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

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

[1] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Germany

[2] CIBSS – Centre for Integrative Biological Signalling Studies, University of Freiburg, Germany

[3] Institute of Anatomy and Cell Biology, Department Molecular Embryology, Faculty of Medicine, University of Freiburg, Germany

[4] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Germany

Correspondence to maria.krissmer@uniklinik-freiburg.de
maren.hackenberg@uniklinik-freiburg.de

February 26, 2026

Abstract

Pre-trained language models promise to enrich analyses of single-cell data with additional layers of information leveraging large text corpora. Yet, it is still unclear how to achieve optimal alignment with the primary quantitative single-cell data. To address this, we construct text-based training datasets from both scRNA-seq data and biomedical literature targeted to the experimental setting at hand. We then jointly train language models on both information sources to learn a common, literature-enriched representation. Our examples on functionality, disease associations, and temporal trajectories show the potential of knowledge-augmented embeddings as a generalizable and interpretable strategy for enriching single-cell analysis pipelines.

Keywords: *Single-cell RNA sequencing, Encoder-only language models, Text-based embeddings, Biological knowledge integration*

1 Introduction

In single-cell data analysis, an increasing number of foundation models has recently become available [1], i.e., models that have been pre-trained on a huge amount of data. This includes models trained directly on quantitative single-cell profiles (e.g.,[2, 3]), and language models trained on biomedical literature, (e.g.,[4, 5, 6]), which contains rich contextual knowledge about cellular programs and gene function.

While both types of foundation models have been predominantly investigated as primary end-to-end alternatives to classical quantitative single-cell analysis workflows [7, 8, 9, 10, 11], they also offer an opportunity to complement the original data and augment standard analyses. In particular, language models trained on scientific literature could add layers of qualitative biological information to the original single-cell data by embedding qualitative descriptors, such as functional programs, disease associations, or developmental stage information, in a joint lower-dimensional representation. Yet, it remains unclear how to best leverage such textual knowledge to enrich analyses of a specific dataset of interest [11].

A natural approach for using language models with single-cell data is to convert single-cell omics data to "cell sentences" [8, 12], which represent each cell as a ranked list of its most highly expressed genes. We hypothesize that this approach also enables the integration of dataset-specific metadata available from increasingly complex and multi-layered datasets [13, 14, 15] and related contextual information retrieved from scientific literature. Yet, to achieve this, text- and expression-derived information needs to be combined. Here, we explore a contrastive alignment strategy to learn a joint embedding space in which expression profiles and literature-encoded knowledge are directly comparable.

This alignment is achieved through the training of relatively small encoder-only language models [16, 17], which combine a number of potential advantages for the task at hand. They are flexible in terms of task-specific fine-tuning, allow for straightforward integration of all text-based data sources and can be trained to generate semantically meaningful representations that are inherently compatible with well established downstream analysis of single-cell data.

Current research for the application of pre-trained language models for single-cell data focuses mainly on large language models (LLMs) that can be employed for diverse primary analysis tasks such as cell-type annotation, batch integration, and perturbation prediction. For example, C2S-Scale [8] uses LLMs like Pythia [18], which are additionally trained on cell sentences to support both traditional single-cell tasks and novel applications like dataset interpretation via natural language queries. Other approaches leverage general-purpose LLMs (e.g., GPT-3.5) to create cell embeddings based on expression-weighted gene vectors [9, 19], or for cluster-based annotation [7]. These approaches

highlight the flexibility of applying language models for modeling gene expression data, enabling adaptation for diverse tasks and integration with various data types. Yet, it has not been comprehensively shown that such generalist LLMs necessarily outperform smaller, task-specific models [11].

Approaches that combine the strengths of both quantitative and text-based data representation strategies leverage language models for text-based interaction with the quantitative data. For example, joint embedding spaces shared between transcriptomic models such as Geneformer [3] or scGPT [2] and literature-trained models like BioBERT [20] integrate raw expression data with textual knowledge by allowing user queries to be matched directly to cells via semantic similarity [4, 21, 22]. 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 [11].

Therefore, we propose a more targeted approach for taking information from biomedical literature into account, specifically focusing on smaller task-specific language models for adding layers of information to the original single-cell data. In particular, we train small encoder-only language models [16, 17] to align gene expression profiles with biomedical literature, learning a shared semantic embedding space that integrates both data-derived and text-derived information. To build text-based datasets with a consistent structural configuration from both scRNA-seq data- and literature-derived information, we construct cell sentences and collect biomedical literature retrieved from PubMed [23] based on available metadata. Each cell sentence is paired with a similar and a dissimilar sentence to form a triplet [24, 25], and corresponding triplets are constructed from PubMed titles and abstracts. This triplet structure allows for efficient alignment of literature-derived and scRNA-seq-based cell sentences using contrastive loss functions to capture semantic relationships in the embedding [26, 27].

We demonstrate how this unified representation, where single-cell profiles, annotations, and textual descriptions are embedded within a shared coordinate system, provides a context-aware and interpretable representation. By projecting biological concepts expressed in natural language into this representation, we demonstrate how the trained models can transfer contextual knowledge from literature to datasets, and enable flexible annotation, e.g. for exploring cellular states across functional, disease-related, and developmental contexts.

2 Results

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

To integrate single-cell gene expression profiles and complementary biomedical knowledge, we developed a novel approach that embeds single-cell profiles and literature-derived information into a shared semantic space (Fig. 1). This enables direct comparison between individual cells and biomedical concepts expressed in textual form, thus adding layers of contextual information such as function, disease association, or developmental state to the data.

Our approach builds on scRNA-seq datasets that have been processed by a conventional analysis workflow (Fig. 1 left) with annotated cell types, e.g., based on cluster-specific differential gene expression analysis [28, 29]. Starting from the gene expression count matrix(Fig. 1, Step 1), we transformed the top expressed highly-variable genes of each cell into a list of gene symbols and concatenate them to cell sentences. In addition to the gene list, metadata can be added to the cell sentence to allow for incorporating additional, dataset-specific contextual information (e.g., on donor or cell types).

In parallel, we built a literature-derived training dataset by retrieving titles and abstracts of research articles stored in the National Center for Biotechnology Information (NCBI) PubMed database [23] that are relevant to the organism, cell types, or developmental stages represented in the scRNA-seq data. This allows for capturing complementary biological knowledge not explicitly encoded in the gene expression profiles, such as cell-type functionality, disease relationships, or developmental trajectories. We thus obtained a gene expression-derived and a literature-derived dataset in textual format, facilitating direct embedding by a language model into a shared space.

To align embeddings of both datasets based on their semantic relationship, we used a label-aware triplet strategy for fine-tuning a language model. Specifically, we constructed triplets within the expression-derived and the literature-derived datasets, consisting of an anchor, a positive and a negative sample (Fig. 1, Step 1, bottom). Cell-type annotations were used as labels for triplet generation, i.e., positive samples were chosen randomly from all data samples with the same label, while data samples with a different label but high cosine similarity to the anchor were chosen as negatives (Methods Section 5.6). On these triplets, we fine-tuned a small encoder-only language model with a Siamese-BERT architecture [17] (Methods Section 5.5), which is specifically designed to efficiently compute similarity of multiple inputs (Fig. 1, Step 2). Training alternates epoch-wise between the gene expression– and literature-derived datasets (Methods Sections 5.8) and employs a multiple negatives ranking loss function, designed to capture semantic relationships within triplets by pulling positive samples towards the anchor while pushing the negative samples further apart, thus encouraging em-

