Adding layers of information to scRNA-seq data using pre-trained language models

Sonia Maria Krißmer[1,2], Jonatan Menger[1,3], Johan Rollin[4], Tanja Vogel[4], Harald Binder[1,3,5], Maren Hackenberg[1,5]

[1] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Germany
[2] Swiss Tropical and Public Health Institute, Basel, Switzerland
[3] CIBSS – Centre for Integrative Biological Signalling Studies, University of Freiburg, Germany
[4] Institute of Anatomy and Cell Biology, Department Molecular Embryology, Faculty of Medicine, University of Freiburg, Germany
[5] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Germany

Correspondence to maren.hackenberg@uniklinik-freiburg.de

August 23, 2025

Abstract

Single-cell technologies generate increasingly complex and multi-layered datasets, increasing the need for analysis workflows that incorporate additional biological information. Pretrained language models, with access to large corpora of biomedical literature, promise to provide such additional context to complement data-based analyses, yet recent approaches largely focus on data-intrinsic tasks. Here we propose a framework for context-aware enrichment of single-cell RNA sequencing data by aligning data-derived and literature-derived representations in a shared embedding space. We represent cells as sentences derived from ranked gene expression and metadata, and construct structurally parallel datasets from PubMed titles and abstracts. Lightweight encoder-only language models are trained jointly on both sources to learn a common embedding space, thus integrating additional layers of information from biomedical literature. Analyzing the joint embedding space, we show that biomedical literature can be meaningfully aligned with single-cell profiles to enrich standard analysis workflows. The trained models achieve robust annotation, capture functional states such as cytotoxicity, and reveal disease associations from literature-aligned embeddings. In developmental data, incorporating temporal metadata enables capturing temporal transitions consistent with cell lineage trajectories, demonstrating the potential of knowledge-augmented embeddings as a generalizable and interpretable strategy for extending single-cell analysis pipelines.
Keywords: *Single-cell RNA sequencing, Encoder-only language models, Text-based embeddings, Biological knowledge integration*

1 Introduction

Recent advancements in the field of single-cell biology allow for the generation of increasingly more complex and multi-layered data at single-cell resolution [1], for example, comprising different time points or disease statuses [2, 3, 4]. Foundation models have the ability to learn generalized representations that encode comprehensive biological knowledge from being pre-trained on such heterogeneous, large-scale datasets [5]. In particular, language models pre-trained on large text-corpora can be specialized on scientific literature, which makes them promising for context-aware enrichment of multi-layered single-cell analysis [6, 7, 8]. Yet, recent adaptations of single-cell foundation models

focus primarily on data-intrinsic tasks [e.g., 9, 10], but a principled strategy for an in-depth analysis of single-cell datasets in the context of relevant scientific literature is still missing.

Therefore, we propose a framework for training specialist language models for integrating and exploring additional layers of information from scientific literature or metadata to complement and enrich analysis of a single-cell RNA-seq (scRNA-seq) dataset. To integrate scRNA-seq data and literature knowledge for use in a unified modeling framework, we build text-based datasets with a consistent structural configuration from both information sources, such that they can be used in a joint training strategy for learning a common embedding space where they are meaningfully aligned. Specifically, we employ small encoder-only language models, which are well-suited for capturing semantic relationships [11, 12] and can generate meaningful embeddings of both scRNA-seq data and literature-derived knowledge using text as joint modality. This enables the analysis of scRNA-seq data in the context of scientific literature.

These language models are based on a transformer architecture. Originally developed for natural language processing, transformers are designed to process sequential data with inherent order, like words in a sentence, which contrasts with the unordered structure of omics data [5]. To make scRNA-seq data compatible for processing with transformer-based models, two main strategies have emerged [5]. The first approach adapts transformer architectures to handle quantitative gene expression directly. For example, models like scGPT incorporate expression values as an additional input layer, analogous to positional embeddings in language models [9, 5, 13]. Metadata is added through extra tokens to preserve relevant context.

Alternatively, omics data can be reformatted into a sequential structure without modifying the model architecture. In this case, a cell is represented as an ordered list of its most highly expressed genes, so-called "cell sentences." Models like Geneformer use gene ranking to build these sequences [10], while reconstructed expression values can be approximated through inverse-rank weighting schemes [14]. Although this strategy sacrifices some biological resolution [5], it enables compatibility with large-language models (LLMs) [for an overview, see, e.g., 15] and thus facilitates integration of quantitative expression data and text-based metadata (e.g., tissue, time point, cell type) without architectural changes. These strategies are implemented across different transformer variants. While decoder-only models dominate in LLM-based natural language processing [e.g., GPT-4 and DeepSeek 16, 17], single-cell foundation models use different architectures: encoder-only, decoder-only, or encoder–decoder hybrids [5].

Once pretrained, these models can be fine-tuned for diverse tasks such as cell type annotation, perturbation prediction, and batch integration. For instance, C2S-Scale uses decoder-only models like Pythia trained on cell sentences to support both traditional single-cell tasks and novel applications like dataset interpretation via natural language queries [18]. Other models leverage gene embeddings from general-purpose LLMs (e.g., GPT-3.5) to create cell embeddings based on expression-weighted gene vectors [19, 20]. Recent work aims to combine the strengths of both data representation strategies. For example, embedding spaces shared between transcriptomic models like Geneformer or scGPT and literature-trained models like BioBERT [21] allow user queries to be matched directly to cells via semantic similarity [22, 6, 23], thus integrating raw expression data with textual knowledge.

These approaches highlight the flexibility of applying language models for modeling gene expression data, enabling fine-tuning for diverse tasks and integration with various data types. However, scientific literature is typically only incorporated implicitly during general pre-training, and models do not yet leverage literature-based knowledge specific to the dataset of interest for downstream analysis and interpretation.

Here we present a modeling approach that allows for the integration of contextual background information from biomedical scientific literature to add layers of information to single-cell datasets of interest. To build text-based datasets with a consistent structural configuration from both information sources, we represent high-dimensional scRNA-seq data as cell sentences including available metadata for single cells and build triplets of cell sentences, where each sentence is paired with a similar and a dissimilar sentence, respectively [24, 25] . Based on available metadata, we collect biomedical literature retrieved from PubMed [26] to specifically enrich the dataset of interest, and build corresponding triplets from PubMed titles and abstracts. This allows for employing encoder-only language models with a Siamese-BERT (SBERT) architecture [12]. These models use contrastive losses

during fine-tuning [27], which has been shown to effectively capture semantic relationships between sentences in a high-dimensional embedding space. This is crucial for aligning scientific literature with the gene expression data in a semantically meaningful way [28]. We showcase the capabilities of the trained models to align scRNA-seq data with the text-based metadata and scientific literature by investigating cell type functionality annotation, disease relations and temporal pattern in the context of lineage trajectories. Our work highlights the unique advantages of enriching a conventional data-centered scRNA-seq analysis with additional layers of information via pre-trained language models, to explore relationships of genes, cells and cell types with other layers of interest.

2 Results

2.1 A unified workflow to embed single-cell profiles in the context of biomedical knowledge

We developed a novel approach for adding layers of information such as cell type functionality, disease, or temporal trajectory relationships to high-dimensional scRNA-seq data using Siamese-BERT models [12]. Specifically, we generate embeddings of gene expression data via a training strategy on literature-derived and scRNA-seq-derived datasets using text as a joint modality. This is enabled through a transformation of high-dimensional gene expression vectors of each cell into a ranked gene list, such that a cell's gene expression is represented as textual information and can thus be directly embedded with a language model without depending on an additional embedding model. The approach builds on the availability of labeled sc-RNA datasets, that have been processed by a conventional analysis workflow comprising dimensionality reduction and clustering for capturing cell clusters, that can be annotated, e.g., based on cluster-specific differential gene expression analysis [29, 1]. This workflow is illustrated on the left in Figure 1. Our proposed framework allows for enriching such already processed datasets with layers of additional information, e.g., on cell type functionality, temporal patterns, or disease relationships, as illustrated on the right in Figure 1.

