Joint inference of discrete cell types and continuous type-specific variability in single-cell datasets with MMIDAS

Yeganeh Marghi [1,*], **Rohan Gala** [1], **Fahimeh Baftizadeh** [1], **and Uygar Sümbül** [1,2,*]

[1]Allen Institute, 615 Westlake Ave N, Seattle, WA, USA
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA

[*]Correspondence: yeganeh.marghi@alleninstitute.org, uygars@alleninstitute.org

Reproducible definition and identification of cell types is essential to enable investigations into their biological function, and understanding their relevance in the context of development, disease and evolution. Current approaches model variability in data as continuous latent factors, followed by clustering as a separate step, or immediately apply clustering on the data. Clusters obtained in this manner are considered as putative cell types in atlas-scale efforts such as those for mammalian brains. We show that such approaches can suffer from qualitative mistakes in identifying cell types robustly, particularly when the number of such cell types is in the hundreds or even thousands. Here, we propose an unsupervised method, MMIDAS (Mixture Model Inference with Discrete-coupled AutoencoderS), which combines a generalized mixture model with a multi-armed deep neural network, to jointly infer the discrete type and continuous type-specific variability. We develop this framework in a way that can be applied to analysis of both uni-modal and multi-modal datasets. Using four recent datasets of brain cells spanning different technologies, species, and conditions, we demonstrate that MMIDAS significantly outperforms state-of-the-art models in inferring interpretable discrete and continuous representations of cellular identity, and uncovers novel biological insights. Our unsupervised framework can thus help researchers identify more robust cell types, study cell type-dependent continuous variability, interpret such latent factors in the feature domain, and study multi-modal datasets.

Keywords: cellular diversity, discrete cell types, continuous type-specific variability, cell states, mixture model, variational inference.

Introduction

Understanding the nature and extent of cellular diversity in the brain is key to unraveling the complexity of neural circuits, their connectivity and roles in behavior, in health and disease.[1, 2] Despite recent advances in high-throughput measurement technologies and related computational approaches, reproducible analysis of high-dimensional single-cell data remains a challenge due to the interplay between hundreds of cell types and multiple sources of continuous biological variability.

Cellular phenotypes form a dynamical landscape.[3] In the adult brain, neurons reside within stable "wells of attraction" of this landscape.[4, 3] Here, the well that a cell belongs to represents a discrete source of variability, giving rise to the notion of cell type identity[5, 2] that is accessible via physiological properties, neurotransmitter phenotypes, overall gene expression, and morphological features.[6, 7, 8, 9] On the other hand, the location within the well is a continuous source of variability.[10, 11] Since these two sources jointly determine the phenotype, computational approaches that focus only on one of these sources may infer a qualitatively different characterization of the overall landscape. Moreover, individual wells can have complicated shapes as manifested by within-type variability in molecular, physiological, and anatomical features (e.g., gene expression levels, connectivity strengths), exacerbating the problem.[10, 11, 12, 13] Statistically, such a collection of discrete wells with non-canonical shapes can be modeled by a generalized mixture model.

Here, we introduce Mixture Model Inference with Discrete-coupled AutoencoderS (MMIDAS), a mixture model-based method designed to jointly infer interpretable discrete (categorical) and continuous latent factors. We achieve this by utilizing discrete-coupled autoencoders, where the different autoencoding neural networks infer the parameters of their own mixture models from augmented copies of a given sample. We find that encouraging the networks to agree on their categorical assignments via the objective function avoids the notorious mode collapse problem common in generative models,[14] even with a high-dimensional discrete space. MMIDAS breaks from existing methods in two key aspects. First, while numerous deep generative models infer low-dimensional continuous representations,[15, 16, 17, 18, 19] post-hoc clustering on these latent variables is suboptimal, especially in scenarios involving a large number of cell types. This is because the first step of dimensionality reduction is typically not informed about the subsequent clustering objective. Instead, MMIDAS jointly performs clustering and dimensionality reduction in an end-to-end differentiable variational formulation.

Second, to interpret within-cluster variability, conventional approaches first cluster cells into discrete types and then characterize the continuous diversity within each cluster separately.[11, 12] Yet, the discrete identity and remaining continuous variability are intimately related to each other. Ignoring these dependencies can produce spurious cluster definitions, which can, in turn, produce flawed characterizations of the remaining within-cluster variability. To address this, MMIDAS offers a unified and interpretable solution by jointly inferring discrete and continuous variabilities and avoids the need for separate modeling for each cluster. This not only simplifies the analysis but also uncovers intricate relationships that might be missed otherwise.

Unlike existing methods,[20, 21, 22] MMIDAS is applicable to both uni-modal and multi-modal data. Importantly, it is a fully unsupervised method and does not require any priors on the relative abundances of cell types, labelled data, or batch information.[23] We benchmark MMIDAS on a diverse selection of datasets including two uni-modal single-cell RNA sequencing (scRNA-seq) datasets profiling multiple regions of the mouse cortex obtained by different platforms,[24, 12] a multi-modal dataset with both electrophysiological and transcriptomic profiles of cortical neurons in mice,[8] and a recent scRNA-seq dataset from the human middle temporal gyrus (MTG) exploring changes in gene expression in disease, the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD).[25] We demonstrate that the mixture models learned by MMIDAS discover accurate and interpretable discrete and continuous factors of cellular diversity. It proposes better cell type definitions, as demonstrated by significantly improved clustering and reproducibility metrics. Furthermore, we show that MMIDAS offers novel insights into the molecular mechanisms governing the cellular landscape. The inferred continuous variability encodes biological processes such as cell metabolism and disease stages, which is achieved by identifying genes whose expression controls the continuous within-type variability.

Results

A. Coupled mixture variational model

MMIDAS is a collective learning framework that uses multiple variational autoencoding neural networks (VAE) to jointly infer interpretable and accurate discrete (categorical) and continuous factors of variability in the presence of a highsional discrete space (Fig. 1a). MMIDAS consists of a set of independent and architecturally identical VAEs. Each VAE "arm" parameterizes cell type identity (discrete diversity) and continuous type-dependent variability (continuous diversity) (Fig. 1b), and their parameters are tuned by efficient and scalable stochastic gradient descent.

In unsupervised VAE-based models, parameter optimization can result in underestimation of discrete variability, known as the mode collapse problem,[14] where the network ignores a subset of discrete latent factors, by overloading the available continuous latents. To avoid this problem, MMIDAS optimizes the parameters of all VAEs simultaneously using a consensus constraint on the discrete factors of variability across the networks, which regularizes the mixture representations at the time of training.

While each arm has its own mixture representation, all arms cooperate to learn a consensus category, $\mathbf{c}$, for a given cell, representing the discrete diversity encoding the cell type. This assignment lives in a simplex whose dimensionality, K, denotes the expected number of cell types in a given dataset (Fig. 1c). The within-type deviation of the cell's phenotype around the cell type representative is captured by the continuous latent factor, $\mathbf{s}$ (Fig. 1c). While the true number of types and their relative abundances are unknown, MMIDAS utilizes a pruning technique on an over-parameterized simplex to automatically identify the number of discrete cell types in the dataset (Fig. 1d). The underlying principle is that it is harder to reach a consensus over spurious types (Methods).

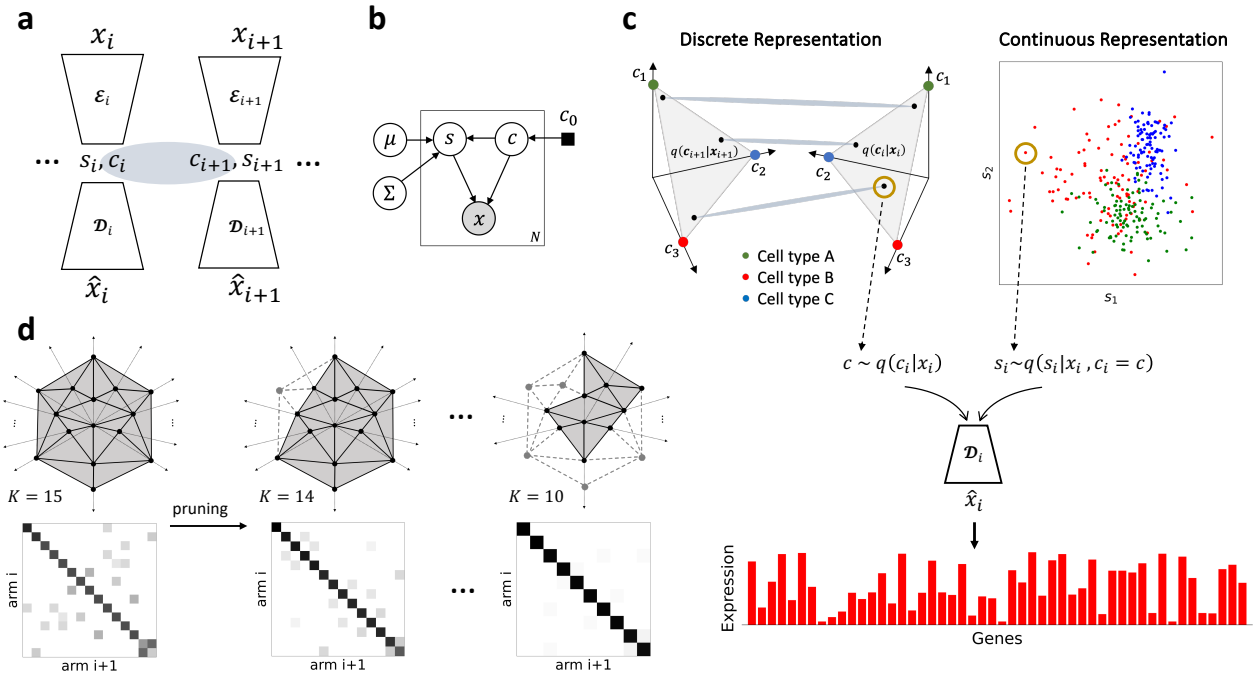

Figure 1. MMIDAS, a multi-arm autoencoder framework for mixture model inference. (a) Schematic of MMIDAS. Individual arms receive non-identical noisy copies of a given sample $\mathbf{x}$, i.e., $\{\ldots, \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots\}$, belonging to a neighborhood of $\mathbf{x}$, to learn mixture representations, i.e., $\{\ldots, q(\mathbf{c}_i, \mathbf{s}_i), q(\mathbf{c}_{i+1}, \mathbf{s}_{i+1}), \ldots\}$. VAE arms cooperate to learn categorical abundances, $p(\mathbf{c})$. (b) Graphical model of each VAE, depicting categorical variable (type), $\mathbf{c}$ and type dependence of the continuous variable (state), $\mathbf{s}$. (c) Discrete and continuous representations of cellular heterogeneity in the latent space. The discrete cell type is obtained by mapping cells to a K-simplex, where K denotes the number of categories. MMIDAS regularizes the discrete distribution, $q(\mathbf{c}_i|\mathbf{x}_i)$, via a consensus constraint in the simplex. The continuous type-dependent variability of the cell is obtained by mapping the profile to a lower-dimensional continuous space, given the discrete type identity. Accurate reconstruction is achieved by jointly decoding the contributions of the discrete and continuous factors of variability, e.g. transcriptomic profile (bar plot for a red cell) through the decoder of the i-th arm, D_i. (d) Schematic of the pruning process in an over-parameterized simplex. During training, the model calculates the consensus measure between pairs of arms for each vertex. During pruning, vertices with the lowest consensus measures are pruned. For the sake of visualization, simplices are plotted in a 2-dimensional space.

We use the Gumbel-Softmax distribution,[26] a continuous relaxation of the categorical distribution, which allows for a differentiable approximation of sampling from a categorical distribution in the simplex (Methods). The consensus constraint is defined based on a distance metric according to Aitchison geometry in the simplex,[27] which avoids the mode collapse problem (Methods and Supplementary Note 4). Theoretical analysis (Methods and Supplementary Note 3) and experimental results that we present in the next section show that this approach enhances accuracy, robustness, and interpretability of the inferred factors without requiring prior information on the relative abundances of the categories.

Before demonstrating MMIDAS on single-cell datasets, we first study a benchmark machine learning dataset, the MNIST handwritten digits dataset (Supplementary Note 5 and Supplementary Note 7). Although MNIST is not as complex as single-cell datasets, we have included it in our evaluation to enable interpretability and facilitate the assessment of MMIDAS results with known ground truth. Our results indicate that MMIDAS outperforms earlier mixture model-based methods, and its computational cost remains comparable to that of the baselines (Fig. S1, Table S1, and Table S2). Supplementary Note 8 and Fig. S4 demonstrate that MMIDAS's performance is minimally affected by variations in the consensus (coupling) factor. Thus, we do not fine-tune this hyperparameter in any of our studies.

