The Specious Art of Single-Cell Genomics

Tara Chari[1], Joeyta Banerjee[2], and Lior Pachter[1,2,*]

[1]Division of Biology and Biological Engineering,
California Institute of Technology, Pasadena, California

[2]Department of Computing and Mathematical Sciences,
California Institute of Technology, Pasadena, California

[*]Address correspondence to Lior Pachter (lpachter@caltech.edu)

September 21, 2021

Abstract

Dimensionality reduction is standard practice for filtering noise and identifying relevant dimensions in large-scale data analyses. In biology, single-cell expression studies almost always begin with reduction to two or three dimensions to produce ‘all-in-one’ visuals of the data that are amenable to the human eye, and these are subsequently used for qualitative and quantitative analysis of cell relationships. However, there is little theoretical support for this practice. We examine the theoretical and practical implications of low-dimensional embedding of single-cell data, and find extensive distortions incurred on the global and local properties of biological patterns relative to the high-dimensional, ambient space. In lieu of this, we propose semi-supervised dimensionality reduction to higher dimension, and show that such *targeted* reduction guided by the metadata associated with single-cell experiments provides useful latent space representations for hypothesis-driven biological discovery.

Introduction

The high-dimensionality of “big data” genomics datasets is considered a practical nuisance as it requires dimensionality reduction to filter noise, enable tractable computation, and to facilitate exploratory analysis. Trial and error application of common techniques has resulted in a currently accepted workflow combining initial dimensionality reduction to a few dozen dimensions using principal component analysis (PCA) with further non-linear reduction to two dimensions using t-SNE [1] or UMAP [2–5].

The methods used are largely unsupervised, with dimensionality reduction applied directly to the data modality of interest, consisting of expression, genetic, or other genomic data, without regard to metadata that can include cell type in the case of single-cell experiments, spatial information, environmental conditions, etc. Instead, data is reduced to two dimensions with metadata layered on the image to confirm hypotheses and to visually detect interesting patterns or anomalies.

While it is often claimed that quantitative properties of genomics data are recapitulated in such two-dimensional spaces [4, 5], there is surprisingly little supporting theory for this. For example, while the popular t-SNE and UMAP methods were intended to faithfully represent local and global structure of high-dimensional data in two or three dimensions, there is evidence they fail in this regard [3, 6], and theorems providing guarantees on the embeddings rely on numerous assumptions unlikely to hold in practice and ignore the coupling of PCA to non-linear methods [7]. Yet, in single-cell gene expression analysis, PCA pre-conditioned t-SNE and UMAP visuals are used extensively to infer or confirm relationships between cells where apparent closeness in two-dimensional spaces are assumed to translate to transcriptional compatibility in higher dimensions. Such two-dimensional embeddings are used in qualitative and quantitative manners to 'validate' transcriptional similarities between clusters [3, 7] and assess the 'quality' of data integration [8–10], to construct representations for RNA velocity embedding and trajectory inference calculations [11–13], and to direct focus of downstream analysis or experimentation [4, 12, 14, 15].

Dimensionality reduction methods are used across many disciplines, but single-cell gene expression analyses are particularly well-suited to explore the properties of various low-dimensional reduction methods. Analyses of single-cell RNA-seq data with biological ground truth can lead to testable hypotheses and rigorous benchmarks of predictions [16]. However, the organization of tissues as composites of distinct cell types is not well understood [17], and thus meaningful reduction of single-cell data for facilitating analyses is of great relevance, creating a pressing need to assess the validity of omnipresent techniques such as t-SNE and UMAP. On a more fundamental level, estimates of the intrinsic dimension of transcriptomes may yield insights into the nature of transcriptional programs.

We therefore investigate dimensionality reduction for single-cell gene expression, focusing on fundamental obstacles in embedding high-dimensional data in two dimensions and on the scale of subsequent distortions. Our results lead us to consider semi-supervised reduction into higher dimensions as a way to circumvent problems with current approaches.

