Enhancing Single-Cell Data Integration and Discovery through Knowledge Graph Extraction from scGPT Embeddings

Anne Cooley, Oleg Stroganov, Dan Rozelle
Rancho Biosciences, LLC

Abstract

The exponential growth of single-cell sequencing data necessitates innovative approaches for integration, analysis, and knowledge discovery. This study presents a novel method for extracting knowledge graphs from single-cell embeddings, utilizing the Rancho dataset from the Single Cell Data Science (SCDS) consortium, which comprises 70 million deeply curated cells. We employed scGPT, a large language model tailored for single-cell analysis, to embed a subset of 10 million cells from the SCDS dataset. The embeddings were projected onto a low-dimensional space using UMAP and subsequently clustered. Each cell cluster was represented as a node in a knowledge graph, with predicates such as "similar to," "subset of," and "superset of" connecting clusters across different datasets.

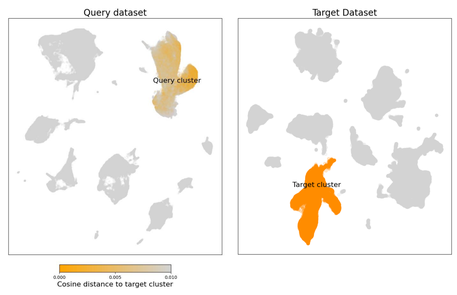
Leveraging the deep annotation at the sample level, we linked cell sets to specific diseases, tissues, and perturbations/conditions in which they were enriched. Additionally, cell sets were connected to overexpressed genes. The resulting knowledge graph effectively bridged diseases and perturbations to genes through cell sets, providing a high-resolution representation of the data. This approach significantly enhances the value of single-cell data by enabling the discovery of cell populations linked to specific phenotypes. Our method demonstrates the power of combining large language models with knowledge graph techniques to improve data integration and facilitate novel insights in single-cell biology. This framework has the potential to accelerate discoveries in various fields, including disease mechanisms, drug responses, and cellular heterogeneity.

Construction of the knowledge graph

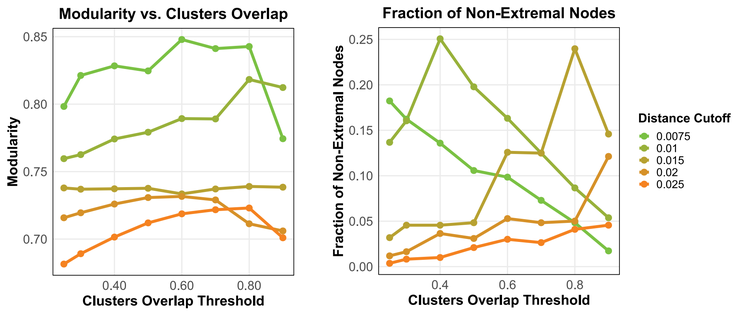
SINGLE CELL Data Science Consortium

1. 71 AI-ready datasets from Rancho's SCDS (11.2 million cells, 1324 donors, 3095 samples) were selected. The dataset was mostly focused on liver samples.
2. Gene expression embeddings (512 dimensions) were calculated with scGPT and loaded into a vector database for fast retrieval.

3. For every dataset UMAP 2D projection of scGPT embeddings was calculated. UMAP projections were clustered using DBSCAN.
4. Pairwise distance matrix between each cell cluster across all datasets was constructed. Two parameters were used to construct the matrix:
   - distance cutoff (maximal distance in embedding space so that cells are considered "the same")
   - cluster overlap – fraction of cells in a cluster that are within distance cutoff from another cluster



To achieve balanced clusters in a knowledge graph, the overlap threshold and distance cutoff can be optimized using metrics such as modularity—defined as the fraction of edges within a group compared to the expected fraction if edges were randomly distributed—or the fraction of non-extremal connected components. Sub-networks are robust when using distance parameters in the range of 0.0075 to 0.01 and overlap thresholds between 0.25 and 0.6.