To achieve this, we start from the gene expression count matrix also used for conventional analysis (Figure 1, 'Data Preparation'), and transform the top expressed highly-variable genes of each cell, ranked by their expression, into a list of genes with decreasing expression values. For each cell, ranked gene lists were truncated at a length of 50. This cutoff was determined from an evaluation of the base model, which showed that restricting the lists to the top 50 ranked genes provides a practical balance between representation accuracy and computational efficiency (results not shown). The truncated lists were then transformed into cell sentences. In addition to the gene list, metadata can be added to the cell sentence, e.g., about the cell type or donor, for training the models with additional context information, depending on the specific application. For later evaluation of the model, cell sentences from a test split of the dataset consisting of the gene lists only are kept aside.

To generate literature-based training datasets, we focus on articles related to cell types or individual cells in our scRNA-seq dataset. Specifically, we use a search term strategy on literature stored in the National Center for Biotechnology Information (NCBI) PubMed database [26] to identify articles related to the organism, cell types, or embryonic time points of interest, and extract their titles and abstracts. This allows for capturing additional information related to the dataset, such as cell type functionality (*functionality layer*) and disease relationships (*disease layer*) or time-dependent information such as lineage trajectories (*time layer*), illustrated in Figure 1.

The aim is to train a model that aligns the embeddings of the cell sentences and literature information, based on their semantic relationship rather than the original source of information, i.e., such that they generate similar embeddings for single-cells and their respective textual descriptions such as cell type labels and functionality descriptions. To achieve this, we transform the generated cell sentences and corresponding NCBI search results into label-aware triplets consisting of an anchor, a positive and a negative sample, where the cell type serves as the label. Triplets are collected within the cell-derived and literature-derived dataset with a a label-aware random-positive and hard negative mining strategy, which ensures high triplet quality, illustrated in Figure 1, 'Label-aware Hard-negative Mining', and described in detail in Methods Section 5.6. On these triplets, we fine-tune an encoder-only language

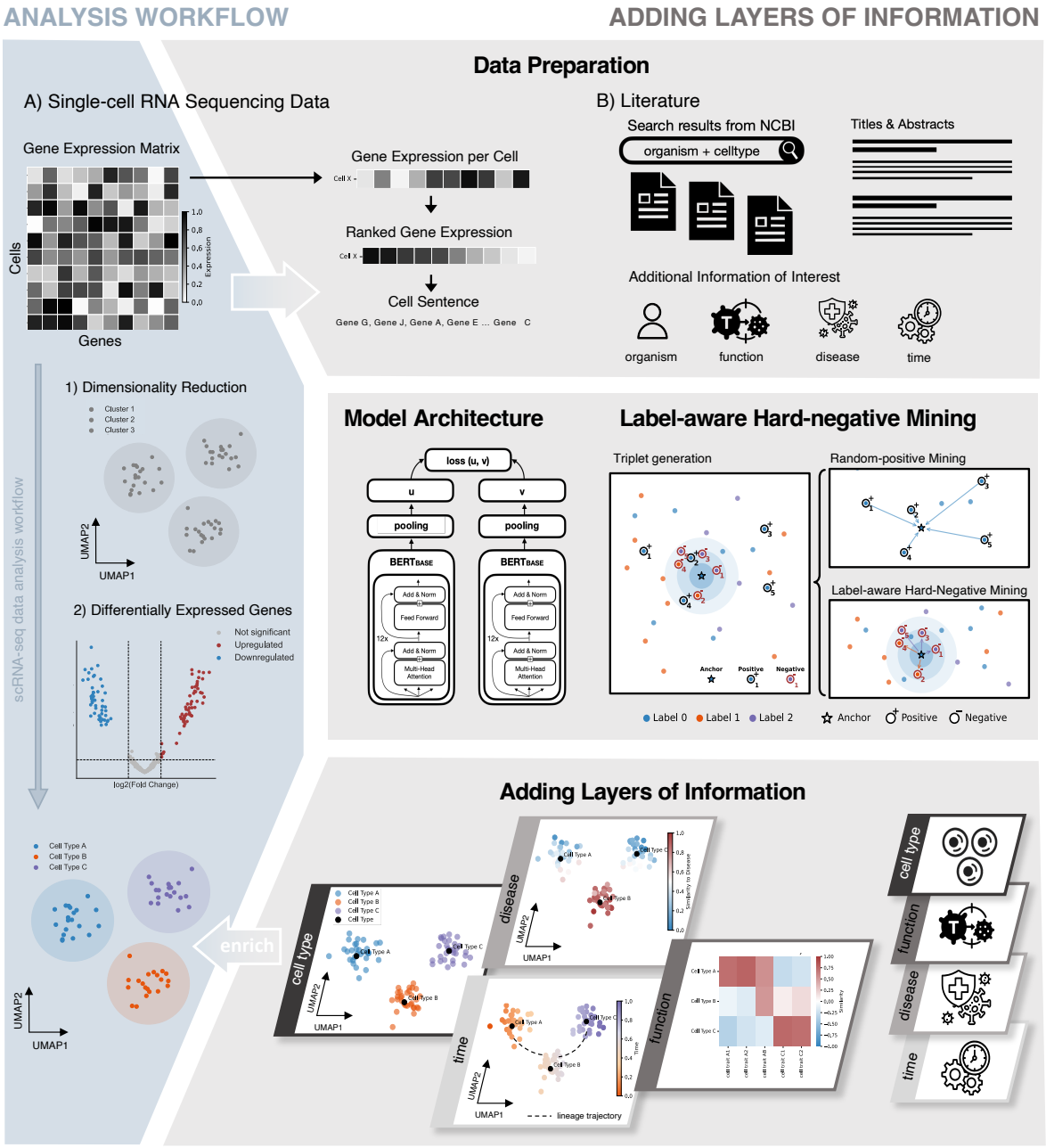


Figure 1: Overview of the methodology for adding layers of information to scRNA-seq datasets. A gene expression dataset is processed through a quantitative analysis pipeline typically comprising dimensionality reduction steps and differential gene expression for annotation of cell type clusters. For adding layers of information we translate single-cell RNA-sequencing data into cell sentences comprising of the top highly expressed genes per cell for generating the cell-derived dataset. We additionally perform a literature search on the known or suspected cell types in the dataset capturing any additional information about the cell types stored in the literature such as cell type functionality (function), disease relationships (disease) or lineage trajectories (time) for generating the literature-derived dataset. Using label-aware hard-negative-mining and random-positive mining we construct triplets for both datasets for training our models comprised of a Siamese-BERT [12] architecture. Aligning the embeddings of cell type labels with single-cell data and the additional information captured in the literature we evaluate additional layers of information in the context of our scRNA-seq dataset.

model using epoch-wise alternating training between the datasets (Methods Section 5.8). As training objective, we use a so-called 'multiple negatives ranking loss' function. This allows for capturing semantic relationships within the input data triplets, by pulling the embedding of the positive sample

closer to the anchor, while pushing the negative further apart (for details, see Methods Section 5.7). Specifically, we employ small encoder-only language models with a Siamese-BERT architecture (see Methods Section 5.5), which is specifically designed for the efficient computation of the similarity of multiple inputs, shown in Figure 1, 'Model Architecture'. This training set-up encourages the model to learn a joint embedding space where different layers of information, such as cell type labels and expression-derived ranked gene lists representing single cells, are aligned, rather than clustered by data modality (cell type label vs single cell). For example, the embedding of a cell type label 'Naive CD4 T Cell' should be close to the embedding of the ranked gene-lists of the naive CD4 T cells rather than being close to the embeddings of other cell type labels

After fine-tuning, we extract the output of the last hidden layer, the embedding of the input data learned by the model and apply UMAP for dimensionality reduction and visualization. We investigate the alignment of the cell sentences, composed of the top 50 expressed genes, with the cell type labels showcasing how the model can distinguish and identify cells of different cell types based on their gene expression (Section 2.2). Via the PubMed search strategy explained above, we can enrich the representation by adding further layers of information based on scientific literature or metadata, and specifically report results on investigating cell type functionality and disease status in Section 2.3, and capturing temporal transitions in Section 2.4.