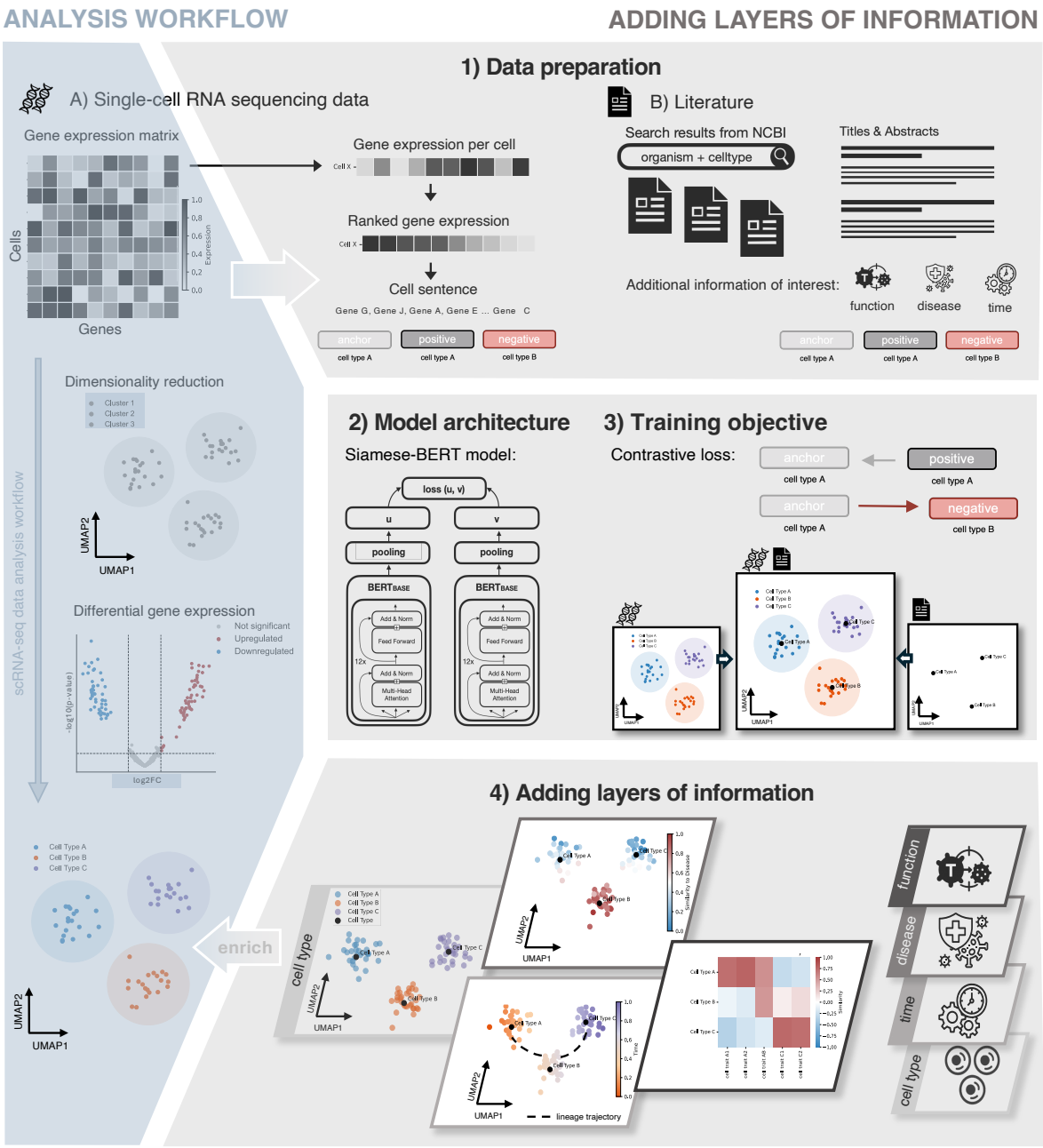


Figure 1: Overview of the approach for adding contextual layers of information to scRNA-seq datasets. Left: A conventional single-cell RNA-seq workflow processes a gene expression matrix through a quantitative pipeline comprising dimensionality reduction and differential expression analysis for annotation of cell-type clusters. **1)** In our approach, we translate single-cell profiles into cell sentences by ranking highly expressed genes and appending metadata for generating the cell-derived dataset (A). In parallel, we build a literature-derived dataset by retrieving PubMed titles and abstracts related to the known or suspected cell types, capturing complementary information such as cell-type functionality (*function*), disease relationships (*disease*) or lineage trajectories (*time*). For both datasets, we construct label-aware triplets. **2)** On the triplets, we train models with a Siamese-BERT [17] architecture, using **3)** contrastive fine-tuning to align the embeddings of single cells, cell-type labels and literature-derived information. **4)** The resulting joint embedding space enables context-aware analysis of scRNA-seq data with additional layers of biological information captured in the biomedical literature.

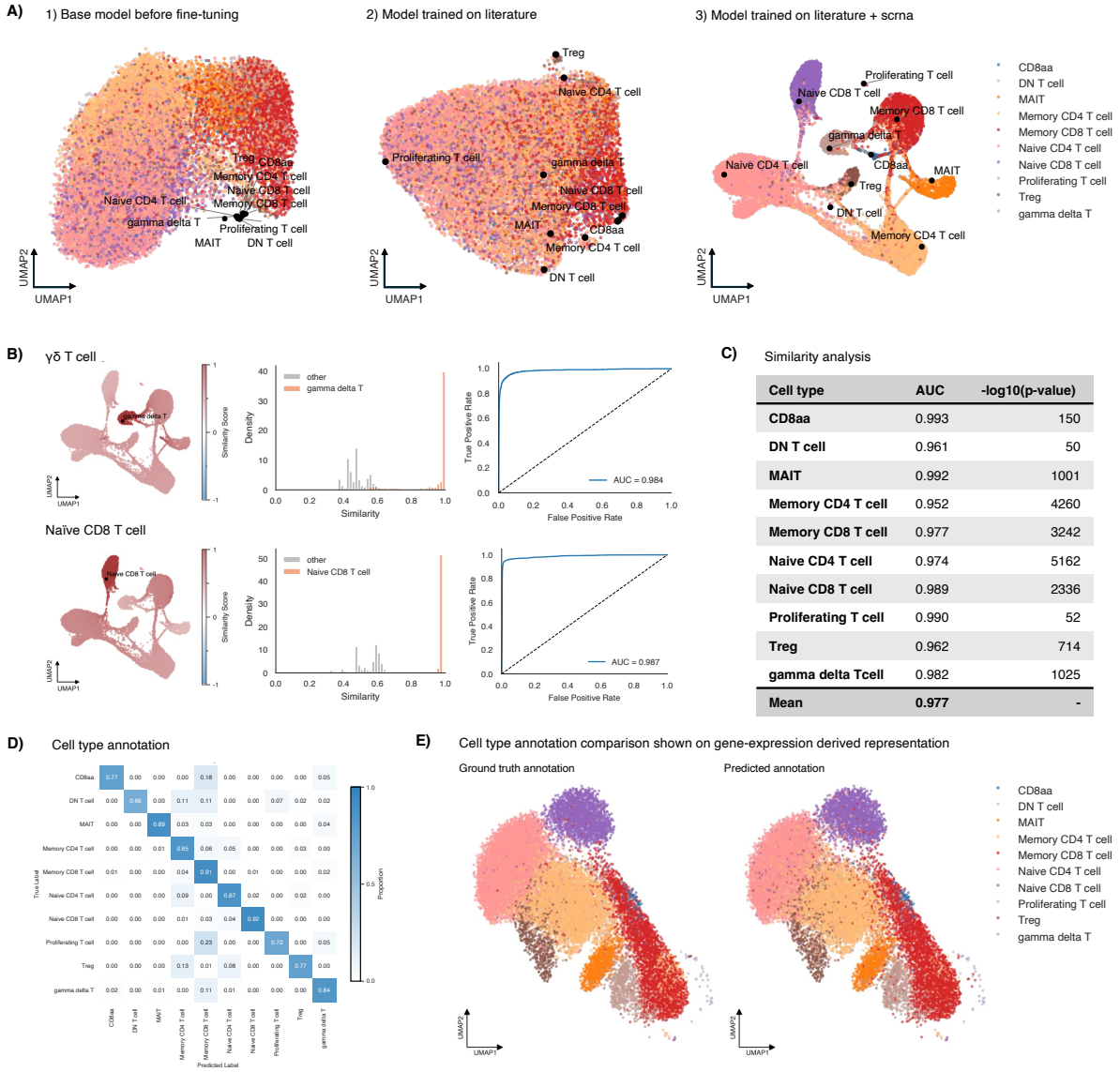
beddings of biologically related entities to cluster together while separating dissimilar ones. (Fig. 1, Step 3; Methods Section 5.7). This results in a joint embedding space where complementary textual information and cell sentences are aligned in a biological meaningful way by design (Fig. 1, Step 3). After fine-tuning, we visualized the learned embedding using UMAP [30].

