Exploring the applications of Topic Modeling in scMultimodal data

Aditya Gorla

Statistics 254

March 20, 2022

Abstract

The advent of multimodal single cell data has opened new avenues to explore single cells biology. Here we explore the application of a Latent Dirichlet Allocation (LDA) model to multimodal single cell data. We see that naive horizontal integration is not viable, at least with LDA. However, we do see strong results when LDA is applied to multimodal, scRNA and scATAC, 10X Multiome data. We notice that lower dimensional representation of the topic model can accurately separate cell types. Finally, we also see that the topics themselves seem to capture biologically relevant information and, in the future, could be used to learn the underlying pathways which lead to immune cell activation.

Introduction

The advent of single-cell profiling has allowed us to get a deeper mechanistic understand of processes which lead to the vast diversity of cell types we observe in nature [1]. Most of exploration done in the single cell space over the last decade has been powered by single cell RNA-seq (scRNA) which gives us insight into the differences in the gene expression at an individual cell level. However, in the last few years, new platforms have been developed to measure chromatin accessibility (scATAC-seq), histone modification (scCHiP-seq), surface proteins (CITE-seq), and other features at a single cell resolution [2,3,4]. These new methods offer complementary information to scRNA which can be combined to provide a deeper view of the molecular profiles of cells and the mechanistic process which differentiate their identity. However, scRNA and scATAC data are vastly different in their dimensionality, signal type, scarcity and cannot be naively integrated and co-analyzed using exiting scRNA methods. Recently, new methods are being proposed to tackle the challenge of integrative analysis of single cell multimodal (scMultimodal) data. Methods such as scAI, MOFA+ and scJoint have been developed to perform integrated scMultimodal data analysis using factor analysis, joint clustering and neural networks, respectively. In this paper we explore the applications of Topic Modeling in scMultimodal data [5,6,7]. Specifically, we explore the application of Latent Dirichlet Allocation (LDA) to combined scRNA and scATAC data [8]. LDA is a highly popular latent variable graphical model widely used on natural language processing (NLP). In NLP, LDA models the relationship between words in documents (features) and different documents (observations) using latent variables (topics). One reason for its wide popularity is the highly interpretable nature of its topic output which often cluster words based on topic similarity and the topic similarity can be used to cluster documents with similar topic distributions. This framework has a very intuitive mapping to single cell biology: the observations are cells, the features are scRNA and scATAC features for each cell and each topic would be a combination of a set genes and chromatin accessibility which could be used to find similar cell types. Additionally, given the generative nature of this model, one can use metric such as LogLikelihood of the model to select the optimal number of topics which best describe a given dataset. Given all of the, it is no surprise that LDA has previously been used to discover coaccessible enhancers and stable cell states from scATAC data in cisTopic [9]. However, the focus of this paper is using LDA to model scMultimodal data which far more complicated but could also be far more informative than analyzing any single data modality.

Therefore, in this paper we explore the application of LDA to horizontally (same features) integrated scRNA/scATAC data and paired scRNA+ATAC data (10X Multiome). Next, we visually compare the cell clustering accuracy of LDA with unimodal data and multimodal data. We then analyze the characteristics of the topics learned from multimodal data. Finally, we inspect the relation between topics and cell type to understand if the topics capture biologically relevant information.

Methods

LDA Model and Implementation

We use the a basic LDA model which is derived from the graphical model described in Fig.1 [10]. In General LDA estimates a probability distribution over the topics for each cell and the probability distribution over the feature (RNA and/or ATAC) over each of the topics. In other words, each topics is composed of a mixture of certain features and each cell is described by the mixture of all topics. LDA algorithms estimate the posterior probability distribution each topic being assigned to each feature, however, this posterior is not analytically tracked, therefor is solved using MCMC approaches. We use a pre-built LDA algorithm called WarpLDA [11]. It is, to our knowledge, the most efficient implement of LDA publicly available with O(1) time complexity and memory complexity of O(K) per cell, where K is the number of topics. Internally, it uses a highly efficient implementation of MCMC Metropolis-Hastings [12]. In general MCMC samplers are used to asymptotically estimate the intractable posterior distribution of topic assignments through repeated sampling (after a burn-in) which can then be used to estimate the topic-cell feature-topic distributions of interest. In WarpLDA users only need to specify the number of topics, the topic-cell, topic-feature priors, max iterations and stopping convergence threshold. Across all experiments we set number of topics (K) to 25 and use the prior recommended by Griffiths and Steyvers, where the topic-cell and topic-feature priors are set to 50/K and 0.1, respectively [13]. Please not that LDA provide both a model fit LogLikelihood and Perplexity, the max and min of which, respectively, can be used to select an optimal K value. We did not perform a grid search to find optimal K due to time constraints and used a fixed K of 25 for all experiments to ensure comparability of results across multiple datasets. Finally, we set the number of iterations to 150, which empirically seemed sufficient, and a stopping threshold of 0.0001. That being said MCMC is asymptotically exact and a larger number of iterations produces more accurate results, if time permits one should always use a larger number of iterations (for burn-in). The WarpLDA algorithm is implemented in the R package “text2vec”.

Data Analysis

Datasets

In the study we use multiple different dataset. The first dataset we construct is the horizontally integrated Mouse Brain dataset. We construct this data by combining data from scRNA-seq dataset from GEO (GSE110823), where the annotated count matrix (10X Genomics protocol) from Zeisel et al. (http://mousebrain. org; L5_all.loom), and the preprocessed ATAC-seq data (gene activity scores) from Luecken et al. (https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_-_integration_task_datasets_Immune_and_pancreas_/12420968 ; small_atac_gene_activity.h5ad) [14,15]. Here, we first find the set of all shared genes between the two data sets: 1975. Next, we match the cell types since the two datasets label cells to varying specificity. The scRNA has a finer grain annotation, therefor we merge cells to their inferred common cell type labels to match those in the scATAC data. For example, Spinal cord excitatory neurons and Di- and mesencephalon excitatory neurons labels in scRNA are merged into a single Excitatory Neurons label. After this filtering, we are left with the following cell types: Astrocytes, Cerebellar Granule Cells, Endothelial Cells, Excitatory Neurons, Inhibitory Neurons, Microglia, Oligodendrocytes. Next we balance the cell counts in the both the datasets. The scRNA database has 105,312 cells while scATAC 11,270 cells. To prevent major effects from data source imbalance, we randomly subsample the scRNA data down to 11,270 cells. We now concat the 2 dataset along the shared feature, which results in a matrix with 22,540 unique cells and 1,975 shared genes. We also extract and combine the per cell annotation from both datasets which results in the following metadata for each cell: cellID, cellLabel, batch and data source. This final matrix, with associated metadata, is what we shall now refer to as the ‘Mouse Brain dataset’.

Next, we construct the Mouse Atlas dataset. This dataset is constructed by horizontally integrating the scRNA Tabula Muris mouse data (https://tabula-muris.ds.czbiohub.org/) and scATAC Mouse Atlas data (https://atlas.gs.washington.edu/mouse-atac/; gene activity scores) [16,17]. We first find the 17,165 shared genes between both the datasets. Next, we rename cell labels in the scATAC data to match those in the scRNA data. This is quite a manual process where over 15 shared cell types were relabeled. The relabeling logic was the same as with the Mouse Brain dataset construction. For example, immature B cell, early pro-B cell and late pro-B cell were merged into a single label B-cell. It is far too expansive to describe the re-labeling for each cell, so we will not describe it in further detail in this paper but the exact R script used to do this can be provided upon request. At end of this, there are about 62 unique cell types when both the datasets are combined. Please note that only just over 20 of the 62 cell types are present in both datasets. Due to memory constrains we were only able to use 41,079 out of the nearly 100,000 cells in the Tabula Muris database, which we believe is more than sufficient. We were able to use the 65,975 out of the 80,920 cells from the scATAC database, after dropping unknown, Testes and Cerebellum cells. We the proceeded to combine the these two databases along their shared features. This results in a matrix with 107,054 unique cells and 17,165 shared genes. We extract and combine the per cell annotation from both datasets which results in the following metadata for each cell: cellID, cellLabel, tissue and data source. This final matrix, and the associated metadata, shall now be referred to as the 'Mouse Atlas dataset'.

The final dataset we construct is the 10X Multiome dataset containing only major immune cells. We obtain the preprocessed scRNA and scATAC data from the NureIPS 2021 Open problems in single cell analysis challenge in GEO (GSE194122) [18]. To be exact we download the 10X Multiome BMMC processed training dataset. This dataset has 42,492 unique cells with 13,431 genes and 116,490 ATAC peaks. Note that ATAC data also has version where the 116,490 peaks are mapped down to 19,039 gene activity scores. This dataset has 22 unique cell types. Given the paired nature of the data, we simply combine both the RNA and ATAC data along the shared cells. This results in a matrix with 42,492 shared cells cells and 129,921 unique features. We extract and combine the per cell annotation from both datasets which results in the following metadata for each cell: cellID, cellLabel, batch and cell phase. This final matrix, and the associated metadata, shall now be referred to as the '10X Multiome dataset'. We also create parallel dataset here using the gene activity score (AS) instead of the raw ATAC peaks. This matrix has 42,492 shared cells cells and 32,470 unique features. This final matrix shall now be referred to as the '10X Multiome AS dataset'.

Finally, all data metrics are binarized to reduce memory usage so they can run on a local machine with 16G RAM. We believe binarizing the data still retains useful information as demonstrated by the high fidelity of the results in scJoint which binarizes all inputs [7]. However, it would be very inserting to see how results change when given a quantitative matrix instead of binarized data in the future.