Results

Distortion by Projection to Low Dimensions

To understand the effects of standard two-dimensional reduction techniques on single-cell RNA-seq data, we began by examining the extent of distortion in embedded cells. A key result on linear low-distortion embedding of points in Euclidean space is the Johnson-Lindenstrauss Lemma [18], which provides a sufficiency condition for low-distortion dimensionality reduction: the preservation of pairwise distances of m points within a factor of $1 \pm \epsilon$ can be accomplished with order $log(m)/\epsilon^2$ dimensions. While this shows that dimensionality reduction can preserve the structure of high dimensional data, the number of required dimensions is much higher than two or three. Using the constant factor from [19], distortion of pairwise distances within 20% for a dataset of 10,000 points can be achieved with at least 1,842 dimensions.

To further understand the extent of distortion in two dimensions, we first focused on a particularly difficult case: the embedding of equidistant points. It is impossible to embed greater than $n+1$ equidistant points in $\mathbb{R}^k$ for $k \leq n$ (Supplementary Note 1), but even relaxing the equidistance constraint to near-equidistance, where for any three points in $\mathbb{R}^n$ a pair will be at unit distance, one can only accommodate seven points in $\mathbb{R}^2$ or ten in $\mathbb{R}^3$ [20]. Even settling for near-equidistance is impossible: the ratio of the maximum distance, D, to minimum distance, d, among n points in two dimensions grows as $O(\sqrt{n})$ [21] (Supplementary Note 2). Moreover, the distortion of equidistant points can be particularly acute with PCA, often used to "pre-condition" data. PCA of equidistant points is tantamount to applying a random projection (Supplementary Note 3), and as a result projected points display numerous mirages of structure in two dimensions (Supplementary Fig. 1).

We then asked whether equidistant points are present in biological data, and if dimensionality reduction methods such as t-SNE and UMAP applied to PCA preconditioned data induce distortion. We scanned the Seurat-integrated [8] ex- and in-utero mouse embryo dataset (at the E10.5 stage) from [22] for cells with pairwise distances close to the same value (see Methods). This data was chosen because the structure of the two-dimensional embedding was used as part of the validation for ex-utero cultured mouse embryos precisely recapitulating in-utero development. To limit the search space, we examined cells only within the Chondrocytes and Osteoblasts (Fig. 1a) (see Methods). We found visible distortion of equidistant points in 1,511,502 distinct 'near and equidistant' groups of cells (Fig. 1b) and 1,020,120 distinct 'far and equidistant' groups (Fig. 1c), with both appearing similarly clustered or distributed in two-dimensions, despite their differing properties in ambient space. We measured changes in variance within these groups, ranging from 3 to 9 equidistant cells, as well as distortion of distances, i.e. the ratio of the maximum to minimum (max/min) pairwise distances (Supplementary Fig. 2a,e). For 'near and equidistant' cells, variance of pairwise distances in the UMAP space increased, on average, 135- to 1,040-fold in comparison to the high dimensional variance. Distortion of distances increased 25- to 95-fold (Supplementary Fig. 2b,c). Though large, distortion in the higher 15-dimensional PCA space was less pronounced, with a 214-fold increase in variance and a 4.5-fold increase in distortion (Supplementary Fig. 2b,c). This was also the case for higher dimensional PCA spaces (Supplementary Fig. 3). 'Far and equidistant' cells showed similar trends, with 321- to 1029-fold larger variance and 22- to 39-fold larger distortion ratios (Supplementary Fig. 2f,g). We then examined 3,763,130 groups of equidistant cells with distances centered around the mode of all pairwise distances. We found groups ranging from 3 to 10 cells per group (Supplementary Fig. 4a,b), and found sizeable changes for these large numbers of groups: 221-fold to 1,226-fold increases in variance and 25-fold to 68-fold distance distortion ratios (Supplementary Fig. 4a,b).