B. scRNA-seq analysis of mouse cortical regions

When analyzing uni-modal scRNA-seq datasets, each VAE arm receives a noisy copy of a cell's profile, on which they try to infer a shared discrete factor through a variational optimization (Methods). In the unsupervised setting, the noisy copies are generated by "type-preserving" augmentation, which generates independent and identically distributed copies of data within a small neighborhood around the original sample, thus remaining in the well of attraction and preserving the cell's type identity with high probability (Methods, Supplementary Note 6, Fig. S2b,c, and Fig. S3).

Transcriptomic cell type identification. We first study cellular diversity in a Smart-seq dataset of the mouse primary visual (VISp) and anterior lateral motor (ALM) cortices with 115 transcriptomic types (t-types) according to the reference hierarchical clustering scheme (Fig. 2, Fig. S5a).[24] In this dataset, MMIDAS with a pair of VAE arms (Supplementary Note 10), uncovers 92 distinct transcriptomic categories using a pair of VAE arms(Fig. 2a). At the time of inference on a test set, the two VAEs agree with each other on 94% of their categorical assignments, which we define as the consensus score (Fig. 2b, Methods). These results were obtained by optimizing an initial, over-parameterized 120-simplex and a 2-dimensional continuous space (Supplementary Note 9). Mutual information analysis shows strong alignment between categories inferred by MMIDAS and the reference t-types (Fig. 2a). The MMIDAS categories, however, uncover a more reproducible and separable discrete representation of the underlying data as quantified by the classification performance on 10-fold cross-validation (Fig. 2c) and the average Silhouette score (Fig. 2d), which takes values between -1 and 1, where 1 denotes perfect clusterability. To isolate the improvement due to MMIDAS, Fig. 2c and Fig. 2d also show the results due to merging of the t-types according to the reference hierarchy until 92 clusters are obtained.

We next analyze the expression landscape of the mouse cortex as profiled by the 10x platform, where the reference taxonomy reports 113 glutamatergic and 97 GABAergic t-types[12] (Fig. 3, Fig. S5b,c). In contrast, MMIDAS infers 97 glutamatergic and 70 GABAergic categories (Fig. 3a,b). As before, Fig. 3c,d,e and Fig. S9 quantify that MMIDAS offers a significantly more reproducible and identifiable characterization of the discrete neuronal diversity compared to the reference taxonomy. Indeed, the superiority of MMIDAS is even more evident in analyzing scRNA-seq data obtained by the 10x platform, whose throughput, noise level, and gene dropout rate all tend to be higher.[28]

Inferring genes contributing to continuous diversity. While the discrete representation uncovers relevant transcriptomic categories, the latent representation of continuous type-specific variability captures the intrinsic structure of each type. Fig. 2e demonstrates the continuous representation, a 2-dimensional latent space of all cells in the Smart-seq dataset, highlighting the type-dependent variability for a few example types. As expected, clusterability of the overall dataset based on the continuous latent variable appears much lower since this information is captured by the categorical-like variable $\mathbf{c}$. Yet, the remaining continuous variability, $\mathbf{s}$ given $\mathbf{c}$, is also highly structured and displays different dominant axes and shifts across categories.

The continuous representation can help to characterize the contribution of each gene to both the discrete and continuous heterogeneity. To study correlates of the continuous representation with gene expression, we introduce a traversal analysis. We hold the discrete variable at a constant value (thereby fix the cell type), move along the span of the continuous representation, and evaluate the change in reconstructed expression values for each gene (Methods). To quantify this change and enable comparisons across cell types, we introduce $V_s(g)$, the normalized

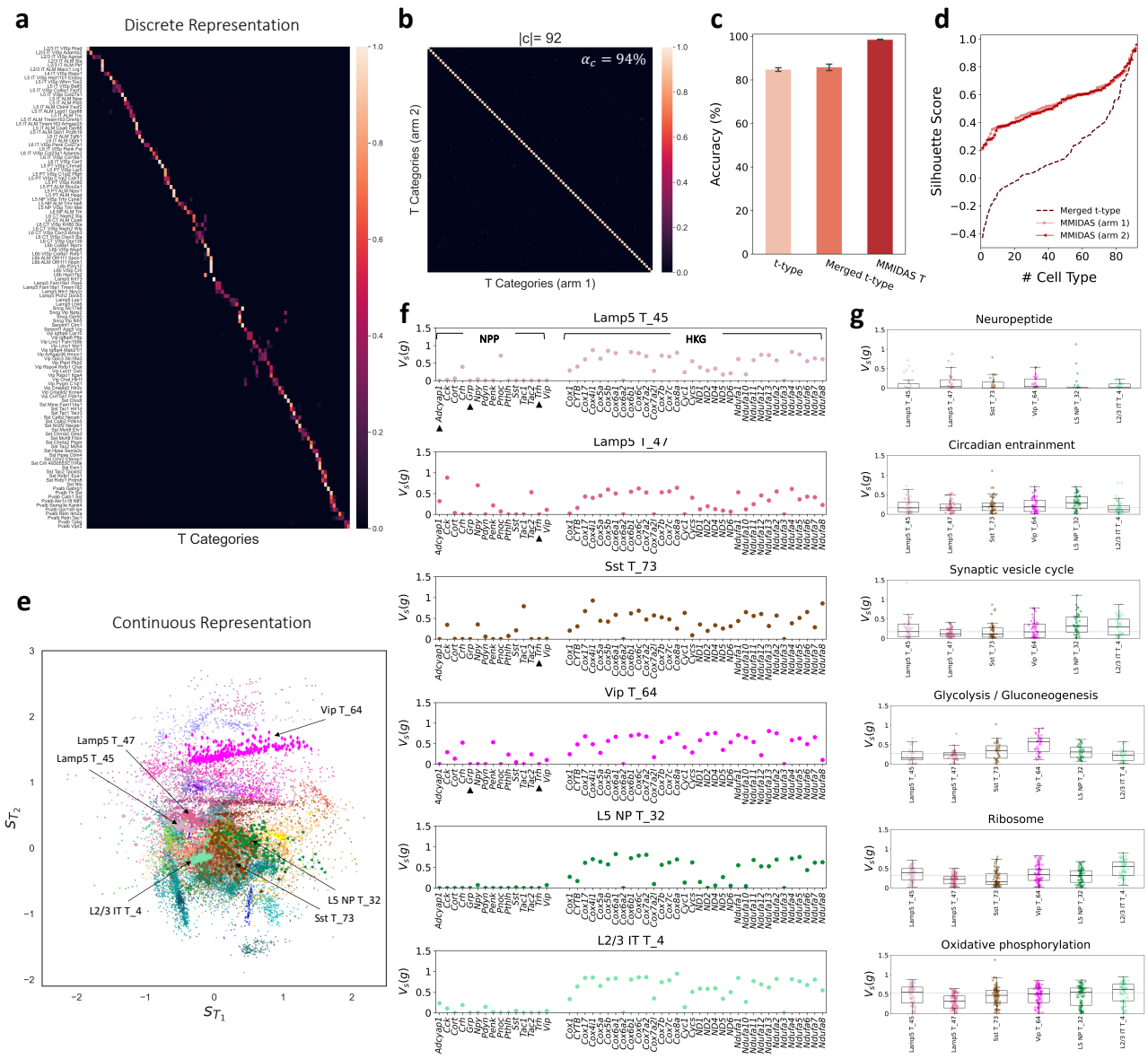

Figure 2. Characterizing discrete and continuous diversity in the adult mouse VISp and ALM neurons sequenced with the Smart-seq platform. (a) Mutual information scores between 115 reference t-types[24] (y-axis) and 92 consensus categories inferred by MMIDAS (x-axis). (b) Confusion matrix between discrete representations of a pair of VAE arms. Consensus score, α_c (Methods), is 94%. (c) Balanced classification accuracy based on 10-fold cross-validation (mean and standard deviation) for three groups of class labels: t-types denote the 115 reference types, merged t-types are t-type labels merged to 92 classes according to the reference hierarchy,[24] MMIDAS categories are the 92 categories inferred by MMIDAS. (d) Average per-cluster Silhouette scores based on test cells for merged t-types and MMIDAS T categories. (e) 2-dimensional representation of within-cluster continuous diversity, color coded based on transcriptomic types inferred by MMIDAS, highlighting the distributions for 6 example T categories. The representation does not encode cell type information (classification accuracy 33%). (f) Variation in gene expression, $V_s(g)$ (Methods), based on the inferred continuous variable. For a given transcriptomic type, we alter the continuous variable s by traversing along the dominant principal component. Each data point illustrates the resulting variation for a subset of neuropeptide precursor (NPP) genes associated with neuropeptide signaling, and a subset of housekeeping genes (HKG) contributing to oxidative phosphorylation. Genes with zero expression are marked with ▲. (g) Variation in gene expression across a subset of transcriptomic types for pathway-associated gene modules. Each subfigure illustrates box plots of overall variation in gene subsets contributing to five distinct pathways: Neuropeptide signaling (15 NPP genes and 10 NP-GPCR genes), circadian entrainment (50 genes), synaptic vesicle cycling (SynapCyc, 49 genes), Glycolysis / Gluconeogenesis (31 genes), Ribosome function (73 genes), and oxidative phosphorylation (OxPhos, 88 genes). The horizontal line within each box marks the median value and the dashed line represents the overall median across the indicated cell types. Subfigures are arranged from the lowest to the highest variation. Pathway-associated gene modules do not overlap, except for SynapCyc and OxPhos, which share 15 genes.