Taken together, the approach allows for enriching representations of single-cell profiles by adding layers of complementary contextual information, such as cell type, function, disease, or time (Fig. 1, Step 4). After assessing the alignment of cell sentences and cell-type labels to confirm that embeddings reflect known cell identities, we illustrate how the learned embeddings capture richer, literature-informed relationships, enabling annotation based on unseen functional descriptions, discovery of disease associations, and recovery of developmental trajectories when including temporal metadata.

2.2 Designing the joint training on literature and gene expression data to be consistent with cell-type labels

With our label-aware training strategy, we aim to preserve consistency and ensure label alignment across gene-expression- and literature-derived training data. To evaluate the effectiveness of this approach, we investigated the alignment of cell sentences with their cell-type labels on a T cell-specific subset of a primary blood mononuclear cell atlas, specifically the Human Immune Health atlas from the Allen Institute (HIAI dataset) [13]. The embedding of the base language model before fine-tuning showed poor separation of cell subtypes in a UMAP representation, where embeddings of cell-type labels formed a separate cluster distant from the corresponding single-cell embeddings (Fig. 2 **A** 1)). Next, we trained a model on a literature-based dataset only, using the cell-type annotation in the HIAI dataset as queries. We observed improved alignment between cell-type labels and the corresponding single-cell profiles but no clear clustering by subtype (Fig. 2 **A** 2)). Finally, fine-tuning the model on both literature- and expression-based datasets resulted in an embedding with distinct cell clusters by subtype, with their respective cell-type label embedded within those clusters (Fig. 2 **A** 3)). This indicates that fine-tuning was effective for differentiating cells of different types based on their top expressed genes. Additionally, we observed meta-clusters of CD4+ and CD8+ T cells and a tendency of cells with similar functionality to also cluster together in the embedding, such as memory CD8+ T cells and γδ T cells, both characterized through strong cytotoxic capabilities [31, 32]. This is supported by assessment of marker gene expression in the model embeddings (see Additional file 1, Fig. S2), confirming that the model embeddings capture biologically meaningful marker gene expression patterns while preserving within-cell-type heterogeneity.

To quantify alignment between cell embeddings and their corresponding cell-type label, we computed the cosine similarity between each cell embedding and the embeddings of candidate cell-type



Cell type	AUC	-log10(p-value)
CD8aa	0.993	150
DN T cell	0.961	50
MAIT	0.992	1001
Memory CD4 T cell	0.952	4260
Memory CD8 T cell	0.977	3242
Naive CD4 T cell	0.974	5162
Naive CD8 T cell	0.989	2336
Proliferating T cell	0.990	52
Treg	0.962	714
gamma delta Tcell	0.982	1025
Mean	**0.977**	**-**

Figure 2: Joint training on cell-derived and literature-derived datasets enables encoder-only language models to capture biologically meaningful structure in joint text-based embeddings. A) UMAP representation of the *HIAI* dataset based on 1) base model embeddings, 2) model trained solely on literature-derived dataset, 3) model trained on cell- and literature-derived dataset. Cells are colored by ground truth label from the *HIAI* dataset while the respective cell type labels are shown in black. **B)** Cosine similarity between cell embeddings and cell-type label embeddings is used to assess alignment. For each cell type, distributions of similarity scores are compared between cells of that type and all other cells. Receiver operating characteristic (ROC) curves are used to evaluate how well similarity scores separate cells of a given type from others using threshold decision tests. **C)** Overview of similarity analysis metrics showing the area under the ROC curve (AUC) and Mann-Whitney U test p-values for all cell types. **D)** Cosine similarity of cells to labels can be used for cell-type annotation: assigning each cell the cell-type label with the highest similarity yields a confusion matrix comparing predicted and true cell types. **E)** UMAP representations colored by the predicted vs true cell-type annotations illustrate the alignment between the learned embedding structure and the ground truth.

labels. We compared similarity scores using threshold decision tests for the separation of cells of the respective cell type from all other cells based on the similarity score and evaluated them with receiver operating characteristic (ROC) curves. We observed that cells consistently showed highest similarity to their correct label, as illustrated for γδ and naïve CD8+ T cells in Fig. 2 **B** (metrics reported for all other cell types in Fig. 2 **C**).

Finally, we assessed cell-type annotation of single cells on held-out test data by assigning to each cell the cell-type label with the highest cosine similarity (Fig. 2 **D, E**). Averaged over all cell types, we obtained an accuracy of 82.0 %, demonstrating the model's ability to capture meaningful relationships between cell expression profiles and textual labels. We observed lower prediction accuracy within functionally similar cell types, such as γδ T cells and memory CD8 T cells, or regulatory T cells with other CD4+ T cell subtypes [31], reflecting biologically plausible overlap between closely related functional states.

2.3 Capturing cell-type functionality programs unsupervised and at single-cell level

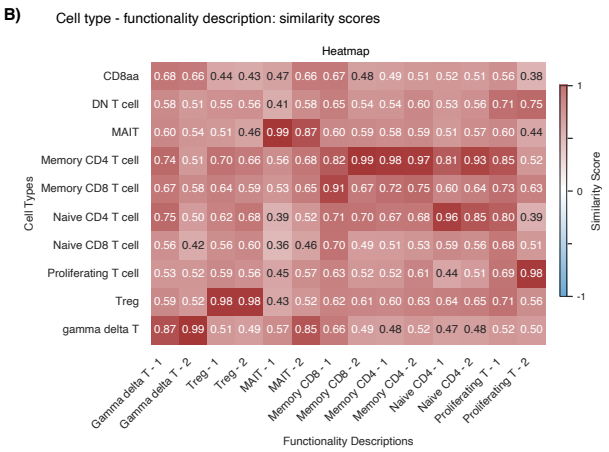
Training our approach jointly on literature- and cell-derived datasets enables analyses with respect to broader biological knowledge captured in the scientific literature. On the HIAI T cell dataset, we investigated whether the model captures cell-type functionalities reflected in literature-derived descriptions.

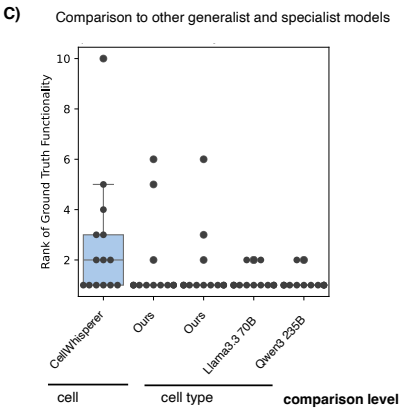
We used expert-curated functionality descriptions (Fig. 3 **A**) for each cell type and computed cosine similarity between these descriptions and the different cell-type labels. We refer to this as cell type-level comparison. The similarity scores shown in Fig. 3 **B** indicate that our approach accurately captured different cell-type functionalities. We also observed higher similarity between functionality descriptions that share biological context. For example, MAIT cells can be described to 'bridge innate and adaptive immunity with semi-invariant TCR and rapid cytokine output.' [33], a function also characteristic of the Vδ2 subset of γδ T cells [31]. This is reflected by higher similarity scores between functionality descriptions of MAITs for γδ T cells compared to all other unrelated T cell subtypes, indicating the model's ability to capture shared functional programs beyond cell-type labels.

