bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data

Yifan Zhao[1], Huiyu Cai[2], Zuobai Zhang[3], Jian Tang[4,*] and Yue Li[1,*]

[1]School of Computer Science, McGill University, Montreal, Canada
[2]Department of Machine Intelligence, Peking University, Beijing, China
[3]School of Computer Science, Fudan University, Shanghai, China
[4]HEC Montreal, Montreal, Canada
[1]School of Computer Science, McGill University, Montreal, Canada
[*]Correspondence to jian.tang@hec.ca, yueli@cs.mcgill.ca

Abstract

The advent of single-cell RNA sequencing (scRNA-seq) technologies has revolutionized transcriptomic studies. However, integrative analysis of scRNA-seq data remains a challenge largely due to batch effects. We present single-cell Embedded Topic Model (scETM), an unsupervised deep generative model that recapitulates known cell types by inferring the latent cell topic mixtures via a variational autoencoder. scETM is scalable to over 10^6 cells and enables effective knowledge transfer across datasets. scETM also offers high interpretability and allows the incorporation of prior pathway knowledge into the gene embeddings. The scETM-inferred topics show enrichment in cell-type-specific and disease-related pathways.

Background

Advances in high-throughput sequencing technologies [1] provide an unprecedented opportunity to profile the individual cells' transcriptome across various biological and pathological conditions, and have spurred the creation of several atlas projects [2–5]. Emerged as a key application of scRNA-seq data, unsupervised clustering allows for cell-type identification in a data-driven manner. Flexible, scalable, and interpretable computational methods are crucial for exploiting the full potential of the wealth of single-cell datasets and translating the transcription profiles into biological insights. Despite considerable progress made on clustering method development for scRNA-seq data analysis [6–16], several challenges remain.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

First, compared to bulk RNA-seq, scRNA-seq data commonly exhibit higher noise levels and drop-out rates, where the data only captures a small fraction of a cell's transcriptome [17]. Changes in gene expression due to experimental design, often referred to as batch effects [18], can have a large impact on clustering [12, 18–20]. If not properly addressed, these technical artefacts may mask true biological signals in cell clustering.

Second, the partitioning of the cell population alone is insufficient to produce biological interpretation. The annotations of the cell clusters require extensive manual literature search in practice and the annotation quality may be dependent on users' domain knowledge [20]. Therefore, an interpretable and flexible model is needed. In the current work, we consider model interpretability as whether the model parameters can be directly used to associate the input features with latent factors or target outcome. In particular, latent topic models are a popular approach in mining genomic data [21, 22] as one can use them to infer the topic distribution for both the samples and genomic features by problematically decomposing the samples-by-features matrix into samples-by-topics and topics by features, respectively. However, their values in modeling scRNA-seq data have not been fully realized [23].

Third, model transferability is an important consideration. We consider a model as *transferable* if the learned knowledge manifested as the model parameters could benefit future data modeling. In the context of scRNA-seq data analysis, it translates to learning feature representations from one or more large-scale annotated reference datasets and applying the learned representations to a query dataset without annotation. As the number and size of scRNA-seq datasets continue to increase, there is an increasingly high demand for efficient exploitation and knowledge transfer from the existing reference datasets.

Several recent methods have attempted to address these challenges. Seurat [7] uses canonical correlation analysis to project cells onto a common embedding, then identifies, filters, scores and weights anchor cell pairs between batches to perform data integration. Harmony [24] iterates between maximum diversity clustering and a linear batch correction based on the mixture-of-experts model. Scanorama [10] performs all-to-all dataset matching by querying nearest neighbors of a cell among all remaining batches, after which it merges the batches with a Gaussian kernel to form a single cell panorama. These methods are often not scalable to cope with the entire genes-by-cells data matrices, or are vulnerable to the noise inherent to scRNA-seq read count data; hence they rely on feature (gene) selection and/or dimensionality reduction methods. They are also non-transferable, meaning the knowledge learned from one dataset cannot be easily transferred through model parameters to benefit the modeling of another dataset. LIGER [9] uses integrative non-negative matrix factorization to jointly factorize multiple scRNA-seq matrices across conditions using genes as the common axis, linking cells from different conditions by a common set of latent factors also known as metagenes. Although relying on Seurat's preprocessing pipeline, LIGER is weakly transferable in the sense that the global metagenes-by-genes matrix can be transferred when modeling new datasets, whereas in the case of Seurat both the correlation components and the anchor cell pairs must be recomputed.

Deep learning approaches, especially autoencoders, have demonstrated promising perfor-

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mance in scRNA-seq data modeling. scAlign [15] and MARS [25] encode cells with non-linear embeddings using autoencoders, which is naturally transferable across datasets. While scAlign minimizes the distance between the pairwise cell similarity at the embedding and original space, MARS looks for latent landmarks from known cell types to infer cells of unknown type. Variational autoencoders (VAE) [26] is an efficient probabilistic framework known to better account for noise compared to conventional autoencoders. scVAE-GM [11] changed the prior distribution of the latent variables in the VAE from Gaussian to Gaussian mixture, adding a categorical latent variable that clusters cells. Single-cell variational inference (scVI), another VAE-based method, models library size and takes into account batch effect in generating cell embeddings [6]. A key drawback for autoencoders for modeling scRNA-seq data is the lack of interpretability. These approaches often require posthoc analyses to interpret the learned model parameters and associate condition and cell-type-specific gene signatures. To improve interpretability, a linear decoded VAE (hereafter referred to as scVI-LD) was proposed and included in the scVI software package [14].

In this paper, we present *single-cell Embedded Topic Model* (scETM), a generative topic model that facilitates integrative analysis of large-scale single-cell transcriptomic data. Our key contribution is the novel, efficient and scalable Bayesian inference framework which utilizes a transferable neural-network-based encoder while having an interpretable linear decoder. scETM simultaneously learns a set of highly interpretable cell embeddings, gene embeddings, topic embeddings, and batch effect embeddings from scRNA-seq data. The flexibility and expressiveness of the encoder network enable us to model extremely large raw scRNA-seq datasets. By the tri-factorization design, we are able to incorporate existing pathway information into gene embeddings during the model training to further improve interpretability, which is a salient feature compared to the related methods such as scVI-LD. This incorporation allows scETM to simultaneously discover interpretable cellular signatures and gene markers while integrating scRNA-seq data across conditions, subjects and experimental studies. We demonstrate that scETM offers state-of-the-art performance across a diverse range of datasets with desirable runtime and memory requirements. We also show scETM's capability of effective knowledge transfer across datasets with different sequencing technologies and even cross-species. We then use scETM to discover biologically meaningful gene expression signatures and to differentiate known cell types as well as pathological conditions. We analyze scETM-inferred topics and show that several topics are enriched in cell-type-specific or disease-related pathways. Finally, we directly incorporate known pathway-gene relationships (pathway gene sets) into scETM in the form of gene embeddings, and use the learned pathway-topic embedding to show the pathway-informed scETM (p-scETM)'s capability of learning biologically and pathologically meaningful information.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

scETM model overview

Topic models are natural for scRNA-seq data modeling. Each sampled cell transcriptome can be viewed as a bag of genes, and its cell type identity could be inferred from its topic proportions. Each topic is a distribution over genes that would capture certain aspect of cell functions. We choose Embedded Topic Model (ETM) [27] as the backbone of our model, as it inherits the benefits of topic models, and is especially effective for handling large and heavy-tailed vocabularies. The amortized inference process of ETM is very similar to that of VAEs, while the data modeling of the former is much more interpretable. We model the cells-by-genes read-count matrix by factorizing it into a cells-by-topics matrix $\boldsymbol{\theta}$ and a topics-by-genes matrix $\boldsymbol{\beta}$, which is further decomposed into topics-by-embedding $\boldsymbol{\alpha}$ and embedding-by-genes $\boldsymbol{\rho}$ matrices (**Fig.** 1a,b). This tri-factorization design allows for the simultaneous embedding of cells, topics, and genes into low-dimensional spaces, and exploring their relations in a highly interpretable way through automatically inferred latent topics. To account for biases across conditions or subjects, we introduce an optional batch correction parameter $\boldsymbol{\lambda}$ which acts as an intercept term in the categorical softmax function to relieve the burden of modeling batch variations from the cell topic mixture $\boldsymbol{\theta}_d$. We infer the topic mixture $\boldsymbol{\theta}$ of a cell (also referred to as the cell embedding) via a two-layer fully-connected neural network (**Fig.** 1c) using VAE [26]. Details are described in **Methods**.

Clustering

We benchmarked scETM, along with seven state-of-the-art single-cell clustering or integrative analysis methods – scVI [6], scVI-LD [14], Seurat (integrated) [7], scVAE-GM [11], Scanorama [10], Harmony [24] and LIGER [9], on five published datasets, namely Mouse Pancreatic Islet (MP) [28], Human Pancreatic Islet (HP) [7], *Tabula Muris* (TM) [3], Alzheimer's Disease dataset (AD), and Major Depressive Disorder dataset (MDD) [29]. Across all datasets, scETM performs on par with the state-of-the-art methods in terms of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) (**Table** 1; **Supp. Table** S1). Specifically, it has the best clustering performance among the transferable and interpretable models. scETM stably yields competitive results, while others methods fluctuate across the five datasets. Overall, Harmony and Seurat have slightly higher ARIs than scETM, with trade-offs of model transferability, interpretability, and/or scalability (more details next).

To further verify the clustering performance and validate our evaluation metrics, we visualized the cell embeddings using Uniform Manifold Approximation and Projection (UMAP) [30] (**Fig.** S2). This result demonstrates that scETM effectively captures cell-type-specific information, while accounting for artefacts arising from individual or technological variations. scETM is also robust to hyperparameter changes, requiring very few or no hyperparameter tuning efforts when applied to unseen datasets (**Supp. Table** S2). We also performed a comprehensive

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ablation analysis to validate our model choices. The ablation experiment demonstrates the necessity of key model components, such as the batch effect correction λ and batch normalization in the encoder. Normalizing gene expression as the input to the encoder also improves the performance (**Supp. Table** S3).

Scalability

A key advantage of scETM is its high scalability and efficiency. We demonstrate this by comparing the run time, memory usage, and clustering performance of the state-of-the-art models using their recommended pipelines when integrating a merged dataset consisting of MDD and AD (**Fig.** 2; see Efficiency and scalability benchmark of the existing methods section). Because of the simple model design and efficient implementation (sparse matrix representation, multi-threaded data retrieval, to name a few tricks), scETM has the shortest run time among all deep-learning based models. Specifically, on the largest dataset (148,247 cells), it runs 3-4 times faster than scVI and scVI-LD, and over 10 times faster than scVAE-GM. Notably, although The architectures are not exactly the same in these deep models, the run time is not heavily dependent on the architecture choices but rather on the implementation. Harmony and Scanorama are the only methods faster than scETM, yet they both operate on no more than a hundred principal components, while scETM can operate on all genes for better model transferability and interpretability.

Because of stochastic variational inference [26, 31, 32] and minibatch parameter update, scETM takes almost constant run-time memory with respect to the sample size, with the increase attributed to the data loader. In contrast, the memory requirement of Seurat increases rapidly with the number of cells, due to the vast numbers of plausible anchor cell pairs in the two brain datasets. In accord with the results above, scETM consistently yields first-class clustering results, whereas Harmony and Scanorama show sub-optimal performance when dataset sizes vary. UMAP visual inspection of scVAE embeddings suggests that scVAE likely suffers from under-correction of batch effects (**Supp. Fig.** S3). The sudden drop of Liger's clustering performance in the largest benchmark dataset may be due to overfitting because of the frequentist numerical optimization of the least square objective in Liger in contrast to the Bayesian inference in ours and other approaches.

Transfer learning across single-cell datasets

A prominent feature of scETM is that its parameters, hence knowledge of modeling scRNA-seq data, are transferable across datasets. As an example, we trained an scETM model on the fluorescence-activated-cell-sorting-based *Tabula Muris* dataset (TM-FACS) from a multi-organ mouse single-cell atlas, and evaluated it using the MP data, which only contains mouse pancreatic islet cells (see Transfer learning with scETM section). Though the two datasets were obtained using different sequencing technologies, the model yields an encouragingly high ARI score of 0.94, considering that the ARI score is 0.95 if the model is directly trained on MP

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**Fig.** 3a,b). Interestingly, the TM-FACS-pretrained model puts B cells, T cells and macrophages far away from other clusters and separates B cells and T cells from macrophages, which is not observed in the model directly trained on MP.