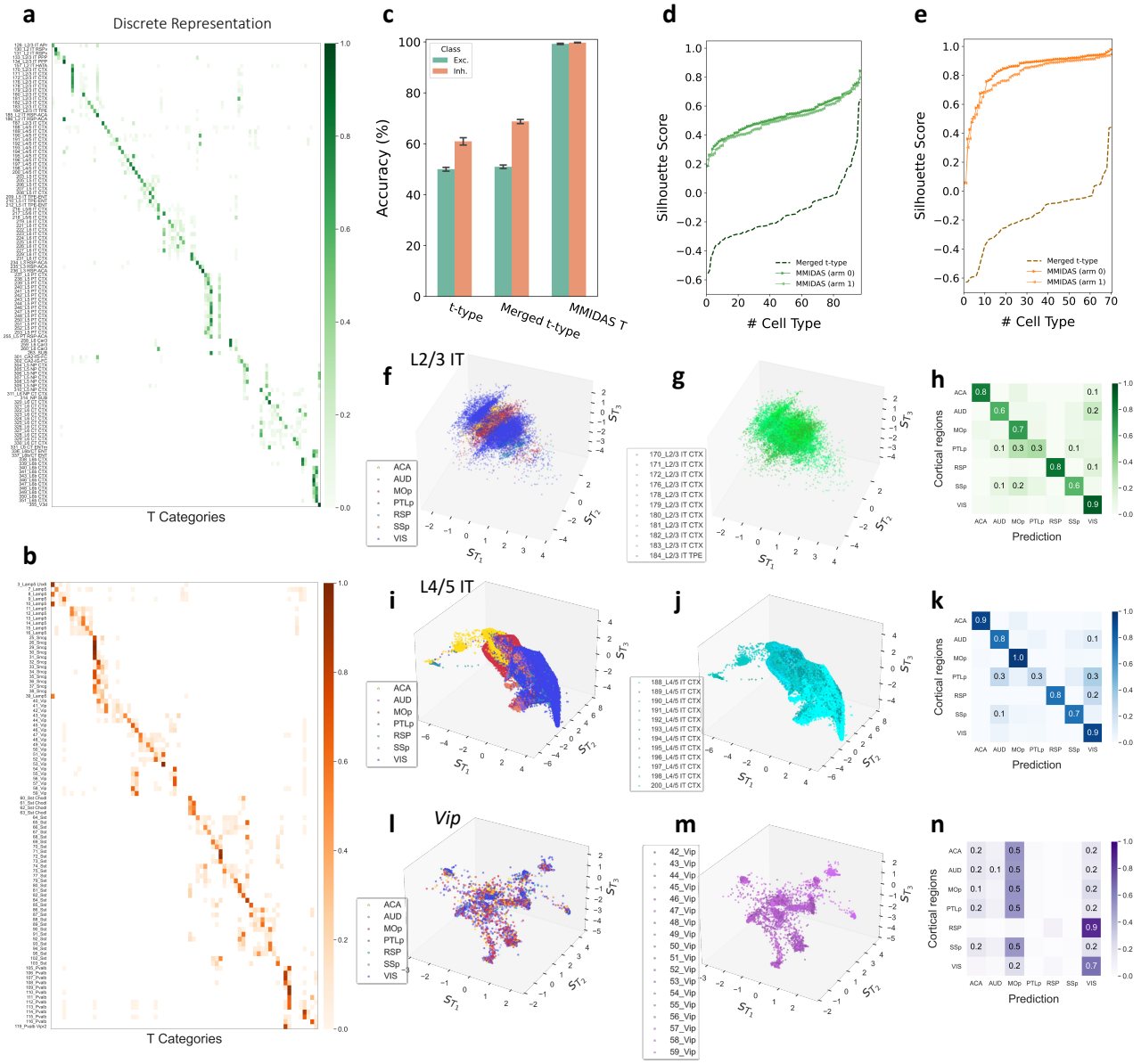

Figure 3. Characterizing discrete and continuous diversity in the adult mouse ACA, AUD, MOp, PTLp, RSP, SSp, VIS neurons sequenced with the 10xv2 platform. (a) Mutual information scores between 113 excitatory (glutamatergic) reference t-types[12] (y-axis) and 97 consensus categories inferred by MMIDAS (x-axis). (b) Mutual information scores between 92 inhibitory (GABAergic) reference t-types[12](y-axis) and 70 consensus categories inferred by MMIDAS (x-axis). (c) Balanced classification accuracy based on 10-fold cross-validation (mean and standard deviation) for three groups of class labels: t-types denote the 205 reference types, merged t-types are t-type labels merged to 167 classes according to the reference hierarchy,[12] MMIDAS T are the 167 transcriptomic categories inferred by MMIDAS. (d) Average per-cluster Silhouette scores based on test cells for glutamatergic merged t-types and MMIDAS categories. (e) Same as d, but for GABAergic clusters. (f) Latent representation of the continuous diversity for L2/3 IT cells, color coded by cortical regions. The representation learned by MMIDAS encodes spatial gradients in the cortex. (g) Same as (f), but color coded by their t-types, showing that the latent representation does not encode cell type information (classification accuracy 23%, Fig. S10a). (h) Confusion matrix for predicting cortical regions, with accuracy 68%, using the inferred continuous representation of L2/3 IT neurons. (i) Continuous latent representation of L4/5 IT cells, color coded by cortical regions. (j) Same as (i), but color coded by their t-types, showing that the latent representation does not encode cell type information (classification accuracy 43%, Fig. S10b). (k) Confusion matrix for predicting cortical regions, with accuracy 77%, using the inferred continuous representation of L4/5 IT neurons. (l,m,n) Same as f,g,h, respectively, but for *Vip* cells, showing that the latent representation encodes neither spatial gradients (accuracy 22%) nor cell type information (accuracy 32%, Fig. S10c) for *Vip* cells.

change in the expression of gene g triggered by traversing the continuous variable $\mathbf{s}$ along the primary principal component of the continuous representations of cells in the given type (Methods).

Fig. 2f and Fig. 2g showcase the results of this traversal analysis for six MMIDAS T categories within the *Lamp5*, *Sst*, *Vip*, L5 NP, and L2/3 IT subclasses. In Fig. 2f, the x-axis lists two gene groups: the first comprises 15 neuropeptide precursor (NPP) genes contributing to neuropeptide signaling pathways, and the second consists of a set of 35 housekeeping genes (HKGs) primarily involved in cellular oxygen consumption within the oxidative phosphorylation pathway. We observe that NPP genes, which exhibit highly specific expression across different neuronal types,[29] display lower variability under the continuous variable traversal. In contrast, HKGs, integral to fundamental cellular maintenance functions,[30] exhibit greater variation (Fig. 2f). Consistent with previous studies, this implies that NPP genes play a more substantial role in the discrete heterogeneity rather than the continuous one. HKGs from the *Cox*, *Ndufa*, and *ND* families, which are related to the electron transport chain and oxidative phosphorylation, are far more variable within cell types, suggesting there may be modules of genes related to a notion of cellular state that the continuous variable captures. Our results also suggest that the expression of activity-regulated genes (e.g. *Cox*) is cell type-dependent so that these genes respond differently to neuronal activation based on the underlying cell type (compare sub-figures in Fig. 2f).

To investigate this further, we selected gene modules implicated in a broad range of pathways,[31] that *a priori* are rather generic cellular processes not expected to govern cellular identity (Fig. 2g). We observe that different pathways drive continuous variability across types. For example, variability in synaptic vesicle cycle pathway genes is the most in cells of type L5 NP T_32, whereas that in glycolysis/gluconeogenesis pathway genes is much higher in cell type Vip T_46 (among the types shown in Fig. 2g). A more comprehensive picture is provided in the Supplementary Figures S6, S7, and S8. The state variables inferred by MMIDAS can thus be viewed as a handle to explore and identify cell type specific gene modules that may be driving variability within populations.

Spatial gradient analysis of neuron types across regions of the isocortex. We next study whether the continuous type-dependent factor can inform about spatial organization. Fig. 3f and Fig. 3i demonstrate such encoding of a contiguous set of cortical regions – anterior cingulate area (ACA), auditory cortex (AUD), primary motor cortex (MOp), posterior parietal associative area (PTLp), retrosplenial cortex (RSP), primary somatosensory area (SSp), and visual cortex (VIS) – by the continuous latent factor, $\mathbf{s}$, for L2/3 IT and L4/5 IT neurons in the mouse 10X dataset, consistent with previous literature.[12] The power of the continuous representation in predicting spatial location is more evident in Fig.3h and Fig.3k, which quantify the separability of cortical areas based on inferred $\mathbf{s} \sim q(\mathbf{s}|\mathbf{c}, \mathbf{x})$ (PTLp is a relatively small region with a smaller number of cells in this dataset, which can explain its confusion with other regions). We contrast this with a glutamatergic subclass, the *Vip* cells, for which MMIDAS did not identify such spatial gene gradients across the cortical sheet (Fig. 3l,n). Thus, the continuous type-dependent latent factor inferred by MMIDAS can effectively encode spatial gradients of gene expression, when they exist.

C. Multimodal single-cell analysis

Simultaneous capture of a diverse range of cellular features from individual cells is key to attaining a definitive characterization of neuronal diversity.[1, 32, 33, 8, 34, 9, 22, 2] A primary challenge in analyzing multimodal single-cell datasets is disentangling factors of variability that are shared among modalities (e.g., cell types) from those that are unique to specific modalities, without access to (and the potential bias of) labeled data. MMIDAS resolves this issue by enabling joint inference of both (shared) discrete and (modality-specific) continuous variabilities. In the context of multimodal data analysis, individual VAE arms learn mixture representations for their respective input modalities (i.e., one VAE arm per modality) while being constrained to learn similar discrete representations across modalities (Fig. 4a).

We study transcriptomic and electrophysiological diversity in a large dataset of mouse visual cortical neurons[8] (Sorensen, Gouwens, Wang *et al.*, unpublished) (Fig. 4). These two observation modalities were profiled by the Patch-seq technique,[32, 35, 8] which combines scRNA-seq (T data) with whole-cell patch clamp electrophysiology (E data). In this dataset, MMIDAS uncovers 60 shared TE categories (Fig. 4b,c) with a consensus score of 81% (Fig. 4b) between the T and E modalities. As in the previous datasets, the classification and Silhouette analyses for each modality indicate that consensus TE categories inferred by MMIDAS offer a more accurate and reproducible discrete representation than the t-type or the merged t-type labels (Fig. 4d,e,f). Noting that the reference transcriptomic taxonomy suggested 97 t-types (and MMIDAS inferred 92 categories in the analysis of the VISp Smart-seq dataset), we observe that the presence of another data modality, E data, influences cell type definitions and leads to fewer cell types.[33, 8, 22] On the other hand, the pronounced alignment between TE categories and t-type labels(Fig. 4c) suggests that a diagonally-dominant, square-like confusion matrix is potentially attainable between transcriptomic and morpho-electric characterizations.

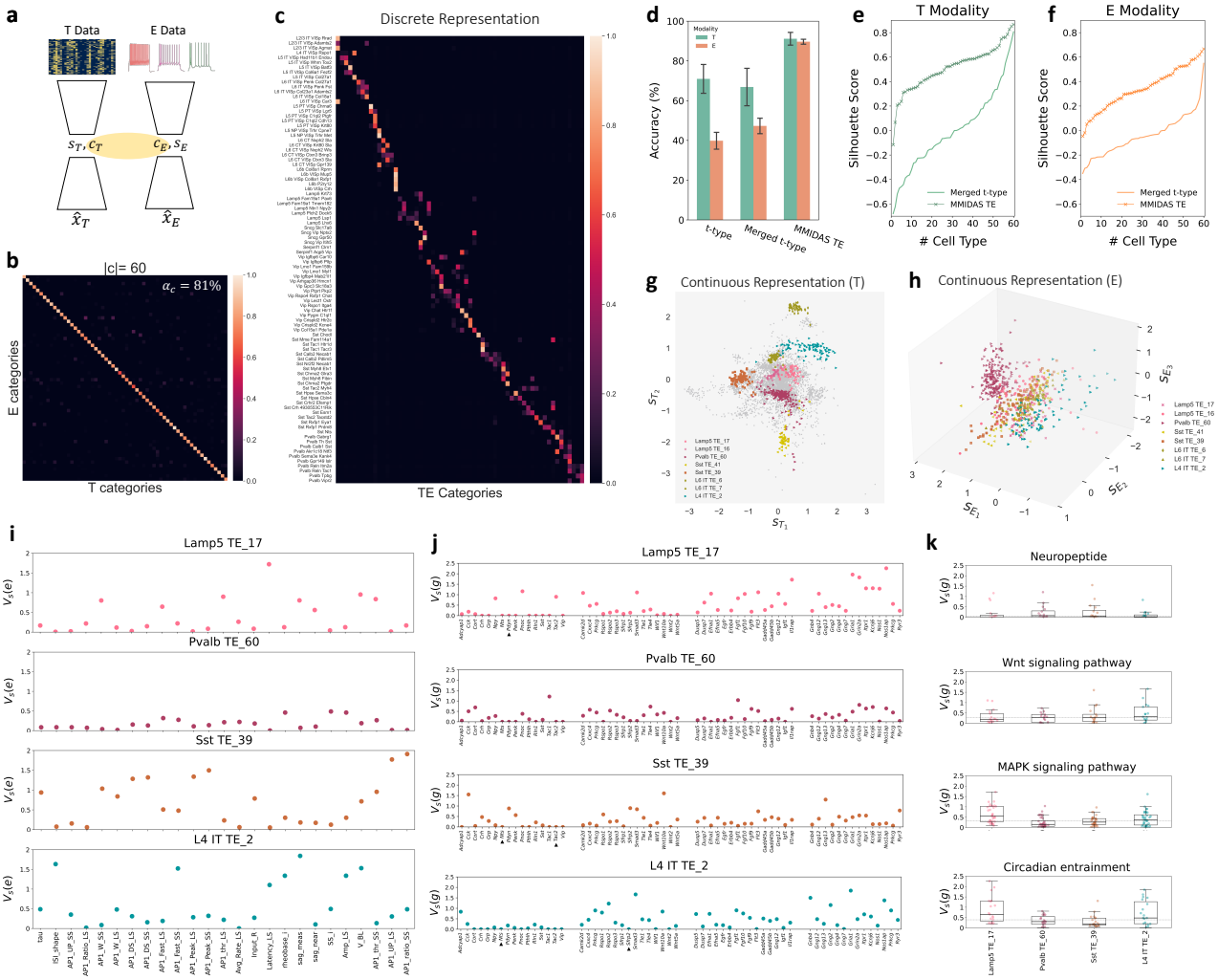

Figure 4. Multimodal characterization of discrete and continuous diversity in adult mouse VISp neurons. (a) Schematic of MMIDAS showing the discrete-coupled VAE architecture in the presence of two data modalities, i.e. transcriptomics (T) and electrophysiology (E), profiled with Patch-seq.[8] The consensus constraint encourages the T and E arms to achieve identical discrete representations, $\mathbf{c}_T$ and $\mathbf{c}_E$, while independently learning their own latent continuous factors, $\mathbf{s}_T$ and $\mathbf{s}_E$. (b) Consensus matrix between discrete representations of a pair of VAE-arms, at the time of inference with a consensus score (α_c) of 81%. (c) Mutual information scores between the 92 reference t-types[8] (y-axis) and 60 transcriptomic-electrophysiology (TE) categories inferred by MMIDAS (x-axis). (d) Balanced classification accuracy based on 10-fold cross-validation (mean and standard deviation) for three groups of class labels: t-types denote the 92 reference types, merged t-types are t-type labels merged to 70 classes according to the hierarchy in,[12] MMIDAS categories are the 70 TE categories inferred by MMIDAS. (e) Average per-cluster Silhouette scores based on transcriptomic profiles of test cells for merged t-types and MMIDAS categories. (f) Same as e, but based on electrophysiology profiles. (g,h) Latent continuous representations of within-cluster diversity for transcriptomic and electrophysiology modalities, outlining the continuous diversities of eight MMIDAS TE categories. (i,j) Continuous traversal analysis of electrophysiological features. For a given transcriptomic type, the continuous variable $\mathbf{s}$ is changed by traversing along the primary principal component. (i) Total variation of $\mathbf{s}_E$ for 25 IPFX features in the patch-seq dataset. (j) Gene expression variation for 16 NPP genes associated with neuropeptide signaling, as well as 45 genes suggested as contributor for Wnt signaling, MAPK signaling, and circadian entrainment pathways. (k) Gene expression variations across transcriptomic types for pathway-associated gene modules. The box plots display total variation of gene modules contributing to different pathways across four TE types. Here, we have neuropeptide signaling involving a set of 35 NPP and NP-GPCR genes, Wnt signaling with 17 genes, MAPK signaling with 36 genes, and circadian entrainment with 19 genes. The dashed line displays the overall median of gene expression variation across the indicated TE types. The gene module involved in the neuropeptide pathway exhibit the least variation, while one contributing to circadian entrainment display the highest variation.