Analysis and Visualization

All analysis, LDA execution and visualization were performed in a series of R (v4.1.2, aarch64-apple-darwin20) scripts. We record the following outputs from LDA and use for further analysis: LogLikelihood, perplexity, topic-cell distributions, topic-feature distributions and topic components. We can use each of the topic cell/feature distribution matrices, along with metadata, to build heatmaps to visualize the topic weights given to each cell and feature. And clustering along metadata feature such as cell type, source, batch effects, etc. All heatmaps were constructing using "ggplot2" and "ComplexHeatmaps" packages. Next, we use the topic-cell distribution matrices to perform tSNE, UMAP and diffusion map visualization and analysis. We can also use the distribution matrices to analyze the topics. We can analyze the topic-cell type proportions by taking the weighted sum normalized contributions of each topic in each cell for every subtype. We can further explore the weighted contribution of each feature type in each topic by the weighted sum normalized contributions each feature type (RNA or ATAC) for each

topic. We further combine the two previous results to understand the feature type contribution to each cell type. All visualizations were generated using ggplot2, unless otherwise stated. Please not that all the data curation, pre-processing, analysis and visualization code is available upon request.

Results

Horizontally integrated Mouse Brain dataset

We first run LDA on the binarized, horizontally integrated Mouse Brain dataset. The UMAP of the topic-cell output with multiple metadata overlays are visualized in Fig.2. In Fig.2(A), we see that cells are completely stratified based on the source of their data type. In Fig.2(B), we further see that the cells from scATAC are moderately separated based on their cell types. However, we see no separation in the cell from the scRNA (GEX), in fact they look perfectly mixed. Finally, in Fig.2(C), we see that the cells from scATAC suffer from major batch effects based on their study of origin. While this was an initial exploratory application of LDA to integrated data, we are still very surprised by the major batch effects based on data type and source. It is not entirely clear what is causing these effects. However, we posit that LDA is trying to capture the top structure in the data which in this cases happens to originate from the major batch effects in the scATAC data. Thus, it might just be capturing latent factors to model this structure. Additionally, its also likely that there's far more homogeneity in scRNA cells than we originally thought, so LDA is failing to capture any relevant structure. We further investigate this by examining the raw topic-cell probability distribution matrix visualized in Fig.3. Here we see that no single topic models a significant distribution across both scRNA and scATAC cells, which we find very surprising. Again, we're not able to say for certain why this is happening but do posit the aforementioned theory on the major structure in the scATAC data which could lead to such an effect. Given all this we are not able to make any conclusive claims about the applicability of LDA to the horizontally integrated data.

Mouse Atlas dataset

We now run LDA on the horizontally integrated Mouse Atlas dataset. We run tSNE on the topic-cell distribution matrix to visualize the cell type clustering in Fig.4. Here, we see fairly good separation and clustering of the of the over 60 unique cell types in this dataset. This is the first result we see which seems to indicate LDA might be applicable to multimodal single-cell datasets. We then examining the raw topic-cell probability distribution matrix visualized in Fig.5., the results here are less encouraging. We do observe some topic distributions (eg. Tropic 15 and 11) spanning across cells from both scATAC and scRNA. We also see reasonable clustering of similar cell types. However, we note that applying hierarchical clustering of the cells sees them separate into 2 major groups along the source of data type. This is not ideal behaviors as it means that the major structure being captured by the LDA topics is data source rather than the cell types themselves, however, they do also cluster by cell types as well. This major structure originating from the data source is better visualized by the tSNE in Fig.6. Given, the results from the Mouse Brain and Mouse Atlas data we, can tentatively conclude that applying LDA naively to the horizontally integrated datasets in not appropriate. One way to correct for this issue might be through modifying the LDA model priors (specifically the feature-topic priors) to ensure that they weight data from both sources evenly. Although, it's not clear how this might effect the final cell type clustering and the interpretability of the topics. There might also be other approaches such as deriving a different latent variable model that can accurately account for this issue, but this is out of the scope for this paper.

10X Multiome dataset

Given the results from the last two datasets, we decide to now explore the application of LDA to multimodal data based in the 10X Genomics Multiome sequencing. In the 10X Multiome

dataset all cells have both scRNA and scATAC data which should eliminate the major batch effects observed from data source type. Indeed in Fig.7, we see that the topic distributions are not stratified along any obvious sources and that there's quite a few topics whose distributions span across all cells. We also see that there is good mixing across all batches (sampling location and individual) and the cells are hierarchically clustered mostly along their true cell types. We can also see this results visualized in the topic-cell UMAPs in Fig.8. Here we observe quite a good separation between CD4/8 and NK cells. We can revisualizes Fig.8 as Fig.9 to only highlighting certain cells of interest, CD4 and CD8. Interesting doing so allows us to see clearly a good separation between the activated CD4 (Fig.9(A)) and CD8 (Fig.9(B)) cells, while the naive CD4/8 are clustered closely together. Additionally, we also capture trends like a trajectory where activated CD4 come before activated CD8 cells. All this is in line with our biological understanding where the naive CD4/8 cells are quite similar and that CD4 promote the induction of a robust primary CD8 T cell response through numerous mechanisms [19]. We also are able to capture the close relationship between CD14 and CD16 cells [20].

Given all this, we now contract the scMultimodal results to those obtained by applying LDA to each of the data modalities separately. We take the combined 10X Multiome dataset and partition it into 2 datasets, with one containing scRNA and scATAC data, then apply LDA to each of these separately. The LDA results for each of these are visualized as UMAPs in Fig.10. We observer lesser separation between the CD4/8 naive cells and CD4/8 activated cells in each of the single modality data. We also see a more significant separation between the CD14/16 mono cells relative all other cells in the cluster when using the multimodal data. Given this, it seems that LDA might be performing a wighted integration of the different modalities to retain the most informative structure in the data which helps explain cell type heterogeneity. Well shall investigate this claim further in the following section where we analyze the topics.

We also derive a diffusion map from the LDA results and plot it in Fig.11. We again see here a good separation between multiple cell types and see 3 distinct extremums in the star shape. A closer inspect of the diffusion map reveals that multiple types of progenitor and naive immune cells are located at the bottom left and the middle. These progenitor/naive cells mature and differentiate in two different paths, as indicated the vertical and horizontal red arrows. Specifically, the bottom right of the plot show the terminal states of NK and Dendritic cells. The top left main contains the terminal states of B1, CD4 and CD8 cells. These tends seems to line up with our board understanding of the immune cell differentiation pathway. Barring a more fine grained analysis of these results by an immuno-biologist, we can tentatively say that LDA does seem to capture biologically relevant information from the multimodal data which an be used to infer broad trends in cell differentiation. All of this, give us a better understand of why we see better cell clustering in the previously discussed UMAP and that this is likely a results of LDA capturing relevant biology, not simply some nuisance covariate in the data.

Investigating the Topics

In Fig.12, we visualize the per modality contribution to each of the 25 topics. First, we notice the most topics are heavily biased to a single modalities, with most topics mostly focusing on scATAC data. However, Topics 6, 14, 15, 18 and 19 are mainly focused on scRNA data. Second, some topics like 23 and 24 are approximately evenly split between scATAC and scRNA features. All this seems to indicate the LDA seems to find the scATAC data more informatively than scRNA data. However, it is important to keep in mind that only 10% of the all the feature are from scRNA with the rest being scATAC. So, it is not possible to rule our the fact that this increased focus on scATAC feature is not simply an artifact of the feature imbalance. It is worth noting that this distribution of feature is not in line with expectation if LDA was naively weighting all features equally. As matter of fact that avenger scRNA contribution to all topics is much more than 10%. So we believe LDA is capturing significant information from both modalities. Furthermore, based on or analysis of Fig.10, we know that there is a large amount of shared information across both modalities. Given this and Fig.12, we think LDA might extracting unique information from each modality and is putting most of the weight on scATAC features, given the

feature imbalance, when the signals are similar in both modalities. In the future, we could perform a more detailed analysis to validate this theory by performing controlled simulations and feature ablation analysis.

In Fig.13, we visualize the per modality contribution to each of the cell types. We notice the the per modality contribution is approximately the same across all cell. Furthermore, scRNA contributes about 30% across all cell types which is noticeably higher than its 10% share of the feature set. This further bolsters our claim that LDA is extracting relevant and unique information from both modalities cable of explaining cell types from both modalities even under the large feature set imbalance.

We finally visualize the per topic contribution to each cell type in Fig.14. First, we notice that no two cell type share an identical topic profile. This means that cell type are uniquely distinguishable based on their topic profile. This is further validation that LDA capture biological relevant information which can be used to cluster cells in the previously discussed UMAPs. This further implies that give a hold out set of cell with the same features, we could apply the exiting feature topic matrix to this data and classify cell based on their resultant topic profiles. This is a very interesting theory to test in the future. Second, we notice that the topic profiles of similar cells are closer to each other and distant cell types are quite different. This might seem like a trivial first order observation but serves as indication of the level of relevant biological information captured by the topics. Note that the topic profiles for naive vs activated CD4 cells are identical, save topic 1. We see a up regulation of topic 1 features explains the transition of CD4 cell from naive to activated. Furthermore, we note that topic 1 is mostly informed by ATAC features which indicates a change in the histone accessibility being very importing in the activation of CD4 cells. We see this exact same trend in CD8's transition from naive to activated, which could indicate that this a conserved biological pathway in immune cell activation. Next, we also note that CD4 and CD8 naive cell profiles are quite similar which is inline with out biological understanding of these cells types. Given this preliminary analysis of the topic profiles, it is our opinion that LDA topics do indeed capture significant biological information which can be used to understand cell state and cell type transitions, in post-hoc analysis. We, unfortunately, have not been able to perform deeper analysis in this paper. But strongly believe that analysis of the biology captured by the topics can help recapitulate know cells activation/transition pathways and uncover new ones.