Our transfer learning can also be cross-species. Using the gene orthologs between human and mouse, scETM trained on the human pancreas (HP) dataset achieves a 0.79 ARI on MP (**Fig.** 3c), which surpassed scVI (0.39) and scVI-LD (0.49) on the same HP-MP transfer learning task. The performance is even better than scVAE-GM (0.66) trained directly on MP. This improvement is attributable to the gene embedding learning and the explicit batch effect correction in our scETM model. To assess scETM's capability to embed unseen query cells and similar reference cells together in the embedding space, we trained a k-Nearest Neighbors classifier on the HP embeddings generated by the HP-pretrained scETM model, and evaluated it on the MP embeddings generated by the same HP-pretrained model, which was not trained on the MP data. The classifier achieves 79.8% accuracy in MP cell type prediction, demonstrating the capability of automatically annotating query scRNA-seq datasets using the pretrained scETM model on the reference scRNA-seq data. We expect that these results can be improved by further tuning of the model on the unannotated query data, or by the use of a compatible transfer learning framework such as MARS [25].

Gene set enrichment analysis of scETM topics

We next investigated whether the scETM-inferred topics are biologically relevant in terms of known human gene pathways. We conducted pathway enrichment analysis using pathDIP4, a data portal that integrates 24 major pathway databases [33]. For each topic, we selected the top 30 genes based on topic intensity as the input gene set and identified significantly enriched pathways based on a hypergeometric test with false discover rate (FDR) below 0.05 [34]. We found that several topics learned from the human pancreatic islet dataset are significantly enriched in pathways relevant to pancreas functions, including insulin signalling pathway, fat digestion and absorption, starch and sucrose metabolism, etc (**Supp. Table** S4). Topic learned from AD and MDD datasets are also enriched in brain function-related pathways: about 64% and 37% of the topics respect to AD and MDD have significant hits with neuronal system pathways (**Supp. Table** S5, S6).

Interestingly, several topics are also enriched for disease-relevant pathways. In AD, the top 30 genes from topic 18 are enriched for Alzheimer's Disease pathway itself, and the top 30 genes from topic 75 are highly enriched in amyloid fiber formation (**Fig.** 4a). Notably, amyloid fibrils are widely known to be associated with aging and AD, and β-amyloid plaques are among the major characteristics of AD brains [35–37]. We also found that the top 30 genes in topic 15 are enriched in the GABA synthesis pathway (FDR < 0.001), which is known to have an important role in AD pathogenesis [38, 39]. In MDD, topic 7 is enriched for neurodegenerative diseases such as Parkinson's, Alzheimer's and Huntington's disease; topic 94 is enriched in toll-like receptor (TLR) pathway (FDR < 0.05), which is known to be associated with MDD severity [40, 41] (**Fig.** S1a).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Differential scETM topics in disease conditions and cell types

We sought to use the scETM topics to differentiate pathological conditions. We separated the cells derived from the AD subjects from the cells derived from the control subjects. We then performed two-sided t-tests to evaluate whether the two cell groups exhibit significant differences in terms of their topic expression (see Differential analysis of topic expression section). Here we consider the topic expression for each cell as the *metagene* expression because we projected the original gene expression of each cell onto the topic embedding (**Fig.** 4b), and we also observed that each of these topics is highly selective of a small fraction of the genes (**Fig.** 4a). We found that topic 12 and 58 are differentially expressed in the AD cells and control cells (**Fig.** 4c, d; t-test p-value ≈ 0). Interestingly, topic 58 is highly enriched for mitochondrial genes. Indeed, it is known that β-amyloids selectively build up in the mitochondria in the cells of AD-affected brains [42]. The MDD topics 52, 68, 77, 82, 86 also exhibit differential expressions between the suicidal and healthy populations (**Supp. Fig.** S1c) and interesting neurological pathway enrichment (**Supp. Table** S6).
We also identified several cell-type-specific scETM topics from the AD and MDD datasets. In AD, as shown by both the cell embedding heatmap and the differential expression analysis (**Fig.** 4b), topics (or metagenes) 19, 50, 97 are up-regulated in oligodendrocytes, endothelial cells, and oligodendrocyte progenitor cells (OPCs), respectively (t-test Bonferroni q-value ≈ 0; **Fig.** 4, **Supp. Fig.** S4). Interestingly, two subpopulations of cells that exhibit high expression of topics 12 and 58 colocalize within the oligodendrocytes and one of the excitatory subclusters, respectively (**Supp. Fig.** S5). Meanwhile, a clear separation of AD and controls within those two subpopulations is present in the cell embedding. Among the AD cells, there is also a strong enrichment for the female subjects, which is consistent with the original finding [43]. For MDD, topics 1, 20 and 72 are up-regulated in astrocytes, oligodendrocytes and OPCs, respectively (**Supp. Fig.** S1c). This is consistent with the positive correlations observed in the heatmap (**Supp. Fig.** S1b).

Pathway-informed scETM topics

In the above analysis of the MDD dataset, we found that several topics are dominated by long non-coding RNAs (lincRNAs) (**Fig.** S1a). While previous studies have suggested that lincRNAs can be cell-type-specific [44], it remains difficult to interpret them [45]. This prompted us to incorporate the known pathway information in the form of gene embedding. In particular, we fixed the gene embedding ρ to a pathways-by-genes matrix obtained from the pathDIP4 pathway database (see Incorporation of pathway knowledge section) [33, 46] and learn only the pathways by topics embedding α, which provides a direct interpretation of disease-pathways associations. We tested our pathway-informed scETMs (p-scETM) on the HP, AD and MDD datasets. Without compromising the clustering performance (**Supp. Table** S8), p-scETM learned functionally meaningful topics as shown by the pathway-topic embedding α (**Fig.** 5). In the topic α-embedding inferred by p-scETM trained on HP, we found 6 topics with top pathways related

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to insulin signaling, and 6 topics related to nutrient digestion and metabolism (**Fig.** 5a). In the MDD α embedding, we found 8 topics with top pathways known as therapeutic targets for MDD treatment (**Fig.** 5b, **Supp. Table** S9) [47–55], 2 topics with top pathways related to MDD pathogenesis [53, 56], and 3 topics with top pathways correlated with MDD [57–59]. Notably, the top pathway in topic 40, “beta-2 adrenergic receptor signaling", is also statistically enriched (p=0.021) in MDD genome-wide association studies (GWAS) [60].
In the AD topic α-embedding, we found 6 topics with top pathways related to AD treatment and 1 topic related to AD pathogenesis (**Fig.** 5c, **Supp. Table** S10) [38, 39, 61–65]. Importantly, “Alzheimer disease-amyloid secretase" pathway, which is directly related to AD pathogenesis [66], is the seventh-highest expressed pathway in topic 9. Therefore, the p-scETM inferred topics are highly related with not only the primary tissue types but also the disease of interests, although overall the former case tends to be the predominant signals we observe in our analyses.

Discussion

As scRNA-seq technologies become increasingly affordable and accessible, large-scale datasets have emerged. This challenges traditional statistical approaches and calls for robust, reliable and scalable representation learning methods to mine the latent biological knowledge from the vast amount of scRNA-seq data. To address this challenge, we developed scETM and demonstrated its state-of-the-art performance on the unsupervised clustering task across diverse datasets. scETM demonstrates excellent capabilities of batch effect correction and knowledge transfer across datasets. Many integration methods require running on both reference and query datasets to perform posthoc analyses such as joint clustering and label transfer [7, 9, 10, 24]. In contrast, our method enables a direct knowledge transfer of the reference-based pretrained parameters in annotating a new dataset, which is more efficient than the existing methods. Recently proposed by [23], single-cell Hierarchical Poisson Factor (scHPF) model applies hierarchical Poisson factorization to discover interpretable gene expression signatures in an attempt to address the interpretability challenge. However, compared to our model, scHPF lacks the flexibility in learning the gene embedding and incorporating existing pathway knowledge, and is not designed to account for batch effects. Moreover, scETM has the benefits of both interpretability in the linear decoder and flexibility in the neural network encoder. Our qualitative experiments show that scETM topics preserve cell functional and state-specific biological signals within single-cell transcriptome profiles. By seamlessly incorporating the known pathway information in the gene embedding, p-scETM finds biologically and pathologically important pathways without the need for posthoc analyses. Together, with the scalability and interpretability, scETM serves as a useful tool for large-scale single-cell transcriptome analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

scETM data generative process

To model scRNA-seq data distribution, we take a topic-modeling approach [67]. In our framework, each cell is considered as a “document", each scRNA-seq read as a “token" in the document, and the gene that gives rise to the read is considered as a “word" from the vocabulary of size V. We assume that each cell can be represented as a mixture of latent cell types, which are commonly referred to as the latent topics. The original LDA model [67] defines a fixed set of K independent Dirichlet distributions $\boldsymbol{\beta}$ over a vocabulary of size V. Following the ETM model, here we decompose the unnormalized topic distribution $K \times V$ $\boldsymbol{\beta}^*$ into the topic embedding $\boldsymbol{\alpha} \in \mathbb{R}^{K \times L}$ and gene embedding $\boldsymbol{\rho} \in \mathbb{R}^{L \times V}$, where L denotes the size of the embedding space. Therefore, the unnormalized probability of a gene belonging to a topic is proportional to the dot product between the embeddings of the topic and gene. Formally, the data generating process of each scRNA-seq profile d is:

1. Draw a latent cell type proportion $\boldsymbol{\theta}_d$ for a cell d from logistic normal $\boldsymbol{\theta}_d \sim \mathcal{LN}(0, \mathbf{I})$:

$$\boldsymbol{\delta}_d \sim \mathcal{N}(0, \boldsymbol{I}), \quad \boldsymbol{\theta}_d = \text{softmax}(\boldsymbol{\delta}_d) = \frac{\exp(\delta_{d,k})}{\sum_{k=1}^{K} \exp(\delta_{d,k})} \tag{1}$$

2. For each gene g in cell d, draw its expression from a categorical distribution:

$$y_{d,g} \sim \text{Cat}(r_{d,g}) = \prod_{i=1}^{N_d} r_{d,g}^{[w_{i,d}=g]} = r_{d,g}^{\sum_i [w_{i,d}=g]} = r_{d,g}^{y_{d,g}} \tag{2}$$

Here N_d is the library size of cell d, $w_{i,d}$ is the index of the gene that gives rise to the ith read in cell d (i.e., $[w_{i,d} = g]$), and $y_{d,g}$ is the total read counts of gene g in cell d. The transcription rate $r_{d,g}$ is parameterized as follows:

$$r_{d,g} = \frac{\exp(\hat{r}_{d,g})}{\sum_{g'} \exp(\hat{r}_{d,g'})}, \quad \hat{r}_{d,g} = \boldsymbol{\theta}_d \boldsymbol{\alpha} \boldsymbol{\rho}_g + \lambda_{s(d),g} \tag{3}$$

Here $\boldsymbol{\theta}_d$ is the $1 \times K$ cell topic mixture for cell d, $\boldsymbol{\alpha}$ is the global $K \times L$ cell topic embedding variable, $\boldsymbol{\rho}_g$ is a $L \times 1$ gene-specific transcriptomic embedding, and $\lambda_{s(d),g}$ is the batch-dependent and gene-specific scalar effect, where $s(d)$ indicates the batch index for cell d. Notably, to model the sparsity of gene expression in each cell (i.e., only a small fraction of the genes have non-zero expression), we use the softmax function to normalize the transcription rate over all of the genes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

scETM model inference

In scETM, we treat the latent cell type mixture $\boldsymbol{\theta}_d$ for each cell d as the only latent variable. We treat the topic embedding $\boldsymbol{\alpha}$, the gene-specific transcriptomic embedding $\boldsymbol{\rho}$, and the batch-effect $\boldsymbol{\lambda}$ as point estimates. Let $\mathbf{Y}$ be the $D \times V$ gene expression matrix for D cells and V genes. The posterior distribution of the latent variables $p(\boldsymbol{\Theta}|\mathbf{Y})$ is intractable. Hence, we took a variational inference approach using a proposed distribution $q(\boldsymbol{\delta}_d)$ to approximate the true posterior. Specifically, we define the following proposed distribution: $q(\boldsymbol{\delta} \mid \mathbf{y}) = \prod_d q(\delta_d|\mathbf{y}_d)$, where $q(\boldsymbol{\delta}_d|\mathbf{y}_d) = \boldsymbol{\mu}_d + \text{diag}(\boldsymbol{\sigma}_d)\mathcal{N}(0, I)$ and $[\boldsymbol{\mu}_d, \log \boldsymbol{\sigma}_d^2] = \mathrm{NNET}(\tilde{\mathbf{y}}_d; \mathbf{W}_\theta)$. Here $\tilde{\mathbf{y}}_d$ is the normalized gene expression as the read counts for each gene divided by the total reads in cell d. The function $\mathrm{NNET}(\mathbf{v}; \mathbf{W})$ is a two-layer feed-forward neural network used to estimate the sufficient statistics of the proposed distribution for the cell topic mixture $\boldsymbol{\theta}_d$.
To learn the above variational parameters $\mathbf{W}_\theta$, we optimize the evidence lower bound (ELBO) of the log likelihood, which is equivalent to minimizing the Kullback-Leibler (KL) divergence between the true posterior and the proposed distribution: $\mathrm{ELBO} = \mathbb{E}_q[\log p(\mathbf{Y}|\Theta)] - \mathrm{KL}\left[q(\Theta|\mathbf{Y})||p(\Theta)\right]$. The Bayesian model is learned by maximizing the reconstruction likelihood with regularization in the form of KL divergence of the proposed distribution from the prior. For computational efficiency, we optimize ELBO with respect to the variational parameters by amortized variational inference [26, 31, 32]. Specifically, we draw a sample of the latent variables from $q(\boldsymbol{\delta} \mid \boldsymbol{y})$ for a minibatch of cells from reparameterized Gaussian proposed distribution $q(\boldsymbol{\delta} \mid \mathbf{y})$ [26], which has the mean and variance determined by the NNET functions. We then use those draws as the noisy estimates of the variational expectation for the ELBO. The optimization is then carried out by back-propagating the ELBO gradients into the variational parameters.