Fig. 4g and Fig. 4h illustrate the continuous representations corresponding to a subset of categories for each data modality. We examine the relationship between the features of each modality and their corresponding continuous representations through the continuous traversal analysis. Fig. 4i,j,k showcase the findings of this analysis for four

MMIDAS TE categories within the *Lamp5*, *Pvalb*, *Sst*, and L4 IT subclasses. Each individual panel in Fig. 4i illustrates the variation of electrophysiological features resulting from traversing $\mathbf{s}_E$ along the primary principal axis (associated with the category) within the continuous representation of the E modality. Here, the x-axis corresponds to a set of 25 intrinsic physiological features (IPFX features – see the Datasets section in Methods). We observe that, unlike most other features, the average firing rate (Avg_Rate_LS) does not seem to be modulated within individual categories. Moreover, the intrinsic electrophysiology of some categories seem to be highly stereotyped (e.g., Pvalb TE_60), while others display much higher within-type variability (e.g., Sst TE_39, L4 IT TE_2).

Fig. 4j and Fig. 4k present the outcomes of continuous traversal (for the same TE types) obtained by traversing $\mathbf{s}_T$ along the principal axis for the T modality. In Fig. 4j, the x-axis shows four gene sets: 16 NPP genes contributing to neuropeptide signaling, 15 genes linked to Wnt signaling, 15 genes related to MAPK signaling, and 15 genes involved in circadian entrainment. Among their many functions, the Wnt- and MAPK signaling pathways are thought to regulate neuronal homeostasis and synaptic plasticity.[36, 37] Consistent with our earlier findings (Fig. 2f,g), NPP genes display lower variability under changes in the continuous latent factor relative to the remaining gene subsets (Fig. 4j). We further refine our focus on pathways comprising a minimum of 15 genes in this dataset (Fig. 4k). As in the mouse Smart-seq dataset, the level of variability is a function of both the pathways and the TE types. For instance, the variation in the gene set related to Wnt signaling is most pronounced in cell type L4 IT TE_2, whereas the variation in genes associated with MAPK signaling is notably higher in cell type Lamp5 TE_17 (among the considered types).

D. Cellular heterogeneity in Alzheimer's disease

Many neurological diseases, such as Alzheimer's disease (AD), display progressive pathology and clinical symptoms. Such progression impacts cells differentially,[38, 25] although imperfect understanding of phenotypic changes in cell types as a function of disease progression presents a challenge. As a further complication, disease progress may include changes to the relative abundance of cell types as well as within-type continuous transitions over the gene expression and pathology landscapes. We address these challenges by exploring the recent Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) dataset,[25] which includes scRNA-seq profiles from the middle temporal gyrus (MTG) of 84 donors spanning the spectrum of AD pathology (Methods).

We train a MMIDAS model with two arms on three subclasses of neurons in this dataset; the IT, *Sst*, and *Pvalb* neurons. The original SEA-AD study[25] describes 42 transcriptomic categories for this set (13 IT, 16 *Sst*, 13 *Pvalb*), referred to as "supertypes" and obtained by mapping transcriptomic data to the BICCN neurotypical reference data[39] and annotating cells via scANVI.[23] In contrast, MMIDAS infers 30 categories (14 IT, 9 *Sst*, 7 *Pvalb*), explaining GABAergic variability with a smaller number of discrete types. Instead, the continuous variability around those (fewer) discrete types contributes to the observed transcriptomic diversity. Beyond this difference, the mutual information scores (Fig. 5a,b) suggest a reasonable correspondence between the inferred categories and the SEA-AD supertypes. We also observe more confusion between the categorical assignments of L2/3 IT neurons and some *Sst* neurons. Similar to our previous results, the classification and Silhouette analyses of transcriptomic types indicate that consensus T categories inferred by MMIDAS are more accurate and reproducible than the suptertypes or the merged supertypes (Fig. 4c,d).

We next study changes to the discrete and continuous factors of variability across the spectrum of AD pathology. For this analysis, we use the disease pseudo-progression score (DPS), which is a number between 0 and 1, increasing with disease progression and calculated per donor, capturing neuropathological changes in the MTG associated with AD.[25] By calculating the expected categorical assignment probabilities, $P_t(\mathbf{c})$, over cells belonging to a donor, we observe declines in the relative abundances of L2/3 IT and *Sst* neurons (i.e., their categorical probabilities) during early pathology (Fig. 5e, Fig. S11), suggesting selective death of those neurons during early stages of AD. In contrast, L4 IT and *Pvalb* neurons display reductions to their relative abundances during later stages of the disease (Fig. 5e, Fig. S11).[25] The rate of cellular loss in Alzheimer's disease can be summarized by the change in $P_t(\mathbf{c})$ between the early and late stages of the disease. Notably, a considerable amount (35%) of the transcriptomic types demonstrate a significant reduction ($p < 0.005$, Welch's t-test) as they transition from the early to the late stage of the disease.

Lastly, we conduct a non-linear regression analysis to explore the relationship between the continuous latent variables inferred by MMIDAS and the DPS (Methods, Fig. S11a). Fig 5f and Fig 5h reveal novel correlations between type-dependent continuous variables and the disease stage for IT neurons in the MTG. Similarly, Fig. S11c,d,e,f demonstrate that the within-type, continuous factor of variability, $\mathbf{s}$, can also serve as an accurate predictor of the BRAAK score (staging of AD-associated neurofibrillary pathology) for IT neurons. In contrast, the continuous variability of GABAergic neurons does not exhibit significant predictive power over the DPS ($R2 < 0.5$, Fig 5f,h).

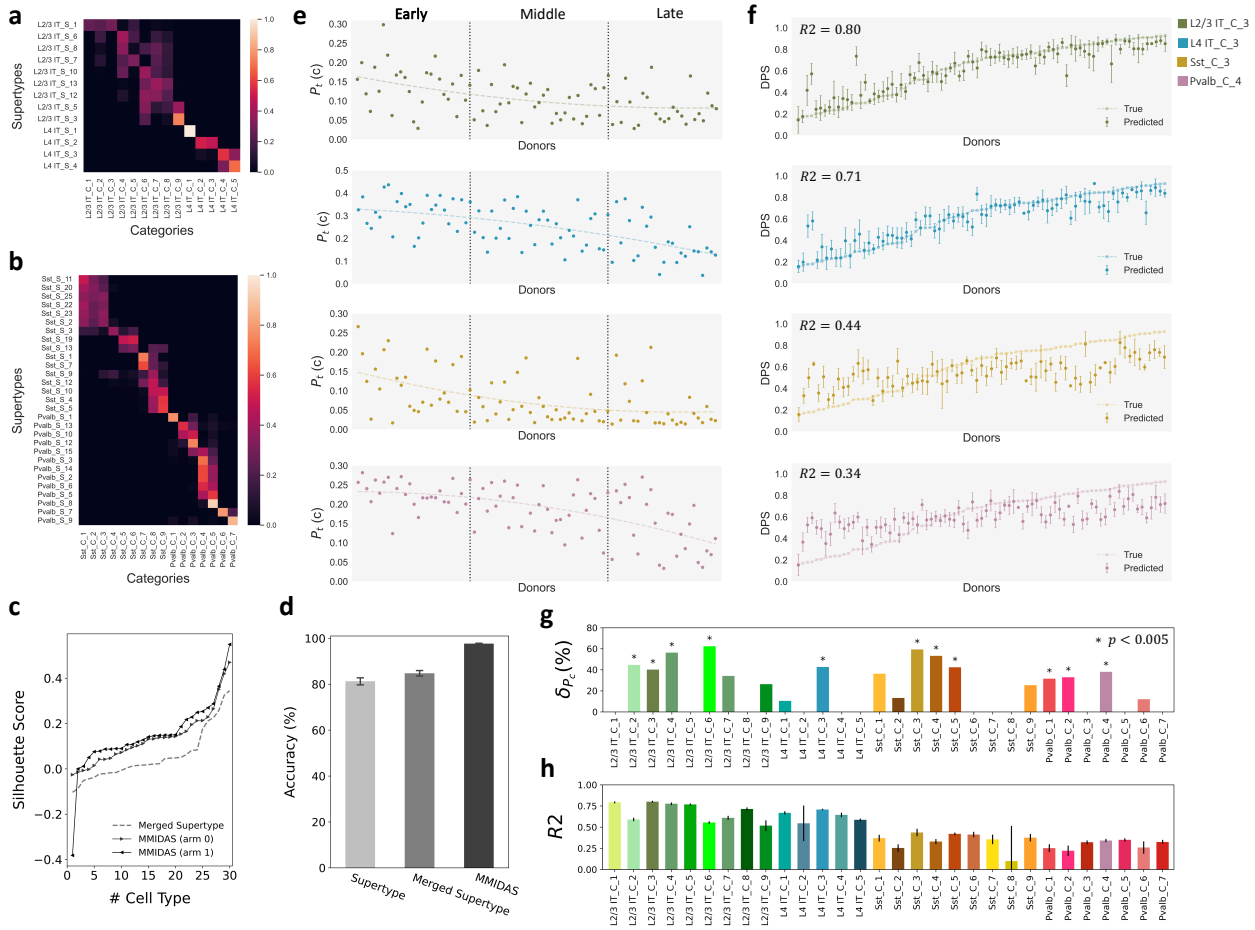

Figure 5. Characterizing discrete and continuous diversity in human MTG scRNA-seq data in Alzheimer's disease. (a) Mutual information scores between 13 IT supertypes obtained with scANVI[23, 25] (y-axis) and 14 consensus categories inferred by MMIDAS (x-axis). (b) Mutual information scores between 29 *Sst* and *Pvalb* supertypes[25] (y-axis) and 16 consensus categories inferred by MMIDAS (x-axis). (c) Balanced classification accuracy based on 10-fold cross-validation (mean and standard deviation) for three groups of class labels: supertypes denote the 42 reference types, merged supertypes are supertypes labels merged to 30 classes according to the reference hierarchy,[25] MMIDAS categories are the 30 categories inferred by MMIDAS. (d) Average per-cluster Silhouette scores based on test cells for merged supertypes and MMIDAS T categories. (e) Predicting relative abundance of cell types inferred by MMIDAS ($P_t(\mathbf{c})$ for category $\mathbf{c}$ at (pesudo) time t) from the pseudo-progression score of donors. x-axis same as in (c). Vertical lines separate early, middle, late stages of disease. Abundances of L2/3 IT, *Sst*, and *Pvalb* neurons decrease in the early stage while that of L4 IT neurons decrease in the late stage. (f) Predicting disease progression pseudotime (DPS)[25] from the continuous representation inferred by MMIDAS. Dots/bars denote mean/standard deviation across 10-fold cross validation. IT cells, but not *Sst* and *Pvalb* cells, have high coefficient of determination, $R2$, which provides a measure of how well the continuous representation could explain the variance in the pseudo disease progression. Donors ordered by their pseudo-progression score obtained from MTG neuropathology[25] in the x-axis. (g) Rate of cellular loss in AD. δ_{P_c} denotes the proportion of changes of $P_t(\mathbf{c})$, between early and late stages of the disease (Methods). $\sim 35\%$ of the transcriptomic types exhibit significant reduction as they progress from the early to the late stage of AD. p-values are obtained by Welch's t-test. (h) The $R2$ scores between the continuous representation and DPS for all inferred cell types are obtained through 10-fold cross-validation (mean and standard deviation). Notably, the continuous representation of glutamatergic cells has $R2 > 0.5$, whereas GABAergic cells exhibit less $R2$ values.