Importantly, this extent of distortion is not specific to the data in [22]. We separately analyzed a (10x sequenced) mouse ventromedial hypothalamus (VMH) neuron dataset [23], consisting of a large number of cell types and a rich experimental design. Within the Estrogen Receptor 1-expressing cells (Esr1_6) we uncovered 10,375,096 distinct groups of equidistant cells, with up to 14 cells per group (Supplementary Fig. 5a,b). We found variance increases ranging from 42-fold to 77-fold in UMAP space, and 443-fold to 625-fold in t-SNE space (Supplementary Fig. 5c). Distortion ratios ranged from 104- to 154-fold (Supplementary Fig. 5d). Zooming out to cell types, distortion was also present in distances between their equidistant *centroids* (Supplementary Fig. 5f), with 3.1- to 4.3-fold distortion in both UMAP and t-SNE space (Supplementary Fig. 5g). These 110 distinct groups of cell type centroids encompassed up to 7 types, highlighting that even large-scale structure of cell types is not accurately reflected with t-SNE or UMAP. For both datasets, we also identified large distortion ratios within groups of nearest neighbors. The distortion ratios of distances between each cell's 10 nearest neighbors ranged from 6.5-fold to 7.2-fold that of the high-dimensional space for the integrated embryo data (Supplementary Fig. 2d,h, 4e), and 17-fold to 64-fold in the VMH data (Supplementary Fig. 5e), but with minimal distortion for both by PCA (Supplementary Fig. 6). Though it is difficult to retain the correct nearest neighbors even in higher dimensions [6, 18],

