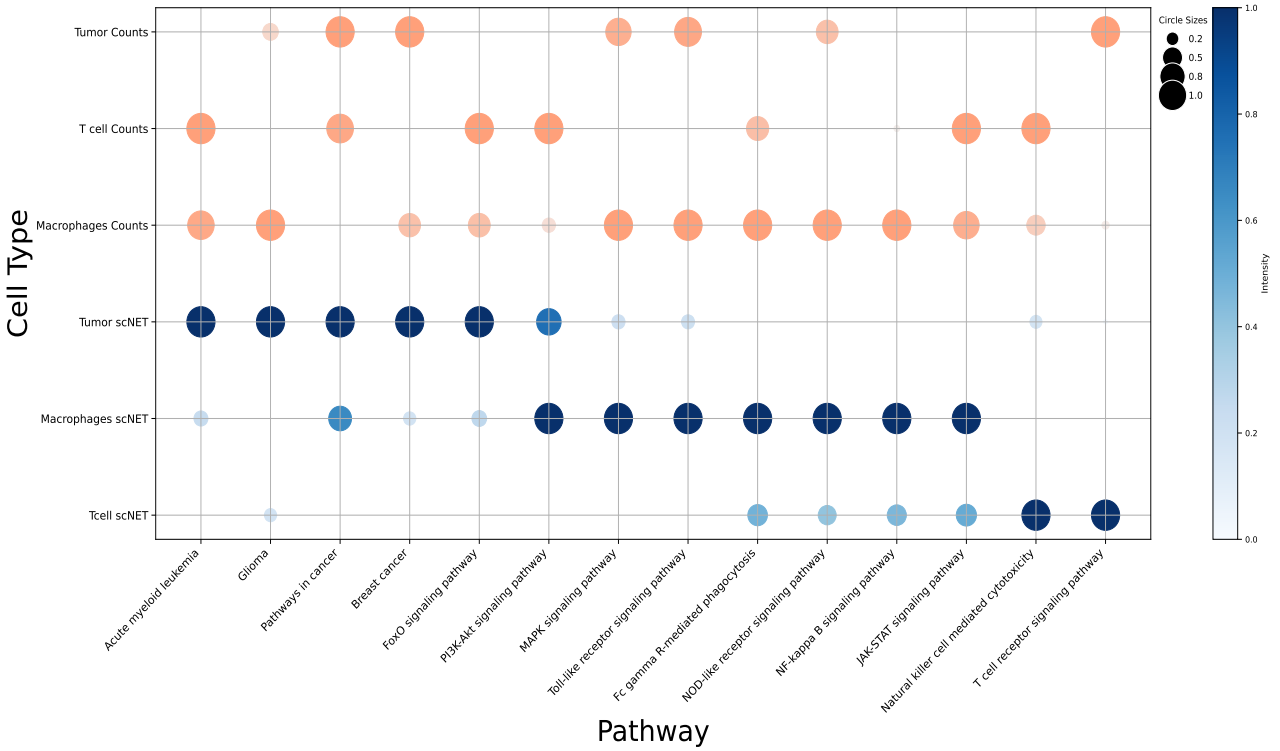
Importantly, we wanted to ensure that scNET's embeddings do not simply learn the graph topology in isolation but instead integrate information from both the PPI network and gene expression data. To test whether scNET's learning process depends on gene expression or merely overfits to the network topology, we conducted two additional cross-validation experiments using the mouse visual cortex dataset. In the first experiment, we added Gaussian random noise to the gene expression matrix, reducing the true biological signal in the data. In the second experiment, we randomly permuted the rows of the expression matrix, eliminating all true co-expression correlations.

As shown in Supplementary Figure 5 (right), both experiments resulted in a significant drop in AUROC scores on the test set. After adding random noise, the average AUROC decreased to approximately 0.61, and after row permutation, the AUROC dropped further to around 0.5. These results indicate that scNET does not simply learn the input network's topology; rather, it captures meaningful co-expression patterns present in the data.

It is important to note that scNET is not specifically designed or optimized to discover novel PPIs. Instead, this validation demonstrates that scNET effectively learns the network topology in conjunction with gene expression data, resulting in an integrated embedding space. In the following sections, we highlight the biological advantages and validity of this embedding space.