scETM implementation details

We implemented scETM using the PyTorch library. We chose the encoder to be a 2-layer neural network, with hidden sizes of $(256, 128)$, ReLU activations [68], 1D batch normalization [69], and 0.1 dropout rate between layers. We set the gene embedding dimension to 300, and the number of topics to 100. We optimize our model with Adam Optimizer and a 0.02 learning rate. To prevent over-regularization, we start with zero weight penalty on the KL divergence and linearly increase the weight of the KL divergence in the ELBO loss function during the first 300 epochs. We show that our model is robust to changes in the above hyperparameters (**Supp. Table** S2). During the evaluation, we used the variational mean of the unnormalized topic mixture $\boldsymbol{\mu}_d$ as the scETM cell embedding for cell d. With a minibatch size of 2000, scETM typically needs 5k-20k training steps to converge.

Transfer learning with scETM

We trained scETM on the MP dataset and visualized the cell embedding using UMAP (**Fig.** 3a). For TM-FACS, we first subset the genes of both TM-FACS and MP to their intersection (13263

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes). We then trained scETM on the processed TM-FACS and evaluated it on MP (**Fig.** 3b). For HP, we first matched the orthologous genes (12603 genes) based on the Mouse Genome Informatics database [70, 71]. We then trained and evaluated scETM on the aligned gene sets (**Fig.** 3c). The k was set to 5 for the k-NN classifier trained on the reference embeddings and the reference cell types and used to predict cell types of the query cells.

Differential analysis of topic expression

We aimed to identify topics that are differentially associated with known cell type labels or disease conditions. For each label (e.g., AD positive), we first separated the cells into cells with the label and cells without the label. We then performed two-sided t-tests to evaluate whether the cells with the label exhibit significantly higher or lower topic expression relative to the cells without the label. Here we used the Gaussian topic expression (i.e., $\boldsymbol{\delta}$) without the softmax transformation because it is more suitable to the normality assumption of the t-test. We determine a topic to be differentially expressed (DE) if the Bonferroni corrected p-value is lower than 0.01 (i.e., q-value $<$ 0.01). **Supp. Table** S7 summarizes the number of DE topics we identified for each cell type and disease conditions from the AD and MDD data.

Incorporation of pathway knowledge

We downloaded the pathDIP4 pathway database from [46]. Pathway gene sets containing fewer than five genes were removed. We represent the pathway knowledge as a pathways-by-genes $\boldsymbol{\rho}$ matrix, where $\rho_{ij} = 1$ if gene set i contains gene j, and $\rho_{ij} = 0$ otherwise. For the fixed-rho version of p-scETM, we fix the gene embedding matrix $\boldsymbol{\rho}$ to the pathways-by-genes matrix.

Clustering performance benchmark of the existing methods

We assessed the performance of each method by three metrics: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). ARI [72] and NMI are widely-used representatives of two families of clustering agreement measures, pair-counting and information theoretic measures, respectively. A high ARI or NMI indicates a high degree of agreement for a given clustering result against the ground-truth cell type labels.
All embedding plots were generated using the Python scanpy package [16]. We use UMAP [30] to reduce the dimension of the embeddings to 2 for visualization, and Louvain [73] and Leiden [74] to cluster the cell embeddings. During clustering, we tried multiple resolution values and reported the result with the highest ARI for each method. We ran all methods under their default pipeline settings (see Experimental details of other scRNA-seq methods), and we use batch correction option whenever applicable to account for batch effects. All results are obtained on a compute cluster with Intel Gold 6148 Skylake CPUs and Nvidia V100 GPUs. We limit each experiment to use 8 CPU cores, 128 GB RAM and 1 GPU.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Efficiency and scalability benchmark of the existing methods

To create a benchmark dataset for evaluating the run time of each method, we merged MDD and AD, keeping genes that appear in both datasets. We then selected 3000 most variable genes using scanpy's `highly_variable_genes(n_top_genes=3000, flavor='seurat_v3')` function, and randomly sampled 28,000, 14,000, 70,000 and 148,247 (all) cells to create our benchmark datasets. The memory requirements reported in **Fig.** 2 were obtained by reading the `VmRSS` entry in `/proc/[pid]/status` at the end of each process. We kept the same experimental settings (RAM size, number of GPUs, etc) as in the Clustering performance benchmark of the existing methods section above.

Funding

YL is supported by New Frontier Research Fund - Exploration (NFRFE-2019-00980). YZ is supported by Jacqueline Johnson Desoer Science Undergraduate Research Award (SURA).

Availability of data and materials

scETM source codes as well as the benchmarking workflows have been deposited at the GitHub repository (`https://github.com/li-lab-mcgill/scETM`). The datasets analysed during the current study are from publicly available repositories or data portals. The acquisition and quality control steps for all datasets are included in the supplementary information.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Authors' contributions

YL and JT conceived of the study. YZ, HC, YL analyzed and interpreted the data, wrote the manuscript, and wrote the code for scETM. HC optimized and completed the final scETM code. ZZ ran some initial experiments. YL and JT supervised the project. All authors approved the final manuscript.

References

1. Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

2. Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, et al. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

3. Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367, 2018.

4. Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.

5. Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature News*, 550(7677):451, 2017.

6. Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

7. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

8. Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

9. Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

10. Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11. Christopher H Grønbech, Maximillian F Vording, Pascal N Timshel, Capser K Sønderby, Tune H Pers, and Ole Winther. scvae: Variational auto-encoders for single-cell gene expression datas. *bioRxiv*, page 318295, 2018.

12. Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

13. Zhe Sun, Li Chen, Hongyi Xin, Yale Jiang, Qianhui Huang, Anthony R Cillo, Tracy Tabib, Jay K Kolls, Tullia C Bruno, Robert Lafyatis, et al. A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. *Nature communications*, 10(1):1–10, 2019.

14. Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

15. Nelson Johansen and Gerald Quon. scalign: a tool for alignment, integration, and rare cell identification from scrna-seq data. *Genome biology*, 20(1):1–21, 2019.

16. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

17. Peng Qiu. Embracing the dropouts in single-cell rna-seq analysis. *Nature communications*, 11(1):1–9, 2020.

18. Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. Assessment of batch-correction methods for scrna-seq data with a new test metric. *BioRxiv*, page 200345, 2017.

19. Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7:39921, 2017.

20. Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

21. Daniel Backenroth, Zihuai He, Krzysztof Kiryluk, Valentina Boeva, Lynn Pethukova, Ekta Khurana, Angela Christiano, Joseph D Buxbaum, and Iuliana Ionita-Laza. FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *The American Journal of Human Genetics*, 102(5):920–942, 2018.

22. Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, 16(5):1–14, April 2019.

23. Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, et al. De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular systems biology*, 15(2):e8557, 2019.

24. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, pages 1–8, 2019.

25. Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature Methods*, pages 1–7, 2020.

26. Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv.org*, December 2013.

27. Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*, 2019.

28. Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

29. Corina Nagy, Malosree Maitra, Arnaud Tanti, Matthew Suderman, Jean-Francois Théroux, Maria Antonietta Davoli, Kelly Perlman, Volodymyr Yerko, Yu Chang Wang, Shreejoy J Tripathy, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nature Neuroscience*, pages 1–11, 2020.

30. L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

31. R Ranganath, S Gerrish, D Blei Artificial Intelligence Statistics, , and 2014. Black box variational inference. *jmlr.org*.

32. M D Hoffman, D M Blei, C Wang, and J W Paisley. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)*, 2013.

33. Sara Rahmati, Mark Abovsky, Chiara Pastrello, Max Kotlyar, Richard Lu, Christian A Cumbaa, Proton Rahman, Vinod Chandran, and Igor Jurisica. pathdip 4: an extended pathway annotations and enrichment analysis resource for human, model organisms and domesticated species. *Nucleic acids research*, 48(D1):D479–D488, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

34. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

35. Dennis J Selkoe. The molecular pathology of alzheimer's disease. *Neuron*, 6(4):487–498, 1991.

36. M Fändrich. On the structural definition of amyloid fibrils and other polypeptide aggregates. *Cellular and Molecular Life Sciences*, 64(16):2066–2078, 2007.

37. Sian-Yang Ow and Dave E Dunstan. A brief overview of amyloids and alzheimer's disease. *Protein Science*, 23(10):1315–1331, 2014.

38. Yanfang Li, Hao Sun, Zhicai Chen, Huaxi Xu, Guojun Bu, and Hui Zheng. Implications of gabaergic neurotransmission in alzheimer's disease. *Frontiers in aging neuroscience*, 8:31, 2016.

39. Karan Govindpani, Beatriz Calvo-Flores Guzmán, Chitra Vinnakota, Henry J Waldvogel, Richard L Faull, and Andrea Kwakowsky. Towards a better understanding of gabaergic remodeling in alzheimer's disease. *International journal of molecular sciences*, 18(8):1813, 2017.

40. Yi-Yung Hung, Hong-Yo Kang, Kai-Wei Huang, and Tiao-Lai Huang. Association between toll-like receptors expression and major depressive disorder. *Psychiatry research*, 220(1-2):283–286, 2014.

41. B García Bueno, JR Caso, José Luis Muñoz Madrigal, and Juan Carlos Leza. Innate immune receptor toll-like receptor 4 signalling in neuropsychiatric diseases. *Neuroscience & Biobehavioral Reviews*, 64:134–147, 2016.

42. Xi Chen and Shi Du Yan. Mitochondrial aβ a potential cause of metabolic dysfunction in alzheimer's disease. *IUBMB life*, 58(12):686–694, 2006.

43. Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

44. Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927, 2011.

45. Elena Perenthaler, Soheil Yousefi, Eva Niggl, and Stefan Barakat. Beyond the exome: the non-coding genome and enhancers in malformations of cortical development. *Frontiers in cellular neuroscience*, 13:352, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46. pathdip 4.0 database. `http://ophid.utoronto.ca/pathDIP/Download.jsp`. accessed 23 oct 2020.

47. Domenico De Berardis, Laura Orsolini, Nicola Serroni, Gabriella Girinelli, Felice Iasevoli, Carmine Tomasetti, Andrea de Bartolomeis, Monica Mazza, Alessandro Valchera, Michele Fornaro, et al. A comprehensive review on the efficacy of s-adenosyl-l-methionine in major depressive disorder. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, 15(1):35–44, 2016.

48. Loris A Chahl. Trp channels and psychiatric disorders. pages 987–1009. Springer, 2011.

49. J Craig Nelson, Carolyn M Mazure, and Peter I Jatlow. Desipramine treatment of major depression in patients over 75 years of age. *Journal of clinical psychopharmacology*, 15(2):99–105, 1995.

50. J Craig Nelson. Use of desipramine in depressed inpatients. *The Journal of Clinical Psychiatry*, 1984.

51. Julie A Blendy. The role of creb in depression and antidepressant treatment. *Biological psychiatry*, 59(12):1144–1150, 2006.

52. Jakob M Koch, Susanne Kell, Dunja Hinze-Selch, and Josef B Aldenhoff. Changes in creb-phosphorylation during recovery from major depression. *Journal of psychiatric research*, 36(6):369–375, 2002.

53. John Q Wang and Limin Mao. The erk pathway: molecular mechanisms and treatment of depression. *Molecular neurobiology*, 56(9):6197–6205, 2019.