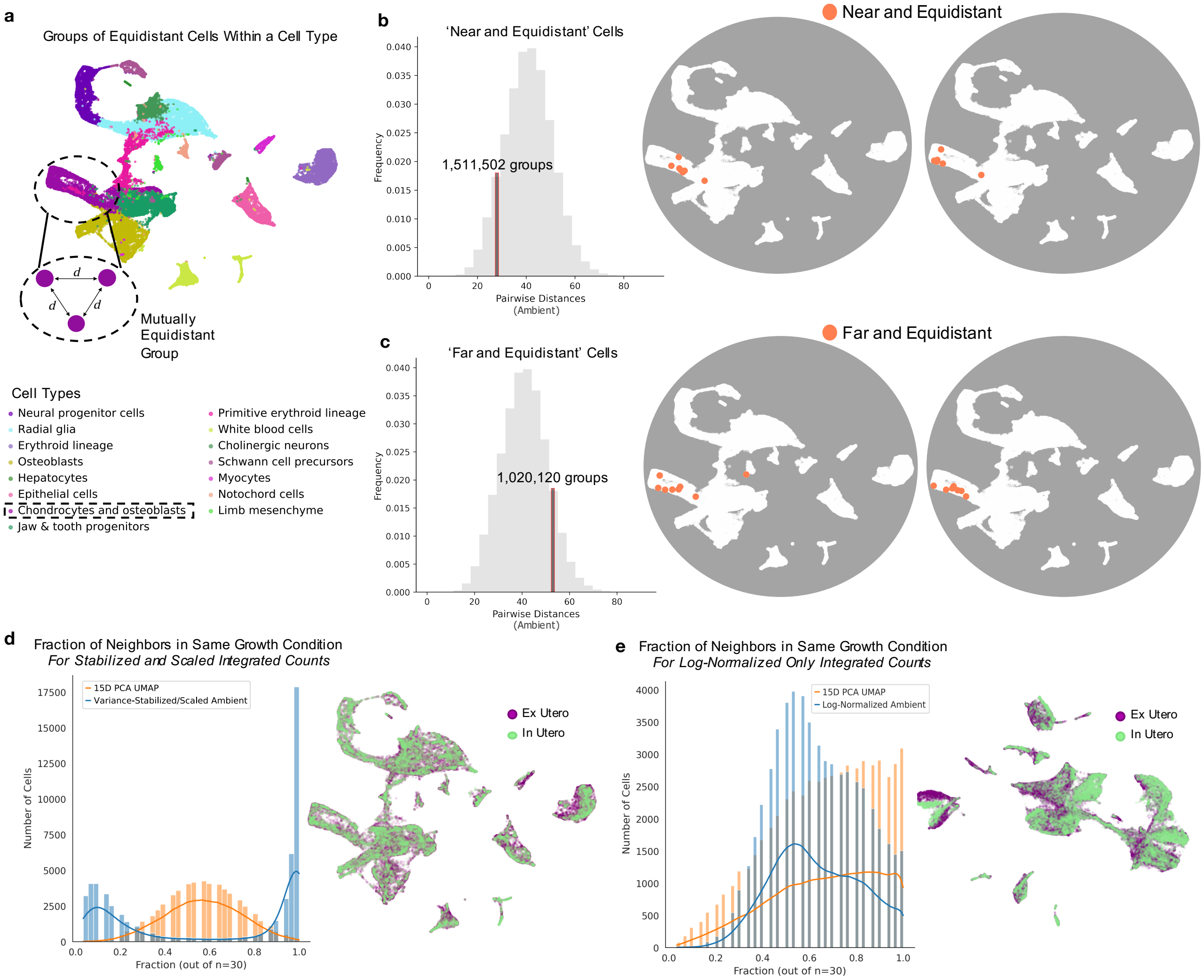

Figure 1: Distortion in Two-Dimensional Embeddings for Integrated Ex- and In-Utero Embryo E10.5 **a)** Determination of groups of mutually equidistant cells, from the ambient, high-dimensional space, in a given cell type. **b)** Selection of 'near and equidistant' groups from **a** and their respective positions in the generated UMAP. **c)** Selection of 'far and equidistant' groups from **a** and their respective positions in the generated UMAP. **d)** Distributions for number of 30 nearest neighbors per cell in the same growth condition (ex- or in-utero) for the variance-stabilized and scaled integrated E10.5 embryo data (post-log normalization) and PCA/UMAP embedding following the original study. UMAP embedding shown on the right, colored by growth condition. **e)** Distributions for number of 30 nearest neighbors per cell in the same growth condition (ex- or in-utero) for the log-normalized integrated E10.5 embryo data and PCA/UMAP embedding following the original study. UMAP embedding shown on the right. [Code]

t-SNE/UMAP embeddings displayed large Jaccard distances (dissimilarities) from the neighbors in ambient dimension, with an average consistently above 0.7 and increasing dissimilarity with the size of the dataset. The distortion of neighbors is even worse for two-dimensional PCA, which together is consistent with other findings on the poor preservation of local neighborhoods by both PCA and t-SNE/UMAP non-linear reduction methods [3, 6] (Supplementary Fig. 7). A similar level of dissimilarity was maintained between the t-SNE/UMAP spaces and the PCA-reduced spaces each was coupled to (second column, Supplementary Fig. 7). Thus, contrary to conventional wisdom, inaccurate neighborhood representations occur locally and globally across these projections.

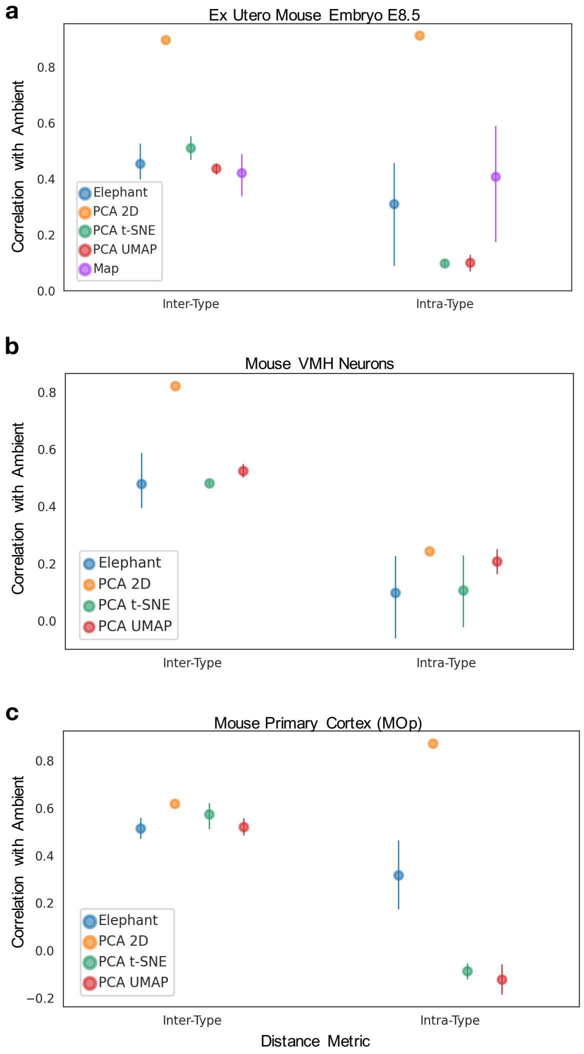

Figure 2: Correlation Benchmarks for Two-Dimensional Embeddings. a) Comparison of correlation metrics for world map and 'von Neumann' Picasso embeddings of the ex-utero E8.5 Mouse embryo dataset, with the other 2D embedding baselines. Bars denote the 95% C.I. **b)** Comparison of correlation metrics for the 'von Neumann' Picasso embedding of the mouse VMH neurons (SMART-Seq) dataset, with the other 2D embedding baselines. Bars denote the 95% C.I. **c)** Comparison of correlation metrics for the 'von Neumann' Picasso embedding of the MOp dataset, with the other 2D embedding baselines. Bars denote the 95% C.I. [Code a] [Code b] [Code c]