Discussion

In the paper we investigate the application of LDA to multimodal single cell data consisting of scATAC and scRNA. We apply LDA on a couple of horizontally integrated dataset and a multimodal single cell data. We notice that LDA, in its current implementation, is not applicant to naively, horizontally integrated data. Under this data regime we notice that topics seems to be stratified based on the cell's mode of origin. We extrapolate and say that gene expression and gene activitie scores, from ATAC, might be capturing orthogonal information and can't be simply mapped from one data type to another. Next, we applied LDA to 10X Multiome multimodal data and see quite good cell type cluster separation and good generalizability of the topics across the two data modalities, even with the large feature imbalance. We see better clustering performance in the multimodal data relative to either of the single data modalities. We additionally, do not notice any significant batch effects and can also can capture biologically meaningful cells trajectories. We also investigate the topics and note that topics capture signal from both data modes. We also note that most cell types seems to have approximately similar modality contributions. We finally, show that the topic profiles are unique to each cell type and likely contain information which can elucidate how cells differentiate and activate. Overall, in the multimodal data regime we were able to show that topics do capture biologically relevant information which is capable of distinguishing cells types and further investigation might reveal biological pathways which contribute to cell type identity and activation.

There are quite a few future direction for this project. First, one can try a range of K (topic) values and selected the optimal K which minimizes perplexity and re-analyze the results check if

we get better results. Second, one can further investigate the RNA/ATAC features in each topic to understand if they are capturing know and/or novel biological pathways which explain cell development or differentiation. We could also try to apply the topic-feature weights to a hold out set of cells to see if we can use the current topic profiles to accurately cluster the holdout cells. Third, we could try to use quantitative data instead of binarized data to see if we get a performance improvement. Finally, we see that a basic LDA model is not appropriate for horizontally integrated dataset but LDA is a very specific parameterization of a broad class of latent variable graphical models. One could in the future try to develop a new latent factor model which can be applied to horizontally integrated data. This model would need to solve for joint topic wights while treating the scATAC and scRNA inputs matrices as conditionally independent. We were not able to explore such a model in this paper but do believe it is entirely feasible to build such a model and could be very promising.

References

[1] Hao, Yuhan, et al. “Integrated Analysis of Multimodal Single-Cell Data.” *Cell*, vol. 184, no. 13, 2021, https://doi.org/10.1016/j.cell.2021.04.048.

[2] Buenrostro, Jason D., et al. “Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation.” *Nature*, vol. 523, no. 7561, 2015, pp. 486–490., https://doi.org/10.1038/nature14590.

[3] Grosselin, Kevin, et al. “High-Throughput Single-Cell Chip-Seq Identifies Heterogeneity of Chromatin States in Breast Cancer.” *Nature Genetics*, vol. 51, no. 6, 2019, pp. 1060–1066., https://doi.org/10.1038/s41588-019-0424-9.

[4] Stoeckius, Marlon, et al. “Simultaneous Epitope and Transcriptome Measurement in Single Cells.” *Nature Methods*, vol. 14, no. 9, 2017, pp. 865–868., https://doi.org/10.1038/nmeth.4380.

[5] Jin, Suoqin, et al. “SCAI: An Unsupervised Approach for the Integrative Analysis of Parallel Single-Cell Transcriptomic and Epigenomic Profiles.” *Genome Biology*, vol. 21, no. 1, 2020, https://doi.org/10.1186/s13059-020-1932-8.

[6] Argelaguet, Ricard, et al. “MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data.” *Genome Biology*, vol. 21, no. 1, 2020, https://doi.org/10.1186/s13059-020-02015-1.

[7] Lin, Yingxin, et al. “scJOINT Integrates Atlas-Scale Single-Cell RNA-Seq and ATAC-Seq Data with Transfer Learning.” *Nature Biotechnology*, 2022, https://doi.org/10.1038/s41587-021-01161-6.

[8] Blei, David, et al. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research*, vol. 3, 3 Jan. 2003, pp. 93–1022., https://doi.org/https://dl.acm.org/doi/10.5555/944919.944937.

[9] Bravo González-Blas, Carmen, et al. “CisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-Seq Data.” *Nature Methods*, vol. 16, no. 5, 2019, pp. 397–400., https://doi.org/10.1038/s41592-019-0367-1.

[10] Hong, Tae-Ho, et al. “Multi-Topic Sentiment Analysis Using LDA for Online Review.” *The Journal of Information Systems*, vol. 27, no. 1, 한국정보시스템학회, Mar. 2018, pp. 89–110, doi:10.5859/KAIS.2018.27.1.89.

[11] Chen, Jianfei, et al. “WarpLDA: a Cache Efficient O(1) Algorithm for Latent Dirichlet Allocation.” *ArXiv.org*, 2 Mar. 2016, https://doi.org/https://doi.org/10.48550/arXiv.1510.08628.

[12] Robert, Christian P. “The Metropolis-Hastings Algorithm.” *ArXiv.org*, 27 Jan. 2016, https://doi.org/10.48550/arXiv.1504.01896.

[13] Griffiths, Thomas L., and Mark Steyvers. “Finding Scientific Topics.” *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl_1, 2004, pp. 5228–5235., https://doi.org/10.1073/pnas.0307752101.

[14] Zeisel, Amit, et al. “Molecular Architecture of the Mouse Nervous System.” *Cell*, vol. 174, no. 4, 2018, https://doi.org/10.1016/j.cell.2018.06.021.

[15] Luecken, Malte D., et al. “Benchmarking Atlas-Level Data Integration in Single-Cell Genomics.” *Nature Methods*, vol. 19, no. 1, 2021, pp. 41–50., https://doi.org/10.1038/s41592-021-01336-8.

[16] “Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris.” *Nature*, vol. 562, no. 7727, 2018, pp. 367–372., https://doi.org/10.1038/s41586-018-0590-4.

[17] Cusanovich, Darren A., et al. “A Single-Cell Atlas of in Vivo Mammalian Chromatin Accessibility.” *Cell*, vol. 174, no. 5, 2018, https://doi.org/10.1016/j.cell.2018.06.052.
[18] Luecken, Malte D, et al. “A Sandbox for Prediction and Integration of DNA, RNA, and Proteins...” *NeurIPS 2021 Datasets and Benchmarks Track (Round 2)*, 23 Aug. 2021, https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/158f3069a435b314a80bdcb024f8e422-Paper-round2.pdf.
[19] Laidlaw, Brian J., et al. “The Multifaceted Role of CD4+ T Cells in CD8+ T Cell Memory.” *Nature Reviews Immunology*, vol. 16, no. 2, 2016, pp. 102–111., https://doi.org/10.1038/nri.2015.10.
[20] Ong, Siew-Min, et al. “A Novel, Five-Marker Alternative to CD16–CD14 Gating to Identify the Three Human Monocyte Subsets.” *Frontiers in Immunology*, vol. 10, 2019, https://doi.org/10.3389/fimmu.2019.01761.

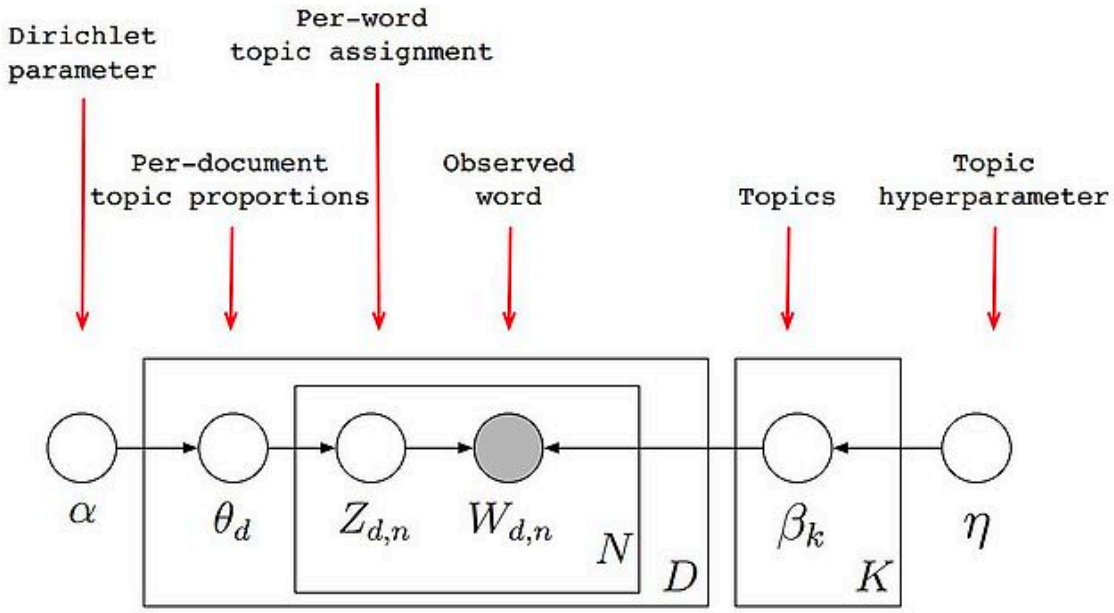

Fig.1: A graphical representation of the Latent Dirichlet Allocation model for NLP in plate notation. Image source [1].

Fig.2: (A) UMAP visualization of the LDA Mouse Brain dataset results with the cell source information overlayed. **(B)** UMAP visualization of the LDA Mouse Brain dataset results with the cell type information overlayed.