Supplementary Figure 5: scNET Captures PPI Network Information. AUROC scores of PPI prediction on five datasets. The last two columns (highlighted in orange) display performance metrics after introducing random noise or permuting the gene expression data.



Supplementary Figure 6. Evaluation of Gene Latent Space Dynamics. Evaluation of known immune and cancer pathways. For each pathway, the maximum enrichment ($-\log$(P-value)) across all gene clusters within each embedding space is presented. The enrichment scores have been scaled using MinMax scaler.

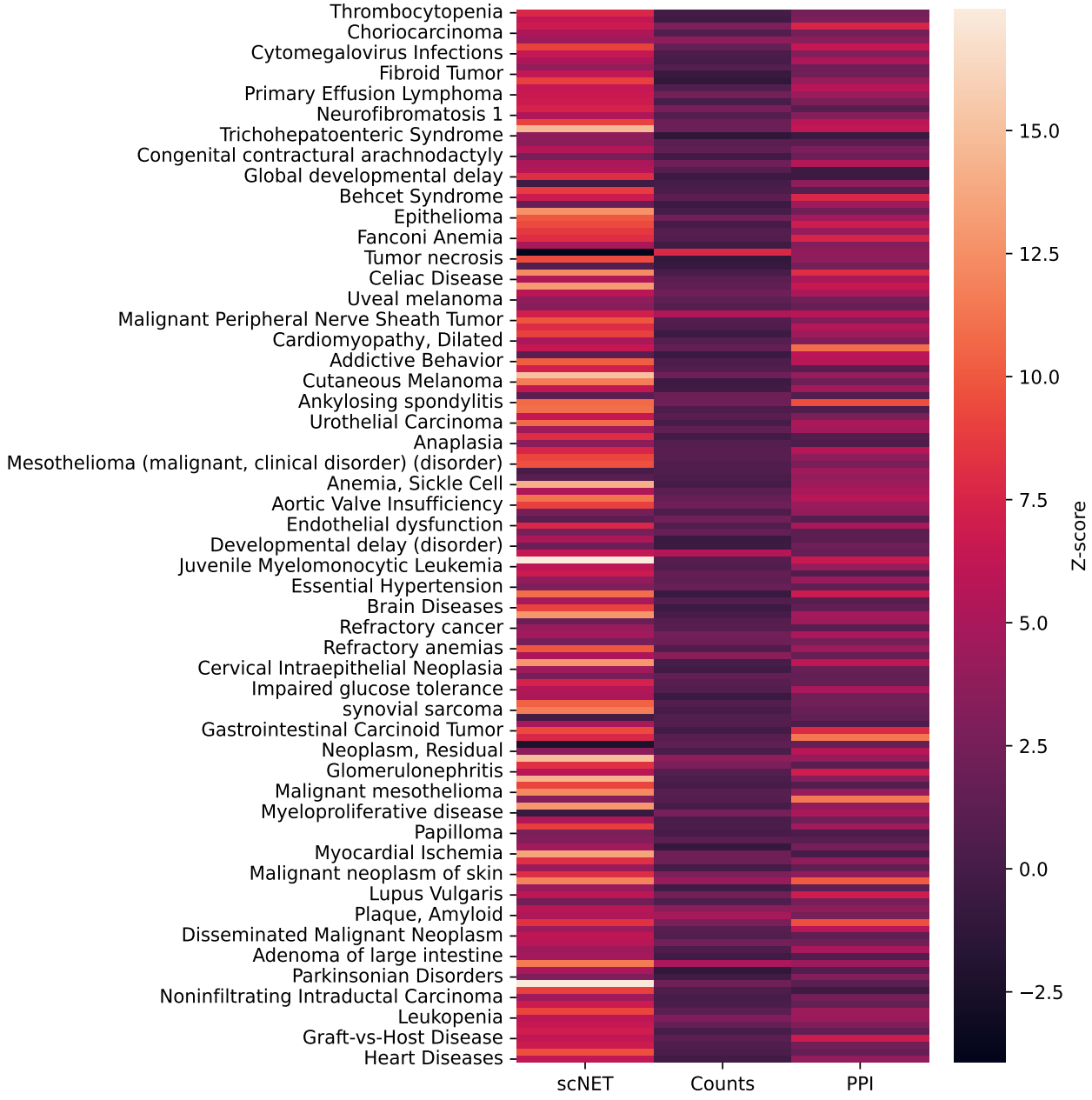
3 Gene Embedding Dynamics of Different Cell Populations