54. Emmanuelle Goubert, Marc Altvater, Marie-Noelle Rovira, Ilgam Khalilov, Morgane Mazzarino, Anne Sebastiani, Michael KE Schaefer, Claudio Rivera, and Christophe Pellegrino. Bumetanide prevents brain trauma-induced depressive-like behavior. *Frontiers in molecular neuroscience*, 12:12, 2019.

55. Elliot Ehrich, Ryan Turncliff, Yangchun Du, Richard Leigh-Pemberton, Emilio Fernandez, Reese Jones, and Maurizio Fava. Evaluation of opioid modulation in major depressive disorder. *Neuropsychopharmacology*, 40(6):1448–1455, 2015.

56. Matthew N Hill, Gregory E Miller, W-S Vanessa Ho, Boris B Gorzalka, and Cecilia J Hillard. Serum endocannabinoid content is altered in females with depressive disorders: a preliminary report. *Pharmacopsychiatry*, 41(2):48, 2008.

57. Richard Day, Patricia A Ganz, and Joseph P Costantino. Tamoxifen and depression: more evidence from the national surgical adjuvant breast and bowel project's breast cancer prevention (p-1) randomized study. *Journal of the National Cancer Institute*, 93(21):1615–1623, 2001.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

58. Olga Szczesniak, Knut A Hestad, Jon Fredrik Hanssen, and Knut Rudi. Isovaleric acid in stool correlates with human depression. *Nutritional neuroscience*, 19(7):279–283, 2016.

59. Naomi Breslau, M Marlyne Kilbey, and Patricia Andreski. Nicotine dependence, major depression, and anxiety in young adults. *Archives of general psychiatry*, 48(12):1069–1074, 1991.

60. Anqi Qiu, Mojun Shen, Claudia Buss, Yap-Seng Chong, Kenneth Kwek, Seang-Mei Saw, Peter D Gluckman, Pathik D Wadhwa, Sonja Entringer, Martin Styner, et al. Effects of antenatal maternal depressive symptoms and socio-economic status on neonatal brain development are modulated by genetic risk. *Cerebral Cortex*, 27(5):3080–3092, 2017.

61. Giancarlo V De Ferrari and Nibaldo C Inestrosa. Wnt signaling function in alzheimer's disease. *Brain research reviews*, 33(1):1–12, 2000.

62. Boris Decourt, Debomoy K Lahiri, and Marwan N Sabbagh. Targeting tumor necrosis factor alpha for alzheimer's disease. *Current Alzheimer Research*, 14(4):412–425, 2017.

63. Sumiti Saharan and Pravat K Mandal. The emerging role of glutathione in alzheimer's disease. *Journal of Alzheimer's Disease*, 40(3):519–529, 2014.

64. Jennifer C Palmer, Rachel Barker, Patrick G Kehoe, and Seth Love. Endothelin-1 is elevated in alzheimer's disease and upregulated by amyloid-β. *Journal of Alzheimer's Disease*, 29(4):853–861, 2012.

65. Johanna Michael, Julia Marschallinger, and Ludwig Aigner. The leukotriene signaling pathway: a druggable target in alzheimer's disease. *Drug discovery today*, 24(2):505–516, 2019.

66. RM Damian Holsinger, Catriona A McLean, Konrad Beyreuther, Colin L Masters, and Geneviève Evin. Increased expression of the amyloid precursor β-secretase in alzheimer's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 51(6):783–786, 2002.

67. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

68. Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings.

69. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

70. Cynthia L Smith and Janan T Eppig. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):390–399, 2009.

71. Mouse genome informatics database. `http://www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt`. accessed 30 nov 2020.

72. Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

73. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

74. Vincent Traag, L. Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 03 2019.

75. Seurat3.0 finding integration vectors: long vectors not supported yet number 1029. `https://github.com/satijalab/seurat/issues/1029`. accessed 5 jan 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

Figure 1: **scETM model overview.** (a) Probabilistic graphical model of scETM. We model the scRNA-profile read count matrix $y_{d,g}$ in cell d and gene g across S subjects or studies by a multinomial distribution with the rate parameterized by cell topic mixture θ, topic embedding α, gene embedding ρ, and batch effects λ. (b) Matrix factorization view of scETM. (c) Encoder architecture for inferring the cell topic mixture θ.

Figure 2: **Benchmark of the efficiency and scalability of seven scRNA-seq clustering algorithms.** The line styles in the plot indicate model inputs. The number of genes was fixed to 3000 in this experiment. See Efficiency and scalability benchmark of the existing methods section for experimental details.

Figure 3: **scETM enables effective transfer learning across single-cell datasets.** The embeddings of cells in the Mouse Pancreatic islet (MP) dataset inferred by three scETM models trained on (a) MP, (b) TM-FACS, (c) HP, respectively.

Figure 4: **scETM topic embeddings of the Alzheimer's Disease snRNA-seq data. (a)** Gene topics heatmap of top 10 genes in each topic based on topic intensity. We annotated the top genes by the significantly enriched AD-related pathways per topic (rows). For visualization purposes, we divided the topic values by the maximum absolute value within the same topic. Only select topics are shown. **(b)** Topics intensity of cells (n=10,000) sub-sampled from the AD dataset. Topic intensities shown here are the Gaussian mean before applying softmax. Only the select topics with the sum of absolute values greater than 1500 across all sampled cells are shown. The three color bars show disease conditions, cell types, and batch identifiers (i.e., subject IDs). **(c)** Differential expression analysis of topics across the 8 cell types and 2 clinical conditions. Z-scores of the two-sided t-tests were shown. Asterisks indicate Bonferroni q-value < 0.05 for one-sided t-test of up-regulated topics in each cell-type and two-sided t-test for disease-relevant topics.

Figure 5: **p-scETM pathway-topics embeddings of three datasets** (a) The pathway-topics heatmap of top 5 pathways in selected topics, inferred by a p-scETM model trained on HP. Pathways related to pancreas function, insulin signalling and digestion are highlighted. (b) The pathway-topics heatmap of top 5 pathways in selected topics, inferred by a p-scETM model trained on MDD. Pathways related to MDD pathogenesis and therapeutic targets are highlighted. (c) The pathway-topics heatmap of top 7 pathways in selected topics, inferred by a p-scETM model trained on AD. Pathways related to AD pathogenesis and therapeutic targets are highlighted.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

	Transferable	Interpretable	MP	HP	TM	AD	MDD
Harmony			0.974	0.955	0.705	0.996	0.787
Scanorama			0.915	0.859	0.566	0.991	0.616
Seurat Integrated			0.954	0.968	0.700	0.997*	0.795
scVAE-GM	✓		0.878	NA	NA	0.997	0.554
scVI	✓		0.884	0.754	0.642	0.910	0.511
LIGER	✓	✓	0.901	0.904	0.579	0.933	0.718
scVI-LD	✓	✓	0.882	0.868	0.596	0.998	0.668
scETM	✓	✓	0.951	0.937	0.706	0.976	0.734

Table 1: **Model properties and unsupervised clustering performance on 5 datasets.** The clustering performance is measured by Adjusted Rand Index (ARI) between ground truth cell types and Leiden [74] clusters. NA is reported for models that did not converge. See Clustering performance benchmark of the existing methods section for experimental details. *Batch integration was turned off to prevent over-correction.

Additional Files

Additional file 1

Supplementary Information, including Tables S1-3,7-11, Figures S1-4, and supplementary methods.

Additional file 2

Table S4: scETM 100-topic enrichment for Human Pancreas scRNA-seq data.
Table S5: scETM 100-topic enrichment for Alzheimer's Disease snRNA-seq data.
Table S6: scETM 100-topic enrichment for Major Depressive Disorder snRNA-seq data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data

Yifan Zhao[1,2,†], **Huiyu Cai**[3,†], **Zuobai Zhang**[4], **Jian Tang**[5,*], **Yue Li**[1,*]

[1]School of Computer Science, McGill University, [2]Present address: Harvard-MIT Health Sciences and Technology, [3]Department of Machine Intelligence, Peking University, [4]School of Computer Science, Fudan University [5]HEC Montreal
[†]Equal contribution [*]Correspondence: jian.tang@hec.ca, yueli@cs.mcgill.ca

Supplementary Methods

.1 Data processing

All of the single-cell datasets used in this study are from publicly available repositories or data portals. We describe below the acquisition and quality control (QC) for each of the datasets used in the current work.

.1.1 Human pancreatic islet

We obtained the human pancreatic islet dataset and the ground truth cell type labels from Satija Lab at the following link: https://satijalab.org/seurat/v3.0/integration.html (accessed 1 Dec 2020), originally deposited by Stuart et al. [7]. This dataset is a compilation of scRNA-seq data from five studies which can be accessed using the following Gene Expression Omnibus (GEO) accession numbers: GSE81076 (CelSeq), GSE85241 (CelSeq2), GSE86469 (Fluidigm C1), E-MTAB-5061 (SMART-Seq2), and GSE84133 (inDrops). A QC step was conducted by [7], and no additional QC was performed. In our benchmarking experiment, we use the different scRNA-seq technologies as the batch variable.

.1.2 Mouse pancreatic islet

We obtained the mouse pancreatic islet data and ground truth cell type labels from GSE84133 (inDrops) without conducting an additional QC step. There are 1,886 mouse cells from two mice of different strains, ICR and C57BL/6 [28]. The cell counts from the two trains are of approximately equal proportions. In our benchmarking experiment, we treated the mouse strain as the batch variable because of the different genetic backgrounds.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

.1.3 Major Depressive Disorder (MDD)

We obtained the 10X Genomics-based MDD snRNA-seq dataset with ground truth cell type labels from GSE144136. A strict QC step was conducted in the original empirical study by [29], where cells with fewer than 110 detected genes were removed. The top 0.5% of cells based on the total number of UMI (unique molecular identifiers) detected in each cell were also excluded because they are likely to be multiplets rather than single nuclei. No additional QC was performed. The MDD dataset consists of 78,886 cells from the dorsolateral prefrontal cortex of 34 male participants. The participants in the control group (n=17) who died due to natural cause and case group (n=17) who died by suicide were matched for age (18–87 years), postmortem interval (12–93h) and brain pH (6–7.01) [29]. The number of cells from each donor is approximately the same.

.1.4 Alzheimer's disease (AD)

We obtained the droplet-based AD snRNA-seq data and the corresponding ground truth cell type labels from Synapse (`https://www.synapse.org/#!Synapse:syn18485175`) under the doi `10.7303/syn18485175`, and the metadata from `https://www.synapse.org/#!Synapse:syn3157322`. A strict QC step based on UMI counts and mitochondrial ratio values was conducted in the original empirical study by Mathys et al. [43]. The AD dataset consists of 70,634 cells from the prefrontal cortex of 48 individuals, both male and female, in the Religious Order Study (ROS) or the Rush Memory and Aging Project (MAP), two longitudinal cohort studies of aging and dementia. The cases group consists of 24 individuals with high levels of β-amyloid and other pathological hallmarks of AD, and the control group consists of 24 individuals who have no or very low β-amyloid or other pathologies.
Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at `www.radc.rush.edu`.

.1.5 *Tabula Muris*

We obtained the *Tabula Muris* dataset with ground truth cell type labels from FigShare (`https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733`) for the Version 2 release [3]. This dataset includes mouse single-cell transcriptome data sequenced by two tech-

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nologies: microfluidic droplet-based, and fluorescence-activated cell sorting (FACS)-based. A QC cutoff was applied in the original empirical study where only cells with at least 500 genes and 50,000 reads are kept. The droplet subset includes data for 422,803 droplets, 55,656 of which passed the QC cutoff. The TM-FACS subset contains data for 53,760 cells, 44,879 of which passed the QC cutoff.

.2 Experimental details of other scRNA-seq methods

Neural-network based models, including scVI, scVI-LD, scVAE-GM and scETM typically need at least 5000 gradient updates to converge. When running on small datasets, the total number of gradient updates per epoch may be very small (even 1). In these cases, we increase the number of epochs T to ensure the model goes through at least 5000 gradient updates, i.e. $T = \max(400, 5000\frac{B}{N})$, where N is the number of cells in the dataset and B is the mini-batch size.

.2.1 Seurat v3

We downloaded Seurat v3 (version 3.1.5) from CRAN [8]. We followed the steps outlined by the integration workflow (`https://satijalab.org/seurat/v3.2/integration.html`) which includes NormalizeData, FindVariableFeatures, FindIntegrationAnchors, and IntegrateData. To make the comparisons more equitable, we set the min.features=0 to avoid exclusion of cells. All other parameters were set as default. We noted that, with batch integration turned on, Seurat reports error in the integration step due to the high number of anchors arising from the 48 individuals (batch variable in AD), which is a known implementation issue with the standard Seurat v3 integration workflow [75]. We therefore turned off the batch integration for AD in the benchmarking experiments (see Clustering performance benchmark of the existing methods and Efficiency and scalability benchmark of the existing methods) and followed the steps described in the Guided Clustering Tutorial (`https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html`).