In order to asses the potential influence of batch effects on our findings, we conducted another set of analyses where we intentionally omit data from a few donors during the training phase. Later, in the testing phase, we utilize the trained model to predict outcomes for the excluded donors (Fig. S11b). These results reveal meaningful predictions that are robust against donor-specific variations and batch effects.

Discussion

In response to an ever-changing environment to serve the organism, the cell's phenotype is in a constant flux. This exacerbates the challenge of analyzing the landscape of cellular diversity,[4] which is perhaps most confounding in the mammalian brain with hundreds, and perhaps thousands,[12] of types of brain cells. In this work, we introduced a novel and precise deep variational inference method, MMIDAS, for generalized mixture models that is especially suitable for high-dimensional data and large number of categories. It utilizes multiple autoencoders that are collaboratively trained to achieve a consensus discrete representation, enabling us to overcome various technical challenges associated with generalized mixture model inference.

We demonstrated MMIDAS on four datasets spanning multiple technologies, brain regions, conditions, and species. MMIDAS models cellular heterogeneity as continuous variability around discrete cluster representatives. We demonstrated that this approach significantly improves the identifiability and clusterability scores of the inferred cell types with respect to those of the reference clusters. Importantly, we also observed that the number of cell types inferred by MMIDAS tends to be significantly smaller than that of the reference taxonomies, and the present work may help contribute to resolving the large number of clusters observed in single cell experiments. Dissecting discrete and continuous factors of variability enables identification of key genes or other features associated with aspects of cellular identity or organism-level observables, such as disease progression. In multimodal settings, MMIDAS allows for the identification of biological factors that influence multiple modalities, as well as aspects that are specific to individual modalities. This integrated analysis provides a comprehensive and accurate description of cellular diversity. In both unimodal and multimodal analysis, if it is desirable to use existing cell type labels, one autoencoder in MMIDAS can be replaced to supply those available labels during training. This way, new cells can be analyzed into their discrete and continuous factors based on those available training labels.

The present approach has certain limitations for data without an underlying categorical basis. MMIDAS assumes that the underlying data manifold can be faithfully represented via a discrete variable that denotes the cluster identity (type), and a continuous variable that denotes the within-cluster variability. Consequently, datasets that deviate from this model are ill-suited for MMIDAS. Nevertheless, contemporary cell type analysis is based on the assumption of fundamental underlying types with variable state, as opposed to a continuum of states, and MMIDAS was developed supporting this approach. Another challenge is the choice of hyperparameters, which MMIDAS shares with other unsupervised approaches. Fig. S4 shows that MMIDAS' performance is robust to the value of the coupling factor within a relatively large range. Supplementary Note 9 discusses the difficulty of model selection for the dimensionalities of the latent factors. Lastly, Table S3 shows that MMIDAS' performance further improves with more VAE arms, although this produces diminishing returns with considerable computational cost.

Brain cells occupy a diverse set of locations in a complicated landscape of cellular phenotypes. As the signals they receive and the homeostatic needs change, cell states move within wells of attraction of that landscape to best help the organism they are a part of. However, identifying the wells of attraction and the shapes of those wells may become ambiguous or ill defined. We believe MMIDAS can help researchers to dissect such perplexing single-cell datasets in a principled and efficient manner.

Bibliography

[1] H. S. Seung and U. Sümbül, "Neuronal cell types and connectivity: lessons from the retina," *Neuron*, vol. 83, no. 6, pp. 1262–1272, 2014.

[2] H. Zeng, "What is a cell type and how to define it?," *Cell*, vol. 185, no. 15, pp. 2739–2755, 2022.

[3] E. A. Mukamel and J. Ngai, "Perspectives on defining cell types in the brain," *Current opinion in neurobiology*, vol. 56, pp. 61–68, 2019.

[4] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[5] G. Fishell and N. Heintz, "The neuron identity problem: form meets function," *Neuron*, vol. 80, no. 3, pp. 602–612, 2013.

[6] U. Sümbül, S. Song, K. McCulloch, M. Becker, B. Lin, J. R. Sanes, R. H. Masland, and H. S. Seung, "A genetic and computational approach to structurally classify neuronal types," *Nature communications*, vol. 5, no. 1, p. 3512, 2014.

[7] A. Paul, M. Crow, R. Raudales, M. He, J. Gillis, and Z. J. Huang, "Transcriptional architecture of synaptic communication delineates gabaergic neuron identity," *Cell*, vol. 171, no. 3, pp. 522–539, 2017.

[8] N. W. Gouwens, S. A. Sorensen, F. Baftizadeh, A. Budzillo, B. R. Lee, T. Jarsky, L. Alfiler, K. Baker, E. Barkan, K. Berry, *et al.*, "Integrated morphoelectric and transcriptomic classification of cortical gabaergic cells," *Cell*, vol. 183, no. 4, pp. 935–953, 2020.

[9] F. Scala, D. Kobak, M. Bernabucci, Y. Bernaerts, C. R. Cadwell, J. R. Castro, L. Hartmanis, X. Jiang, S. Laturnus, E. Miranda, *et al.*, "Phenotypic variation of transcriptomic cell types in mouse motor cortex," *Nature*, vol. 598, no. 7879, pp. 144–150, 2021.

[10] K. D. Harris, H. Hochgerner, N. G. Skene, L. Magno, L. Katona, C. B. Gonzales, P. Somogyi, N. Kessaris, S. Linnarsson, and J. Hjerling-Leffler, "Classes and continua of hippocampal ca1 inhibitory neurons revealed by single-cell transcriptomics," *PLoS biology*, vol. 16, no. 6, p. e2006387, 2018.

[11] G. Stanley, O. Gokce, R. C. Malenka, T. C. Südhof, and S. R. Quake, "Continuous and discrete neuron types of the adult murine striatum," *Neuron*, vol. 105, no. 4, pp. 688–699, 2020.

[12] Z. Yao, C. T. van Velthoven, T. N. Nguyen, J. Goldy, A. E. Sedeno-Cortes, F. Baftizadeh, D. Bertagnolli, T. Casper, M. Chiang, K. Crichton, *et al.*, "A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation," *Cell*, 2021.

[13] S. Bugeon, J. Duffield, M. Dipoppa, A. Ritoux, I. Prankerd, D. Nicoloutsopoulos, D. Orme, M. Shinn, H. Peng, H. Forrest, *et al.*, "A transcriptomic axis predicts state modulation of cortical interneurons," *Nature*, vol. 607, no. 7918, pp. 330–338, 2022.

[14] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[15] E. Pierson and C. Yau, "Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome biology*, vol. 16, no. 1, pp. 1–10, 2015.

[16] J. Ding, A. Condon, and S. P. Shah, "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[17] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[18] J. Ding and A. Regev, "Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces," *Nature communications*, vol. 12, no. 1, p. 2554, 2021.

[19] D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, and T. Nguyen, "Fast and precise single-cell data analysis using a hierarchical autoencoder," *Nature communications*, vol. 12, no. 1, p. 1029, 2021.

[20] A. Gayoso, Z. Steier, R. Lopez, J. Regier, K. L. Nazor, A. Streets, and N. Yosef, "Joint probabilistic modeling of single-cell multi-omic data with totalvi," *Nature methods*, vol. 18, no. 3, pp. 272–282, 2021.

[21] R. Lopez, B. Li, H. Keren-Shaul, P. Boyeau, M. Kedmi, D. Pilzer, A. Jelinski, I. Yofe, E. David, A. Wagner, *et al.*, "Destvi identifies continuums of cell types in spatial transcriptomics data," *Nature biotechnology*, vol. 40, no. 9, pp. 1360–1369, 2022.

[22] R. Gala, A. Budzillo, F. Baftizadeh, J. Miller, N. Gouwens, A. Arkhipov, G. Murphy, B. Tasic, H. Zeng, M. Hawrylycz, *et al.*, "Consistent cross-modal identification of cortical neurons with coupled autoencoders," *Nature computational science*, vol. 1, no. 2, pp. 120–127, 2021.

[23] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models," *Molecular systems biology*, vol. 17, no. 1, p. e9620, 2021.

[24] B. Tasic, Z. Yao, L. T. Graybuck, K. A. Smith, T. N. Nguyen, D. Bertagnolli, J. Goldy, E. Garren, M. N. Economo, S. Viswanathan, *et al.*, "Shared and distinct transcriptomic cell types across neocortical areas," *Nature*, vol. 563, no. 7729, pp. 72–78, 2018.

[25] M. Gabitto, K. Travaglini, J. Ariza, E. Kaplan, B. Long, V. Rachleff, Y. Ding, J. Mahoney, N. Dee, J. Goldy, *et al.*, "Integrated multimodal cell atlas of alzheimer's disease," *bioRxiv*, 2023.

[26] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[27] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[28] X. Wang, Y. He, Q. Zhang, X. Ren, and Z. Zhang, "Direct comparative analyses of 10x genomics chromium and smart-seq2," *Genomics, proteomics & bioinformatics*, vol. 19, no. 2, pp. 253–266, 2021.

[29] S. J. Smith, U. Sümbül, L. T. Graybuck, F. Collman, S. Seshamani, R. Gala, O. Gliko, L. Elabbady, J. A. Miller, T. E. Bakken, *et al.*, "Single-cell transcriptomic evidence for dense intracortical neuropeptide networks," *elife*, vol. 8, p. e47889, 2019.

[30] C. J. Joshi, W. Ke, A. Drangowska-Way, E. J. O'Rourke, and N. E. Lewis, "What are housekeeping genes?," *PLoS computational biology*, vol. 18, no. 7, p. e1010295, 2022.

[31] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic acids research*, vol. 45, no. D1, pp. D353–D361, 2017.

[32] C. R. Cadwell, A. Palasantza, X. Jiang, P. Berens, Q. Deng, M. Yilmaz, J. Reimer, S. Shen, M. Bethge, K. F. Tolias, *et al.*, "Electrophysiological, transcriptomic and morphologic profiling of single neurons using patch-seq," *Nature biotechnology*, vol. 34, no. 2, pp. 199–203, 2016.

[33] R. Gala, N. Gouwens, Z. Yao, A. Budzillo, O. Penn, B. Tasic, G. Murphy, H. Zeng, and U. Sümbül, "A coupled autoencoder approach for multi-modal analysis of cell types," in *Advances in Neural Information Processing Systems*, pp. 9263–9272, 2019.

[34] D. Kobak, Y. Bernaerts, M. A. Weis, F. Scala, A. S. Tolias, and P. Berens, "Sparse reduced-rank regression for exploratory visualisation of paired multivariate data," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 70, no. 4, pp. 980–1000, 2021.

[35] C. R. Cadwell, F. Scala, S. Li, G. Livrizzi, S. Shen, R. Sandberg, X. Jiang, and A. S. Tolias, "Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using patch-seq," *Nature protocols*, vol. 12, no. 12, pp. 2531–2553, 2017.

[36] N. C. Inestrosa and E. Arenas, "Emerging roles of wnts in the adult nervous system," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 77–86, 2010.

[37] G. M. Thomas and R. L. Huganir, "Mapk cascade signalling and synaptic plasticity," *Nature Reviews Neuroscience*, vol. 5, no. 3, pp. 173–183, 2004.

[38] H. Mathys, Z. Peng, C. A. Boix, M. B. Victor, N. Leary, S. Babu, G. Abdelhady, X. Jiang, A. P. Ng, K. Ghafari, A. K. Kunisky, J. Mantero, K. Galani, V. N. Lohia, G. E. Fortier, Y. Lotfi, J. Ivey, H. P. Brown, P. R. Patel, N. Chakraborty, J. I. Beaudway, E. J. Imhoff, C. F. Keeler, M. M. McChesney, H. H. Patel, S. P. Patel, M. T. Thai, D. A. Bennett, M. Kellis, and L.-H. Tsai, "Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to alzheimer's disease pathology," *Cell*, vol. 186, no. 20, pp. 4365–4385.e27, 2023.