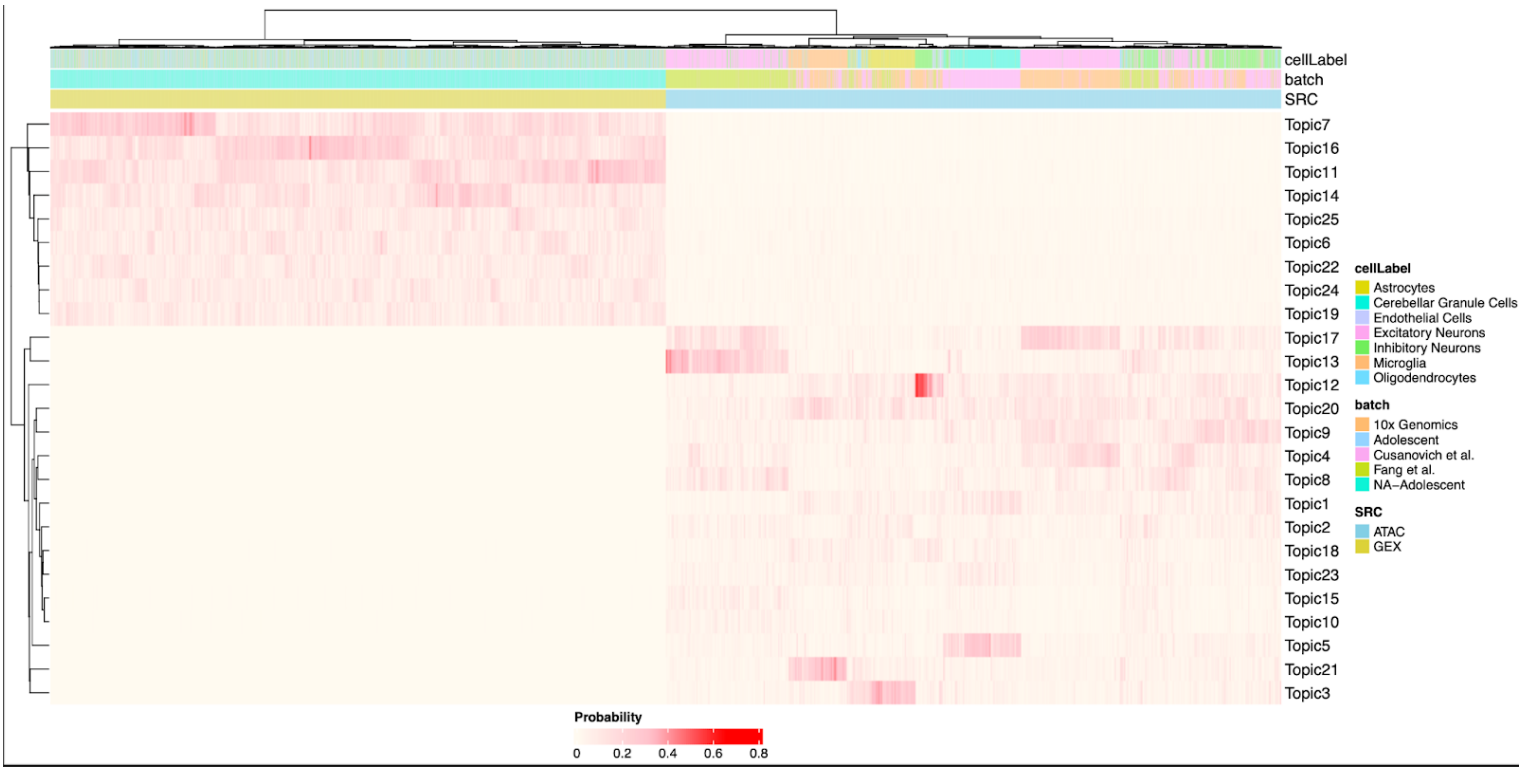

Fig.3: A complex heatmap visualization of the raw topic-cell probability distribution matrix from LDA applied to the Mouse Brain dataset. The row represent topics and the column present cells. We visualize the following per cell metadata: Cell label, source batch and data modality source. The dendrogram are generated by fastcluster hierarchical clustering. Note that GEX means originating from scRNA.

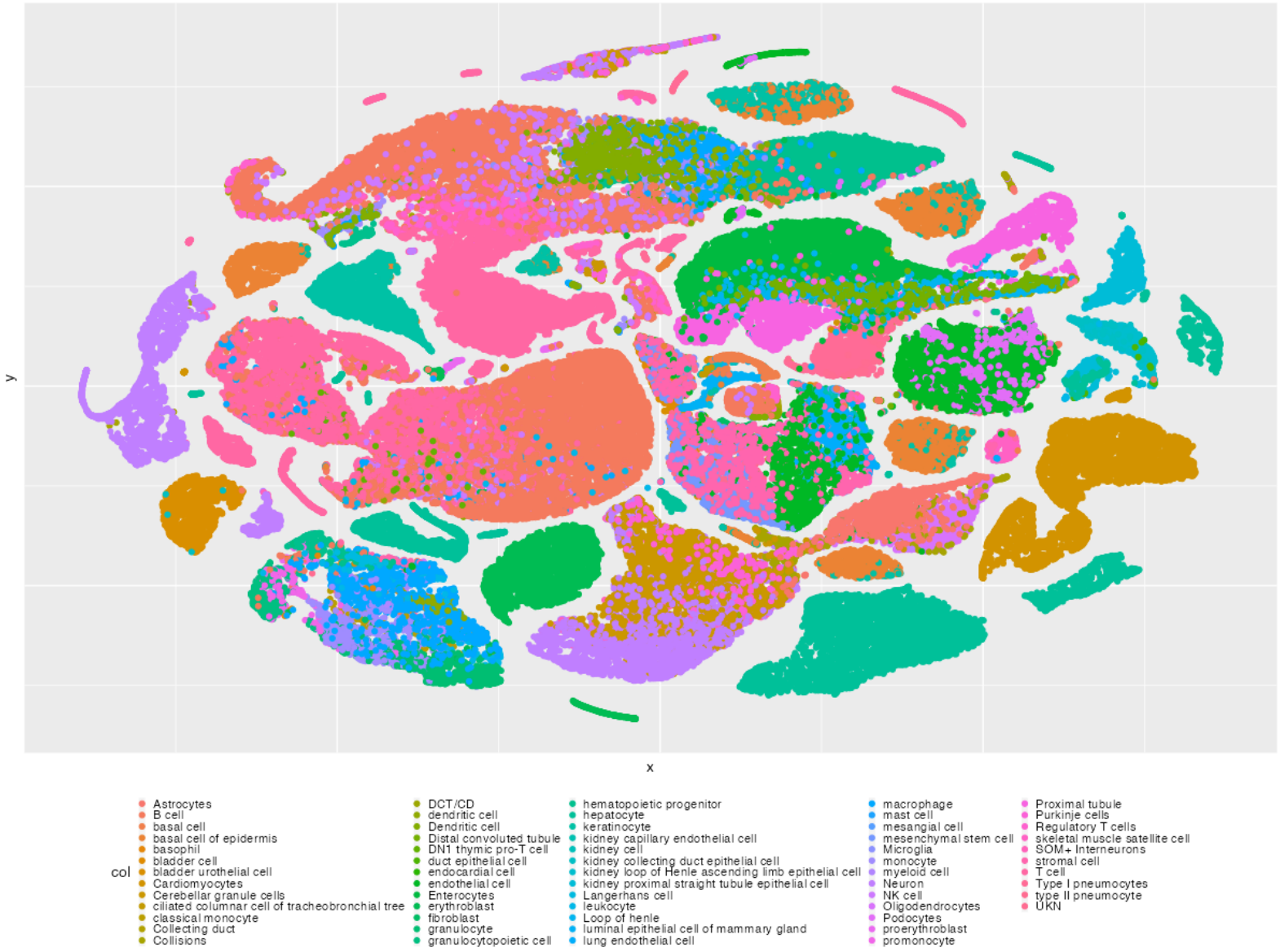

Fig.4: tSNE clustering visualization of the LDA Mouse Atlas dataset results with the cell type information overlayed.