.2.2 Scanorama

We downloaded the source code from GitHub `brianhe/scanorama`. We used the `integrate_scanpy` function for dataset integration and batch correction as suggested by the guided tutorial. All parameters were set as default. The algorithm performs a PCA on the stacked datasets and uses 100 PCs for downstream computation.

.2.3 Harmony

We downloaded the source code from GitHub `slowkow/harmonypy` suggested by the primary repository `immunogenomics/harmony` and followed the preprocessing (normalization and top variable gene selection) described in the publication and the integration steps in the provided

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tutorial. We used the `run_harmony` function to obtain the corrected PCA embeddings and used 50 PCs as input. All other parameters were set as default.

.2.4 LIGER

We used the official implementation provided on the website of the Liger package. For the convenience of implementation, we followed the usage tutorial using Seurat Wrapper to process the raw data and then ran Liger with default parameters.

.2.5 scVI/scVI-LD

We downloaded the implementation from the Github repository `YosefLab/scVI`. We used the default model, which has one layer for both the encoder and decoder (for scVI-LD the decoder is a latent dimensions-by-genes matrix), 128 hidden units, 10 latent dimensions and ZINB distribution for modeling the data. We chose 10^{-3} as the learning rate and trained on each unprocessed dataset for 400 epochs, following the provided tutorials. We change the training batch size to 2000 for faster training. We obtained the cell embeddings via the `get_latent` method.

.2.6 scVAE

We downloaded the implementation from Github repository `scvae/scvae`. We set the hidden units to be $(256, 128)$ for the encoder. The decoder is symmetric to the encoder. Latent dimension was set to 128 to match scETM. We chose 10^{-4} as the learning rate and NB distribution for modeling the data following the authors' recommendation. We trained on each unprocessed dataset for 400 epochs with batch size of 250, including a 200-epoch warm-up for the KL divergence loss. In the scalability benchmark, we disabled the time-consuming per-epoch checkpoints to match other methods. The model did not converge on the Human Pancreatic Islet dataset, where the ELBO went to infinity. It failed to extract meaningful information from the *Tabula Muris* dataset, resulting in an ARI of 0.0.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures

Figure S1: **scETM-topic embeddings learned from the Major Depressive Disorder scRNA-seq data. (a)** Gene embedding heatmap of top 10 genes in selected topics, where genes are ordered based on topic intensity. Only topics with differential expression with respect to cell types or MDD, or with significant enrichment in MDD-related pathways are shown, and these properties are also reflected in the bottom annotations. Rows correspond to topic indices. **(b)** Cell embedding heatmap of cells (n=10,000) sampled from the MDD dataset. Only topics in which the sum of absolute intensity values across all of the sampled cells are above 1,500 are shown. Both rows and columns are clustered using average linkage hierarchical clustering and ordered accordingly. Blue arrows: topics enriched in MDD-related pathways; grey arrows: topics with DE in MDD positive population; red arrows: topics with DE with respect to cell types **(c)** DE analyses across 8 cell types and 2 clinical conditions. Z-scores of t-tests are shown. Rows correspond to topic indices. Red arrows indicate topics with DE with respect to cell types, and grey arrows indicate topics with DE with respect to MDD condtions. Asterisks indicate Bonferroni q-value < 0.05 for one-sided t-test of up-regulated topics in each cell-type and two-sided t-test for disease-relevant topics.

Figure S2: Cell embedding visualization using UMAP on the MP (left) and HP (right) datasets.

Figure S3: UMAP visualization of scETM, Liger and scVAE-GM cell embeddings on the benchmark dataset (MDD-AD) which includes 148247 cells and 3000 genes.

Figure S4: UMAP cell embedding visualization on the AD dataset, colored by differentially expressed topics (or *metagenes*) and ground truth labels for the cell types. Circled cell clusters were discussed in the main text (see Differential scETM topics in disease conditions and cell types section).

Figure S5: UMAP cell embedding visualization on the AD dataset, colored by differentially expressed topics (or *metagenes*) and AD/control or Male/Female labels. Circled cell clusters were discussed in the main text (see Differential scETM topics in disease conditions and cell types section).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Tables

	MP	HP	TM	AD	MDD
Harmony	0.675	0.816	0.765	0.371	0.586
Scanorama	0.845	0.816	0.756	0.969	0.553
Seurat Integrated	0.919	0.945	0.825	NA	0.714
scVAE-GM	0.839	NA	NA	0.989	0.518
scVI	0.846	0.778	0.834	0.920	0.501
LIGER	0.811	0.863	0.667	0.878	0.591
scVI-LD	0.844	0.839	0.792	0.990	0.620
scETM	0.902	0.896	0.832	0.986	0.620

Table S1: Normalized Mutual Information (NMI) between ground truth cell types and leiden clusters on 5 benchmark scRNA-seq datasets. NA is reported for models that did not converge.

	MP	HP
Current model	0.951	0.937
Encoder arch (128)	0.932	0.937
Encoder arch (512, 256, 128)	0.958	0.938
Gene emb. dim. 128	0.956	0.937
Gene emb. dim. 1000	0.953	0.945
10 topics	0.920	0.923
1000 topics	0.959	0.759

Table S2: Robustness analysis of the scETM model. Changing the encoder architecture, gene embedding dimensions and number of topics has limited impact on model performance. We report the average ARI of three repeated trails. scETM was trained on MP for 6000 epochs and on HP for 2000 epochs.

	MP	HP
Current model	0.951	0.937
- BatchNorm in encoder	0.884	0.873
- Input normalization	0.946	0.845
- BatchNorm in encoder - Input normalization	0.000	0.000
- Batch correction module	0.935	0.487

Table S3: Ablation study of the scETM model. We report the average ARI of three repeated trails. scETM was trained on MP for 6000 epochs and on HP for 2000 epochs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S4: scETM 100-topic enrichment for Human Pancreas scRNA-seq data. Only pathway hits with FDR < 0.05 are included. The table is saved in Additional file 2.xls.

Table S5: scETM 100-topic enrichment for Alzheimer's Disease snRNA-seq data. Only pathway hits with FDR < 0.05 are included. The table is saved in Additional file 2.xls.

Table S6: scETM 100-topic enrichment for Major Depressive Disorder snRNA-seq data. Only pathway hits with FDR < 0.05 are included. The table is saved in Additional file 2.xls.

	AD+	AD−	MDD+	MDD−
DE cell-type and topics pairs	237	524	261	464
Cell types identified by DE topics	8	8	8	8
DE topics w.r.t. disease	15	44	65	25
Topics associated with cell types and disease	15	44	65	25

Table S7: Differential expression (DE) analysis summary of topics in AD and MDD data. $+$ indicates up-regulation, and $-$ indicates down-regulation.

	HP	MDD	AD
scETM	0.937	0.734	0.976
p-scETM	0.926	0.753	0.996

Table S8: Adjusted Rand Index (ARI) comparison of scETMs and p-scETMs in three human single cell transcriptomics datasets. Refer to Incorporation of pathway knowledge section for experimental details.

Pathway	**Relevance**	**Reference**
Methionine Metabolism	MDD treatment	[47]
Inflammatory mediator regulation of TRP channels	MDD treatment	[48]
Desipramine Action Pathway	MDD treatment	[49, 50]
2-arachidonoylglycerol_biosynthesis	MDD pathogenesis	[56]
transcription factor creb and its extracellular signals	MDD treatment	[51, 52]
role of erk5 in neuronal survival	MDD pathogenesis and treatment	[53]
Beta2_adrenergic_receptor_signaling	same pathway enrichment found in a MDD study	[60]
Tamoxifen Action Pathway	correlation with MDD onset	[57]
Isovaleric acidemia	correlation with MDD	[58]
Nicotine Action Pathway	correlation with MDD onset	[59]
Bumetanide Action Pathway	MDD treatment	[54]
3-Methylthiofentanyl Action Pathway	MDD treatment	[55]

Table S9: MDD-relevant pathways from the pathway-topic embedding inferred by p-scETM trained on the MDD dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pathway	**Relevance**	**Reference**
Wnt Signaling Pathway and Pluripotency	AD treatment	[61]
Alzheimer_disease_amyloid_secretase	AD pathogenesis	[66]
TNF receptor superfamily (TNFS) members mediating non-canonical NF-kB	AD treatment	[62]
Glutathione synthesis and recycling	AD pathogenesis	[63]
Endothelin_signaling	AD treatment	[64]
GABA-B_receptor_II	AD treatment	[38, 39]
GABA_Transaminase_Deficiency_Metabolite	AD treatment	[38, 39]
Leukotriene_modifiers_pathway_Pharmacodynamics	AD treatment	[65]

Table S10: AD-relevant pathways from the pathway-topic embedding inferred by p-scETM trained on the AD dataset.

Fig. 1

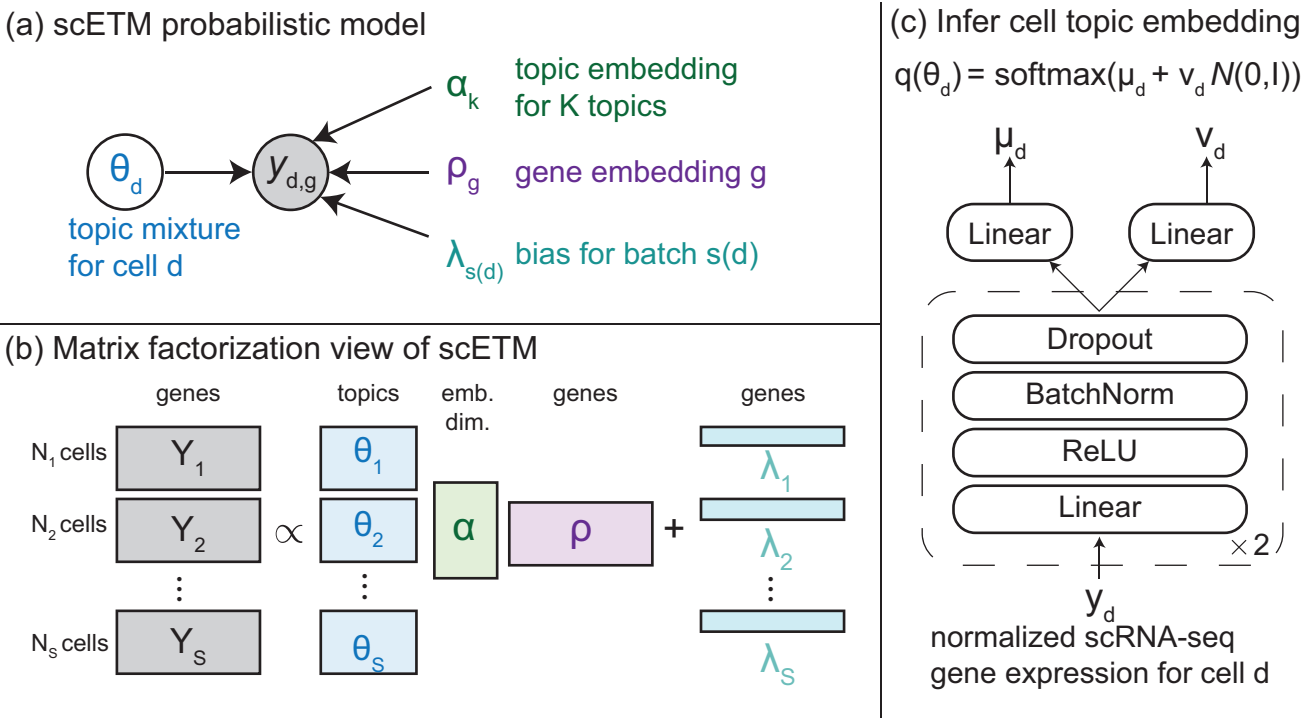

(a) scETM probabilistic model
θ_d
topic mixture for cell d
$y_{d,g}$
α_k
topic embedding for K topics
ρ_g
gene embedding g
$\lambda_{s(d)}$
bias for batch s(d)
(b) Matrix factorization view of scETM
genes
topics
emb. dim.
genes
genes
N_1 cells
N_2 cells
N_S cells
Y_1
Y_2
Y_S
∝
θ_1
θ_2
θ_S
α
ρ
+
λ_1
λ_2
λ_S
(c) Infer cell topic embedding
$q(\theta_d) = \mathrm{softmax}(\mu_d + v_d N(0,I))$
μ_d
v_d
Linear
Linear
Dropout
BatchNorm
ReLU
Linear
×2
y_d
normalized scRNA-seq gene expression for cell d