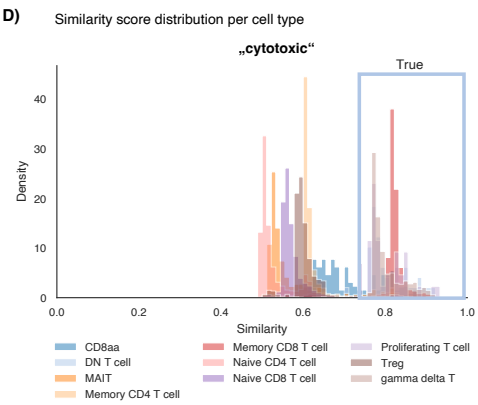
The alignment of cell-type labels and single cells in the embedding space also allows for transferring these functionality descriptions to individual cells: We used cosine similarity between embeddings of individual cells and functionality descriptions to assess how well the similarity scores can separate cells of a given type from all others. We refer to this as cell-level comparison. For example, similarity scores matching the description of γδ T cells separated this subset from all other cells with a ROC-AUC score of 0.984, while MAIT cells with a ROC-AUC of 0.522 could not be distinguished. Ranking the ROC-AUC scores we can assess how the functionality descriptions are aligned with cells of the

A) Functionality descriptions

Cell type	Label	Functionality description
Gamma delta T	**1**	Recognize stress ligands independently of classical MHC molecules.
	2	Recognize lipids, 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.







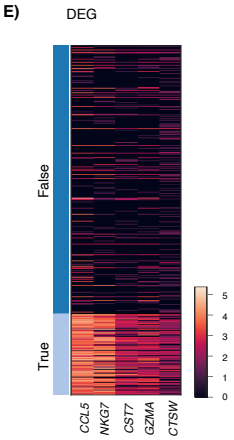


Figure 3: Models trained on cell- and literature-based datasets capture functionality descriptions at both the cell-type and single-cell level. A) Summary table of expert-curated cell-type functionality descriptions. **B)** Heatmap of cosine similarities between model-derived embeddings of functionality descriptions and corresponding cell-type label embeddings. **C)** Comparison of our encoder-only models with existing generalist and specialist models on the single-cell and cell-type label-based level, using rank-based association of functionality descriptions with expert-assigned cell-type labels. **D)** Example application: similarity scores to the functionality annotation 'cytotoxic' identify cytotoxic cells. **E)** Differentially expressed genes (DEGs) of cells annotated as cytotoxic ('True') to non-cytotoxic cells ('False'), confirming enrichment of known cytotoxicity markers.

corresponding cell type, confirming that the model successfully transfers literature-encoded functional knowledge to single-cell embeddings.

To compare our approach to existing single-cell foundation models and general-purpose LLMs, we utilized both comparison strategies (cell-based and cell-type-based). For the cell-level comparison, we compared our approach to CellWhisperer [21], an omics foundation model approach that allows for computing the similarity of cell embeddings to text. For the cell type-level comparison, we compare our approach to general-purpose LLMs, including Llama3.3 70B [34] and Qwen3 235B [35], where we used a pairwise decision strategy with a rank-based evaluation (described in Methods Section 5.10.3) to obtain similarity relations between functionality descriptions and cell types. Both comparison strategies yield association scores (similarity-based for the cell-type-level comparison and ROC-AUC-based for the cell-level comparison). These scores were employed to compare the different approaches based on the assigned ranks of the ground-truth cell-type label associated with each functionality description shown in Fig. 3 **C**. We found that our small encoder-only models achieved comparable performance to LLMs in the cell type-level comparison, and outperformed CellWhisperer in the cell-level comparison strategy.

Next, we employed the contextualized embeddings of the scRNA-seq data to identify overarching functional programs that span multiple cell types, such as cytotoxicity. We calculated similarity scores between each cell embedding and the textual description 'cytotoxic' and classify cells with a similarity > 0.75 as cytotoxic cells (Fig. 3 **D**). Differential gene expression analysis comparing these cells with all others (Fig. 3 **E**) revealed that the top five differentiable expressed genes are well-established cytotoxicity markers [36]. This illustrates that our approach supports functional annotation directly from text, thus enriching single-cell datasets with interpretable biological context.

2.4 Incorporating metadata and literature knowledge reveals disease-associated functional shifts

To assess whether our approach can effectively incorporate both literature-derived knowledge and metadata, we included individuals' cytomegalovirus (CMV) status into the cell sentences and aligned them with CMV-related literature. Fig. 4 **A** shows the CMV status across all cells in the UMAP representation of the original scRNA-seq dataset. We computed similarity scores between each cell embedding and the interaction description 'Increased cytotoxic capabilities in CMV positive individuals' and employed a Mann-Whitney-U test to evaluate whether CMV-positive and CMV-negative cells could be separated within each cell type based on the cell-wise similarity scores (see Methods Section 5.10.5). As shown in Fig. 4 **B**, CMV-positive cells within memory CD4+ T, memory CD8+ T and γδ T cells have significantly higher similarity scores compared to CMV-negative cells. Notably, CD4+

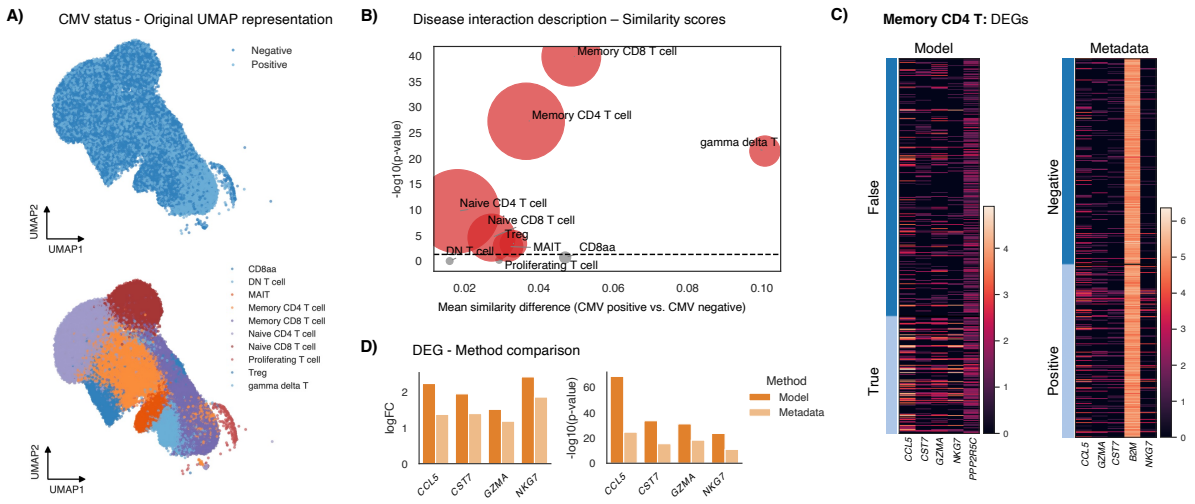


Figure 4: Integrating literature-derived knowledge and metadata enables detection of disease-associated functional alterations in cytomegalovirus (CMV)–positive individuals. A) UMAP representation of the original gene expression data annotated with CMV status metadata and cell-type labels. **B)** Separation of CMV status based on cell-wise cosine similarity to a disease interaction description, where cells are classified as CMV-associated ("True") when exceeding the similarity threshold (see Methods Section 5.10.5). Separation is evaluated for cells of separate cell types using the mean similarity difference and p-value from Mann-Whitney U test. Dot size (red) indicates the number of cells per type. **C)** Differential gene expression (DEG) analysis of disease associated cells 'Memory CD4 T' cells separated either by similarity scores based on the model embeddings or by the metadata labels. **D)** Comparison of log fold changes (logFC) and corresponding log-transformed p-values between the two DEG analyses from the two methods of cell separation.