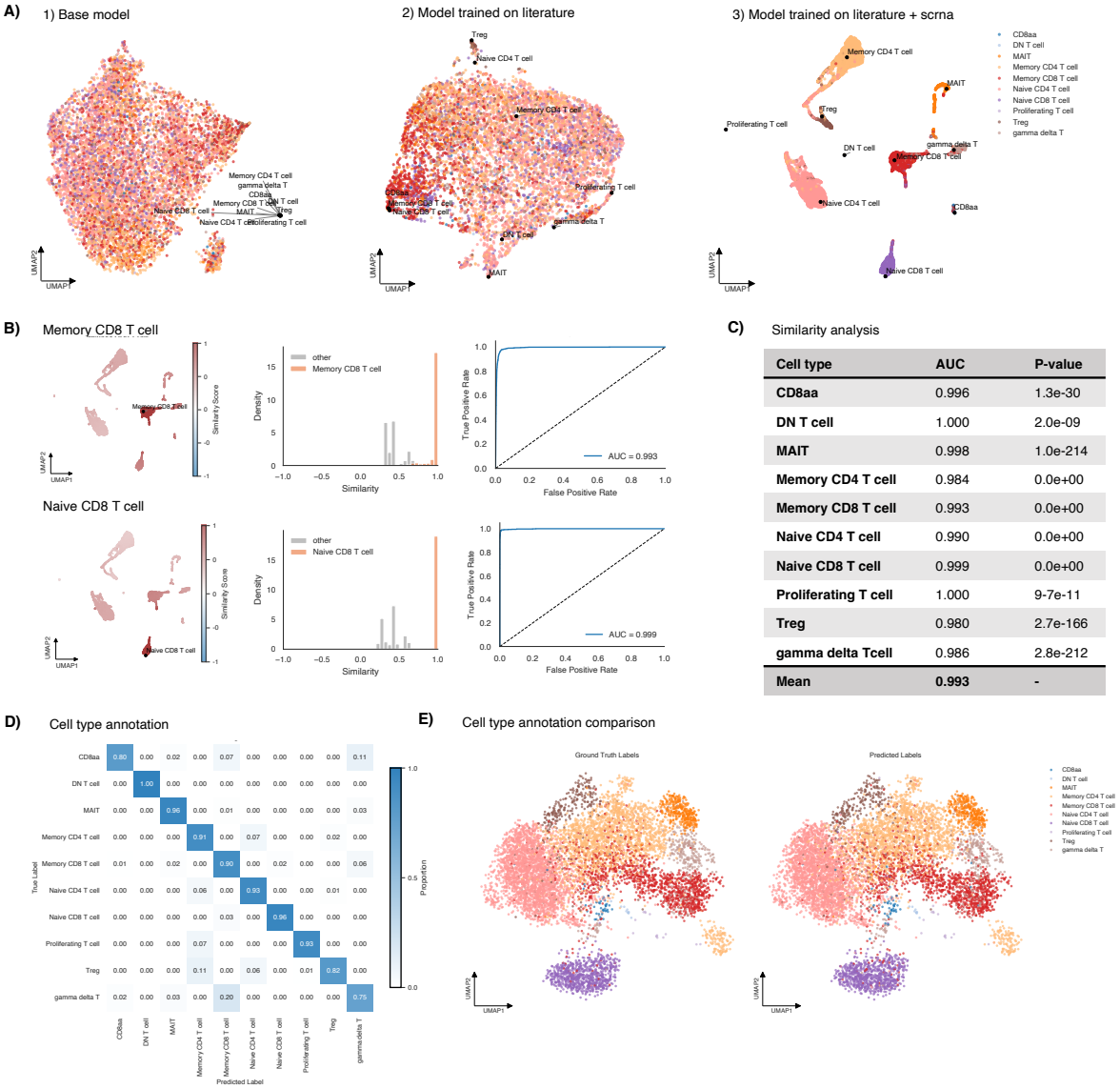
2.2 Joint training on literature and gene expression data improves embeddings and aligns cell type labels

We first train our models on a T cell-specific subset of an atlas of primary blood mononuclear cells (PBMCs), specifically the Human Immune Health atlas from the Allen Institute (HIAI dataset) [2] and first investigate the alignment of the cell sentences with the cell type labels as per our training objective. To verify that training our models is essential for the alignment we inspect the embedding of our base language model before fine-tuning. The low-dimensional UMAP representation of the model embeddings shows poor separation of cell subtypes. Notably, embeddings of cell type labels form a separate cluster distant from the corresponding single-cell embeddings (Figure 2 **A**: '1) Base model').

Next, we train a model on a literature-based dataset only, where we use the cell types in the HIAI dataset as queries. Figure 2 **A**: '2) Model trained on literature' shows improved alignment between cell type labels and the corresponding single-cell profiles, but a clear clustering by subtype is still lacking.

In contrast, fine-tuning the model on both literature-based and expression-based datasets results in an embedding with distinct cell clusters by subtype, with their respective cell type label embedded within those clusters (Figure 2 **A**: '3) Model trained on literature + scrna'). This shows that the fine-tuned model is able to differentiate cells of different cell types based on their top expressed genes. Additionally, we can observe meta-clusters of CD4+ and CD8+ T cells and can observe that cells with similar functionality also cluster closer together. One example includes the close proximity of Memory CD8+ T cells and $\gamma\delta$ T cells in the low-dimensional representation of the embeddings, both characterized through their strong cytotoxic capabilities [30, 31]. Interestingly, we can also see a connection between cell types with a known biological transition such as Naive CD4 T cells, that upon encountering their antigens mature into Memory CD4 T cells [30], suggesting that the models captures the maturation of cells.

To quantify the alignment between cell embeddings and their annotated types, we compute the cosine similarity between each cell embedding and the embeddings of candidate cell type labels. We find that cells consistently show highest similarity to their correct label, as illustrated for Naive and Memory CD8 T cells in Figure 2 **B** (metrics reported for all other cell types in Figure 2 **C**). Finally, we asses the alignment of cell type labels with the single cells using this similarity-based approach for cell type annotation on held-out test data. Here we assign each cell the label with the highest cosine similarity (Figure 2 **D, E**). This yields a balanced accuracy score of 89.7 %, demonstrating the model's ability to capture meaningful relationships between cell expression profiles and textual labels. Here we can observe that the model shows higher prediction errors for functionally similar cell types, such as $\gamma\delta$ T cells and Memory CD8 T cells or Regulatory T cells with other CD4 + T cell subtypes [30], while we can observe little errors for differentiating between functionally different subtypes such as CD4+ and CD8+ T cells.



Cell type	AUC	P-value
CD8aa	0.996	1.3e-30
DN T cell	1.000	2.0e-09
MAIT	0.998	1.0e-214
Memory CD4 T cell	0.984	0.0e+00
Memory CD8 T cell	0.993	0.0e+00
Naive CD4 T cell	0.990	0.0e+00
Naive CD8 T cell	0.999	0.0e+00
Proliferating T cell	1.000	9-7e-11
Treg	0.980	2.7e-166
gamma delta Tcell	0.986	2.8e-212
Mean	**0.993**	**-**

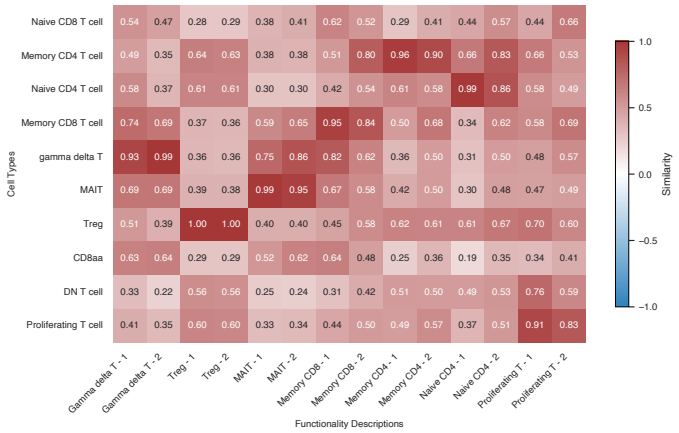
Figure 2: Joint training on cell-derived and literature-derived datasets allows encoder-only language models to capture complex gene interactions using joint text based embeddings. A) Training on both datasets allows the model to cluster cells based on their ranked gene lists and embeds the cell type labels within the cell cluster of the respective cell type. B) Analysis of cosine-similarity of cell type labels of cells of the respective cell type vs all other cells showing similarity values. Similarity scores are compared using threshold decision tests, for the separation of cells of the respective cell types from all other cells using the similarity score, evaluated using the receiver operating characteristic (ROC) C) Overview of analysis metrics showing the AUC and Mann-Whitney U test p-value for all cell types and Mean AUC. Cosine-similarity of cells to cell type labels can be used for cell type annotation, assigning each cell the cell type label with the highest similarity, resulting in the confusion matrix (D) and a comparison of UMAP plots colored by the predicted vs true cell types (E)