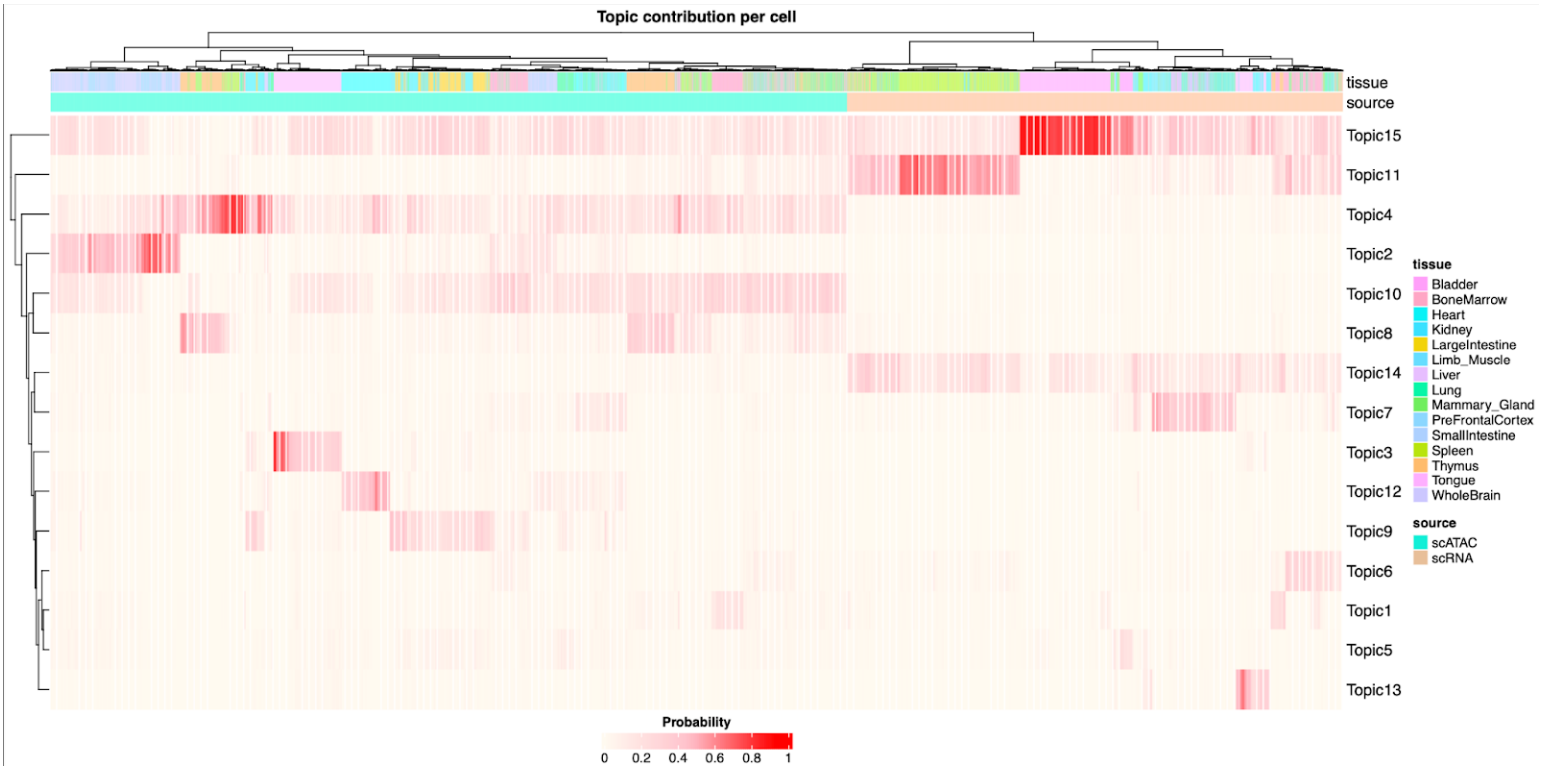

Fig.5: A complex heatmap visualization of the raw topic-cell probability distribution matrix from LDA applied to the Mouse Atlas dataset. The row represent topics and the column present cells. We visualize the following per cell metadata: Tissue of origin and data modality source. The dendrogram are generated by fastcluster hierarchical clustering.

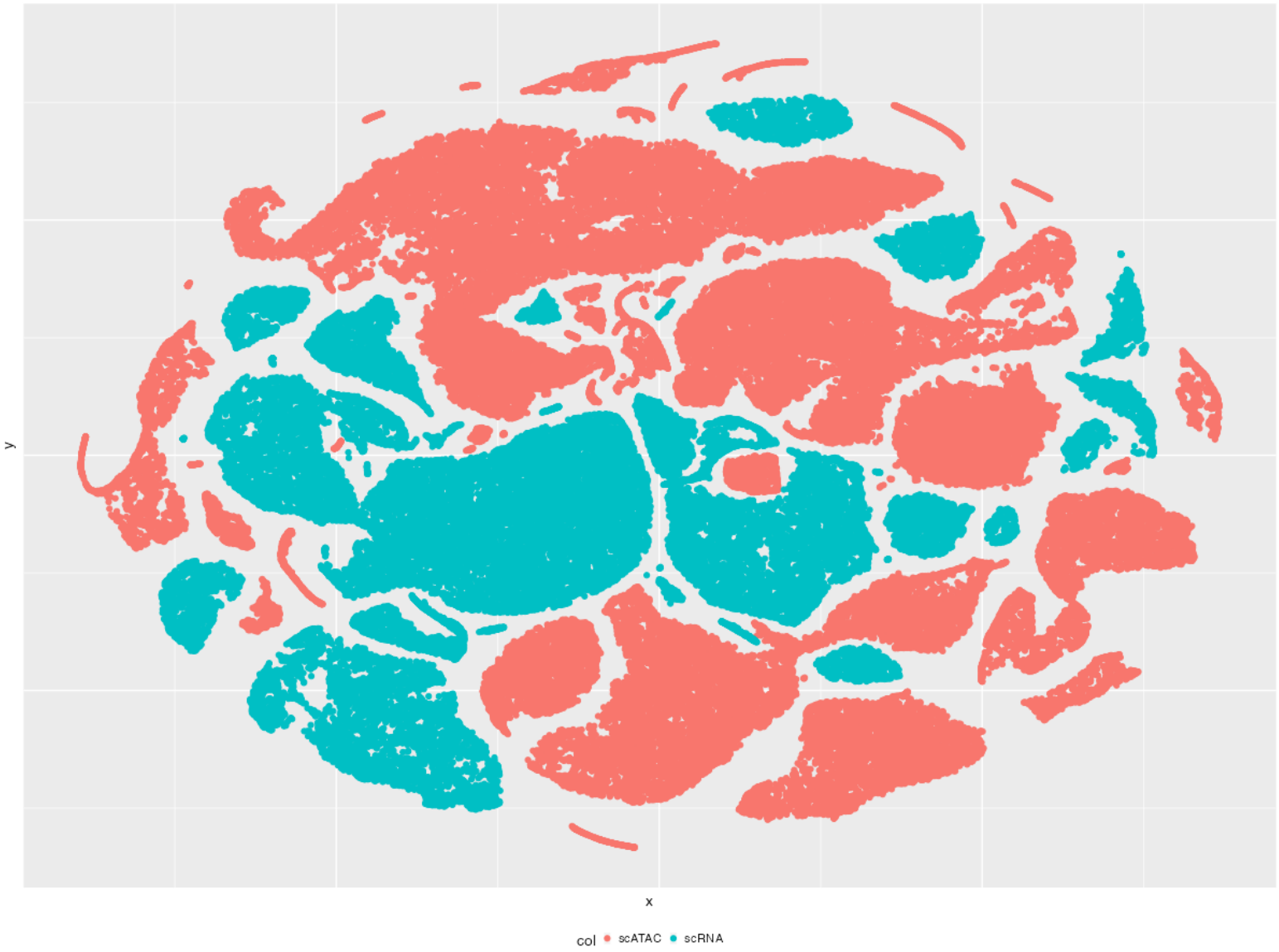

Fig.6: tSNE clustering visualization of the LDA Mouse Atlas dataset results with the cell source information overlayed.

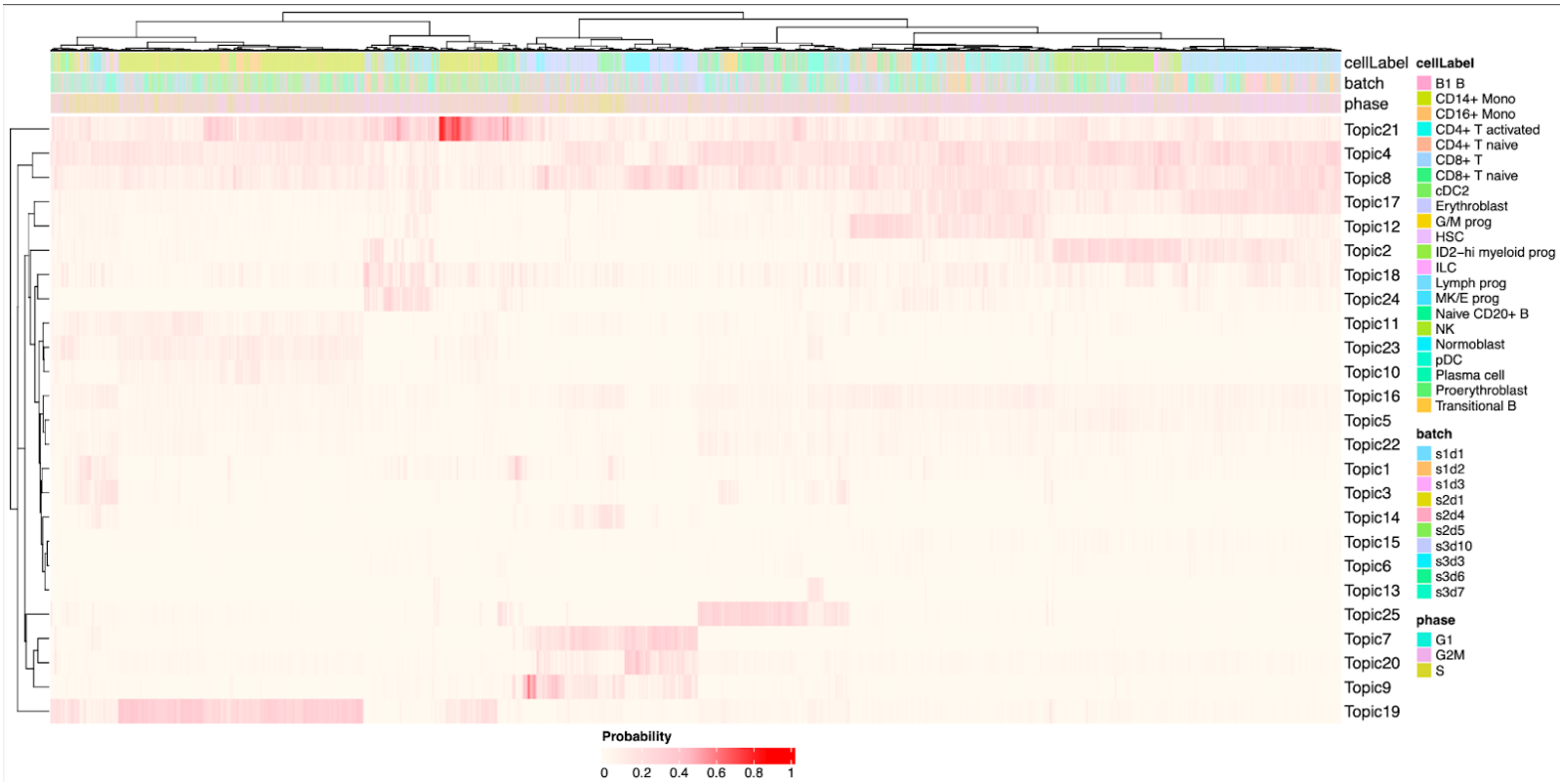

Fig.7: A complex heatmap visualization of the raw topic-cell probability distribution matrix from LDA applied to the 10X Multiome dataset. The row represent topics and the column present cells. We visualize the following per cell metadata: Cell label, Batch (s stands form Site and d standard for Do nor in that site) and inferred cell cycle phase. The dendrogram are generated by fastcluster hierarchical clustering.