T cells are typically non-cytotoxic in healthy individuals but acquire cytotoxic properties upon CMV infection [31], reflecting a known functional shift. To confirm this, we performed differential gene expression analysis for the groups identified as CMV-associated based on either the available metadata annotation, or the separation of similarity scores in the learned model embeddings, where we used Youden's J index to determine an optimal threshold (Fig. 4 **C**). For both group separations, we recovered four jointly up-regulated genes within the top five differentially expressed genes. However, comparison of log-fold changes and corresponding p-values (Fig. 4 **D**) indicated that the model-based separation was able to identify the phenotypically changed cells more clearly, highlighting that incorporating literature-derived knowledge and corresponding metadata allows for identifying disease-related functional changes.

2.5 Embeddings capture temporal transitions aligning with developmental lineage trajectories

We next evaluated the capabilities of our approach to capture temporal metadata information and developmental dynamics. For this, we used an embryonic mouse brain dataset [37], containing a diverse range of cell types sampled at multiple time points of mouse embryonic development from day 7 to day 18. Temporal metadata (embryonic day) was included into the generated semantic cell sentences as described in Section 5.2. To facilitate visualization, we divided the dataset into

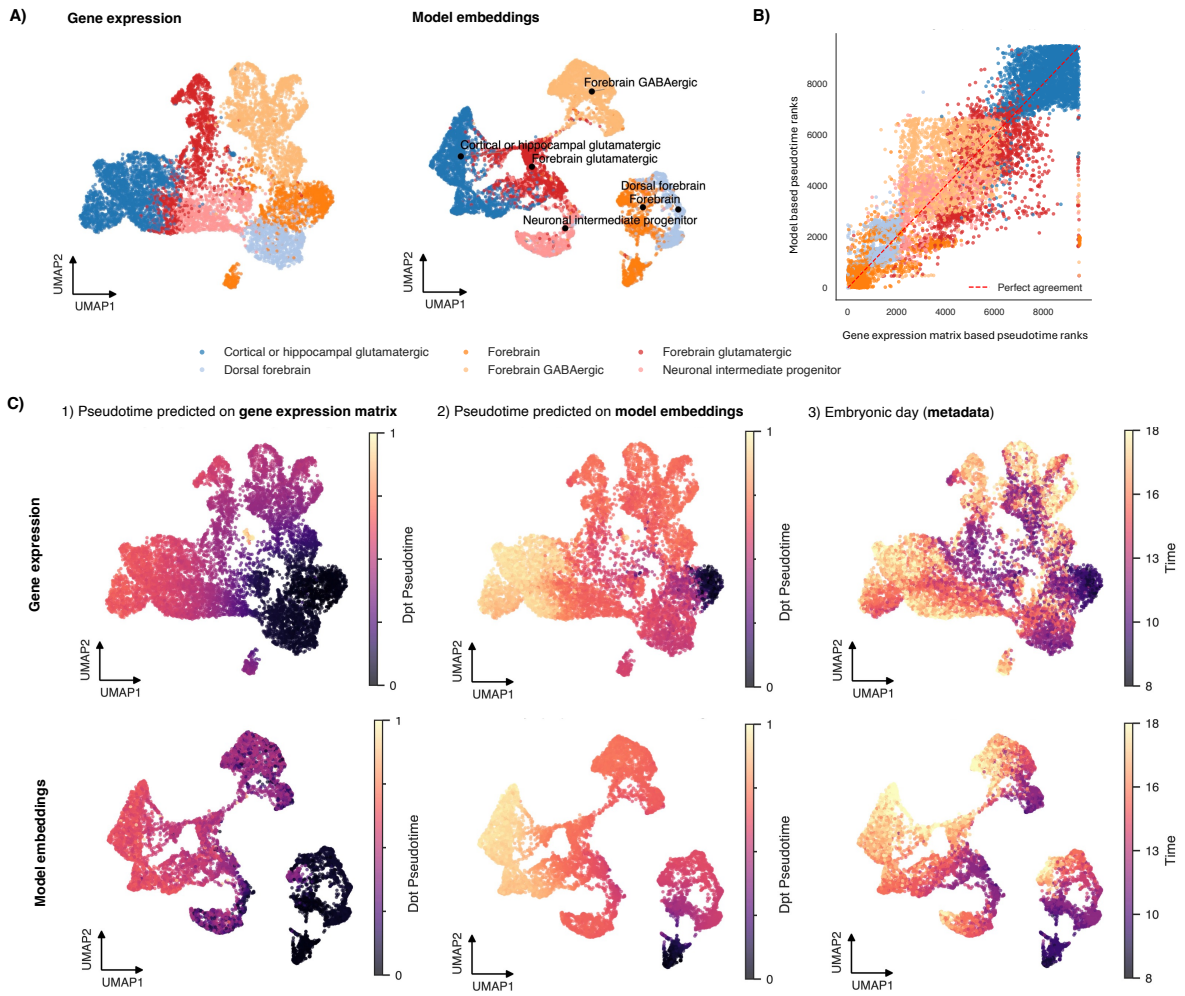


Figure 5: Model trained on datasets incorporating temporal metadata captures developmental patterns. A) Lineage trajectory from 'Dorsal forebrain' to 'Cortical or hippocampal glutamatergic' and 'Forebrain GABAergic' colored by ground truth cell-type label. UMAP plots use model embeddings and gene expression matrix respectively. **B)** Comparison of ranked pseudotime results on model embeddings and gene expression data colored by ground truth cell type **C)** UMAP representations of the original gene expression matrix (top row) and the model embeddings (bottom row) colored by pseudotime values calculated based on 1) model embeddings, 2) gene expression data, and by 3) embryonic day (metadata).

seven subsets. Here, we compared the alignment between the original authors' cell-type annotations (color code) and the model's cell-type predictions (cell label), along with the embryonic time shown in Additional file 1, Fig. S3 **A)**. The fine-tuned model aligned the embeddings of cell-type labels and their corresponding cells with a mean ROC-AUC of 0.979.

Beyond alignment of cell-type labels and corresponding cell profiles, the embeddings revealed temporal trajectories in meta-clusters spanning multiple annotated cell types. Specifically, cells with early embryonic development differentiation states clustered separately from cells of more mature differentiation states (Additional file 1, Fig. S3 **B)**). Together this suggests that with incorporation of temporal information into the model, the embedding resolves both cell-type modality and temporal patterns.

To investigate whether the temporal organization reflected known biological trajectories, we analyzed three exemplary lineage progressions, shown in Additional file 1, Fig. S4. Specifically, we considered the trajectories from epiblast to intermediate meninges cells (Additional file 1, Fig. S4 **A**), the trajectory from hindbrain roofplate cells via hindbrain cells to glutamatergic cells of the cerebellum (Additional file 1, Fig. S4 **C**), and the trajectory from angioblast to endothelial cells and erythroid progenitors to erythrocytes (Additional file 1, Fig. S4 **E**). In all examples, we observed a continuous transition in the model embeddings from the early cells in the respective lineage to the more mature cells, consistent with the respective established developmental lineage. We show that embeddings of cells from earlier time points show higher similarity scores to the embeddings of temporal metadata strings of early time points, with analogous patterns observed for intermediate and later time points, showing that the textual temporal meta-data representations are meaningfully aligned with the single-cell embeddings (Additional file 1, Fig. S4 **B, D, F**).