2.3 Cell type functionality programs and disease associations are captured unsupervised and at single-cell level

Training our model on both literature- and cell-derived datasets allows for further analysis of the scRNA-seq data in the context of knowledge and relationships captured in the scientific literature beyond cell type identities. For the HIAI T cell dataset we are specifically interested in investigating cell type functionality and disease relationships captured in the literature. We use expert-evaluated functionality descriptions for the different cell types of the HIAI T cell dataset and compare the cosine similarity of the descriptions to the different cell type labels (Figure 3 **A**). The results indicate that the model is able to capture the different functionalities of a cell type in relation to the cell type labels,

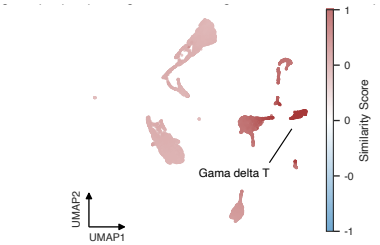
A) Cell type – Functionality

Cell type	Label	Functionality Description
Gamma delta T	**1**	Recognize stress ligands independently of classical MHC molecules.
	2	Recognize lipid, phosphoantigens, and stress ligands via non-conventional pathways.
Treg	**1**	Produce immunosuppressive cytokines such as IL-10 and TGF-β
	2	Stably express FOXP3 to maintain suppressive identity.
MAIT	**1**	Recognize bacterial-derived riboflavin metabolites presented on MR1.
	2	Bridge innate and adaptive immunity with semi-invariant TCR and rapid cytokine output.
Memory CD8	**1**	Rapidly kill target cells upon antigen re-exposure.
	2	Retain long-term survival and recall potential
Memory CD4	**1**	Effector memory subset lacks CCR7 and homes to inflamed tissues.
	2	Central memory subset homes to lymphoid tissues and proliferates upon recall.
Naive CD4	**1**	Serve as unprimed precursors to all helper T cell subsets.
	2	Act as a pool for generating novel responses to previously unseen antigens.
Proliferating T	**1**	Undergo robust IL-2-driven expansion during acute response.
	2	Increased metabolic activity and cell cycle progression.

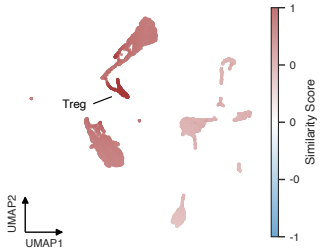


B) Cell – Functionality

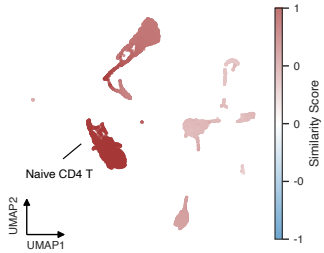
Gamma delta T - 2: Recognize lipid, phosphoantigens, and stress ligands via non-conventional pathways.



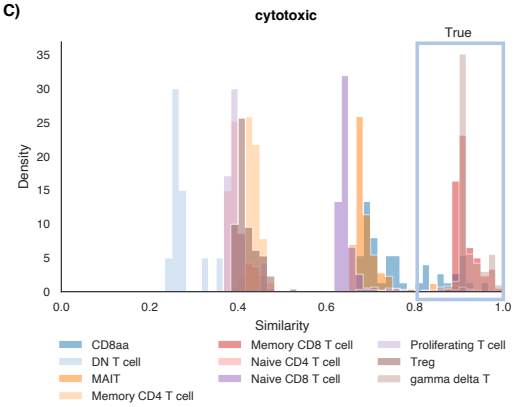
Treg - 1: Produce immunosuppressive cytokines such as IL-10 and TGF-β



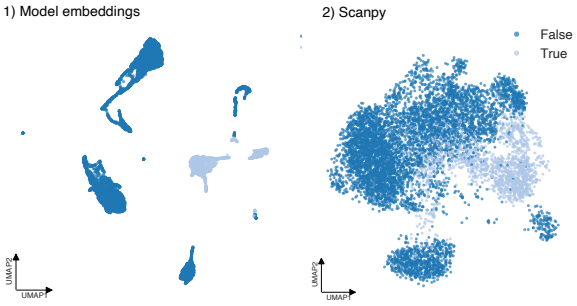
Naive CD4 T – 1: Serve as unprimed precursors to all helper T cell subsets.



C)



D)



E)

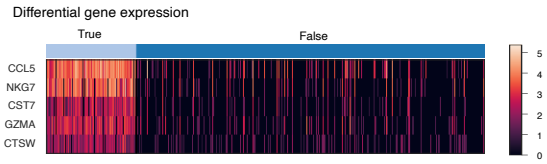


Figure 3: Model trained on cell- and literature-based datasets captures cell type functionality descriptions on cell type and cell level. A) Summary table of expert-derived cell type functionality descriptions, similarity heatmap comparing model-derived embeddings of cell type functionality descriptions in relation to the corresponding cell type labels. B) Single-cells represented through ranked gene-lists show higher similarity to embeddings of functionality descriptions of their respective cell type. C) Similarity scores of single cells to 'cytotoxic' can be used to identify cytotoxic cells in D) in the UMAP representation from the 1) model embeddings and 2) quantitative gene expression. E) Comparing cytotoxic cells 'True' to non-cytotoxic cells 'False' using differential gene expression shows cells with cytotoxic capabilities were selected.

indicated by a high similarity of the functionality descriptions to the respective cell type labels. We additionally observe that the model picks up joint functionality relationships. For example, MAIT cells can be described to 'Bridge innate and adaptive immunity with semi-invariant TCR and rapid cytokine output.' [32]. While this description holds true for all MAIT cells, it also hold true for the Vδ2 subset of γδ T cells [30]. This is captured by the model through a higher similarity of the functionality descriptions of MAITs for γδ T cells compared to all other T cell subtypes, where this description is not correct.

The alignment of cell type labels and single cells in the embedding space additionally allows for the transfer of the functionality descriptions from the cell type labels to single-cells (Figure 3 **B**). We can see that the ranked gene lists of the single cells also show higher similarity to the functionality descriptions related to their respective cell type.

We can employ this for identifying cells of cell type overarching functionalities such as cytotoxicity. We calculate the similarity scores of the single cells represented by their ranked gene list to the description 'cytotoxic' and select those cells with a similarity score greater than 0.8, whichseparates known cytotoxic cell types from other cells (Figure 3 **C**). The selected 'cytotoxic' cells are displayed in Figure 3 **D** where a clear separation can be seen for the model embeddings while the cells do not show a clear separation in the UMAP representation of the quantitative gene expression. Performing differential gene expression analysis on the selected cells (Figure 3 **E**), we can see that the top five genes that are differentiable expressed in the selected cytotoxic cells are known cytotoxicity markers [33]. The expression of the genes is shown using the model and scanpy embeddings in Figure S1, showcasing the expression of cytotoxicity markers in the 'cytotoxic' cells in the model embeddings. This showcases how the alignment of knowledge captured in the literature can be employed for enriching scRNA-seq datasets.

Similar to the cell type functionality description we can analyze how the role of different cell types during different disease states is captured by the model, using expert-evaluated statements and summaries of the role of different cell types with cancerous diseases. The similarity scores of the different descriptions of disease relationships with the cell type labels as well as the cell-derived gene lists confirm that comparable to the functionality descriptions the relation of the different cell types to different disease states is captured during training of the model (Figure 4) This showcases the cross-correlation of gene - cell functionality also within disease states and highlights the benefits of incorporating both literature-based and cell-based data.