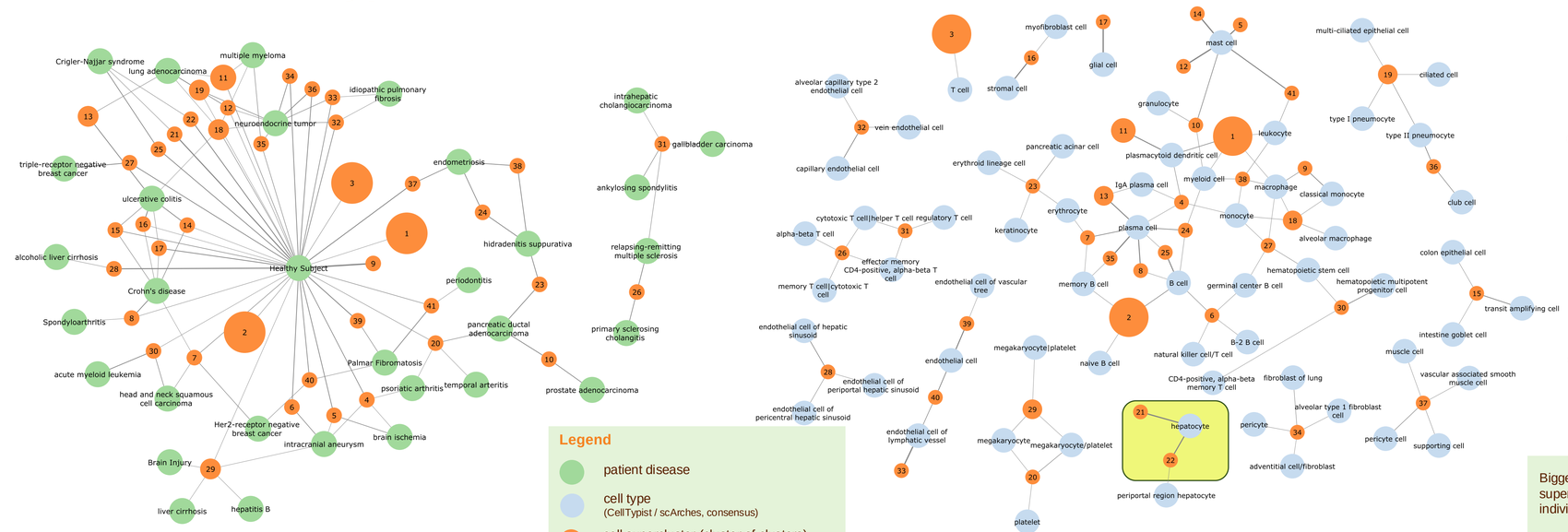


5. Knowledge graph is constructed where nodes are cell clusters, and edges are formed between clusters which are similar and have significant overlap.

   Knowledge graph is further enriched with curated information on patient disease, sample tissue, annotated cell type.

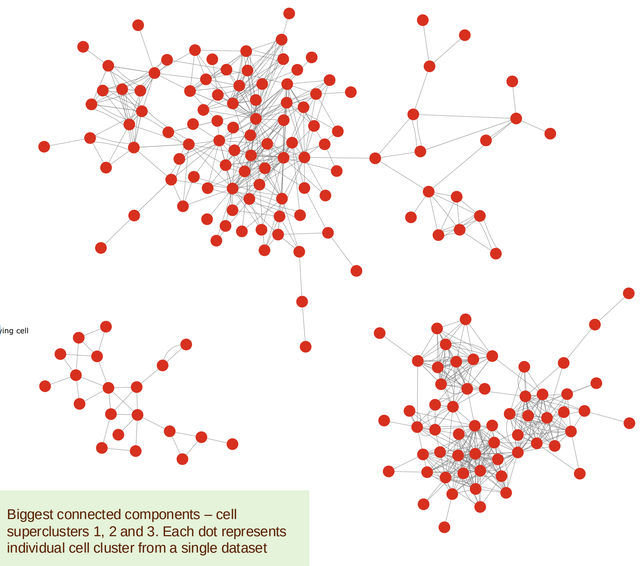
   Graph is further enriched by adding supercluster nodes (connected components), disease, tissue and cell type ontology relationships. Additional information (genes overexpressed in cluster, or conditions where cluster is overrepresented) could be added to address specific analysis needs.

Disease and cell type subnetworks



Legend

- patient disease
- cell type (CellTypist / scArches, consensus)
- cell supercluster (cluster of clusters) size of the circle corresponds to how many datasets are in supercluster
- edge width represents frequency of association

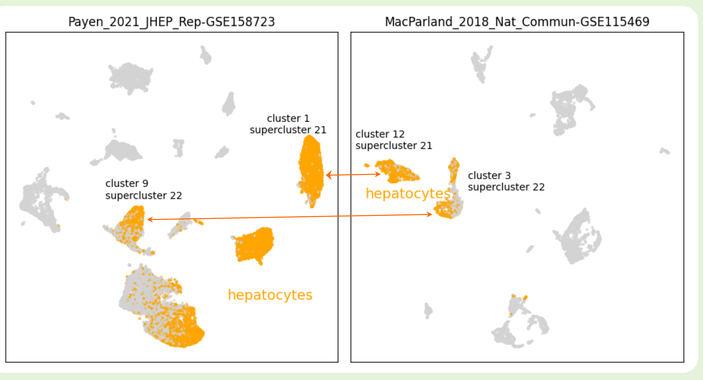


Biggest connected components – cell superclusters 1, 2 and 3. Each dot represents individual cell cluster from a single dataset

Applications

High resolution cell subtype discovery across datasets

Resolution of cell type ontologies and annotations may be insufficient to differentiate cells. For example, a number of different cell clusters are mapped to single cell type "**hepatocyte**" across datasets. The knowledge graph establishes clear connections between hepatocyte subpopulations across different datasets, enabling more biologically relevant analyses.



Knowledge graph analysis identified two distinct superclusters, **21** and **22**, representing different hepatocyte populations. UMAP clustering of cells from these superclusters shows a clear separation, underscoring their distinct biological identities. **ALB** expression, a well-established hepatocyte marker, confirms the hepatocytic nature of the cells in both clusters.

Differential gene expression (DGE) analysis further reveals specific genes distinguishing superclusters **21** and **22**. Notably, transcription factors **RORA** and **ZBTB20** are upregulated in supercluster **21**, with RORA involved in liver metabolism and circadian regulation, and ZBTB20 playing a role in hepatocyte differentiation and liver function.

GSEA analysis highlights metabolic and ubiquitination processes as key factors differentiating the two superclusters, emphasizing the hepatocyte's central role in protein metabolism. This demonstrates how knowledge graph integration across datasets can illuminate unique cellular populations, advancing research into single-cell data.