Fig. 2

Runtime (s)
10000
5000
2000
1500
1000
500
0
10^4
10^5
Number of cells
Memory Requirement (GB)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
10^4
10^5
Number of cells
ARI
0.9
0.8
0.7
0.6
0.5
0.4
0.3
10^4
10^5
Number of cells
Model Name
Harmony
scVI-LD
scVI
scETM
Scanorama
Liger
scVAE
Seurat
Model Input
Principal Components
Top 3k variable genes
All genes

Fig. 3

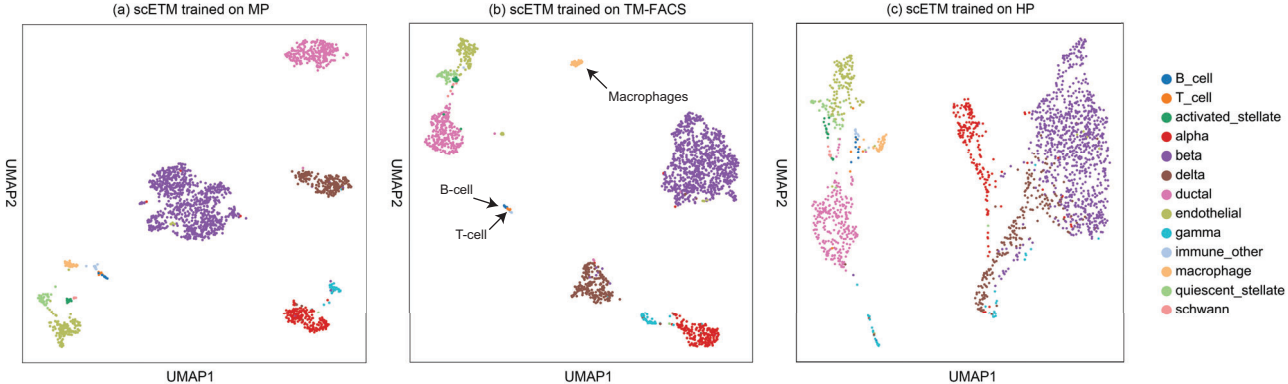

(a) scETM trained on MP
(b) scETM trained on TM-FACS
(c) scETM trained on HP
UMAP1
UMAP2
Macrophages
B-cell
T-cell
B_cell
T_cell
activated_stellate
alpha
beta
delta
ductal
endothelial
gamma
immune_other
macrophage
quiescent_stellate
schwann

Fig. 4

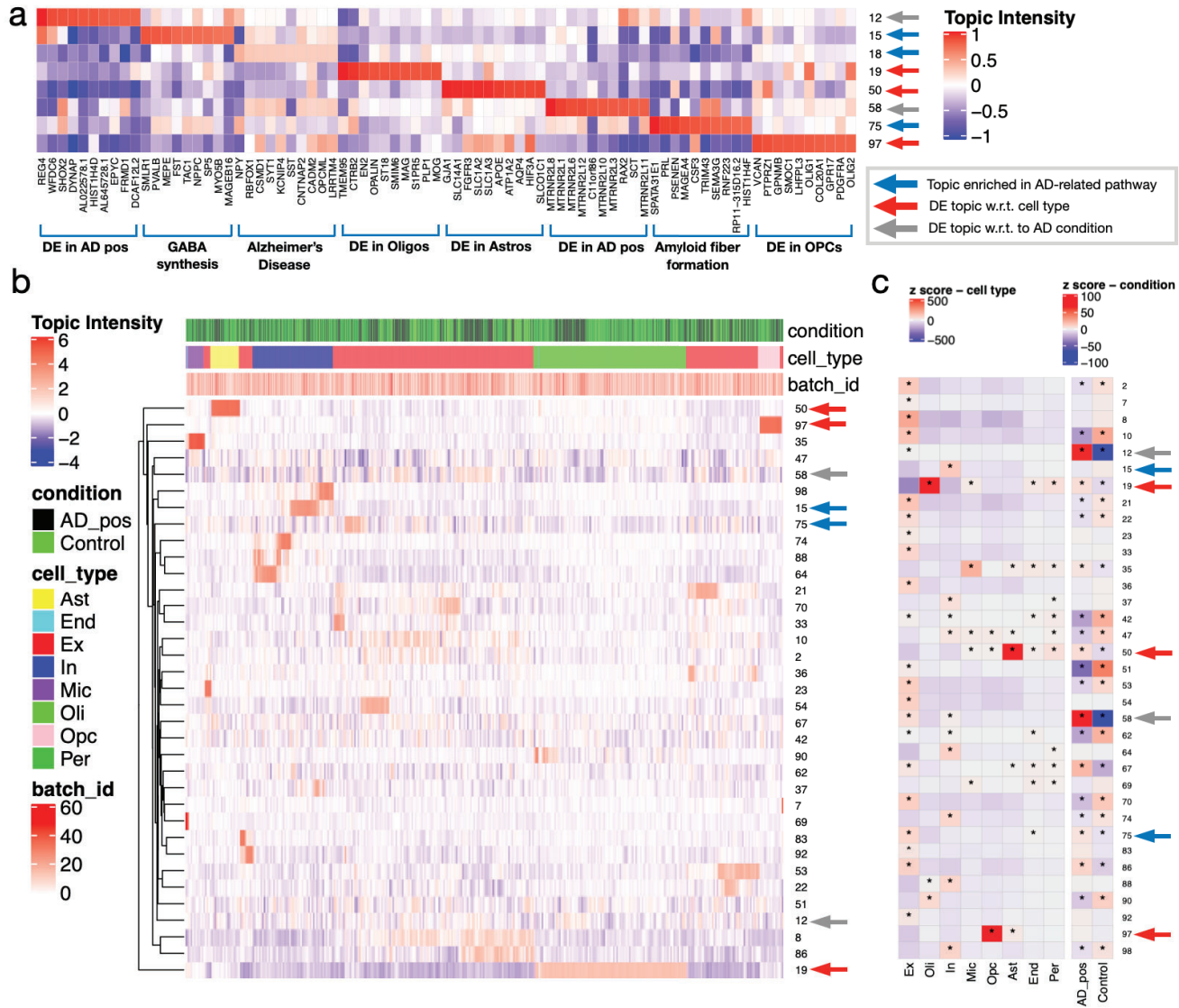

a
Topic Intensity
1
0.5
0
−0.5
−1
12
15
18
19
50
58
75
97
Topic enriched in AD-related pathway
DE topic w.r.t. cell type
DE topic w.r.t. to AD condition
DE in AD pos
GABA synthesis
Alzheimer's Disease
DE in Oligos
DE in Astros
DE in AD pos
Amyloid fiber formation
DE in OPCs
b
Topic Intensity
6
4
2
0
−2
−4
condition
AD_pos
Control
cell_type
Ast
End
Ex
In
Mic
Oli
Opc
Per
batch_id
60
40
20
0
condition
cell_type
batch_id
c
z score – cell type
500
0
−500
z score – condition
100
50
0
−50
−100
Ex
Oli
In
Mic
Opc
Ast
End
Per
AD_pos
Control