[39] N. L. Jorstad, J. Close, N. Johansen, A. M. Yanny, E. R. Barkan, K. J. Travaglini, D. Bertagnolli, J. Campos, T. Casper, K. Crichton, *et al.*, "Transcriptomic cytoarchitecture reveals principles of human neocortex organization," *bioRxiv*, pp. 2022–11, 2022.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[41] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[42] E. Dupont, "Learning disentangled joint continuous and discrete representations," in *Advances in Neural Information Processing Systems*, pp. 710–720, 2018.

[43] Y. Jeong and H. O. Song, "Learning discrete and continuous factors of data via alternating disentanglement," in *International Conference on Machine Learning*, pp. 3091–3099, PMLR, 2019.

[44] T. Minka *et al.*, "Divergence measures and message passing," tech. rep., Citeseer, 2005.

[45] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," *arXiv preprint arXiv:1705.08841*, 2017.

[46] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell rna-seq denoising using a deep count autoencoder," *Nature communications*, vol. 10, no. 1, p. 390, 2019.

[47] S. Hauberg, O. Freifeld, A. B. L. Larsen, J. Fisher, and L. Hansen, "Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation," in *Artificial Intelligence and Statistics*, pp. 342–350, 2016.

[48] A. Jaiswal, R. Y. Wu, W. Abd-Almageed, and P. Natarajan, "Unsupervised adversarial invariance," in *Advances in Neural Information Processing Systems*, pp. 5092–5102, 2018.

[49] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[50] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, pp. 1558–1566, PMLR, 2016.

[51] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[52] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework.," *Iclr*, vol. 2, no. 5, p. 6, 2017.

[53] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal, "Isometric logratio transformations for compositional data analysis," *Mathematical Geology*, vol. 35, no. 3, pp. 279–300, 2003.

[54] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi, "Don't blame the elbo! a linear vae perspective on posterior collapse," in *Advances in Neural Information Processing Systems*, pp. 9403–9413, 2019.

Methods

Mixture model inference with discrete-coupled autoencoders (MMIDAS)

MMIDAS consists of A mixture variational autoencoders (VAE), i.e. A-tuple of independent and architecturally identical autoencoding arms. For an observation $\mathbf{x} \in \mathbb{R}^D$, a VAE learns a generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and a variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $\mathbf{z} \in \mathbb{R}^M$ is a latent variable with a parameterized distribution $p(\mathbf{z})$ and $M \ll D$.[40] Here, a mixture model structure is imposed on the latent variable by introducing a categorical latent variable $\mathbf{c} \in \mathcal{S}^K$, denoting the class label defined in a K-simplex, and a continuous latent variable $\mathbf{s}$, which is conditioned on the discrete variable $\mathbf{c}$. We refer to the continuous variable $\mathbf{s}$ as the *state* or *style* variable interchangeably. $\mathbf{c}$ and $\mathbf{s}$ together form the latent variables of each (mixture) autoencoder.

For completeness, we first derive the evidence lower bound (ELBO)[41] for a single mixture VAE, in which an observation $\mathbf{x}$ can be described by one categorical random variable and one continuous random variable, with conditional dependency between $\mathbf{c}$ and $\mathbf{s}$. The variational approach to inferring the latent variables corresponds to solving the optimization equation

$$q^*(\mathbf{s},\mathbf{c}|\mathbf{x}) = \arg\min_{q(\mathbf{s},\mathbf{c}|\mathbf{x})\in\mathcal{D}} D_{\mathrm{KL}}\left(q(\mathbf{s},\mathbf{c}|\mathbf{x})\|p(\mathbf{s},\mathbf{c}|\mathbf{x})\right) \ , \tag{1}$$

where $\mathcal{D}$ is a family of density functions over the latent variables. Evaluating the objective function requires knowledge of $p(\mathbf{x})$, which is usually unknown. Therefore, we rewrite the divergence term as

$$\begin{aligned}
D_{\mathrm{KL}}\left(q(\mathbf{s},\mathbf{c}|\mathbf{x})||p(\mathbf{s},\mathbf{c}|\mathbf{x})\right) &= \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x})\log\frac{q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x})}{\frac{p(\mathbf{x}|\mathbf{s},\mathbf{c})p(\mathbf{s}|\mathbf{c})p(\mathbf{c})}{p(\mathbf{x})}}\,d\mathbf{s} \\
&= \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x})\log\frac{q(\mathbf{s}|\mathbf{c},\mathbf{x})}{p(\mathbf{s}|\mathbf{c})}\,d\mathbf{s} + \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x})\log\frac{q(\mathbf{c}|\mathbf{x})}{p(\mathbf{c})}\,d\mathbf{s} \\
&\quad + \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|\mathbf{x})\log p(\mathbf{x})\,d\mathbf{s} - \int_{\mathbf{s}}\sum_{\mathbf{c}} q(\mathbf{s}|\mathbf{c},\mathbf{x})q(\mathbf{c}|x)\log p(\mathbf{x}|\mathbf{s},\mathbf{c})\,d\mathbf{s} \\
&= \log p(\mathbf{x}) - \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[\mathbb{E}_{(q(\mathbf{s}|\mathbf{c},\mathbf{x}))}\left[\log p(\mathbf{x}|\mathbf{s},\mathbf{c})\right]\right] \\
&\quad + \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[D_{KL}\left(q(\mathbf{s}|\mathbf{c},\mathbf{x})\|p(\mathbf{s}|\mathbf{c})\right)\right] + \mathbb{E}_{q(\mathbf{s}|\mathbf{c},\mathbf{x})}\left[D_{KL}\left(q(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\right)\right] \\
&= \log p(\mathbf{x}) - \mathcal{L}_{\mathbf{s}}
\end{aligned} \tag{2}$$

Since $\log p(\mathbf{x})$ is not function of the optimization parameters, instead of minimizing Eq. 2, the variational lower bound with the distributions parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$,

$$\mathcal{L}_{\mathbf{s}|\mathbf{c}}(\boldsymbol{\phi},\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{s},\mathbf{c}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{s},\mathbf{c})\right] - \mathbb{E}_{q(\mathbf{c}|\mathbf{x})}\left[D_{KL}\left(q(\mathbf{s}|\mathbf{c},\mathbf{x})\|p(\mathbf{s}|\mathbf{c})\right)\right] - \mathbb{E}_{q(\mathbf{s}|\mathbf{c},\mathbf{x})}\left[D_{KL}\left(q(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\right)\right], \tag{3}$$

can be maximized. The first term in Eq. 3 is referred to as the (negative) "reconstruction loss" or $\mathcal{L}_{rec}$. To generate a sample (conditioned on the observation $\mathbf{x}$), first a discrete latent vector $\mathbf{c}$ is sampled from $q(\mathbf{c}|\mathbf{x})$. Given sample $\mathbf{c}$, a continuous latent vector is sampled from $q(\mathbf{s}|\mathbf{c},\mathbf{x})$. The decoder outputs the parameters of $q(\mathbf{x}|\mathbf{c},\mathbf{s})$, from which we can sample an output vector.

We choose $q(\mathbf{s}|\mathbf{c},\mathbf{x})$ to be a factorized Gaussian, parameterized using the reparametrization trick, and assume that the corresponding prior distribution is also a factorized Gaussian, $\mathbf{s}|\mathbf{c} \sim \mathcal{N}(0,\mathbf{I})$. For the categorical variable, existing VAE-based solutions impose a uniform structure on $p(\mathbf{c})$.[42, 43] However, this assumption is inadequate for single-cell datasets where the abundances of the clusters typically differ significantly. As a result, earlier solutions have limitations in effectively learning interpretable mixture representations with high-dimensional discrete variables in real-world applications.

Additionally, maximizing the ELBO in Eq. 3 creates trivial solutions that underestimate posterior variabilities $\mathbf{s}|\mathbf{c},\mathbf{x}$ and $\mathbf{c}|\mathbf{x}$, known as the mode collapse problem, where the network ignores a subset of latent variables.[44, 41] While recent mixture VAE studies, such as JointVAE[42] and CascadeVAE,[43] attempt to resolve this issue, they remain vulnerable to mode collapse in high-dimensional discrete settings and cluster heterogeneity.

The multi-arm framework of MMIDAS introduces a consensus constraint on the discrete factors of variability across the VAE models (arms) during the training process to address the aforementioned issues. In this framework, individual VAE arms receive a collection of non-identical copies, $\{\mathbf{x}_a, \mathbf{x}_b, \ldots\}$ of the given sample, $\mathbf{x}$. These copies are sampled from a small neighborhood around $\mathbf{x}$ to remain within the same well of attraction and belong to the same

category as $\mathbf{x}$. (See Data augmentation below.) This novel approach serves to regularize the mixture representations by introducing variability during training, thus inferring a more robust and accurate model.

While each arm has its own mixture representation with potentially non-identical parameters, all arms cooperate to learn $q(\mathbf{c}_a|\mathbf{x}_a)$, where $\mathbf{c}_a = \mathbf{c}_b = \cdots$, via a cost function at the time of training. Accordingly, the cost function of a set of VAEs with A arms can be formulated as a collection of constrained variational objectives as follows:

$$\begin{aligned} \max \quad & \mathcal{L}_{\mathbf{s}_1|\mathbf{c}_1}(\boldsymbol{\phi}_1, \boldsymbol{\theta}_1) + \cdots + \mathcal{L}_{\mathbf{s}_A|\mathbf{c}_A}(\boldsymbol{\phi}_A, \boldsymbol{\theta}_A) \\ & \text{s.t. } \mathbf{c}_1 = \cdots = \mathbf{c}_A \end{aligned} \tag{4}$$

where $\mathcal{L}_{\mathbf{s}_a|\mathbf{c}_a}(\boldsymbol{\phi}_a, \boldsymbol{\theta}_a)$ is the variational loss for a single arm a (Eq. 3). Propositions 1 and 2 (Supplementary Note 3) show that the shared categorical assignment inferred from $q(\mathbf{c}|\mathbf{x}_1, \cdots, \mathbf{x}_A)$, under the $\mathbf{c} = \mathbf{c}_1 = \cdots = \mathbf{c}_A$ constraint of the multi-arm framework improves the accuracy of the categorical assignment on expectation, and having more arms increases the expected log posterior for the true categorical latent variable unless it is already at its maximum. Additionally, our theoretical results show that the required number of arms satisfying Eq. S35 is a function of the categorical distribution and the likelihood (Eq. S39, Supplementary Note 3). In the particular case of uniformly abundant categories, one pair of coupled arms is enough to satisfy Eq. S35 (Corollary 1, Supplementary Note 3).

We note that MMIDAS is an unsupervised method. As such, it does not require weak supervision, unlike.[23, 45] Instead, it relies on representations that are invariant under non-identical copies of observations. Moreover, unlike ,[17, 22] the multi-arm framework is not restricted to the continuous space.

Reconstruction error. MMIDAS uses two different noise models for handling single-cell data in this paper:

(i) Gaussian Noise Model: For log-transformed scRNA-seq data obtained via the Smart-seq technique and electrophysiological data, a multivariate normal distribution with equal variances across features can effectively approximate the single-cell measurements. To achieve this, gene expression values are first normalized to counts per million (CPM) and then transformed using the formula $\log(CPM+1)$. In this case, the corresponding objective that minimizes the reconstruction error in Eq. 4 is the mean squared error (MSE).

(ii) Zero-Inflated Negative Binomial (ZINB) Model: For scRNA-seq data obtained through the 10x platform, we use the ZINB model:[46]

$$f_{\text{ZINB}}(x|\pi, r, p) = \pi I_0(x) + (1-\pi) f_{\text{NB}}(x|r, p) \tag{5}$$

with

$$f_{\text{NB}}(x|r, p) = \binom{x+r-1}{x}(1-p)^x p^r, \tag{6}$$

where f_{ZINB} and f_{NB} are probability mass functions for the count variable x. $I_0(.)$ is an indicator function, which equals 1 when $x = 0$, and π approximates the probability of gene dropout in 10x data. r and p denote the number and probability of success (not being a zero) in the NB distribution, respectively. In this case, MMIDAS uses the negative log likelihood of the ZINB distribution as the reconstruction loss in Eq. 4.