For each pathway, we calculated the maximum enrichment ($-\log$(P-value)) across all gene clusters within each embedding space. We observed distinct maximum enrichment scores for each pathway across different embeddings, reflecting expected variations between cell types (depicted in blue). Specifically, cancer-related pathways such as "Glioma", "Acute Myeloid Leukemia", "Breast Cancer", and "Pathways in Cancer" exhibited the highest enrichment in the tumor gene embedding but showed substantially lower enrichment in the immune cell embeddings.

Furthermore, cell-specific pathways demonstrated the highest scores in their corresponding cell type embeddings. For example, the "T Cell Receptor Signaling" pathway exhibited the highest score in the T cell embedding, while the "Toll-Like Receptor Signaling" pathway showed the highest score in the Macrophage embedding. More general immune pathways, such as "NF-kappa B" and "JAK-STAT Signaling", demonstrated high scores in both Macrophages and T cells but not in tumor cells.

In comparison, a similar analysis using the original gene counts (depicted in orange) revealed more noisy dynamics, with cancer pathways enriched in immune cells and the "T Cell Receptor" pathway enriched in tumor cells, but relatively less in T cell populations. These findings suggest that scNET embeddings not only preserve gene expression dynamics but also refine them, providing a more accurate representation of cell-specific characteristics.

4 Disease associated gene lists

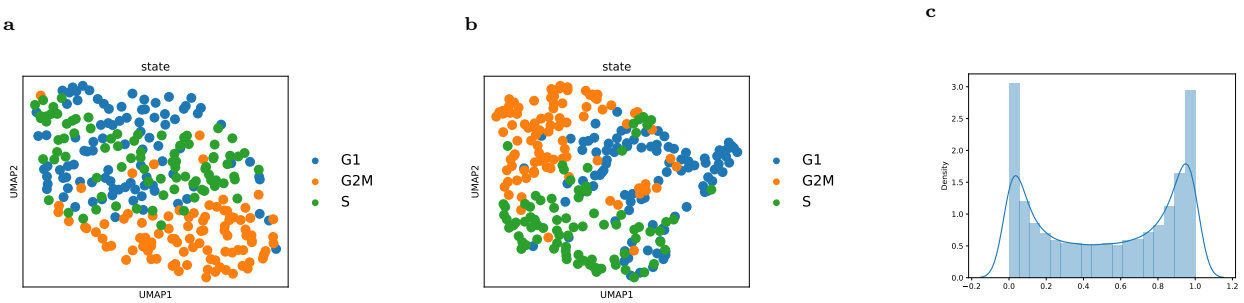


Supplementary Figure 7. Diseases Associated gene list analysis. Database of 230 gene lists associated with diseases was utilized. For each network (PPI, co-embedded, UMI co-express) a set of 30 degree-preserving random permutations was created to serve as a background distribution. Each gene list was split into a training and testing set, and network propagation was used to predict the test set. An AUC score was computed for each network, and the z-score was calculated for each observed network and disease gene list.

5 Cell Cycle Clustering

As an additinol validation of scNET cell embedding advantage, we examined cell state separation in UMAP representations. The original UMAP inadequately distinguished between S and G1 states, merging them into a single cluster. In contrast, UMAPs derived from our models (Supplementary Figure 8,a,b) demonstrated a clearer separation, with the S state cluster appropriately separated from the G1 and G2 clusters. This suggests a more accurate biological representation, even in the global structure. Leiden clustering in our embedding space yielded a maximum ARI of 0.46, surpassing the 0.35 ARI in the original normalized count space.

Next we investigated the distribution of attention coefficients during learning in the KNN graph after 30 epochs (pre-pruning) as shown in A notable two-peak distribution was observed, with one peak near zero (low-quality edges) and another at one (high-quality, informative edges), indicating our model's effective differentiation between edge qualities in the KNN topology.