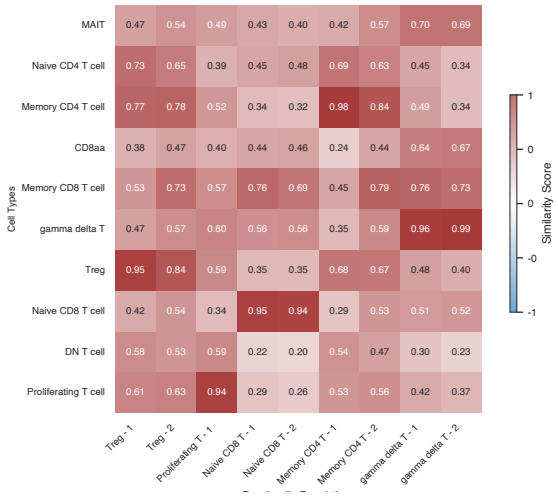
2.4 Embeddings capture temporal transitions aligning with developmental lineage trajectories

We evaluate the capabilities of the models to capture additional information from metadata included in a scRNA-seq dataset, such as the embryonic time point in developmental data. Specifically, we use a second dataset of the developing embryonic mouse brain referred to as LaManno dataset [34]. This dataset includes highly diverse cell types spanning multiple time points of embryonic development from embryonic day 7 to embryonic day 18. We incorporate the temporal metadata into the generated semantic cell sentences as described in Section 5.2. For clearer visualization of the embeddings we have split the dataset into 4 subsets containing cells of around 15 cell types each. The fine-tuned model aligns the embeddings of cell type labels and cells as shown for an exemplary subset in Figure 5 **A** and for all subsets in Figure S2.

The results show that in addition to aligning cell type labels and corresponding cell profiles, the model captures time trajectories in meta-clusters spanning multiple annotated cell types, where cells of structures from the early embryonic development are clustered separately of cells of more refined and mature compositions (Figure 5 **B**). This can be used for the downstream analysis of known lineage trajectories as shown in Figure 5 **C**. Here we show how the model captures the temporal transitions within the trajectory from 'Epiblast', an early cell type giving rise to all three germ layers, to cells of the 'Intermediate meninges', that develop to form the Dura mater and Pia mater of the meninges. We can see the temporal trajectory in the embeddings from Epiblast cells to cells from the intermediate meninges over cells of the 'Early ectoderm', 'Mesenchyme', 'Neural crest' and 'Early fibroblast' in the low-dimensional UMAP representation of the model embeddings (shown in in Figure 5 **D**), which aligns with the known lineage development of the meninges [35].

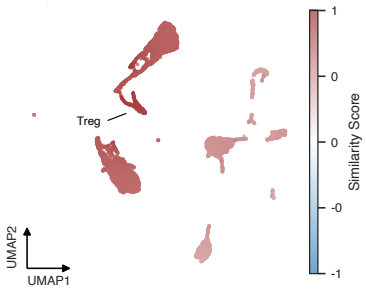
A) Cell – Role in cancerous diseases

Cell type	Label	Description
Treg	**1**	Correlate with poor prognosis in multiple cancer types due to immune suppression.
	2	High infiltration in the tumor microenvironment is associated with reduced survival due to suppression of immune responses.
Proliferating T	**1**	Expansion of clones suggests strong recognition of tumor antigens and TCR engagement.
Naive CD8 T	**1**	Differentiate into tumor-specific cytotoxic cells following priming with neoantigens.
	2	Depend on dendritic cell priming and costimulation to develop effective antitumor function.
Memory CD4 T	**1**	Can become targets of immune checkpoints like PD-1, reducing their helper activity.
	2	Some subsets (Th1) promote immunity, while others (Th2, Th17) can have variable or even pro-tumor effects depending on context.
Gamma delta T	**1**	MHC-I independent recognition of phosphoantigens.
	2	Exhibit anti-tumor function, which can be increased using Zoledronate.

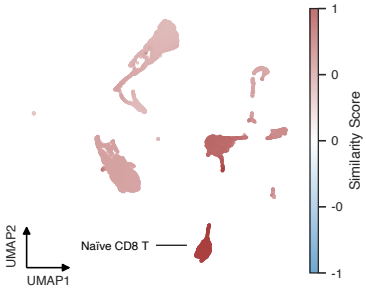


B) Cell – Role in cancerous diseases

Treg - 1: Correlate with poor prognosis in multiple cancer types due to immune suppression.



Naive CD8 T - 1: Differentiate into tumor-specific cytotoxic cells following priming with neoantigens.



Gamma delta T - 1: Exhibit anti-tumor function, which can be increased using Zoledronate.

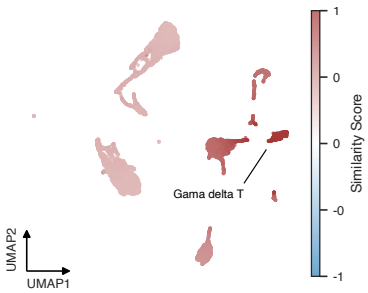


Figure 4: Model trained on cell- and literature-based datasets captures the role of different cell types in cancerous diseases on cell type and cell level. A) Summary table of expert-derived descriptions of the role of different T cell subtypes in cancerous diseases, similarity heatmap comparing model-derived embeddings of disease relationship descriptions in relation to the corresponding cell type labels. B) Single-cells represented through ranked gene-lists show higher similarity to embeddings of the disease relationship descriptions of their respective cell type.

We can additionally analyze the cosine similarity of the single cells of this trajectory with the respective times, shown for E‘7’, E‘9’ and E‘15’ in Figure 5 **E**. Here we can observe that cells of cell types present during the early development such as ‘Epiblast’ cells have a higher similarity score with the time E‘7’, while cells of the ‘Intermediate meninges’ have a higher similarity to E‘15’. The analysis of the lineage trajectories of ‘Erythroid progenitor‘’ cells to mature ‘Erythrocytes’, from ‘Angioblast’ to ‘Endothelial’ cells and from the ‘Hindbrain roof plate’ to glutamatergic neurons of the Cerebellum [36] is shown in Figure S3 **A-F**.

This highlights the potential of incorporating complex layers of information from temporal metadata and biomedical literature for reconstructing lineage relationships.

3 Discussion

We present a framework for integrating scRNA-seq data with biomedical literature by generating joint text-based embeddings using small, encoder-only language models. Through contrastive training on cell-derived and literature-derived data, the model learns to align diverse layers of biological information in a unified semantically meaningful embedding space that captures not only cell identities but also functional programs, disease associations, and developmental trajectories. Our results demonstrate that the approach can meaningfully connect gene expression profiles of individual cells with