Next, we assessed the trajectory from cells of the 'Dorsal forebrain' to mature 'Cortical and hippocampal glutamatergic' neurons (Fig. 5 **A**). We computed diffusion pseudotime on both the model embeddings and the original gene expression matrix (Methods Section 5.10). The ranked pseudotime values showed a strong concordance between the expression-based and model-based representations (Kendall's $\tau = 0.711, -\log_{10} p = 4707$; Fig. 5 **B**). Both are consistent with the chronological order of the embryonic days upon comparison of pseudotime values calculated on the original expression matrix (1), the model embeddings (2) and the embryonic day metadata annotation (3) in UMAP representations of the model embeddings and the original gene expression matrix (Fig. 5 **C**). Interestingly, the model embeddings separate a subset of *Crabp2*+, *Fezf1*+ and *Hes5*- early forebrain neural progenitors previously annotated as 'Forebrain' [38, 39] (gene expression of the mentioned markers in Additional file 1, Fig. S5). This separation is consistent with early time points in the metadata annotation as well as early model-based pseudotime predictions, while the expression-based pseudotime did not separate these early neuronal progenitors. The results highlight that pseudotime inferred from the model embeddings provided a temporal resolution more aligned with the included metadata. This highlights the potential of incorporating temporal metadata and biomedical literature for capturing continuous developmental transitions and hierarchical lineage relationships.

3 Discussion

We have investigated how pre-trained language models can be successfully adapted to enrich classical analysis workflows of single-cell data, more generally exploring language models as a complementary analysis tool. Specifically, we show how an enrichment of single-cell data with biomedical literature is possible using contrastive learning for aligning data sources from cell- and literature-derived

data. Our approach builds on small, encoder-only language models for generating a joint text-based embedding that aligns diverse layers of biological information. We have shown that this embedding provides a semantically meaningful, unified representation of information from both data sources that captures not only cell identities but also functional programs, disease associations, and developmental trajectories.

In particular, we have applied our approach to single-cell RNA-seq datasets including human immune and developing mouse brain data to evaluate its ability to integrate literature-derived context across biological systems. The results demonstrate that the framework can meaningfully connect gene expression profiles of individual cells with textual descriptors of their biological roles and function, producing a semantically coherent space that links data-driven and literature-based representations.

In a T cell dataset, the model accurately aligned cell-type labels with their corresponding gene expression profiles and captured shared functional programs, such as cytotoxicity, consistent with known immune phenotypes. When incorporating metadata describing disease status, the embeddings reflected phenotypic alterations in CMV-positive individuals, recovering known cytotoxic shifts in CD4+ T cells. In developing brain data, the inclusion of temporal metadata enabled the model to reconstruct developmental progressions consistent with established lineage trajectories, demonstrating its capacity to encode continuous biological transitions for studying dynamic processes such as differentiation.

Integrating gene expression data and literature through a shared text modality addresses a limitation in current single-cell foundation models. Previous approaches that use LLMs to embed gene identities or metadata [2, 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 with limited hardware and are adaptable to new datasets with known cell types. Although our framework is based on compact encoder-only models, the results illustrate more broadly that adding contextual annotations represents a promising application for language models, including also large-scale language models. Our results indicate that existing generalist LLMs already capture aspects of biological metadata, such as relationships between cell types and their functionality, suggesting that further exposure to gene expression data could be a promising strategy to improve their applicability to genomic data analysis.

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 requires pre-annotated cell types, which limits its applicability in fully unsupervised settings. Employing cell sentences for converting scRNA-seq data into a text-based format for use in a language model-based framework is a pragmatic choice to allow for flexible integration with other text-based information such as metadata and literature. Yet, this approach reduces the information depth stored in the quantitative single-cell data, and future research could focus on quantifying and reducing this information loss. We have focused on training specialist models, designed for targeted integration of domain-specific knowledge into expression data, where prior labels or context are available. Yet, recent large-scale approaches like CellWhisperer [21] and C2S-Scale [8] show the potential of generalist models trained on large datasets. Training larger base models on broader training datasets and literature corpora with our proposed training strategy 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 through LLMs will be an essential step towards transitioning from task-specific, dataset-bound models to 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 tools for enriching single-cell RNA-seq data with information from biomedical literature. Showcasing how expression profiles and literature-derived knowledge can be aligned in a joint embedding space, our approach enables interpretable analyses across multiple biological layers such as cell identity, function, disease association, and developmental stage across biomedical contexts. This strategy offers a scalable and efficient framework for enriching single-cell analyses employing language models as complementary analysis tool, in contrast to other approaches that place language models at the core of single-cell data analysis. The work thus contributes to ongoing efforts to tailor LLMs 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 around 230,000 T cells [13]. Cell-type annotations are provided at three levels of resolution; we used the intermediate annotation, resulting in 10 distinct T cell subtypes. We preprocessed the data using `scanpy`, according to the workflows described in Heumos, Schaar, et al. [28]. 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 was split into training (50%) and test (50%) sets resulting in around 115,000 cells in both training and test dataset, with the training set further divided into a training (90%) and validation (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 containing around 270,000 cells at 20 embryonic time points E7-E18 La Manno, Gyllborg, et al. [37]. We only included cell types with more than 3 cells in the dataset, resulting in 130 of the 133 original cell types being included in the dataset. We removed cells with less than 100 expressed genes and genes expressed in less than five cells. We performed normalization and log transformation and selected the top 2000 highly variable genes ('seurat_v3'), excluding ribosomal and mitochondrial genes. We split the data into a training and test (10%) set, with the training set subdivided into a training (90%) and validation (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 [12]. Specifically, for each cell we ranked highly variable genes (excluding housekeeping genes) by expression and generated gene lists containing the top 10-200 highest expressed genes.

We assess the variance captured and the set of genes included in cell sentences of varying lengths by comparing the number of principal components (PCs) needed in a principal component analysis (PCA) [40] for explaining 90 % of the variation of the original data. Cell sentences of varying lengths

are transformed into a one-hot encoded matrix, that essentially stores the information of whether each gene was present in the corresponding cell sentence (this transformation discards the rank-based order of the genes, thus resulting in some underestimation of the variation). This one-hot encoded matrix is then used for PCA analogously to the reference gene expression matrix reduced to 2000 highly variable genes. We additionally compare the size of the gene set included in the cell sentences of varying length. The results in Additional file 1, Fig. S1 show that a length of 50 genes includes sufficient variance in a specialist model setting, while obtaining computational efficiency.

In all analyses shown in the manuscript, we therefore used ranked gene lists truncated at a length of 50.

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 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 of gene lists only with sentences including also metadata. For example, in experiments combining gene lists and cell-type information, we used sentences containing only the gene list for 30% of cells, and sentences containing both the gene list and the cell-type label for 70% 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 was converted to gene lists only, without adding metadata.

5.3 Collecting data for training on literature

For enriching scRNA-seq datasets with additional information from scientific literature, we fine-tuned models on PubMed titles and abstracts relevant to the cell types within the datasets. Searches were performed separately for the HIAI and the LaManno dataset using the the NCBI PubMed database, accessed via the Entrez Python package. Specifically, we included as search terms (1) the organism as MeSH search 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 10000 titles and abstracts per query, which allows for pulling the majority of all available publications related to the search term. 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 [16, 41]. 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).

This pre-trained encoder served as the backbone for all model 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 [17], 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 [42]. 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 [17], 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 functions, contrastive loss functions, and multiple negatives ranking (MNR) loss functions, 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 the multiple negatives ranking (MNR) loss as fine-tuning objective, which is based on sample triplets generated as follows:

5.6.1 Triplet generation for scRNA-seq data.

We use each semantic cell sentence $x \in \{x_1, \ldots, x_N\}$ as anchor (see Section 5.2) and sample positive and negative cell sentences to construct $J = 5$ triplets per anchor cell. Specifically, for an anchor cell sentence x, we sampled J positive examples $x_j^+, j = 1, \ldots, J$ from other cells sharing the same cell-type label. Negatives were selected using label-aware hard negative mining: for an anchor x, we identified the J most similar sentences with different cell-type labels, where similarity was measured using cosine similarity in the PubMedBERT base embedding space (see Section 5.7). Each anchor-positive pair (x, x_j^+) for a $j \in \{1, \ldots, J\}$ was paired with one unique hard negative x_j^- for a $j \in \{1, \ldots, J\}$, resulting in J anchor-positive-negative triplets per anchor cell. Formally, the resulting training set is given by

$$\left\{ \left(x, x_j^+, x_j^-\right)_{j=1,\ldots,J} \mid x \in \{x_1, \ldots, x_N\} \right\}.$$

This approach avoids the quadratic expansion of standard hard negative mining, where each anchor-positive pair is assigned k negatives, resulting in k^2 triplets, while maintaining diversity. We implemented hard negative mining using a modified version of the corresponding function implemented within 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.