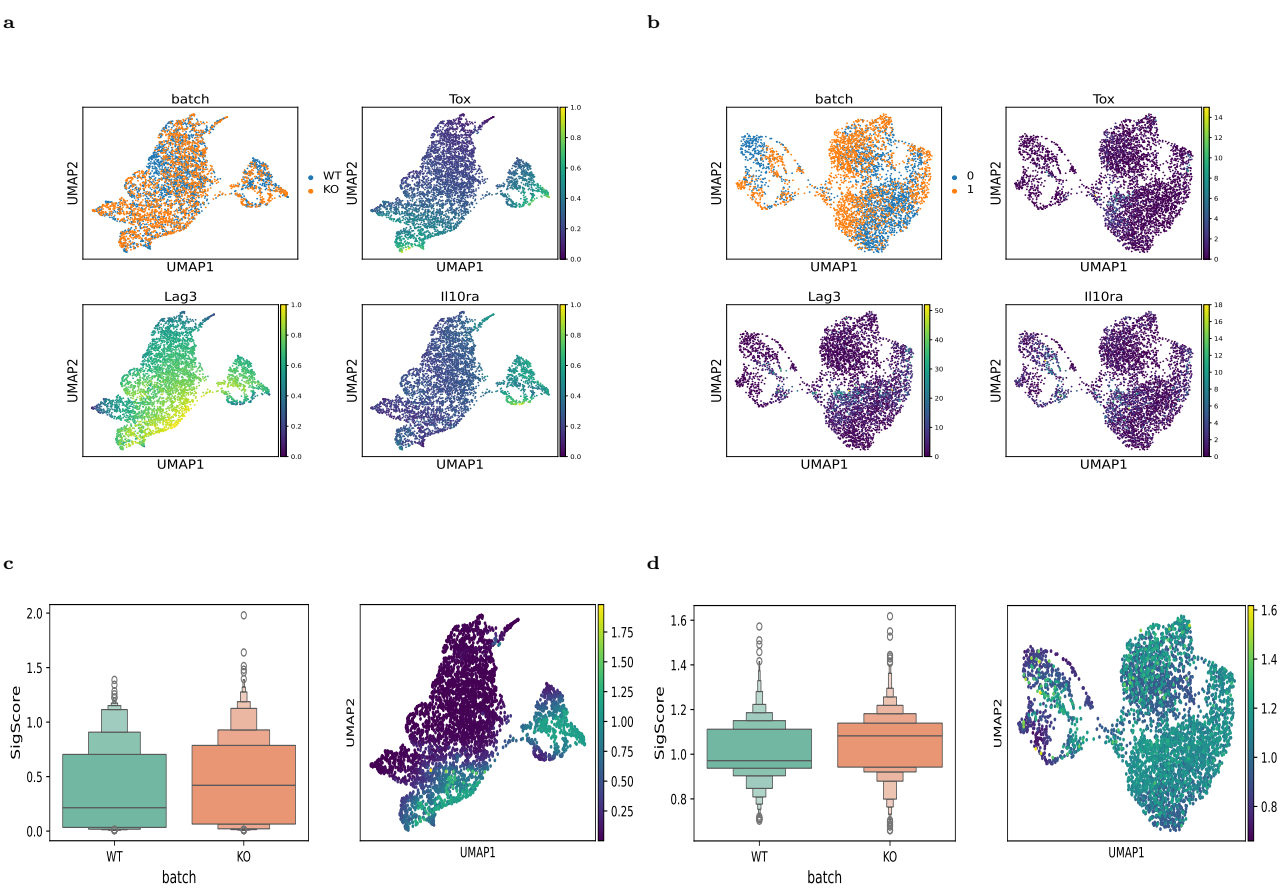
Supplementary Figure 8. Analysis of cell cycle dataset (a) UMAP Representation of the Normalized Count Space: This panel shows theUMAP representation of cells based on normalized count data. (b) UMAP Representation of the scNET Cell Embedding Space: This panel illustrates the UMAP representation derived from the scNET cell embedding, highlighting the spatial distribution of cells. (c) Distribution of Cell Attention Weights at Epoch 30, Prior to the First KNN Pruning: This panel depicts the distribution of cell attention weights in the model at epoch 30, just before the initial K- KNN pruning step.