Fig. 5

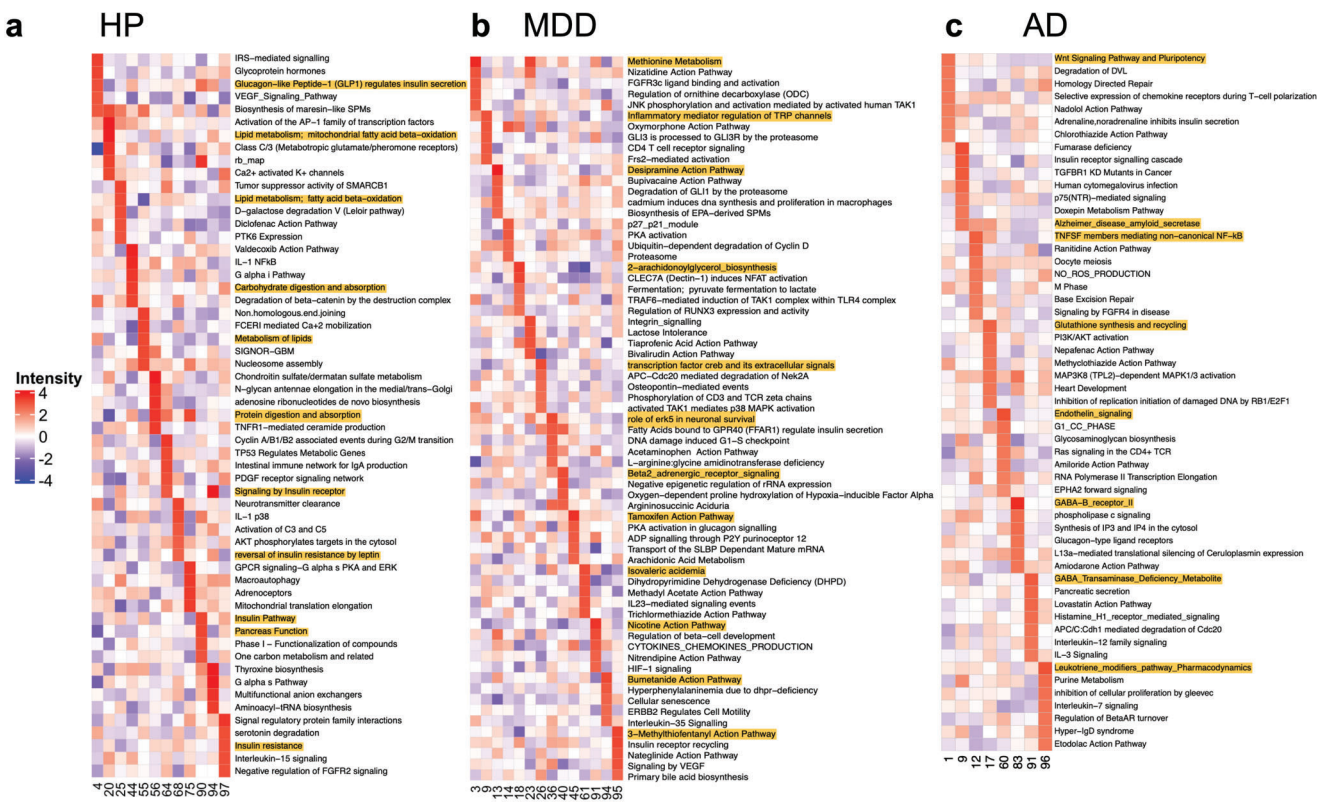

a
HP
IRS−mediated signalling
Glycoprotein hormones
Glucagon−like Peptide−1 (GLP1) regulates insulin secretion
VEGF_Signaling_Pathway
Biosynthesis of maresin−like SPMs
Activation of the AP−1 family of transcription factors
Lipid metabolism; mitochondrial fatty acid beta−oxidation
Class C/3 (Metabotropic glutamate/pheromone receptors)
rb_map
Ca2+ activated K+ channels
Tumor suppressor activity of SMARCB1
Lipid metabolism; fatty acid beta−oxidation
D−galactose degradation V (Leloir pathway)
Diclofenac Action Pathway
PTK6 Expression
Valdecoxib Action Pathway
IL−1 NFkB
G alpha i Pathway
Carbohydrate digestion and absorption
Degradation of beta−catenin by the destruction complex
Non.homologous.end.joining
FCERI mediated Ca+2 mobilization
Metabolism of lipids
SIGNOR−GBM
Nucleosome assembly
Chondroitin sulfate/dermatan sulfate metabolism
N−glycan antennae elongation in the medial/trans−Golgi
adenosine ribonucleotides de novo biosynthesis
Protein digestion and absorption
TNFR1−mediated ceramide production
Cyclin A/B1/B2 associated events during G2/M transition
TP53 Regulates Metabolic Genes
Intestinal immune network for IgA production
PDGF receptor signaling network
Signaling by Insulin receptor
Neurotransmitter clearance
IL−1 p38
Activation of C3 and C5
AKT phosphorylates targets in the cytosol
reversal of insulin resistance by leptin
GPCR signaling−G alpha s PKA and ERK
Macroautophagy
Adrenoceptors
Mitochondrial translation elongation
Insulin Pathway
Pancreas Function
Phase I − Functionalization of compounds
One carbon metabolism and related
Thyroxine biosynthesis
G alpha s Pathway
Multifunctional anion exchangers
Aminoacyl−tRNA biosynthesis
Signal regulatory protein family interactions
serotonin degradation
Insulin resistance
Interleukin−15 signaling
Negative regulation of FGFR2 signaling
4 20 25 44 55 56 64 68 75 90 94 97
Intensity
4
2
0
-2
-4
b
MDD
Methionine Metabolism
Nizatidine Action Pathway
FGFR3c ligand binding and activation
Regulation of ornithine decarboxylase (ODC)
JNK phosphorylation and activation mediated by activated human TAK1
Inflammatory mediator regulation of TRP channels
Oxymorphone Action Pathway
GLI3 is processed to GLI3R by the proteasome
CD4 T cell receptor signaling
Frs2−mediated activation
Desipramine Action Pathway
Bupivacaine Action Pathway
Degradation of GLI1 by the proteasome
cadmium induces dna synthesis and proliferation in macrophages
Biosynthesis of EPA−derived SPMs
p27_p21_module
PKA activation
Ubiquitin−dependent degradation of Cyclin D
Proteasome
2−arachidonoylglycerol_biosynthesis
CLEC7A (Dectin−1) induces NFAT activation
Fermentation; pyruvate fermentation to lactate
TRAF6−mediated induction of TAK1 complex within TLR4 complex
Regulation of RUNX3 expression and activity
Integrin_signalling
Lactose Intolerance
Tiaprofenic Acid Action Pathway
Bivaliruidin Action Pathway
transcription factor creb and its extracellular signals
APC−Cdc20 mediated degradation of Nek2A
Osteopontin−mediated events
Phosphorylation of CD3 and TCR zeta chains
activated TAK1 mediates p38 MAPK activation
role of erk5 in neuronal survival
Fatty Acids bound to GPR40 (FFAR1) regulate insulin secretion
DNA damage induced G1−S checkpoint
Acetaminophen Action Pathway
L−arginine:glycine amidinotransferase deficiency
Beta2_adrenergic_receptor_signaling
Negative epigenetic regulation of rRNA expression
Oxygen−dependent proline hydroxylation of Hypoxia−inducible Factor Alpha
Argininosuccinic Aciduria
Tamoxifen Action Pathway
PKA activation in glucagon signalling
ADP signalling through P2Y purinoceptor 12
Transport of the SLBP Dependant Mature mRNA
Arachidonic Acid Metabolism
Isovaleric acidemia
Dihydropyrimidine Dehydrogenase Deficiency (DHPD)
Methadyl Acetate Action Pathway
IL23−mediated signaling events
Trichlormethiazide Action Pathway
Nicotine Action Pathway
Regulation of beta−cell development
CYTOKINES_CHEMOKINES_PRODUCTION
Nitrendipine Action Pathway
HIF−1 signaling
Bumetanide Action Pathway
Hyperphenylalaninemia due to dhpr−deficiency
Cellular senescence
ERBB2 Regulates Cell Motility
Interleukin−35 Signalling
3−Methylthiofentanyl Action Pathway
Insulin receptor recycling
Nateglinide Action Pathway
Signaling by VEGF
Primary bile acid biosynthesis
3 9 13 14 18 23 26 36 40 57 91 94 95
c
AD
Wnt Signaling Pathway and Pluripotency
Degradation of DVL
Homology Directed Repair
Selective expression of chemokine receptors during T−cell polarization
Nadolol Action Pathway
Adrenaline,noradrenaline inhibits insulin secretion
Chlorothiazide Action Pathway
Fumarase deficiency
Insulin receptor signalling cascade
TGFBR1 KD Mutants in Cancer
Human cytomegalovirus infection
p75(NTR)−mediated signaling
Doxepin Metabolism Pathway
Alzheimer_disease_amyloid_secretase
TNFSF members mediating non−canonical NF−kB
Ranitidine Action Pathway
Oocyte meiosis
NO_ROS_PRODUCTION
M Phase
Base Excision Repair
Signaling by FGFR4 in disease
Glutathione synthesis and recycling
PI3K/AKT activation
Nepafenac Action Pathway
Methyclothiazide Action Pathway
MAP3K8 (TPL2)−dependent MAPK1/3 activation
Heart Development
Inhibition of replication initiation of damaged DNA by RB1/E2F1
Endothelin_signaling
G1_CC_PHASE
Glycosaminoglycan biosynthesis
Ras signaling in the CD4+ TCR
Amiloride Action Pathway
RNA Polymerase II Transcription Elongation
EPHA2 forward signaling
GABA−B_receptor_II
phospholipase c signaling
Synthesis of IP3 and IP4 in the cytosol
Glucagon−type ligand receptors
L13a−mediated translational silencing of Ceruloplasmin expression
Amiodarone Action Pathway
GABA_Transaminase_Deficiency_Metabolite
Pancreatic secretion
Lovastatin Action Pathway
Histamine_H1_receptor_mediated_signaling
APC/C:Cdh1 mediated degradation of Cdc20
Interleukin−12 family signaling
IL−3 Signaling
Leukotriene_modifiers_pathway_Pharmacodynamics
Purine Metabolism
inhibition of cellular proliferation by gleevec
Interleukin−7 signaling
Regulation of BetaAR turnover
Hyper−IgD syndrome
Etodolac Action Pathway
1 9 12 17 60 83 91 96

Fig. S1

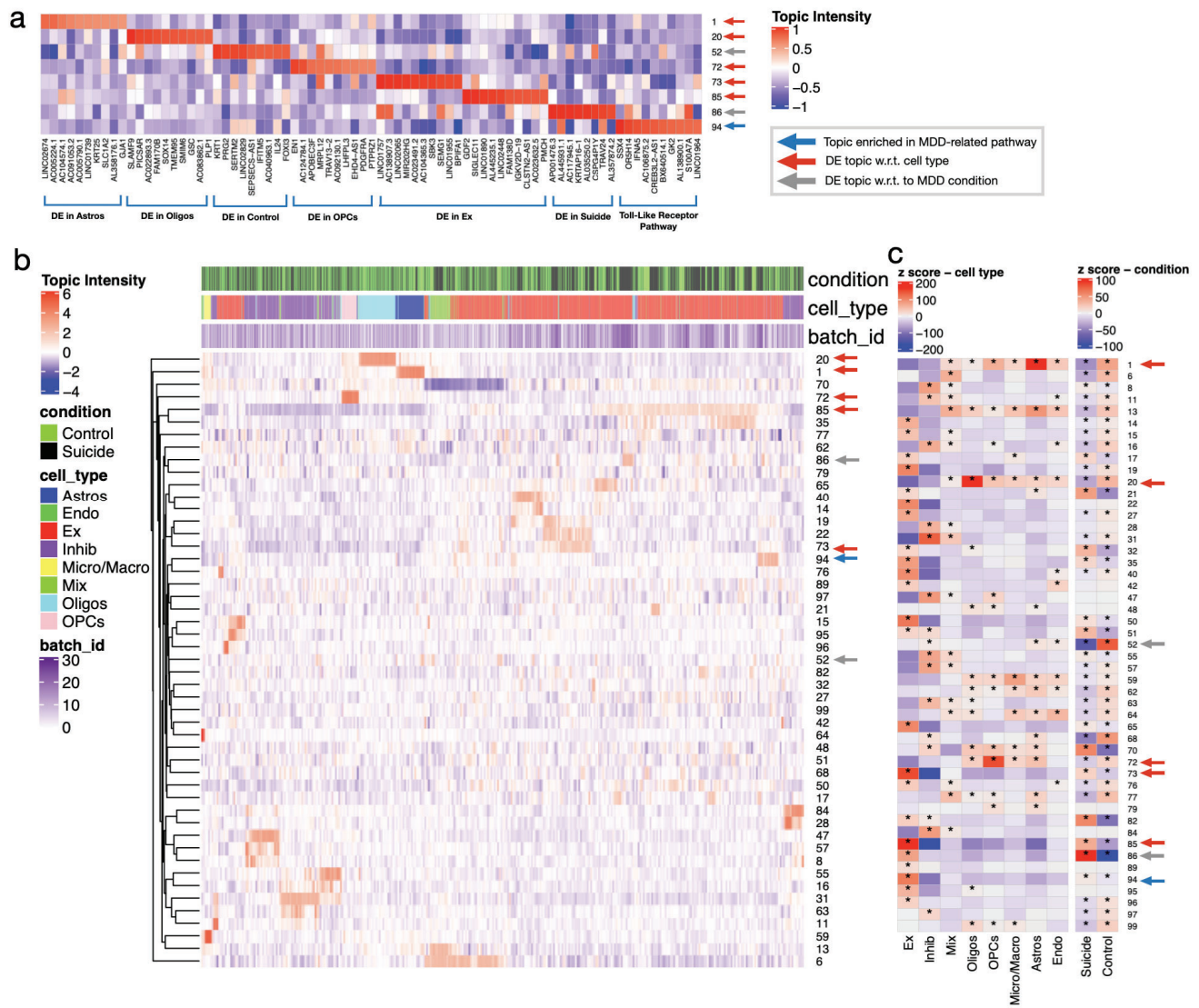

a
Topic Intensity
1
0.5
0
−0.5
−1
Topic enriched in MDD-related pathway
DE topic w.r.t. cell type
DE topic w.r.t. to MDD condition
DE in Astros
DE in Oligos
DE in Control
DE in OPCs
DE in Ex
DE in Suicide
Toll-Like Receptor Pathway
b
Topic Intensity
condition
Control
Suicide
cell_type
Astros
Endo
Ex
Inhib
Micro/Macro
Mix
Oligos
OPCs
batch_id
condition
cell_type
batch_id
c
z score – cell type
z score – condition
Ex
Inhib
Mix
Oligos
OPCs
Micro/Macro
Astros
Endo
Suicide
Control

Fig. S2

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The cop
(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the pr
available under aCC-BY-NC-ND 4.0 International license.