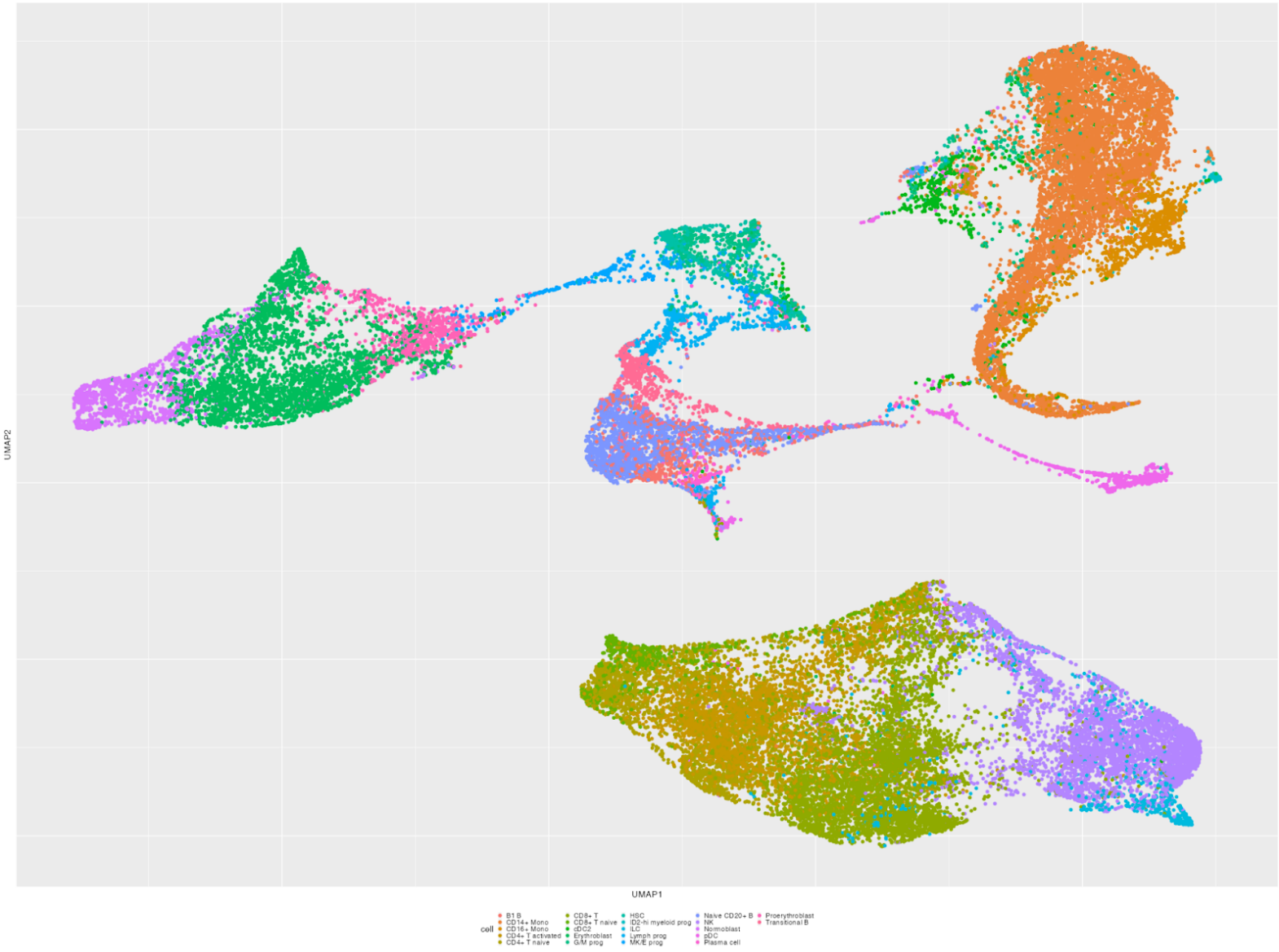

Fig.8: UMAP visualization of the LDA Mouse Atlas dataset results with the cell type information overlayed.

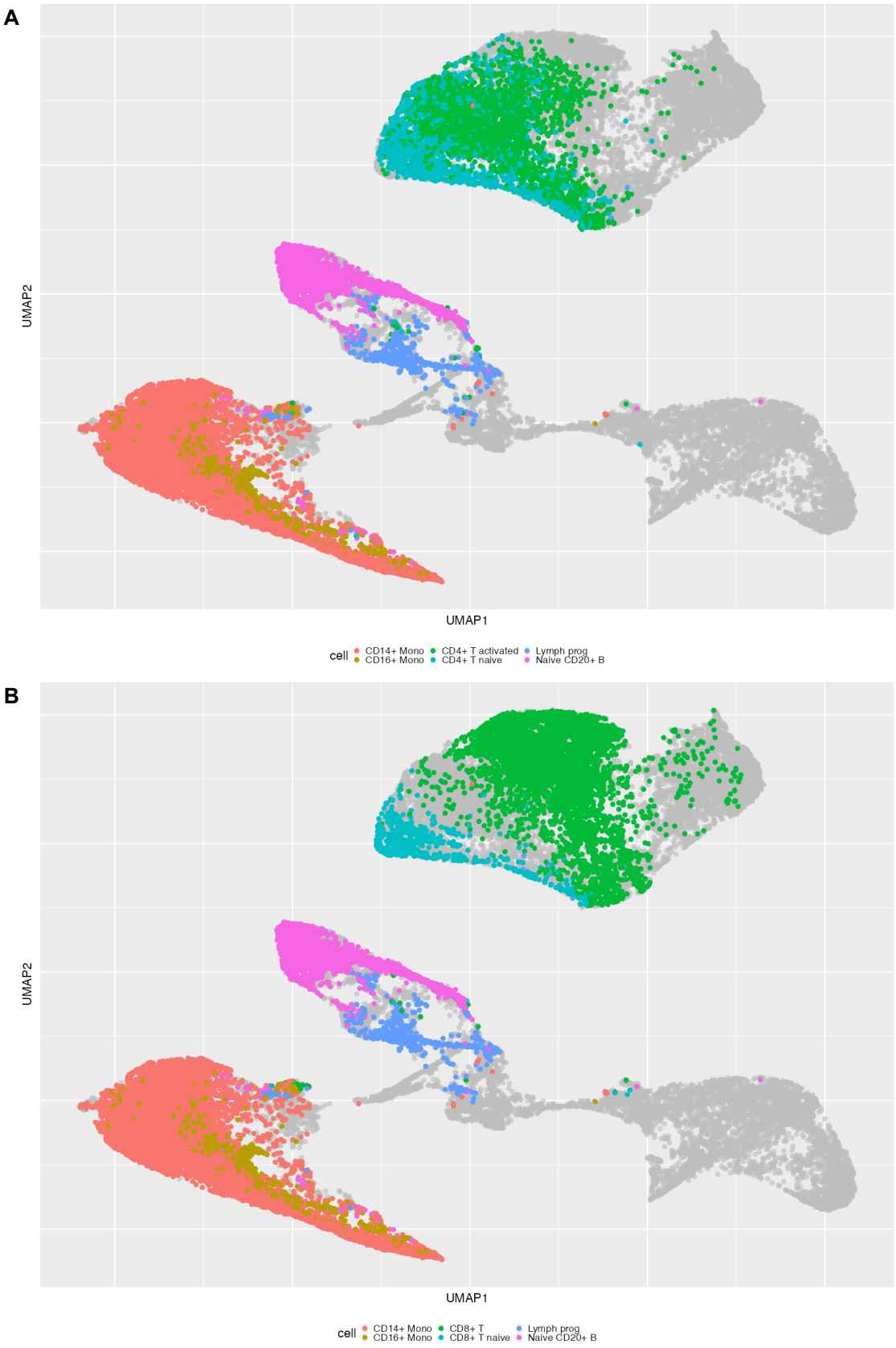

Fig.9: (A) UMAP visualization of the LDA 10X Multiome results with CD4 naive and activated cells highlighted in teal and green, respectively. **(B)** UMAP visualization of the LDA 10X Multiome results with CD8 naive and mature cells highlighted in teal and green, respectively.