6 scNET Enhances Integration and Identifies Subpopulations in Glioma Tumor-Reactive $CD8^+$ T Cells.

To further evaluate the advantages of scNET in standard gene expression analysis, we applied scNET to Glioma tumor-reactive $CD8^+$ T cells [4]. This dataset includes cells from a Glioma microenvironment that underwent MHCII knockout, as well as cells from a wild-type (WT) microenvironment. We used the reconstructed gene expression data to perform standard scRNA-seq analysis.

We observed that scNET better integrates different conditions. In Supplementary Figure 9a, the UMAP representation of the reconstructed data, colored by condition, shows significantly better integration compared to the UMAP of the original counts in Supplementary Figure 9.b, which exhibits a pronounced batch effect. Additionally, in the panels presented Supplementary Figures 9.a and 9.b, we plot the expression levels of Tox, Lag3, and Il10ra, which are differentially expressed between the conditions. In the original counts, we observed high levels of zero-inflation, where a large portion of the cells presented zero reads, while some cells exhibited very high expression levels. However, in the scNET-reconstructed gene expression, these genes are better captured, with strong expression in well-defined regions of the UMAP, indicating specific subpopulations with uniformed expression of these marker genes.

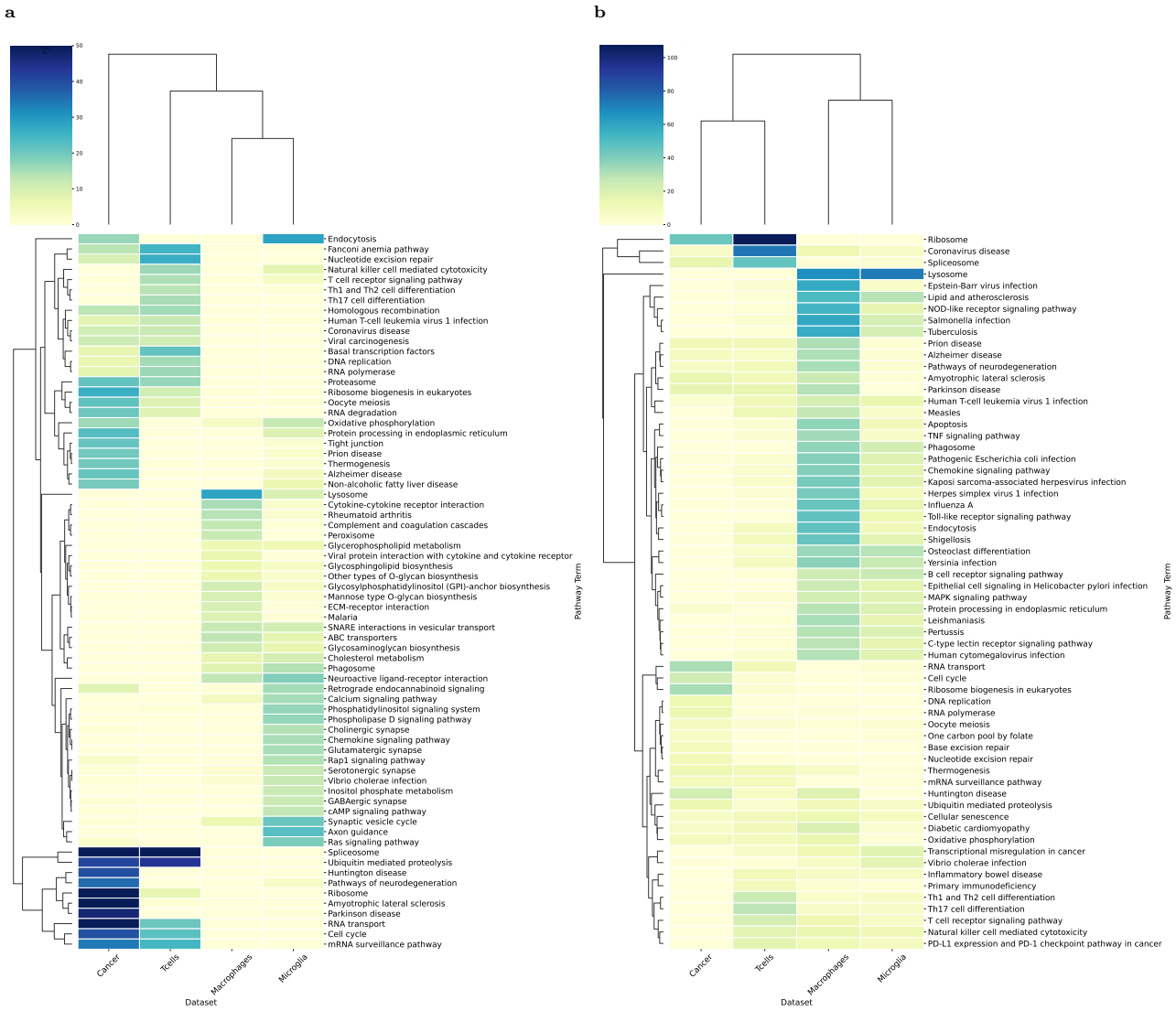
It was previously discovered that cells from the MHCII knockout tumor exhibited an elevated exhaustion phenotype. Therefore, we calculated an exhaustion signature based on known exhaustion genes. In Supplementary Figure 9.c, we show the signature score in scNET-reconstructed gene expression. We observed a significant