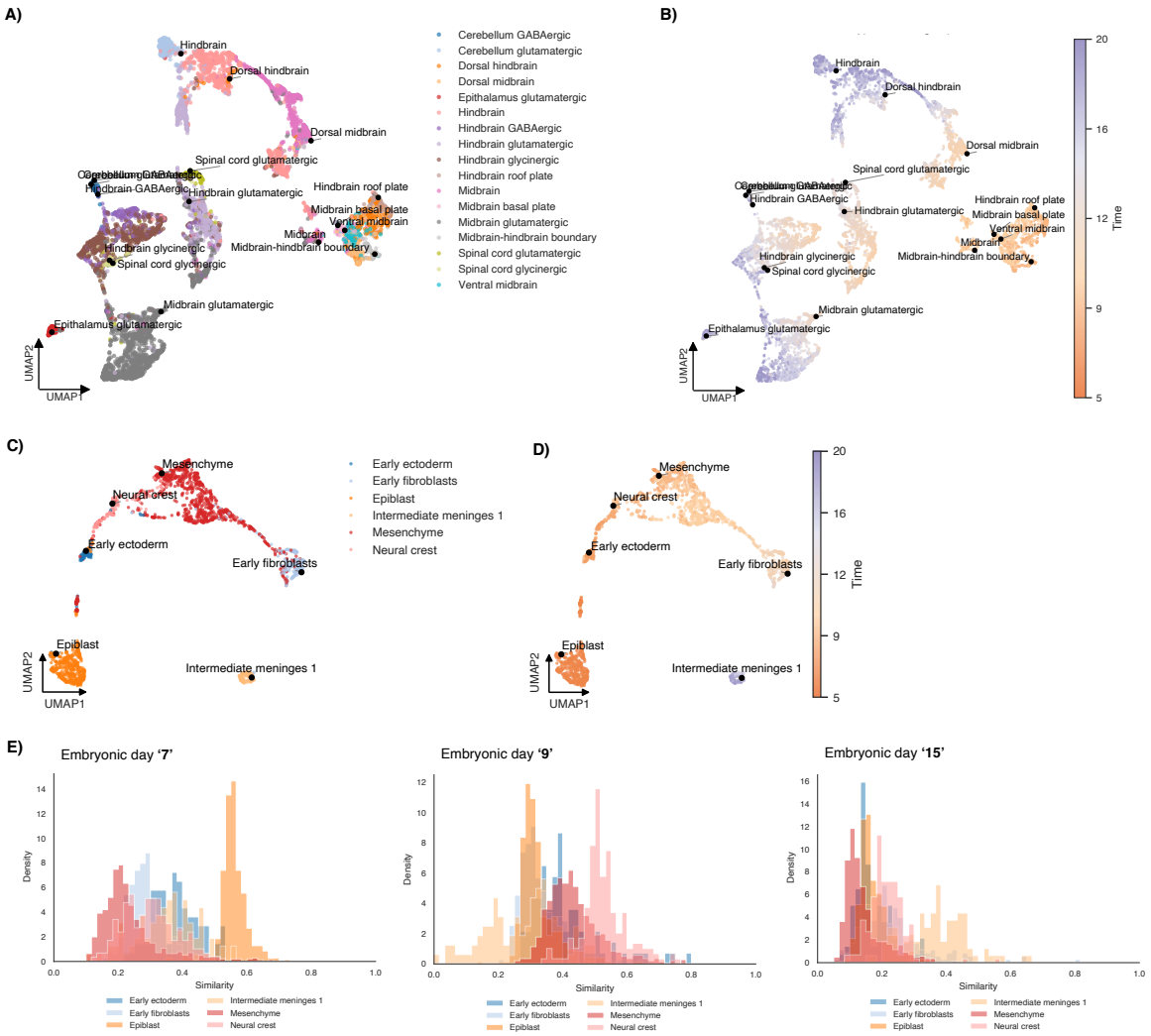


Figure 5: Model trained on single-cell based and literature-based datasets incorporating available metadata of the embryonic time point captures time relationship through meta-clusters related to major developmental stages as well as capturing of single lineage trajectories. A) Clustering of cells based on their ranked gene lists shows distinct clusters with cell type labels embedded within the respective cell clusters, shown on a cell-type filtered subsets of the full dataset. B) Time resolution is captured for all cell types showing a larger meta-cluster for structures in early embryonic development and two meta-clusters for maturer structures. C) Lineage trajectory from 'Epiblast' to 'Intermediate meninges 1' colored by cell type and D) time. E) Cosine similarity of single cells from lineage trajectory to time '7', '9' and '15' colored by cell type label.

textual descriptors of their biological roles and function. Fine-tuning the model on both literature and ranked gene lists leads to embedding spaces where cell type labels cluster with their corresponding cells, and where descriptions of cellular function or disease relevance show strong similarity to appropriate cell types. When incorporating temporal metadata into the cell sentences, the model captures developmental trajectories, highlighting its potential for studying dynamic processes such as differentiation.

The integration of gene expression data and literature through a shared text modality addresses a limitation in current single-cell foundation models. Previous models that use LLMs to embed gene identities or metadata [9, 19] often do not incorporate the rich contextual knowledge in the biomedical literature. By leveraging this information, our approach enables interpretability and hypothesis generation across biological layers, supporting downstream tasks such as context-aware cell type annotation and the transfer of functional and disease-related knowledge into single-cell analyses. Our current implementation relies on relatively lightweight models (110M parameters), which can be efficiently run and fine-tuned even on small machines and are adaptable to new datasets with known cell types.

Nonetheless, the current implementation has important limitations. First, our literature-derived data relies solely on PubMed titles and abstracts, which, while information-dense, omit details found only in full-text articles and may under represent newly described or rare cell types. The keyword-based search strategy can introduce biases, making the model sensitive to specific phrasing (e.g., "immunosuppressive" appearing frequently with "Regulatory T cell"). Additionally, our approach currently depends on pre-annotated cell types, which limits its applicability in fully unsupervised settings. While our approach can be viewed as a specialist model, designed for targeted integration of domain-specific knowledge into expression data, where prior labels or context are available, recent large-scale approaches like CellWhisperer [22] and C2S-Scale [18] show the potential of generalist models trained on large datasets. Training larger base models on broader training datasets and literature corpora could enable robust zero-shot generalization across more diverse tissues and organisms. Additionally, future work should incorporate additional data sources such as full-text articles, curated gene and protein databases, ontologies, and pathway annotations. As single-cell datasets continue to grow in scale and complexity, integrating them with such diverse knowledge sources will be an essential step towards transitioning from task-specific, dataset-bound models toward more versatile foundation models applicable across a broad range of biological contexts.

4 Conclusion

Our approach demonstrates the utility of encoder-only language models as lightweight, specialist tools for enriching single-cell RNA-seq data with information from the biomedical literature. By aligning expression profiles and literature-derived knowledge in a joint embedding space, our approach enables interpretable analyses across multiple biological layers, including cell identity, function, disease association, and developmental stage. This strategy offers a scalable and efficient framework for complementing single-cell analyses to inform functional interpretation and hypothesis generation across biomedical contexts. The work thus contributes to ongoing efforts to tailor large language models to complex biological datasets to enable interpretable and context-aware single-cell-based analyses.

5 Methods

5.1 Preprocessing of single-cell RNA seq data

5.1.1 Human Immune Health Atlas (Allen Institute)

We used a subset of the Human Immune Health Atlas from the Allen Institute (HIAI) containing only T cells (1 million cells) [2]. Cell type annotations are provided at three levels of resolution; we used the intermediate (level 2) annotation, resulting in 10 distinct T cell subtypes. Preprocessing was done using `scanpy`, according to the workflows described in Heumos, Schaar, et al. [29]. Cells were retained if they expressed more than 500 and less than 5000 genes, and genes were retained if detected in at least 50 cells. After normalization and log transformation of the counts, we selected the top 2000 highly variable genes ('seurat_v3' method), excluding ribosomal and mitochondrial genes. Data were split into training (90%) and test (10%) sets, with the training set further divided into a training (90%) and evaluation (10%) set. We refer to this dataset as the *HIAI dataset*.

5.1.2 Developing Mouse Brain dataset (LaManno et al.)

As a second dataset, we used the developing mouse brain atlas (270,000 cells at 20 embryonic time points E7-E18) of La Manno, Gyllborg, et al. [34]. We only considered cell types with more than 500 cells in the dataset. We removed cells with less than 100 expressed genes and genes expressed in less than five cells. We perform normalization and log transformation and selected the top 2000 highly variable genes ('seurat_v3'), excluding ribosomal and mitochondrial genes. As before, we split the data into a training and test (10%) set, with the training set subdivided into a training (90%) and evaluation (10%) set. We refer to this dataset as the *LaManno dataset*.

5.2 Generation of cell sentences

To represent high-dimensional quantitative gene expression profiles in textual form, we followed the Cell2Sentence approach [14]. Specifically, for each cell we ranked highly variable genes (excluding housekeeping genes) by expression and generated gene lists of different lengths (10, 20, 50, 100, 200) containing the highest expressed genes. For the training data, we then used sentence templates to convert gene lists into "semantic" cell sentences that combine the expressed genes with metadata such as cell types, developmental time point, or disease. Templates were generated with AI assistance and included examples such as:

- Gene list only: "*{gene_str} are the top expressed genes in this cell.*"
- Gene list and cell type: "*A cell that expresses the following genes: {gene_str} is likely a {celltype} cell.*"
- Gene list, cell type and time point: "*A {celltype} cell at {time} expresses these genes: {gene_str}.*"

For training with different metadata information, we combined cell sentences containing only the gene list with sentences containing also the metadata. For example, in experiments combining gene lists and cell type information, we used sentences containing only the gene list for 70% of cells, and sentences containing both the gene list and the cell type label for 30% of cells. In experiments combining gene lists, cell type and time metadata, sentences were distributed across four variants: sentences containing the gene list, the cell type label and the time point (50%), sentences containing the gene list and time point (20%), sentences containing the gene list and cell type (20%), and sentences containing the gene list only (10%).

The test data were converted to gene lists only, without adding metadata.