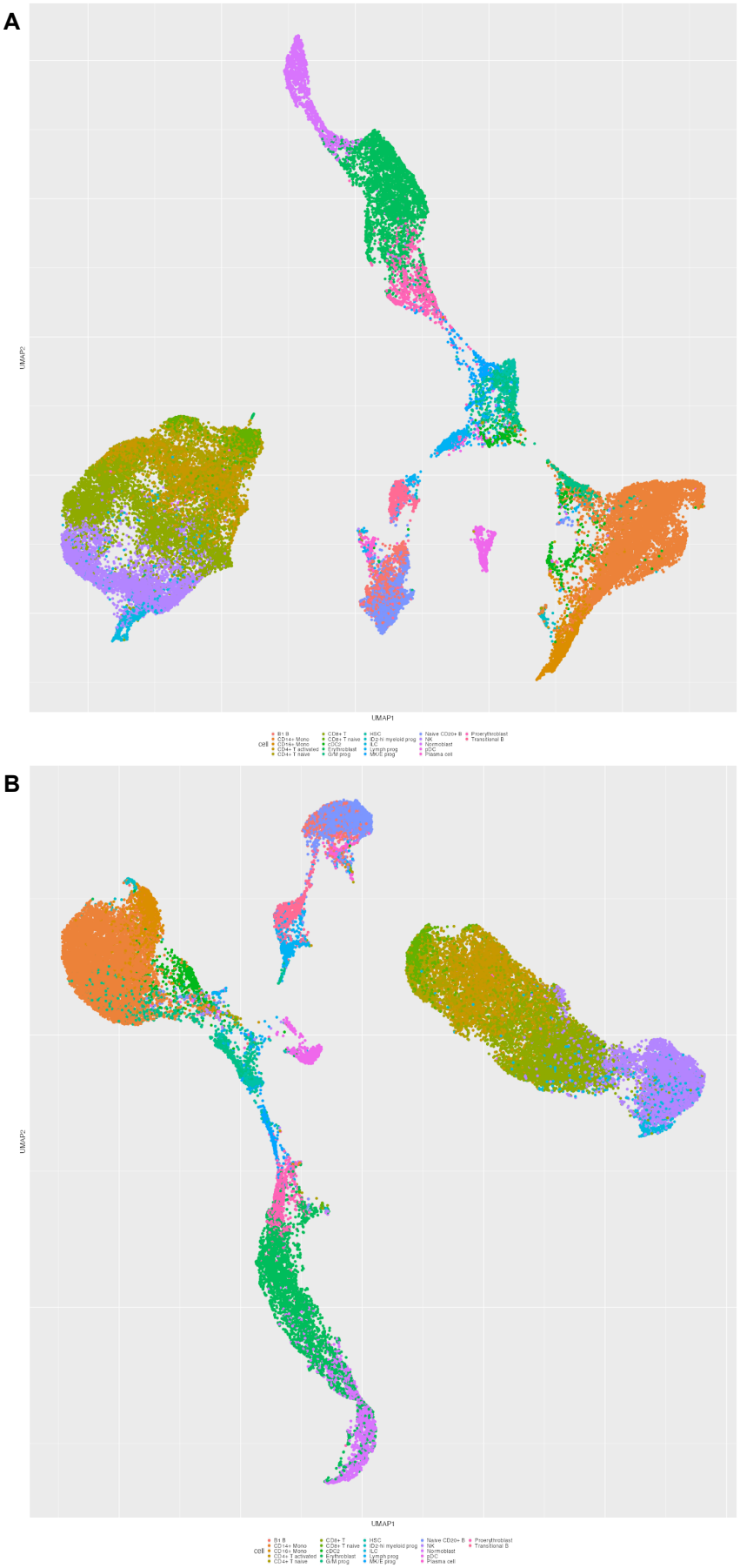

Fig.10: (A) UMAP visualization of the LDA scRNA only 10X Multiome data results with the cell type information overlayed. **(B)** UMAP visualization of the LDA scATAC only 10X Multiome data results with the cell type information overlayed.

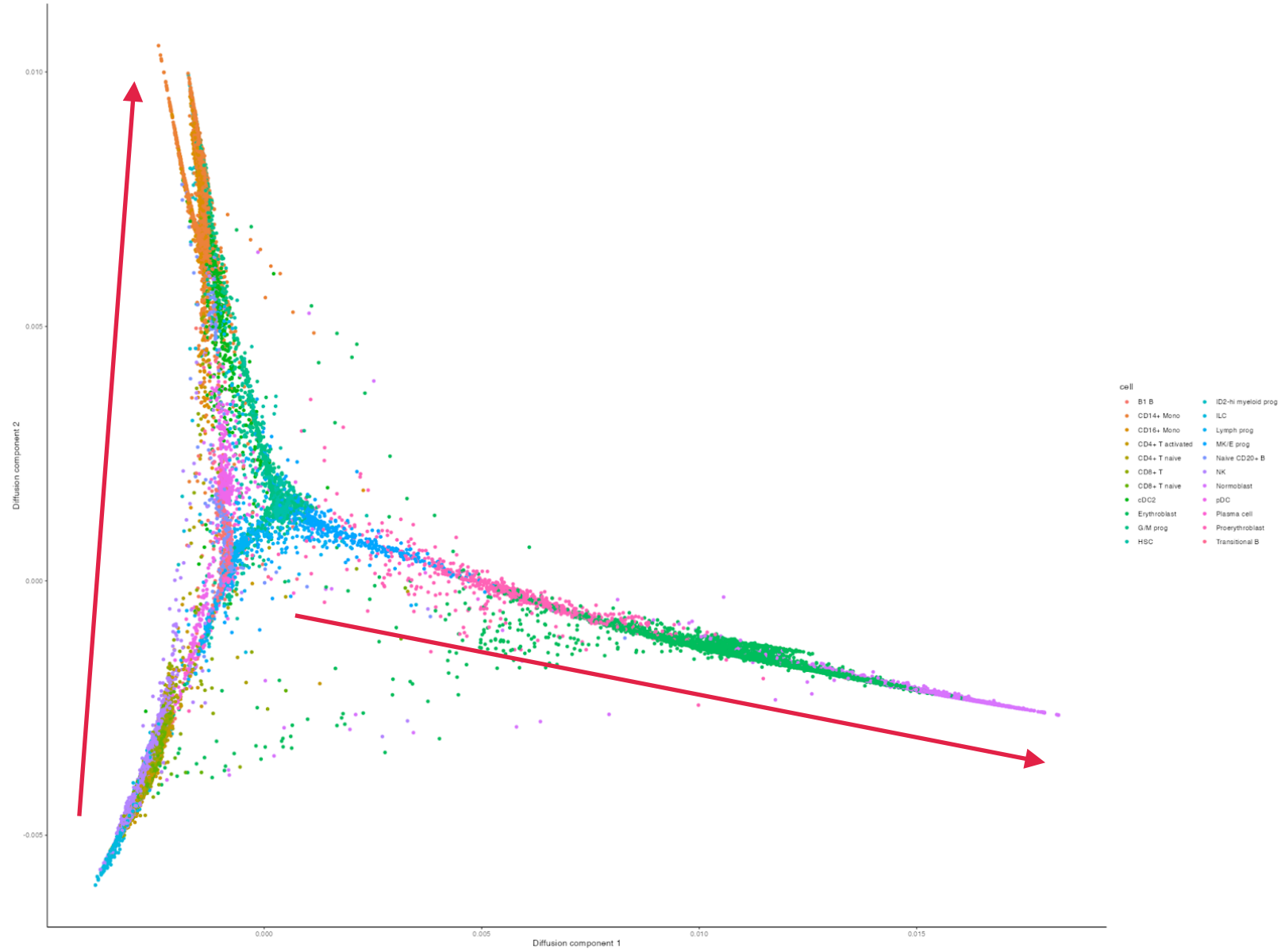

Fig.11: A diffusion map generated from the LDA 10X Multiome dataset results with the cell type information overlayed. The red arrows indicate two possible cell differentiation trajectories. The diffusion maps are generated using the DiffusionMap function in the "Density" R package.