Supplementary Figure 9. Improved Integration and Expression Specificity in Glioma Tumor-Reactive $CD8^+$ T Cells Using scNET. (a) UMAP representation of scNET-reconstructed gene expression data, colored by condition (MHCII knockout vs. wild-type), showing better integration of different conditions compared to the original counts. Additionally, the expression of Lag3, Tox, and Il10ra is clearly captured in specific subpopulations. (b) UMAP representation of the original gene counts, where a pronounced batch effect is observed, along with high levels of zero-inflation in the expression of Lag3, Tox, and Il10ra, resulting in less defined subpopulations. (c) Exhaustion signature scores in the scNET-reconstructed gene expression data highlight a well-defined subpopulation with high signature values, corresponding to the MHCII knockout condition. (d) In contrast, the original counts display significant differences in exhaustion scores between knockout and wild-type populations but do not reveal a specific subpopulation associated with the signature. These findings suggest that scNET enhances the integration of conditions and better captures gene expression dynamics, identifying relevant subpopulations in the data.

difference in scores between the knockout and WT populations, as expected. Moreover, the UMAP projection reveals a well-defined subpopulation with high signature values. In contrast, in the original counts (Supplementary Figure 9.d), while we still observe a significant difference between the populations, there is no specific subpopulation exhibiting a high association with the signature. This result indicates that scNET is able to identify specific subpopulations associated with the signature while preserving the phenotypic differences between the two populations.

7 scNET identify differential enrich pathway in mouse GBM model.



Supplementary Figure 10. Reconstructed gene expression allows for better capture of differential pathway activity in different cell types. (a) Heatmap of the top 20 enriched pathways for different cell types after DE analysis in the original gene expression. (b) Comparison of enriched pathways in P-selectin inhibition treatment CD8+ T cells compared to the control. The first 9 pathways are associated with T cell and immune activation (green), while the remaining pathways are those depleted with respect to T cell-associated genes.

References

[1] Yeini, E., Ofek, P., Pozzi, S., Albeck, N., Ben-Shushan, D., Tiram, G., Golan, S., Kleiner, R., Sheinin, R., Israeli Dangoor, S., *et al.*: P-selectin axis plays a key role in microglia immunophenotype and glioblastoma progression. Nature communications **12**(1), 1912 (2021)

[2] Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O.: Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. Nature biotechnology **33**(2), 155–160 (2015)

[3] Hrvatin, S., Hochbaum, D.R., Nagy, M.A., Cicconet, M., Robertson, K., Cheadle, L., Zilionis, R., Ratner, A., Borges-Monroy, R., Klein, A.M., *et al.*: Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nature neuroscience **21**(1), 120–129 (2018)

[4] Kilian, M., Sheinin, R., Tan, C.L., Friedrich, M., Krämer, C., Kaminitz, A., Sanghvi, K., Lindner, K., Chih, Y.-C., Cichon, F., *et al.*: Mhc class ii-restricted antigen presentation is required to prevent dysfunction of cytotoxic t cells by blood-borne myeloids in brain tumors. Cancer Cell **41**(2), 235–251 (2023)