5.3 Collecting data for training on literature

For enriching each scRNA-seq dataset with additional information from scientific literature, we fine-tuned models on PubMed titles and abstracts relevant to the cell types in the given dataset. Searches were performed separately for the HIAI and the LaMano dataset using the the NCBI PubMed database, accessed via the Entrez Python package. Specifically, we included as search terms (1) the organism as MeSH seach term (human or mouse), (2) the cell types of interest, i.e., the ones present in the considered dataset, and (3) optional additional metadata of interest, such as disease terms. These search terms were combined using AND/NOT operators to generate queries such as:

- cell type: '`{organism}[MeSH]) AND {celltype}`'
- cell type and disease: '`{organism}[MeSH] AND {celltype} AND {disease}`'; '`{organism}[MeSH] AND {celltype} NOT {disease}`'

For each dataset, we retrieved up to 5000 titles and abstracts per query. Abstracts were split into segments of up to 256 tokens for training.

5.4 Base model

All sentence embedding and fine-tuning tasks of our approach were initialized from PubMedBERT (`neuml/pubmedbert-base-embeddings`, available via Hugging Face) as base model, a domain-specific Bidirectional Encoder Representations from Transformers (BERT) model trained on Pubmed titles and abstracts [11, 37]. The base PubMedBERT model is a 12-layer bidirectional Transformer encoder with 110M parameters, and outputs 768-dimensional embeddings.

The original PubMedBERT model has been trained on a large-scale corpus of parsed and indexed PudMed titles and abstracts (via the `paperetl` pipeline), with training pairs created from combinations of titles and corresponding abstracts or semantically similar titles. Training used the Multiple Negatives Ranking Loss (for details, see below in Section 5.7) with mean pooling for token aggregation, which encourages semantically similar pairs to be closer in embedding space. PubMedBERT outperforms general-purpose embedding models on biomedical semantic similarity tasks (e.g., PubMedQA, PubMed Subset, PubMed Summary), showing the advantage of domain-specific fine-tuning for semantic search and text similarity.

This pretrained encoder served as the backbone for all downstream SBERT fine-tuning experiments in this study.

5.5 Siamese Sentence BERT architecture

To enable efficient similarity computations, we used the pre-trained PubMedBERT model within a Sentence-BERT (SBERT) framework [12], a modification of BERT designed for efficient semantic similarity in sentence-pair tasks, such as aligning cell sentences of similar cells. In contrast to standard BERT, SBERT encodes each sentence independently using a siamese or triplet architecture [38]. A mean pooling layer produces a fixed-size (768-dimensional) token embedding, which enables fast pairwise similarity computation using metrics such as cosine similarity. This architecture reduces the number of model passes required for pairwise comparisons [12], making it well suited for large-scale comparison of cell sentences and literature-derived information.

SBERT supports training with contrastive objectives such as Triplet Loss, Contrastive Loss, and Multiple Negatives Ranking (MNR) Loss, which we employed in our training procedure (see Section 5.7 for details). Because the model is designed to cluster semantically similar sentences in embedding space, it is well suited for tasks like semantic alignment and clustering.

All downstream and fine-tuning analyses in this work are therefore based on a PubMedBERT encoder within an SBERT architecture.

5.6 Generation of training data for fine-tuning using label-aware random-positive and label-aware hard-negative-mining

For each dataset, we fine-tuned a PubMedBERT base model to generate embeddings which capture meaningful semantic interactions of genes, cells, cell types and additional information such as disease or time. In particular, we used with the MNR loss as fine-tuning objective, which is based on sample triplets generated as follows:

5.6.1 Triplet generation for scRNA-seq data.

For each semantic cell sentence ('anchor', see Section 5.2), positive examples were randomly sampled from other cells with the same cell type label (five per anchor). Negatives were selected using label-aware hard negative mining: for each anchor sentence, we identified the five most similar sentences with different cell type labels in the PubMedBERT base embedding space. Each anchor-positiv pair was therefore paired with one unique hard negative, resulting in five anchor-positive-negative triplets per anchor cell. This approach avoids the quadratic expansion of standard hard negative mining (where each anchor-positive pair is assigned k negatives, leading to k^2 triplets) while maintaining diversity. We implemented hard negative mining using an a modified version of the `sentence-transformers` package to account for cell type labels and generate triplets with unique negatives to avoid dataset explosion. The label-aware hard negative mining was run in batches of 10,000 cells due to memory constraints.

5.6.2 Triplet generation for PubMed titles and abstracts.

Titles and abstracts retrieved from PubMed were treated analogously. Each title served as an anchor with five positive randomly sampled from titles with the same label, where labels were derived from the PubMed search query. For a query including only organism and cell type, the label was the cell type, while for a query including disease information of the form '`{organism}[MeSH]) AND {celltype} AND {disease}`', the label was '`celltype + disease`', for a query of the form '`{organism}[MeSH]) AND {celltype} NOT {disease}`', the label was '`celltype`'.

For abstracts, each anchor abstract was randomly assigned five positive abstracts returned by the same search query.

Negatives for both titles and abstracts were obtained using label-aware hard negative mining, as described above for the scRNA-seq.

All triplets from both scRNA-seq derived and literature-derived data were combined for model fine-tuning.

5.7 Multiple Negatives Ranking Loss

We fine-tuned all models using the Multiple Negatives Ranking (MNR) loss [27], a contrastive objective related to the InfoNCE loss. Given a batch of anchor-positive pairs, MNR aims to increase the similarity between each anchor and its positive paired sample, while treating all other positives in the batch as implicit negatives:

$$\mathcal{L}_i = -\log \frac{\exp(S_c(x_i, x_i^+))}{\sum_{j=1}^{N} \exp(S_c(x_i, x_j^+))},$$

where $S_c(\cdot, \cdot)$ is the cosine similarity and all positives x_j^+ for $j \neq i$ are treated as in-batch negatives.

From a probabilistic interpretation, the loss function maximizes the log-likelihood of correctly identifying the true positive, where the softmax over all candidates ensures that the similarities are normalized to a probability distribution, i.e., it resembles a softmax classifier over cosine similarities. It is minimized when the similarity between an anchor and its true positive is much larger than to any of the in-batch negatives.

To increase difficulty and improve generalization, we combined these soft in-batch negatives with label-aware hard negatives obtained as described above (see Section 5.6). Given triplets of anchor-positive-negative pairs, obtained as described above in Section 5.6.1, $(x_i, x_i^+, x_i^-)_{i=1,\ldots N}$, we extend the loss as:

$$\mathcal{L}_i = -\log \frac{\exp(S_c(x_i, x_i^+))}{\exp(\sum_{j=1}^{N} \exp(S_c(x_i, x_j^+)) + \sum_{j=1}^{N} \exp(S_c(x_i, x_i^-))}$$

Each anchor–positive pair was therefore trained against both random in-batch negatives and carefully selected near-miss examples.

This combination encourages the model to learn from subtle semantic differences and disambiguate fine-grained relationships, to improve generalization, to pick up also on weaker patterns in metadata and labels and increase robustness to noise. Further, the exposure to a broader range of negative samples helps reduce overfitting.

5.8 Fine-tuning of models using the sentence-transformer package

We fine-tuned the sentence embedding models using the `sentence-transformers` Python package, which provides efficient training routines and tools for contrastive learning with sentence embeddings. Fine-tuning was performed either on literature-derived text alone or jointly on both literature-derived text and cell-derived cell sentences. For joint optimization, we alternated epoch-wise between the two datasets to avoid gradient interference during training.

Training was performed with batch sizes between 128 and 512, depending on the cell sentence length. A batch size of 128 was used for the PubMed abstract data. We trained for 2 to 5 epochs, with 1000 warm-up steps, a weight decay of 0.01, mixed precision (`fp16=True`) to reduce memory usage, and gradient clipping with `max_grad_norm=1.0` to avoid instability.

5.9 Evaluation of gene interactions captured in embedding space

We evaluate our models on basic singe-cell analysis tasks, to evaluate whether complex gene interactions were captured.