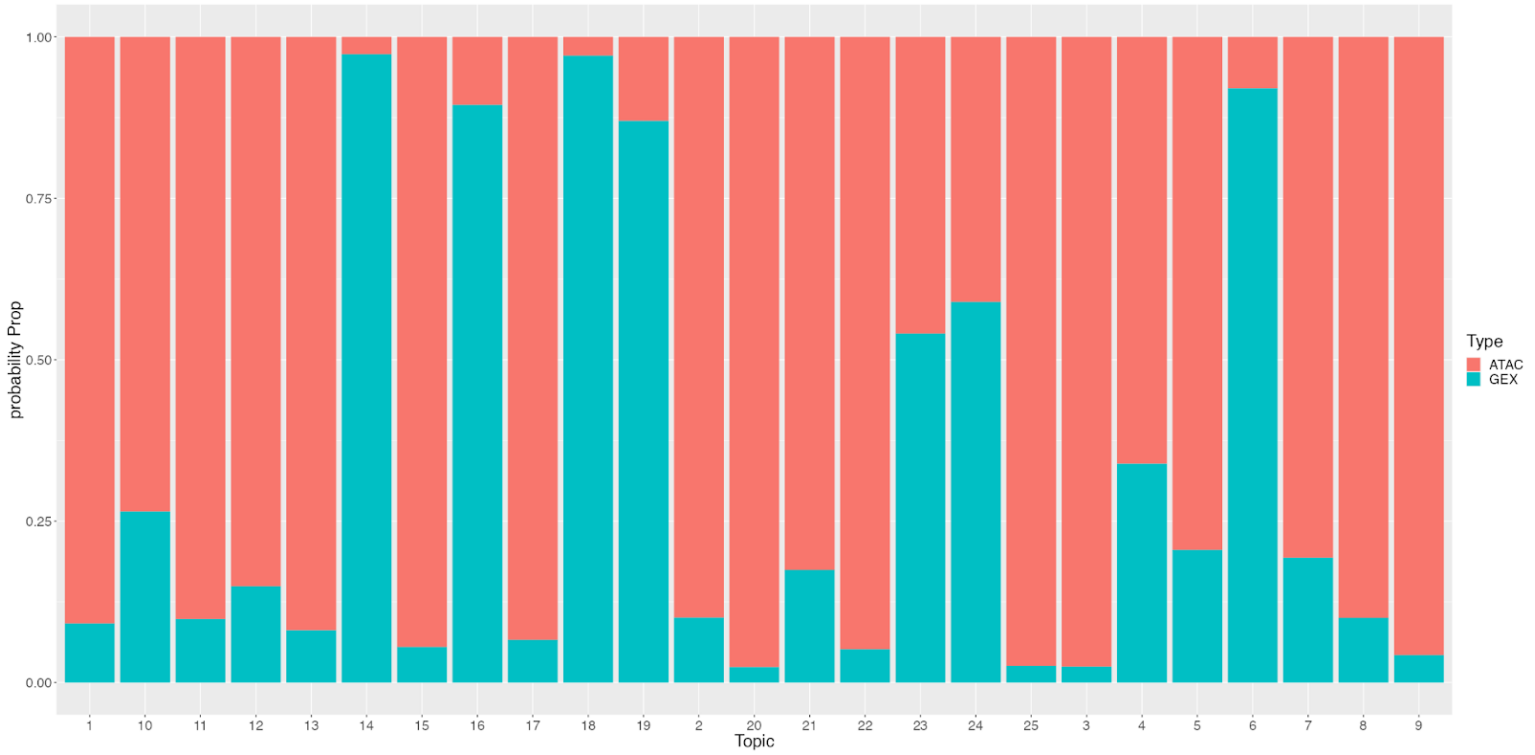

Fig.12: The relative contribution of each modality's information to each topic is visualized as a stacked barplot. Note that GEX means originating from scRNA.

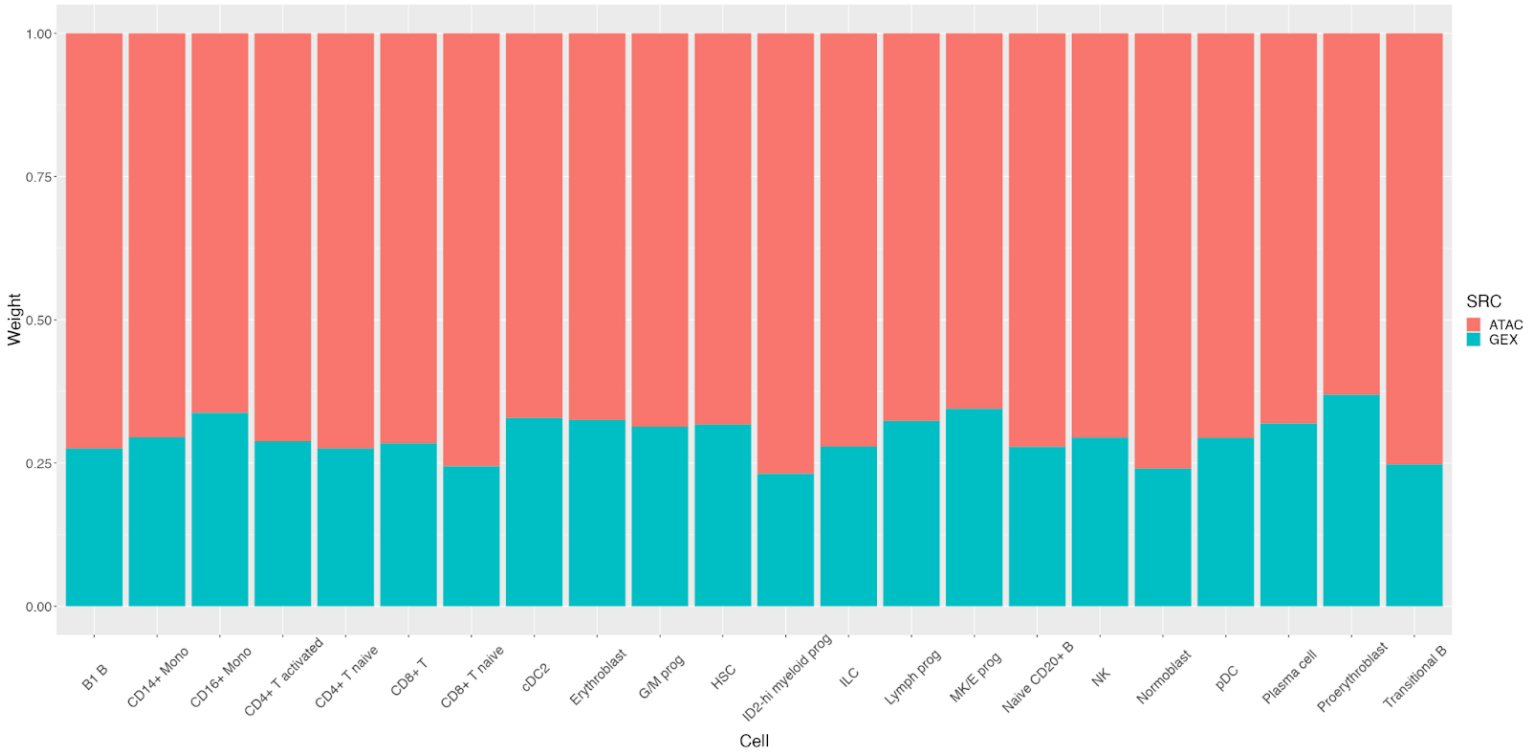

Fig.13: The relative contribution of each modality's information in explaining each cell type is visualized as a stacked barplot. Note that GEX means originating from scRNA.

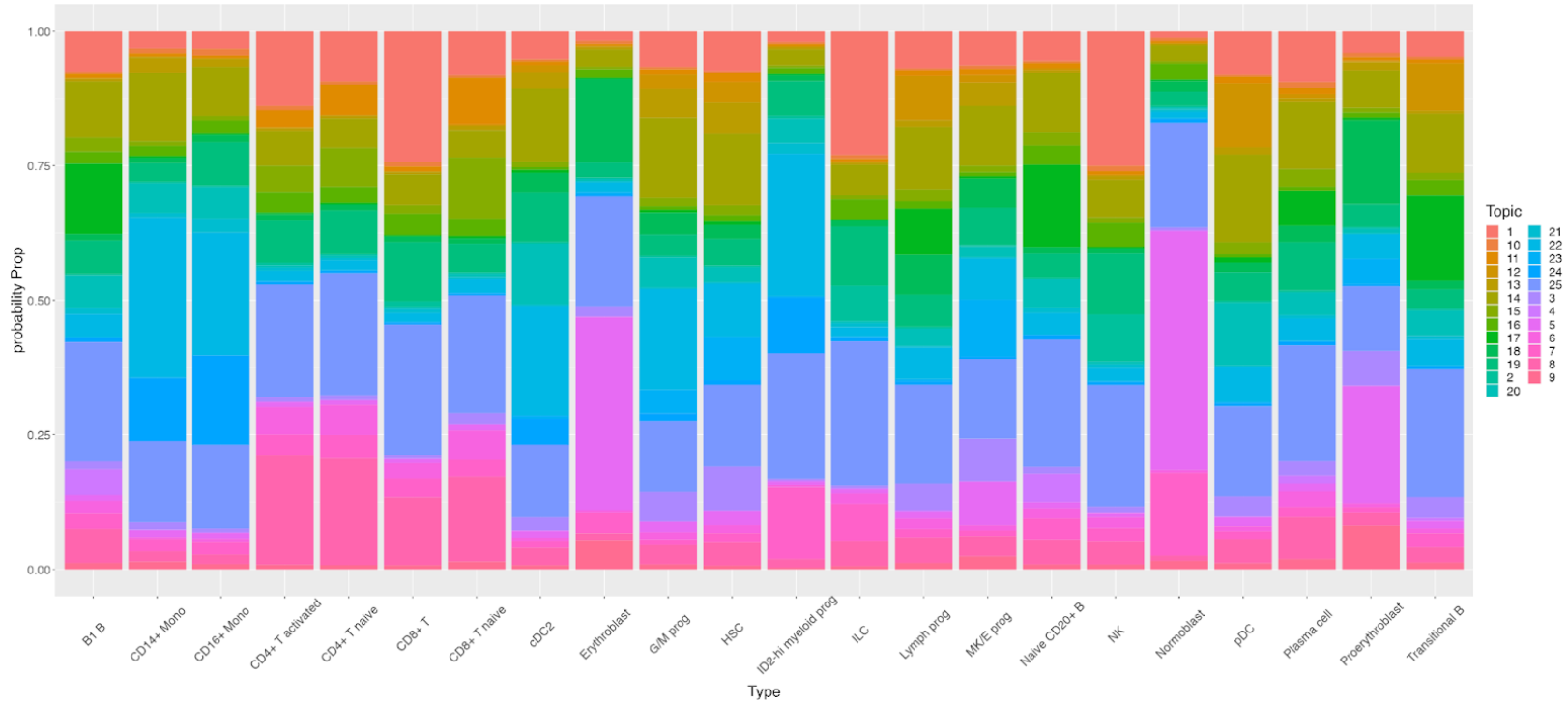

Fig.14: The relative contribution of each topics information in explaining each cell type is visualized as a stacked barplot.