Data augmentation

In MMIDAS, VAE arms receive non-identical observations that share the discrete variational factor. To achieve this in a fully unsupervised setting, we use *type-preserving* data augmentation that generates independent and identically distributed copies of data while preserving its categorical identity.

Augmentation can be considered as a generative process. Hence, we seek a generative model that learns the transformations representing within-class variability in an unsupervised manner. Learning such transformations is not straightforward: while conventional transformations such as rotation, scaling, or translation can serve as type-preserving augmentations for image data, they may not capture the richness of the underlying process. Moreover, such transformations cannot be used when within-class variability is unknown. Alternatives in the literature either rely on the availability of the class label, or are specific to image data.[47, 48] To this end, inspired by DAGAN,[49] we propose an unsupervised augmentation scheme using a VAE-GAN-like architecture,[50] parametrized by a neural network $\mathcal{G}$, which implicitly learns the underlying conditional data distributions. $\mathcal{G}$ should generate type-preserving samples (i.e., samples that remain within the well of attraction of the original observation) $\mathbf{x_a}$ conditioned on the given sample, $\mathbf{x}$. This is achieved by concatenating the low dimensional representation of $\mathbf{x}$ with Gaussian noise $\mathbf{n}$ in $\mathcal{G}$. To prevent $\mathcal{G}$ from disregarding the noise, we formulate the training procedure as the following minmax optimization problem which uses a discriminator network $\mathcal{D}$ as a regularizer:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{V}(\mathcal{D}, \mathcal{G}) - \gamma_1 \mathcal{R}(\mathcal{G}) + \gamma_2 \mathcal{T}_\alpha(\mathcal{G}) + \gamma_3 d(\mathcal{G}) \tag{7}$$

where,

$$\begin{aligned}
\mathcal{V}(\mathcal{D},\mathcal{G}) &= \mathbb{E}_{\mathbf{x}}\left[\log \mathcal{D}(\mathbf{x})\right] + \mathbb{E}_{\mathbf{x}}\left[\log\left(1-\mathcal{D}(\mathbf{x}_{\not{\mathbf{n}}})\right)\right] + \mathbb{E}_{\mathbf{x},\mathbf{n}}\left[\log\left(1-\mathcal{D}(\mathbf{x}_{\mathbf{n}})\right)\right] && \textbf{(8)}\\
\mathcal{R}(\mathcal{G}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z})\right] && \textbf{(9)}\\
\mathcal{T}_{\alpha}(\mathcal{G}) &= \max\left(\|\mathbf{x}-\mathbf{x}_{\not{\mathbf{n}}}\|_2 - \|\mathbf{x}-\mathbf{x}_{\mathbf{n}}\|_2 + \alpha, 0\right) && \textbf{(10)}\\
d(\mathcal{G}) &= d\left(\mathbf{z}_{\not{\mathbf{n}}}, \mathbf{z}_{\mathbf{n}}\right), \quad where\ \mathbf{z}_{\not{\mathbf{n}}} \sim q(\mathbf{z}|\mathbf{x}),\ \mathbf{z}_{\mathbf{n}} \sim q(\mathbf{z}|\mathbf{x},\mathbf{n}). && \textbf{(11)}
\end{aligned}$$

While training, $\mathcal{G}$ generates two samples: $\mathbf{x_n}$ and $\mathbf{x}_{\not{\mathbf{n}}}$. The former denotes $\mathbf{x_a}$, and the latter is a sample generated in the absence of noise. In Eq. 7, $\mathcal{V}$ is the value function for the joint training of the discriminator and generator; $\mathcal{R}$ is the reconstruction loss, which operates only over $\hat{\mathbf{x}}$; $\mathcal{T}_{\alpha}(\mathcal{G})$ is the triplet loss that prevents network $\mathcal{G}$ from disregarding noise and generating identical samples; and $d(\mathcal{G})$ is the Euclidean distance between the latent variables in the absence and presence of noise. $d(\mathcal{G})$ is a regularizer to encourage original and noisy samples to be located close to one another in the latent space. γ_1, γ_2, and γ_3 are hyperparameters, each constrained to the range of 0 to 1. $\mathbf{z}$ denotes the latent representation of network $\mathcal{G}$, not to be confused with the structured latent variables of the mixture VAE arms.

Importantly, the augmenter learns to generate samples in the vicinity of a given sample in the latent space, which is independent of $\mathbf{s}$ and $\mathbf{c}$. The augmenter does not use any label information in any way. We call this a type-preserving augmenter not because label information was utilized during training or label preservation is guaranteed, but because the augmenter generates samples 'similar' to the original $\mathbf{x}$ that typically remain in the same well of attraction (i.e., belong to the same cluster).

In Supplementary Note 3, Remark 1 further discusses an under-exploration scenario in data augmentation, in which the augmented samples are not conditionally independently distributed and are concentrated around the given sample.

Pairwise coupling

In MMIDAS, the mixture representation is obtained through the optimization in Eq. 4. Not only is it challenging to solve the maximization in Eq. 4 due to the equality constraint, but the objective remains a function of $p(\mathbf{c})$ which is unknown, and typically non-uniform. To overcome this, we begin with an equivalent formulation for Eq. 4 by applying a pairwise coupling paradigm as follows (details of derivation in Supplementary Note 1):

$$\begin{aligned}
\max \quad & \sum_{a=1}^{A}(A-1)\Big(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\Big) - \\
& \sum_{a<b}\mathbb{E}_{q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)}\mathbb{E}_{q(\mathbf{s}_b|\mathbf{c}_b,\mathbf{x}_b)}\left[D_{KL}\left(q(\mathbf{c}_a|\mathbf{x}_a)q(\mathbf{c}_b|\mathbf{x}_b)\|p(\mathbf{c}_a,\mathbf{c}_b)\right)\right] \\
& \text{s.t. } \mathbf{c}_a = \mathbf{c}_b \ \ \forall a,b \in [1,A],\ a<b
\end{aligned} \tag{12}$$

We next relax the optimization in Eq. 12 into an unconstrained problem by marginalizing the joint distribution over a mismatch measure between categorical variables (full derivation in Supplementary Note 2):

$$\begin{aligned}
\max \sum_{a=1}^{A}(A-1)\Big(\mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right]\Big) + \\
\sum_{a<b} H(\mathbf{c}_a|\mathbf{x}_a) + H(\mathbf{c}_b|\mathbf{x}_b) - \lambda\mathbb{E}_{q(\mathbf{c}_a,\mathbf{c}_b|\mathbf{x}_a,\mathbf{x}_b)}\left[d^2(\mathbf{c}_a,\mathbf{c}_b)\right]
\end{aligned} \tag{13}$$

In Eq. 13, in addition to entropy-based confidence penalties known as mode collapse regularizers,[51] the distance measure $d(\mathbf{c}_a,\mathbf{c}_b)$ encourages a consensus on the categorical assignment controlled by $\lambda \geq 0$, the coupling hyperparameter.

We refer to the model in Eq. 13 as MMIDAS. In this model, VAE arms try to achieve identical categorical assignments while independently learning their own style variables. In experiments, we set $\lambda = 1$ universally. While the bottleneck architecture already encourages interpretable continuous variables, this formulation can be easily extended to include an additional hyperparameter to promote disentanglement of continuous variables as in β-VAE.[52] Additional analyses to assess the sensitivity of the performance of MMIDAS to its coupling factor can be found in Supplementary Note 8.

In summary, we can cast the optimization in Eq. 13 in an equivalent a collection of constrained variational models as

follows:

$$\max \sum_{a=1}^{A} \mathbb{E}_{q(\mathbf{s}_a,\mathbf{c}_a|\mathbf{x}_a)}\left[\log p(\mathbf{x}_a|\mathbf{s}_a,\mathbf{c}_a)\right] - \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[D_{KL}\left(q(\mathbf{s}_a|\mathbf{c}_a,\mathbf{x}_a)\|p(\mathbf{s}_a|\mathbf{c}_a)\right)\right] + H(\mathbf{c}_a|\mathbf{x}_a)$$
$$\text{s.t. } \mathbb{E}_{q(\mathbf{c}_a|\mathbf{x}_a)}\left[d^2(\mathbf{c}_a,\mathbf{c}_b)\right] < \epsilon \tag{14}$$

where ϵ denotes the strength of the consensus constraint. Here, $\mathbf{c}_b$ indicates the assigned category by any one of the arms, $b \in \{1,\ldots,A\}$, imposing structure on the discrete variable to approximate its prior distribution.

Distance between categorical variables

$d(\mathbf{c}_a,\mathbf{c}_b)$ denotes the distance between a pair of $|\mathbf{c}|$-dimensional un-ordered categorical variables, which are associated with probability vectors with non-negative entries and sum-to-one constraint that form a K-dimensional simplex, where $K = |\mathbf{c}|$. In the real space, a typical choice to compute the distance between two vectors is using Euclidean geometry. However, this geometry is not suitable for probability vectors. Here, we utilize *Aitchison geometry*,[27, 53] which defines a vector space on the simplex. Accordingly, the distance in the simplex, i.e. $d_{SK}(\mathbf{c}_a,\mathbf{c}_b)$ is defined as $d_{SK}(\mathbf{c}_a,\mathbf{c}_b) = \|clr(\mathbf{c}_a) - clr(\mathbf{c}_b)\|_2$, $\forall \mathbf{c}_a,\mathbf{c}_b \in \mathcal{S}^K$, where $clr(\cdot)$ denotes the *isometric centered-log-ratio* transformation in the simplex. This categorical distance satisfies the conditions of a mathematical metric.

Seeking consensus in the simplex

An instance of the mode collapse problem[54] manifests itself in the minimization of $d_{SK}(\mathbf{c}_a,\mathbf{c}_b)$ (Eq. 13): its trivial local optima encourages the network to abuse the discrete latent factor by ignoring many of the available categories. In the extreme case, the representations can collapse onto a single category; $\mathbf{c}_a = \mathbf{c}_b = \mathbf{c}_0$. In this scenario, the continuous variable is compelled to act as a primary latent factor. The model fails to deliver an interpretable mixture representation despite achieving an overall low loss value.

To avoid such undesirable local equilibria while training, we add perturbations to the categorical representation of each arm. If posterior probabilities in the simplex have small dispersions, the perturbed distance calculation overstates the discrepancies. Thus, instead of minimizing $d^2_{SK}(\mathbf{c}_a,\mathbf{c}_b)$, we minimize a perturbed distance $d^2_{\sigma}(\mathbf{c}_a,\mathbf{c}_b) = \sum_k \left(\sigma^{-1}_{a_k}\log c_{a_k} - \sigma^{-1}_{b_k}\log c_{b_k}\right)^2$. This corresponds to the distance between additively perturbed $\mathbf{c}_a$ and $\mathbf{c}_b$ vectors in Aitchison geometry. Here, $\sigma^2_{a_k}$ and $\sigma^2_{b_k}$ indicate the mini-batch variances of the k-th category, for arms a and b. We showed that the perturbed distance $d_{\sigma}(\cdot)$ is bounded by $d_{SK}(\cdot)$ and non-negative values ρ_u, ρ_l (Proposition 6, Supplementary Note 4). Accordingly, when $\mathbf{c}_a$ and $\mathbf{c}_b$ are similar and their spread is not small within the mini-batch, $d_{\sigma}(\mathbf{c}_a,\mathbf{c}_b)$ closely approximates $d_{SK}(\mathbf{c}_a,\mathbf{c}_b)$. Otherwise, it diverges from $d_{SK}(\cdot)$ to avoid mode collapse.

Consensus score. To report the overall agreement between the categorical representations learned by each VAE arm, at the time of inference, we define the consensus score, α_c as follows.

$$\alpha_c = \frac{2}{A(A-1)}\sum_{a=1}^{A}\sum_{a<b}\mathbb{E}_{\mathbf{x}}\left[I_{\mathbf{c}_{a_\mathbf{x}}}(\mathbf{c}_{b_\mathbf{x}})\right], \qquad A \geq 2 \tag{15}$$

where $\mathbf{c}_{a_\mathbf{x}} \sim q(\mathbf{c}_a|\mathbf{x})$, and $I_{\mathbf{c}_{a_\mathbf{x}}}(.)$ is an indicator function, which equals 1 when $\mathbf{c}_{a_\mathbf{x}} = \mathbf{c}_{b_\mathbf{x}}$.An ideal consensus score of 1 indicates that the VAE arms exhibit identical categorical representations for the discrete diversity present in the data.