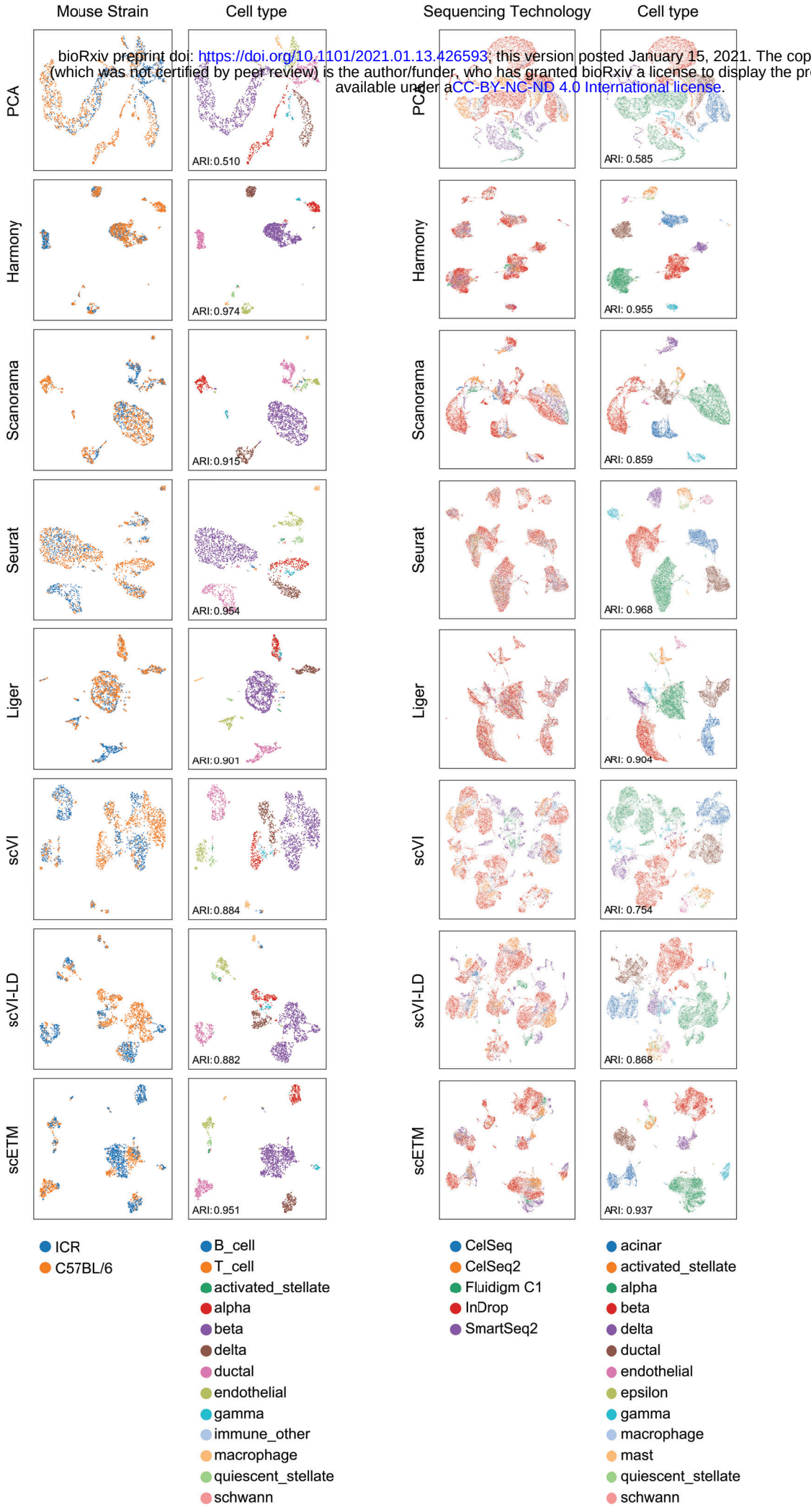
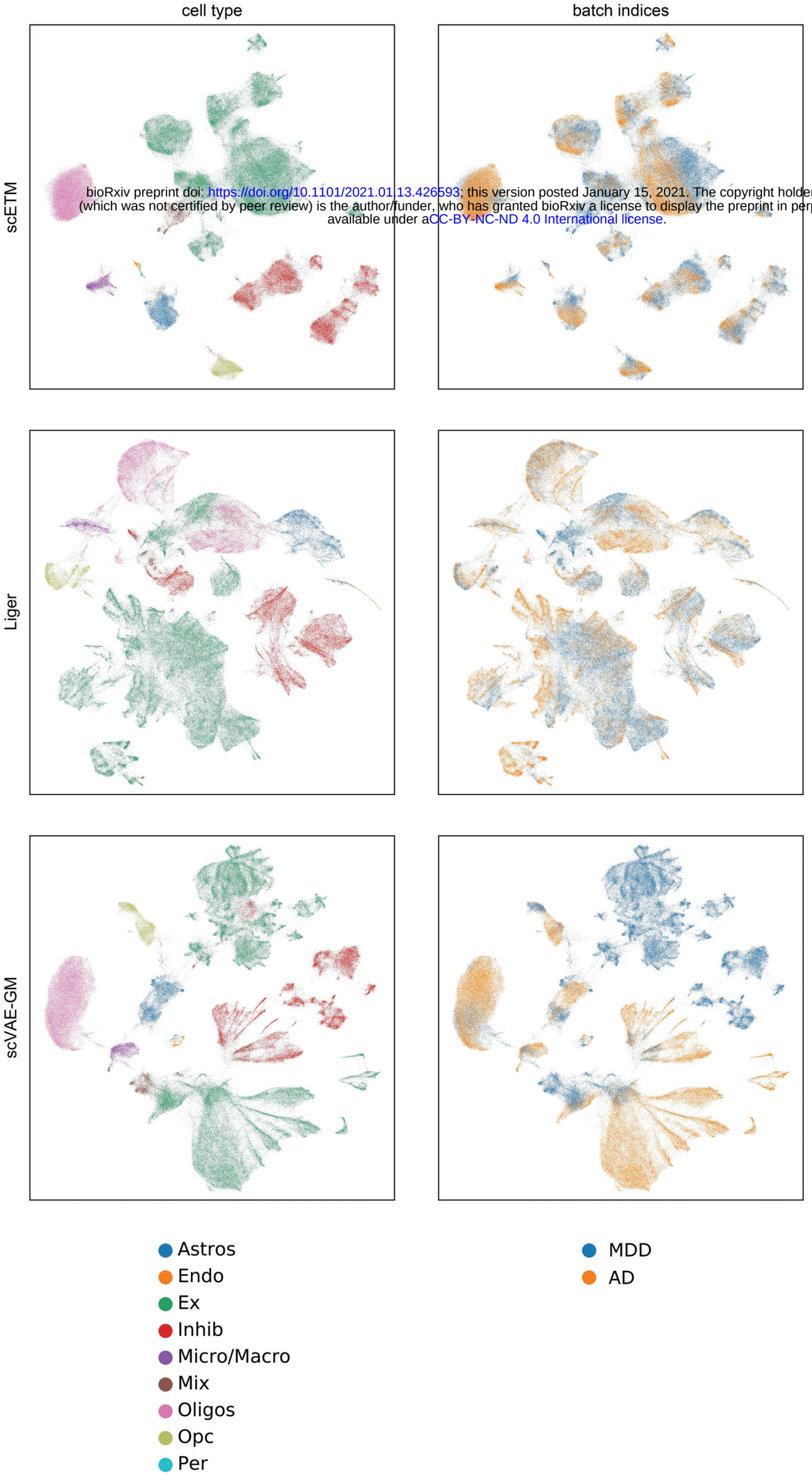


Fig. S3



bioRxiv preprint doi: https://doi.org/10.1101/2021.01.13.426593; this version posted January 15, 2021. The copyright holder (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perp available under aCC-BY-NC-ND 4.0 International license.

Fig. S4

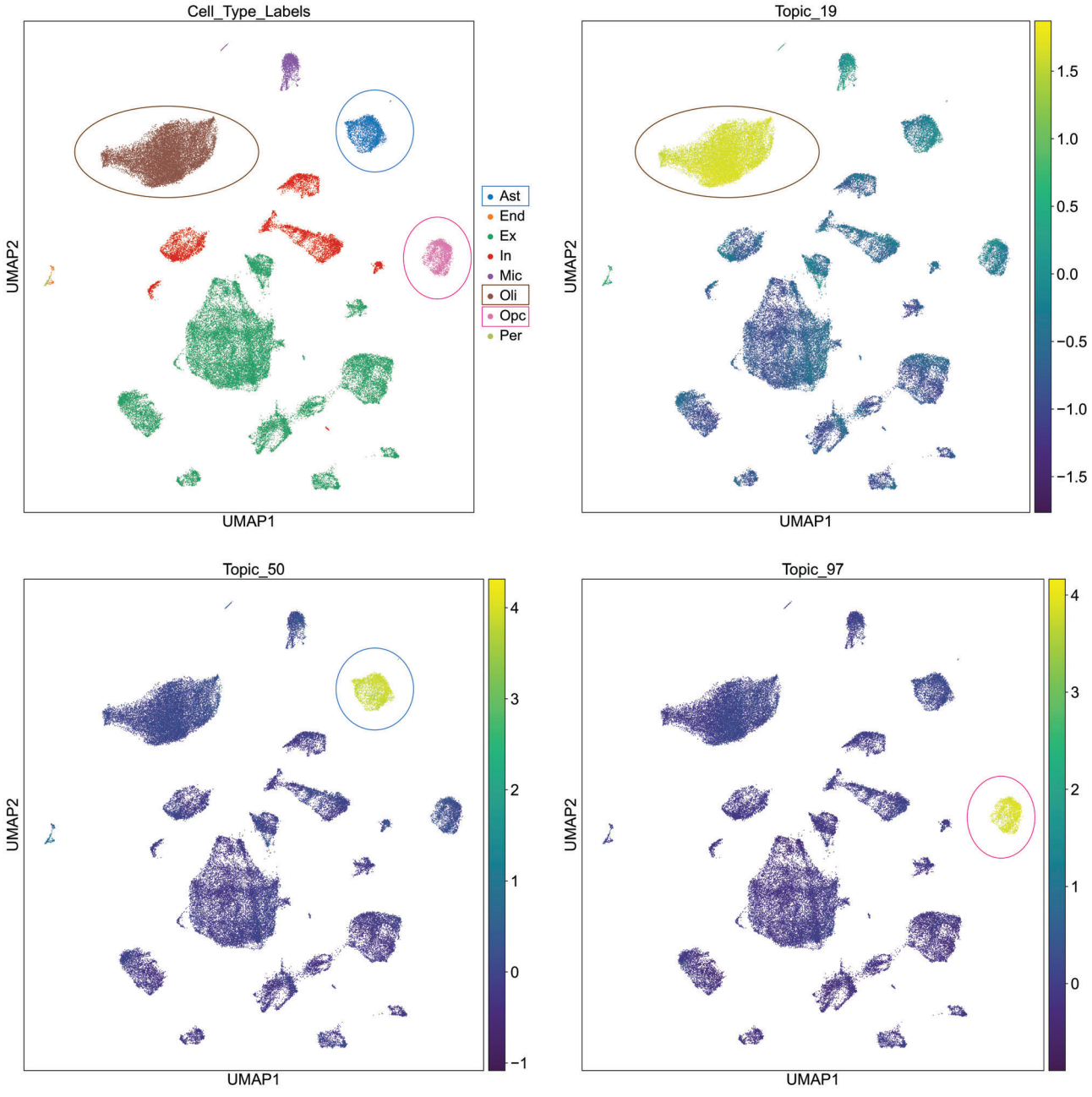
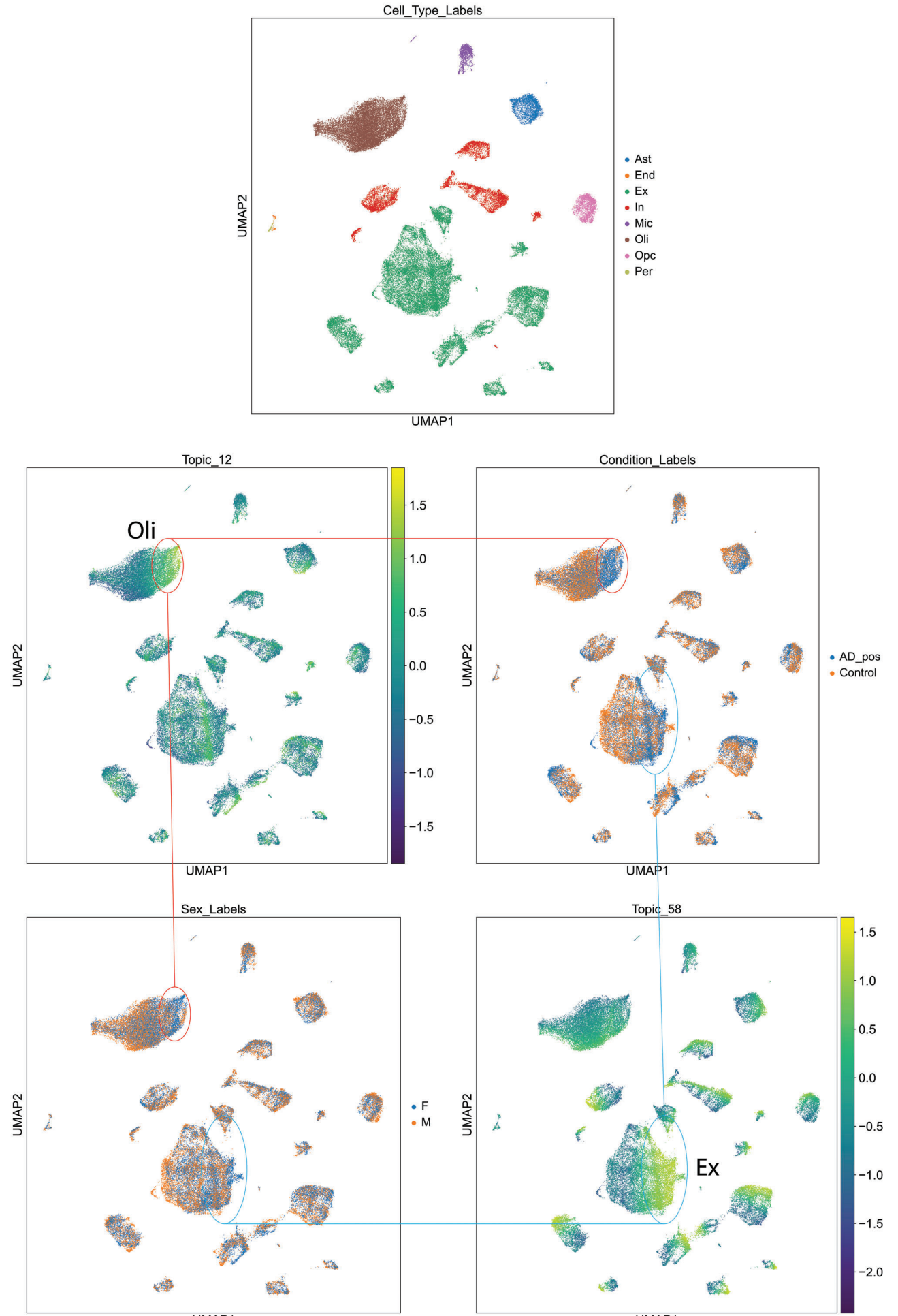

Cell_Type_Labels
Ast
End
Ex
In
Mic
Oli
Opc
Per
UMAP2
UMAP1
Topic_19
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Topic_50
4
3
2
1
0
−1
Topic_97

Fig. S5


Cell_Type_Labels
UMAP2
UMAP1
Ast
End
Ex
In
Mic
Oli
Opc
Per
Topic_12
Oli
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Condition_Labels
AD_pos
Control
Sex_Labels
F
M
Topic_58
Ex
−2.0