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 function [26], a contrastive objective related to the InfoNCE loss. Given a batch of size $M < JN$ of anchor-positive pairs, MNR aims to increase a similarity measure S_c between an anchor x and its positive paired sample $x_j^+, j \in \{1, \ldots, J\}$.The similarity measure is chosen as cosine similarity, defined by

$$S_c(x, x_j^+) = \frac{\langle x \cdot x_j^+ \rangle}{\|x\|_2 \, \|x_j^+\|_2} \in [-1, 1], \quad j \in \{1, \ldots, J\}$$

The MNR loss function treats all other positives in the batch M as implicit negatives: For a given anchor-positive pair (x, x_j^+), where $j \in \{1, \ldots, J\}$, the MNR loss is given by

$$\mathcal{L}_j = -\log \frac{\exp(S_c(x, x_j^+))}{\sum_{m=1; x_m^+ \neq x_j^+}^{M} \exp(S_c(x, x_m^+))},$$

and all positives x_m^+ for $m \neq j$ 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 improve generalization, we combined these soft in-batch negatives with label-aware hard negatives obtained as described above (see Section 5.6). Given a triplet of anchor-positive-negative pairs $(x, x_j^+, x_j^-)_{j=1,\ldots J}$, obtained as described above in Section 5.6.1, we extend the loss for one anchor-positive pair (x, x_j^+) for $j \in \{1, \ldots, J\}$ to:

$$\mathcal{L}_j = -\log \frac{\exp(S_c(x, x_j^+))}{\sum_{m=1; x_m^+ \neq x_j^+}^{M} \exp(S_c(x, x_m^+)) + \sum_{j=1}^{J} \exp(S_c(x, x_j^-)}$$

Each anchor–positive pair was therefore trained against both random in-batch negatives and care-

fully 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 size 128 for both literature- and scRNA-seq derived data. We trained for 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 Cosine similarity of cells with cell-type labels

We computed the cosine similarity S_c between cell embeddings $c_1, \ldots, c_N$ and the embeddings of all cell-type labels $l_1, \ldots, l_L$,

$$S_c(c_i, l_j) = \frac{\langle c_i, l_j \rangle}{\|c_i\|_2 \, \|l_j\|_2}, \quad \text{for all } i = 1, \ldots, N; j = 1, \ldots, L.$$

For assessing the cosine similarity scores of all cells with their respective cell-type label, we obtained the distributions of similarity scores between a cell-type label embedding and all cells assigned to that type, and compared it to the distribution of similarity scores for cells of all other types.

We used the Mann-Whitney U test [43] to evaluate differences in these similarity score distributions between the two groups, and reported the resulting p-values per cell type.

To quantify how well similarity scores distinguish cells of a given type from all others, we calculated the specificity and sensitivity of assigning cell-type labels to cells over continuous similarity thresholds and generated receiver operating characteristic (ROC) curves. We reported the area under the ROC curve (AUC) [44], providing a continuous measure of how accurately similarity-based matching identifies cells of a given cell type.

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.

5.10 Evaluation of contextualization of scRNA-seq data with different layers of information

5.10.1 Contextualization with functionality descriptions

To evaluate whether model embeddings capture functionality- and disease-related context, we employed expert-curated statements describing the general functionality of the cell type. We computed cosine similarities between single cell embeddings 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. We refer to this cell-type label-based assessment as cell type-based comparison level.

Additionally, we computed similarity scores between the embeddings of all cells' ranked gene lists and each functionality description. For a given cell type, we compared the distributions of the similarity scores for cells of the respective cell type with the scores of all other cells. By varying the decision threshold over the similarity scores, we generated ROC curves that quantify how well cells of the target type can be discriminated from the rest of the dataset based on their similarity scores to the given functionality description. The resulting area under the ROC curve thus provides a measure of the association between a functionality description and each cell type, reflecting how distinctively the description matches the cells' embeddings. Ranking the AUC values across cell types for a given description identifies the cell type most strongly associated to that functionality.

We refer to this strategy as the cell-based comparison level.

5.10.2 Cell based comparison level using CellWhisperer

We assessed our cell-based comparison level using CellWhisperer utilizing the internally computed similarity scores of single cells and all functionality descriptions. To generate the embeddings of the single cells of the HIAI data and the query texts, the checkpoints of the published CellWhisperer model (https://medical-epigenomics.org/papers/schaefer2025cellwhisperer/data/models/cellwhisperer_clip_v1.ckpt) were loaded and the raw counts were embedded through their provided helper functions (adata_to_embeds() and score_transcriptomes_vs_texts()). For scoring, the logit scale of the loaded model and a softmax normalization was used. All analyses where conducted with the mmcontext framework [45].

5.10.3 Cell-type-based comparison using generative language models

For comparing our approach to general-purpose generative language models, we used the cell-type-based comparison strategy. To obtain ranked associations of cell types to functionality descriptions from generative LLMs, we employed a pair-wise comparison strategy, where LLMs were tasked with systematically matching cell-type labels to the functionality descriptions. Specifically, we used the following prompt for querying Llama3.3 70B [34] and Qwen3 235B [35] via the OpenRouter interface [46]:

"*You are a biomedical expert specializing in T cell immunology.*
Functional description: {description}
Option A: {celltype_a}
Option B: {celltype_b}
Question: Which of these two T cell types most likely performs the described function? Please respond ONLY with one of the following exactly:
A
B
Do not add any other text."

For each functionality description, all possible ordered pairs of cell-type labels were presented to the model, i.e., each pair was presented in both possible orders to eliminate potential position bias in the model's responses. Each model choice was recorded as a "win" for the selected cell type. Across all pairs, the total number of wins per cell type was normalized by the maximum attainable number of wins and converted into a rank score. These ranks represent the LLM-derived association between functionality descriptions and cell-type labels.

For each functionality description, we recorded the rank position assigned by the model to the expert-assigned "true" cell type for that description (14 in total). Lower ranks indicate better alignment (1 = perfect). We then visualized the distribution of these ranks across descriptions to evaluate performance and compared these rank positions to those obtained with our encoder-based approach.

5.10.4 Assessing cell-type-overarching functional traits

Computed similarity scores as described in Section 5.9 to cell-type-overarching functionalities were used for separation of cells associated to the functionality through manually set cut-offs. Differential gene expression testing as implemented in `scanpy` was then used for confirmation of meaningful group separation.

5.10.5 Evaluation of incorporating metadata with scientific literature

We employed the cell-based comparison level strategy described in Section 5.10 for assessing the similarity of single cells to a disease interaction description. Using Youden's J [47], we determined the optimal similarity score threshold for separating cells from CMV-positive and CMV-negative individuals based on the similarity of their embeddings to the disease interaction description. We used this threshold for splitting our cells into disease-associated and non-disease-associated cells. We then performed differential gene expression testing on associated vs not associated cells using `scanpy`.

5.10.6 Pseudotime anaylsis on model generated embeddings and gene expression matrix

To evaluate how temporal structure is captured when including time as metadata, we ran a pseudotime analysis on the model embeddings and the gene expression matrix. Specifically, we applied the diffusion pseudotime approach [48] on neighborhood graphs constructed on both representations. We compare the computed pseudotime values by assessing the concordance between cell ranks using Kendall's τ [49].