Clustering as mapping to a simplex

Clustering methods have been widely used to identify cell types. In MMIDAS, we define clustering as membership probability of each cell in a simplex. Vertices of the simplex, inferred as the discrete latent representation, formalize the notion of clusters (cell types). This clustering approach can be considered as a non-parametric technique, which offers flexible modeling of the data distribution for complex data structures.

Over-parameterized simplex and network pruning to determine the number of discrete classes

Clustering is a classical ill-defined problem, popularly summarized as the lumpers vs splitters problem. Therefore, clustering methods, including non-parametric techniques, cannot determine the number of clusters truly automatically. Instead, the user either needs to provide an estimate of the count or have some prior knowledge about the underlying data structure, forming parameters or hyperparameters of the method. In the same vein, we propose an over-parameterization technique to determine the number of categories: at the beginning of the training, we initialize the dimensionality of the discrete latent space with a number larger than the expected number of cell types, e.g. if one expects around 100 types, set 200 as an upper bound of the number of clusters. To choose the number of clusters, the model evaluates the importance or contribution of each vertex based on the consensus measure among arms. Those vertices which do not contribute with similar probability across arms are pruned. This iterative process continues until the model reaches a minimum level of consensus (a pre-defined hyperparameter) for all vertices in the simplex.

Comparative analysis on cell types

In the absence of ground truth, evaluating the clustering performance can be challenging. To address this for single-cell analysis, we perform two complementary analyses to quantitatively compare sets of cell types identified by different methods in a given dataset.

In the first analysis, we use a classification approach to obtain a measure of *post hoc* identifiability of the inferred labels. To achieve this, we train a Random Forest classifier and calculate the balanced accuracy score for each set of labels. The balanced accuracy score is the average of the sensitivity and the specificity of the classifier, and is particularly useful in the presence of imbalanced clusters.

In the second analysis, we conduct a Silhouette analysis for each set of cell type labels. The Silhouette score provides a measure of how similar a given sample is to other samples in the same cluster as compared to samples in other clusters. The Silhouette score is a real value between -1 and 1. A Silhouette score of 1 corresponds to ideal clusterability where samples in a cluster are identical to each other and different from samples in other clusters. For each set of cell type labels, we report the average Silhouette score per cell type.

For both of the aforementioned analyses, we calculated accuracy and the Silhouette score for two different latent embeddings of the data: a linear embeddings using PCA with 100 components, and a nonlinear embeddings, which is denoted as $\mathbf{z}$ variable in MMIDAS (depicted in Fig. S10). For each group of cell type labels, we report the highest average performance across embeddings. In instances where the cell type labels are of the t-type labels, the highest scores are consistently attained through the linear embedding. Conversely, in the case of MMIDAS, the maximum scores are consistently achieved through the nonlinear embedding. When considering Merged t-type labels, we find that for Patch-seq and SEA-AD datasets, the nonlinear embedding yields the highest performance, while for the remaining dataset, the linear embedding exhibits the best.

Exploring continuous latent factors

To interpret and investigate the continuous representation obtained from the data, we conduct two sets of analysis, namely traversal analysis and regression analysis.

(i) Traversal analysis: Here, we aim to understand the contribution of the continuous latent factors to the observed profiles, while holding the discrete factor fixed. We achieve this by modifying the continuous variables according to the approximate posterior, $q(\mathbf{s}|\mathbf{c},\mathbf{x})$. We explore the latent space, quantify how changes in the continuous variables affect the data, while keeping the discrete factor constant. For a given cell $\mathbf{x} = [x_1, x_2, \ldots, x_M]^T$ characterized by categorical type $\mathbf{c}$, we change $\mathbf{s}$ along the direction that accounts for the maximum amount of variation. Subsequently, we calculate the normalized variation for each feature using the following formulation.

$$V_{\mathbf{s}|\mathbf{c}}(x_m) = \frac{|D(\mathbf{s}_b, \mathbf{c}|x_m) - D(\mathbf{s}_a, \mathbf{c}|x_m)|}{\sigma_m} \quad \textbf{(16)}$$

where x_m denotes the m-th feature of $\mathbf{x}$ (e.g., m-th gene), $\mathbf{s}_a = \mu_{s|c} - \sigma_{s|c}$, $\mathbf{s}_b = \mu_{s|c} + \sigma_{s|c}$, $\mu_{s|c}$ and $\sigma_{s|c}$ denote the mean and standard deviation of of the continuous variable, respectively. D stands for the decoder function, such that

$\hat{\mathbf{x}} = D(\mathbf{s}, \mathbf{c}|\mathbf{x})$. σ_m denotes the standard deviation of the m-th feature across all cells. We use the notation $V_{\mathbf{s}}(x_m)$ in the paper for ease of reference.

(ii) Regression analysis: The regression analysis serves to explore the intricate relationships between the continuous latent factors and the meta-information associated with the data, such as brain regions or disease progression. In the mouse 10x dataset, we use a Random Forest regression model to quantify the predictability of the cortical regions from the continuous representation. Thus, this study explores the hypothesis that gene expression within a cell type can form spatial gradients in the brain.

For the SEA-AD dataset, we opt for a deep regression approach to better capture the intricate and flexible patterns that might exist between the continuous latent factors and the associated meta-data (Fig. S11a). In order to optimize the regression parameters, we use mean squared error (MSE) minimization, as follows:

$$\min \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{y}_n - g\left(\mathbf{s}_n\right)\right)^2 \ , \ \ \mathbf{s}_n \sim q(\mathbf{s}|\mathbf{c}_n, \mathbf{x}_n) \tag{17}$$

where the variable $\mathbf{y}_n$ represents the metadata associated with the n-th cell, such as the disease pseudo-progression score or the BRAAK score. g is a neural network parametrizing the nonlinear fit between $\mathbf{s}$ and $\mathbf{y}$.

We quantify the quality of the fit between the latent variable and the meta-information with the coefficient of determination, R^2.

Datasets

Mouse Smart-seq dataset. This single-cell dataset was released as part of a transcriptomic cell types study.[24] RNA sequencing of $22,365$ neurons from the anterolateral motor cortex (ALM) and primary visual cortex (VISp) regions of adult mice was performed with the Smart-seq (SSv4) platform. The cell isolation technique utilized for Smart-seq involved fluorescence-activated cell sorting (FACS). The reference transcriptomic taxonomy proposed for this dataset has 115 cortical neuronal types.[24]

To focus on genes relevant to neuronal cell types, genes associated with sex or mitochondria, as well as those showing high expression in non-neuronal cells were removed. Subsequently, a subset of $5,032$ highly variable genes (HVGs) was selected.

Mouse 10x isocortex dataset. This single-cell dataset was made available as part of a comprehensive transcriptomic study of the mouse cortex.[12] The dataset was generated using the 10xv2 RNA sequencing platform and includes $519,093$ glutamatergic (excitatory) neurons and $94,493$ GABAergic (inhibitory) neurons collected from seven cortical regions: anterior cingulate area (ACA), auditory cortex (AUD), primary motor cortex (MOp), retrosplenial cortex (RSP), the posterior parietal associative area (PTLp), primary somatosensory area (SSp), and visual cortex (VIS). The proposed transcriptomic cell type taxonomy for this dataset encompasses 113 excitatory and 92 inhibitory cortical neurons. Similar to the gene selection approach used for the mouse SSv4 dataset, a subset of $10,000$ HVGs was selected as input for MMIDAS.

Patch-seq dataset. This multimodal dataset profiles both transcriptomic and electrophysiological features of neurons in the adult mouse visual cortex. Profiles of inhibitory neurons have already been released by,[8] and those of the excitatory neurons will be available on the Brain map portal (`https://portal.brain-map.org/`). Following patch-clamp recordings to measure intrinsic electrophysiological features in individual cells, scRNA-seq was performed on the same cells using the SSv4 platform.

The reference taxonomy for this dataset includes 92 cell types,[8] which was established by mapping the transcriptomic profiles of neurons to the taxonomy proposed in.[24] The dataset comprises data from a total of $5,165$ cortical neurons, each characterized by both a gene expression vector and a set of electrophysiological features. For gene selection, a two-step filtering procedure was followed. In the first step, genes associated with sex or mitochondria, as well as those exhibiting high expression levels in non-neuronal cells were excluded. In the second step, we removed genes that showed strong associations with the experimental platform, thereby reducing any potential platform-related bias. After applying these steps, we obtained a subset of $1,252$ HVGs for our analysis.

For the electrophysiological modality, following previous studies,[8, 22] we applied principal component analysis (PCA) to each time series. We retained principal components sufficient to explain 99% of the data, resulting in a total of 108 PCs to summarize the electrophysiological time series. We also included an additional 25 measurements of intrinsic physiological features that were computed by the IPFX library (`https://ipfx.readthedocs.io/`). To ensure consistency across experiments, the PCs were scaled to have unit variance, while the remaining IPFX

features were individually normalized to have zero mean and unit norm. The following abbreviations are used for labeling electrophysiological features: action potential (AP), first action potential (AP1), inter-spike interval (ISI), amplitude (Amp), membrane potential (V), current (i), long square (LS), and short square (SS). A more detailed description of each electrophysiology feature can be found in the "Electrophysiology Overview" document available at https://help.brain-map.org/display/celltypes/Documentation.

Seattle Alzheimer's disease dataset (SEA-AD). This dataset is derived from the recently published[25] Seattle Alzheimer's Disease Cell Atlas (SEA-AD) data, which is publicly accessible through the SEA-AD portal (https://SEA-AD.org/). The dataset consists of scRNA-seq data collected via 10xv3 from the Middle Temporal Gyrus (MTG) tissue of the human brain. Additionally, it includes essential meta-information such as cognitive score, BRAAK score, and demographic details.

The cohort consists of 84 donors covering a broad spectrum of Alzheimer's disease (AD) pathology. The dataset also includes a continuous disease pseudo-progression score (DPS) per donor, ranging from low (score 0) to high (score 1), which was derived using a quantitative image-based neuropathology approach, enabling the ordered arrangement of donors based on their burden of AD pathology.

Here, our focus was on investigating four neuronal subclasses that are vulnerable to AD pathology: $299,563$ L2/3 Intratelencephalic (IT) glutamatergic neurons, $152,052$ L4 IT glutamatergic neurons, $45,872$ Somatostatin-expressing (Sst) GABAergic neurons, and $79,976$ Parvalbumin-expressing (Pvalb) GABAergic neurons. The dataset comprises 42 transcriptional types, referred to as "supertypes." These supertypes were identified through an iterative projection of MTG data from young neurotypical reference donors [39] to itself, and by using scANVI[23] to predict the class label, as explained in.[25]

In line with previous datasets, we performed a gene selection process to identify genes relevant to neuronal types. Genes associated with sex or mitochondria, as well as those exhibiting high expression in non-neuronal cells, were excluded. To focus on tissue-specific biology while reducing tissue-specific or technical variability, we independently selected the top 2000 HVGs for each donor. The union of these gene subsets was then utilized for our single-cell analysis. Both the gene selection and single-cell studies were carried out separately for each subclass in this dataset.

KEGG Pathway and gene module database

In the continuous traversal study, we explored total variation across 18 pathways identified in the KEGG database.[31] Each pathway encompasses a gene module formed by the intersection of KEGG-suggested genes and those available in our datasets. Collection of KEGG pathways from the KEGG database[31] along with the corresponding genes can be found at https://github.com/AllenInstitute/MMIDAS/blob/main/KEGG/KEGG.toml.

Data availability

The mouse Smart-seq data Single-cell transcriptomic data are available at the NCBI Gene Expression Omnibus (GEO) under accession GSE115746. The raw and processed sequencing mouse 10x isocortex dataset is deposited in the NeMO Archive for the BRAIN Initiative Cell Census Network, which is available at https://assets.nemoarchive.org/dat-jb2f34y. The inhibitory Patch-seq transcriptomic data are available at https://dandiarchive.org/dandiset/000020/. The excitatory data will be publicly available soon. The SEA-AD dataset is available at https://portal.brain-map.org/explore/seattle-alzheimers-disease/seattle-alzheimers-disease-brain-cell-atlas-download?edit&language=en.

Code availability

Code for MMIDAS implementation and analysis are available at https://github.com/AllenInstitute/MMIDAS.

Acknowledgements

We thank Polina Kosillo for help with cellular pathway analysis and Michael Hawrylycz for helpful feedback on the manuscript. We wish to thank the Allen Institute for Brain Science founder, P. G. Allen, for his vision, encouragement and support. This work was supported by the NIH grants 1RF1MH128778-01 and 1RF1MH125317-01.