5.9.1 Cosine similarity of cells with cell type labels

We computed the cosine similarity S_c between the embedding C of the singlecells $c_1, \ldots, c_n$ and the embedding L of the cell type labels $l_1, \ldots, l_n$,

$$S_c(C, L) := \frac{\sum_{i=1}^{n} c_i l_i}{\sqrt{\sum_{i=1}^{n} c_i^2} \cdot \sqrt{\sum_{i=1}^{n} l_i^2}}$$

For evaluation we compared similarity distributions for all cells of the respective cell type label versus cells of all other types using the Mann-Whitney U test [39] and report the resulting p-values

per cell type. We assess the specificity and sensitivity of assigning cell type labels to cells over continuous similarity thresholds receiver operating characteristic (ROC) curves, and report the area under the curve (AUC) [40].

For cell type annotation, each cell was assigned the cell type label with the highest cosine similarity of the respective embeddings. We evaluate the annotation performance using a balanced accuracy score over all cell types.

5.9.2 Evaluation of different layers of information

To evaluate whether model embeddings capture functionality- and disease-related context, we employ expert-evaluated statements generated with AI support describing a) the general functionality of the cell type and b) the role of the cell type for cancerous diseases. We computed cosine similarities between the cell or cell type label embeddings and the functionality and disease-related descriptions, and compared the resulting scores to the expert-derived matching of the descriptions to the cell types.

To evaluate how temporal structure is captured when including time as metadata, we examined how cells are organized in the overall embeddings based on a UMAP representation. We additionally quantified alignment by computing cosinesimilarity between cell type label embeddings and the textual representation of embryonic time points (e.g., “E7”, “E9”) and assessed the relationship of early versus later developmental stages.

Declarations

Funding

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB 1597 (SMK, HB, MH).

Conflict of interest / Competing interests

The authors declare that they have no competing interests.

Consent for publication

All authors have read and approved the final manuscript and consent to its publication.

Data availability

The study used only publicly available datasets. Specifically, we analyzed a subset of the Human Immune Health Atlas [2], restricted to T cells, and the developing mouse brain dataset from La Manno, Gyllborg, et al. [34]. Detailed preprocessing steps are provided in the Methods section.

Materials availability

No new materials were generated in this study.

Code availability

The code used in this study is available in a GitHub repository, which will be made publicly available at the time of journal submission.

Author contribution

SMK performed the research, conducted the analyses, and wrote the manuscript. JM contributed to the model implementation. JR and TV provided biological expertise and interpretation, in particular for the lineage trajectory section. HB and MH supervised and guided the work, MH co-wrote the manuscript. All authors reviewed and approved the final version.

References

1. Zappia L and Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biology 2021;22:301.
2. Gong Q, Sharma M, et al. Longitudinal Multi-omic Immune Profiling Reveals Age-Related Immune Cell Dynamics in Healthy Adults. 2024.
3. Miller JA, Guillozet-Bongaarts A, et al. Neuropathological and transcriptomic characteristics of the aged brain. eLife 2017;6. Ed. by Nelson SB. Publisher: eLife Sciences Publications, Ltd:e31126.
4. Program CSCB, Abdulla S, et al. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Pages: 2023.10.30.563174 Section: New Results. 2023.
5. Szałata A, Hrovatin K, Becker S, Tejada-Lapuerta A, Cui H, Wang B, and Theis FJ. Transformers in single-cell omics: a review and new perspectives. Nature Methods 2024;21. Publisher: Nature Publishing Group:1430–43.
6. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, and Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare 2022;3. arXiv:2007.15779 [cs]:1–23.
7. Beltagy I, Lo K, and Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. arXiv:1903.10676 [cs]. 2019.
8. Fang L, Chen Q, Wei CH, Lu Z, and Wang K. Bioformer: an efficient transformer language model for biomedical text mining. arXiv:2302.01588 [cs]. 2023.
9. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, and Wang B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods 2024;21:1470–80.
10. Theodoris CV, Xiao L, et al. Transfer learning enables predictions in network biology. Nature 2023;618:616–24.
11. Devlin J, Chang MW, Lee K, and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. 2019.
12. Reimers N and Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs]. 2019.
13. Zeng Y, Xie J, et al. CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. 2024.
14. Levine D, Lévy S, et al. Cell2Sentence: Teaching Large Language Models the Language of Biology. 2023.
15. Chen Q, Hu Y, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. Nature Communications 2025;16. Publisher: Nature Publishing Group:3280.
16. OpenAI, Achiam J, et al. GPT-4 Technical Report. arXiv:2303.08774 [cs]. 2024.
17. Lu H, Liu W, et al. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525 [cs]. 2024.
18. Rizvi SA, Levine D, et al. Scaling Large Language Models for Next-Generation Single-Cell Analysis. 2025.
19. Chen Y and Zou J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. 2023.
20. Liu T, Chen T, Zheng W, Luo X, and Zhao H. scELMo: Embeddings from Language Models are Good Learners for Single-cell Data Analysis. Pages: 2023.12.07.569910 Section: New Results. 2023.
21. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, and Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36. Ed. by Wren J. Publisher: Oxford University Press (OUP):1234–40.

22. Schaefer M, Peneder P, Malzl D, Hakobyan A, Sharma V, Krausgruber T, Menche J, Tomazou EM, and Bock C. JOINT EMBEDDING OF TRANSCRIPTOMES AND TEXT ENABLES INTERACTIVE SINGLE-CELL RNA-SEQ DATA EXPLORATION VIA NATURAL LANGUAGE. 2024.
23. Shi Y, Yang J, Nai C, Li S, Fang J, Wang X, Liu Z, and Zhang Y. Language-Enhanced Representation Learning for Single-Cell Transcriptomics. arXiv:2503.09427 [cs]. 2025.
24. Liu Y, Wei G, Li C, Shen LC, Gasser RB, Song J, Chen D, and Yu DJ. TripletCell: a deep metric learning framework for accurate annotation of cell types at the single-cell level. Briefings in Bioinformatics 2023;24:bbad132.
25. Ma Z, Lu YY, Wang Y, Lin R, Yang Z, Zhang F, and Wang Y. Metric learning for comparing genomic data with triplet network. Briefings in Bioinformatics 2022;23:bbac345.
26. Sayers EW, Beck J, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Research 2025;53:D20–D29.
27. Henderson M, Al-Rfou R, Strope B, Sung Yh, Lukacs L, Guo R, Kumar S, Miklos B, and Kurzweil R. Efficient Natural Language Response Suggestion for Smart Reply. arXiv:1705.00652 [cs]. 2017.
28. Jørgensen TE and Breitung J. Margins in Contrastive Learning: Evaluating Multi-task Retrieval for Sentence Embeddings. 2025.
29. Heumos L, Schaar AC, et al. Best practices for single-cell analysis across modalities. Nature Reviews. Genetics 2023:1–23.
30. Sun L, Su Y, Jiao A, Wang X, and Zhang B. T cells in health and disease. Signal Transduction and Targeted Therapy 2023;8. Publisher: Springer Science and Business Media LLC.
31. Juraske C, Krissmer SM, Teuber ES, Parigiani MA, Strietz J, Wesch D, Kabelitz D, Minguet S, and Schamel WW. Reprogramming of human γδ T cells by expression of an anti-CD19 TCR fusion construct (εTRuC) to enhance tumor killing. Journal of Leukocyte Biology 2024;115:293–305.
32. Jabeen MF and Hinks TSC. MAIT cells and the microbiome. Frontiers in Immunology 2023;14:1127588.
33. Zhang JY, Wang XM, et al. Single-cell landscape of immunological responses in patients with COVID-19. Nature Immunology 2020;21. Publisher: Nature Publishing Group:1107–18.
34. La Manno G, Gyllborg D, et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell 2016;167:566–580.e19.
35. Dasgupta K and Jeong J. Developmental Biology of the Meninges. Genesis (New York, N.Y. : 2000) 2019;57:e23288.
36. Qiu C, Cao J, et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. Nature Genetics 2022;54. Publisher: Nature Publishing Group:328–41.
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I. Attention Is All You Need. arXiv:1706.03762 [cs]. 2023.
38. Schroff F, Kalenichenko D, and Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1503.03832 [cs]. 2015:815–23.
39. Mann HB and Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 1947;18. Publisher: Institute of Mathematical Statistics:50–60.
40. Swets JA. ROC Analysis Applied to the Evaluation of Medical Imaging Techniques. Investigative Radiology 1979;14:109.