Declarations

Funding

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 322977937 – GRK 2344 (JR, TV) and Project-ID 499552394 – SFB 1597 (SMK, HB, MH) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy CIBSS – EXC-2189 – Project ID 390939984 (SMK, JM, HB)

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 [13], restricted to T cells, and the developing mouse brain dataset from La Manno, Gyllborg, et al. [37]. 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 under https://github.com/MariaKrissmer/alias.

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.

Author details

[1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany: Sonia Maria Krißmer, Jonatan Menger, Maren Hackenberg

[2] CIBSS, Centre for Integrative Biological Signalling Studies – University of Freiburg, Germany: Sonia Maria Krißmer, Jonatan Menger, Harald Binder

[3] Institute of Anatomy and Cell Biology, Department Molecular Embryology, Faculty of Medicine, University of Freiburg, Germany: Johan Rollin, Tanja Vogel

[4] Freiburg Center for Data Analysis and Modelling – University of Freiburg, Germany: Harald Binder

References

1. 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 2024:1430–43. DOI: 10.1038/s41592-024-02353-z.
2. 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 2024:1470–80. DOI: 10.1038/s41592-024-02201-0.
3. Theodoris CV, Xiao L, et al. Transfer learning enables predictions in network biology. Nature 2023:616–24. DOI: 10.1038/s41586-023-06139-9.
4. 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 2022:1–23. DOI: 10.1145/3458754.
5. Beltagy I, Lo K, and Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. arXiv 2019:DOI: 10.48550/ARXIV.1903.10676.
6. Fang L, Chen Q, Wei CH, Lu Z, and Wang K. Bioformer: an efficient transformer language model for biomedical text mining. arXiv 2023:DOI: 10.48550/ARXIV.2302.01588.
7. Wu Y and Tang F. scExtract: leveraging large language models for fully automated single-cell RNA-seq data annotation and prior-informed multi-dataset integration. Genome Biology 2025:174. DOI: 10.1186/s13059-025-03639-x.
8. Rizvi SA, Levine D, et al. Scaling Large Language Models for Next-Generation Single-Cell Analysis. bioRxiv 2025:DOI: 10.1101/2025.04.14.648850.
9. Liu T, Chen T, Zheng W, Luo X, Chen Y, and Zhao H. scELMo: Embeddings from Language Models are Good Learners for Single-cell Data Analysis. bioRxiv 2023:DOI: 10.1101/2023.12.07.569910.
10. Zhao S, Zhang J, Wu Y, Luo Y, and Nie Z. LangCell: Language-Cell Pre-training for Cell Identity Understanding. arXiv 2024:DOI: 10.48550/ARXIV.2405.06708.
11. Xie F, Zhao B, Xu S, Wang Z, Moon JJ, and Garmire LX. Overcoming barriers to the wide adoption of single-cell large language models in biomedical research. Nature Biotechnology 2025:1758–62. DOI: 10.1038/s41587-025-02846-y.
12. Levine D, Rizvi SA, et al. Cell2Sentence: Teaching Large Language Models the Language of Biology. bioRxiv 2023:DOI: 10.1101/2023.09.11.557287.
13. Gong Q, Sharma M, et al. Multi-omic profiling reveals age-related immune dynamics in healthy adults. Nature 2025:696–706. DOI: 10.1038/s41586-025-09686-5.
14. Miller JA, Guillozet-Bongaarts A, et al. Neuropathological and transcriptomic characteristics of the aged brain. eLife 2017:e31126. DOI: 10.7554/eLife.31126.

15. CZI Cell Science Program, Abdulla S, et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic Acids Research 2025:D886–D900. DOI: 10.1093/nar/gkae1142.

16. Devlin J, Chang MW, Lee K, and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018:DOI: 10.48550/ARXIV.1810.04805.

17. Reimers N and Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv 2019:DOI: 10.48550/ARXIV.1908.10084.

18. Biderman S, Schoelkopf H, et al. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv 2023:DOI: 10.48550/ARXIV.2304.01373.

19. Chen Y and Zou J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. bioRxiv 2023:DOI: 10.1101/2023.10.16.562533.

20. 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 2020:1234–40. DOI: 10.1093/bioinformatics/btz682.

21. Schaefer M, Peneder P, et al. Multimodal learning enables chat-based exploration of single-cell data. Nature Biotechnology 2025. DOI: 10.1038/s41587-025-02857-9.

22. 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 2025:DOI: 10.48550/ARXIV.2503.09427.

23. Sayers EW, Beck J, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Research 2025:D20–D29. DOI: 10.1093/nar/gkae979.

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:bbad132. DOI: 10.1093/bib/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:bbac345. DOI: 10.1093/bib/bbac345.

26. 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 2017:DOI: 10.48550/ARXIV.1705.00652.

27. Jørgensen TE and Breitung J. Margins in Contrastive Learning: Evaluating Multi-task Retrieval for Sentence Embeddings 2025.

28. Heumos L, Schaar AC, et al. Best practices for single-cell analysis across modalities. Nature Reviews. Genetics 2023:1–23. DOI: 10.1038/s41576-023-00586-w.

29. Zappia L and Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biology 2021:301. DOI: 10.1186/s13059-021-02519-4.

30. McInnes L, Healy J, Saul N, and Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software 2018:861. DOI: 10.21105/joss.00861.

31. Sun L, Su Y, Jiao A, Wang X, and Zhang B. T cells in health and disease. Signal Transduction and Targeted Therapy 2023:235. DOI: 10.1038/s41392-023-01471-y.

32. 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 2024:293–305. DOI: 10.1093/jleuko/qiad128.

33. Jabeen MF and Hinks TSC. MAIT cells and the microbiome. Frontiers in Immunology 2023:1127588. DOI: 10.3389/fimmu.2023.1127588.

34. Touvron H, Lavril T, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv 2023:DOI: 10.48550/ARXIV.2302.13971.

35. Yang A, Li A, et al. Qwen3 Technical Report. arXiv 2025:DOI: 10.48550/arXiv.2505.09388.

36. Zhang JY, Wang XM, et al. Single-cell landscape of immunological responses in patients with COVID-19. Nature Immunology 2020:1107–18. DOI: 10.1038/s41590-020-0762-x.

37. La Manno G, Gyllborg D, et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell 2016:566–580.e19. DOI: 10.1016/j.cell.2016.09.027.

38. Eze UC, Bhaduri A, Haeussler M, Nowakowski TJ, and Kriegstein AR. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. Nature Neuroscience 2021:584–94. DOI: 10.1038/s41593-020-00794-1.

39. Shimizu T, Nakazawa M, Kani S, Bae YK, Shimizu T, Kageyama R, and Hibi M. Zinc finger genes *Fezf1* and *Fezf2* control neuronal differentiation by repressing *Hes5* expression in the forebrain. Development 2010:1875–85. DOI: 10.1242/dev.047167.

40. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, and Tuzhilina E. Principal component analysis. Nature Reviews Methods Primers 2022:100. DOI: 10.1038/s43586-022-00184-w.

41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I. Attention Is All You Need. arXiv 2017:DOI: 10.48550/ARXIV.1706.03762.

42. Schroff F, Kalenichenko D, and Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering 2015. DOI: 10.48550/ARXIV.1503.03832.

43. 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 1947:50–60. DOI: 10.1214/aoms/1177730491.

44. Swets JA. ROC Analysis Applied to the Evaluation of Medical Imaging Techniques. Investigative Radiology 1979:109.

45. Menger J, Krissmer SM, Kreutz C, Binder H, and Hackenberg M. mmContext: an open framework for multimodal contrastive learning of omics and text data 2025.

46. OpenRouter: A Unified API for AI Models.

47. Youden WJ. Index for rating diagnostic tests. Cancer 1950:32–5. DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

48. Haghverdi L, Büttner M, Wolf FA, Buettner F, and Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods 2016:845–8. DOI: 10.1038/nmeth.3971.

49. Kendall MG. A New Measure of Rank Correlation. Biometrika 1938:81. DOI: 10.2307/2332226.