As implied by the distortion of nearest neighbor distances, we find a warping of cell relationships beyond the 'extreme' case of equidistant cells. In the integrated ex- and in-utero E10.5 dataset, we assessed the fraction of each cell's nearest neighbors with the same label as the cell, measuring whether the embeddings accurately reflect the extent of mixing of ex- and in-utero cells in the integrated data (Fig. 1d,e). For the 'Variance-stabilized and Scaled' ambient data, the UMAP embedding to two-dimensions displayed a unimodal distribution with a mode of 0.53 (approximately half of a cell's neighbors share the same condition) while the ambient space reflected a bimodal distribution with extreme modes near 0.1 and 1.0, denoting less 'mixing' of neighbor conditions (Fig. 1d). The 'Log-Normalized' ambient data (counts prior to stabilization and scaling) revealed the opposite trend, with the mode of the UMAP embedding distribution at 1.0, in contrast with the 0.5 mode for the ambient data itself, indicating less 'mixing' in the UMAP space (Fig. 1e). This suggests that methods like UMAP can create illusions of more or less 'mixing' compared to the ambient space. This observation is again concordant with previous results highlighting largely non-overlapping nearest neighbors between ambient and reduced spaces [3, 6].

These drawbacks of two-dimensional embeddings suggest their visual utility as 'faithful' data representations is limited. Nevertheless, one might imagine that (PCA pre-conditioned) t-SNE or UMAP embeddings may yield qualitatively meaningful representations, thanks to the optimizations they perform. To assess whether this is the case, we constructed metrics to quantify the ability of embeddings to reveal relevant biological characteristics i.e. local and global properties which represent implicit measurements often made by eye or inferred from the visuals (see Methods). We measured inter- and intra-distances with respect to biological labels in two-dimensional and ambient spaces, examining the correlation of these distances to assess whether embeddings capture important relative relationships between cells (see Methods) (Supplementary Fig. 8). Inter-distances represent distances between label groups, while intra-distances represent average pairwise distances or internal variance within each group (see Methods). We examined these properties in three datasets, ex-utero cultured mouse embryos (E8.5 stage) [22] (Fig. 2a), (SMART-Seq) mouse VMH neurons [23] (Fig. 2b), and mouse primary motor cortex (MOp) cells [24] (Fig. 2c) once embedded with t-SNE or UMAP, pre-conditioned with PCA. On average the t-SNE and UMAP embeddings respectively displayed low maximum correlations of 0.57 and 0.52 for inter-type (inter-'cell type') distances, and 0.10 and 0.21 for intra-type distances, 2.7-fold lower than the inter-correlations (Fig. 2).

As a control experiment, we developed an autoencoder framework to fit cells from any dataset to an arbitrary shape (defined by user-specified (x, y) coordinates) while preserving cell-to-cell distances (see Methods). The latter was achieved by including in the neural network loss function, a term minimizing reconstruction error between the ambient space and the decoder output (see Methods). This creates two-dimensional representations for the cells that attempt to recapitulate the ambient data under the constraint of approximating a user-specified shape. We call this method Picasso in homage to the eponymous artist's skill in imitating other artistic works. We compared correlations of inter- and intra-distances between Picasso embeddings with those of t-SNE, UMAP and PCA, for each of the three datasets. For the ex-utero dataset, data was fit to the shape of a world map and the 'von Neumann elephant' [25, 26] ('Map' and 'Elephant', Fig. 2a), and to the 'von Neumann elephant' for the VMH and MOp datasets ('Elephant', Fig. 2b,c) . Each Picasso embedding demonstrated comparable metrics to the respective t-SNE and UMAP projections, even

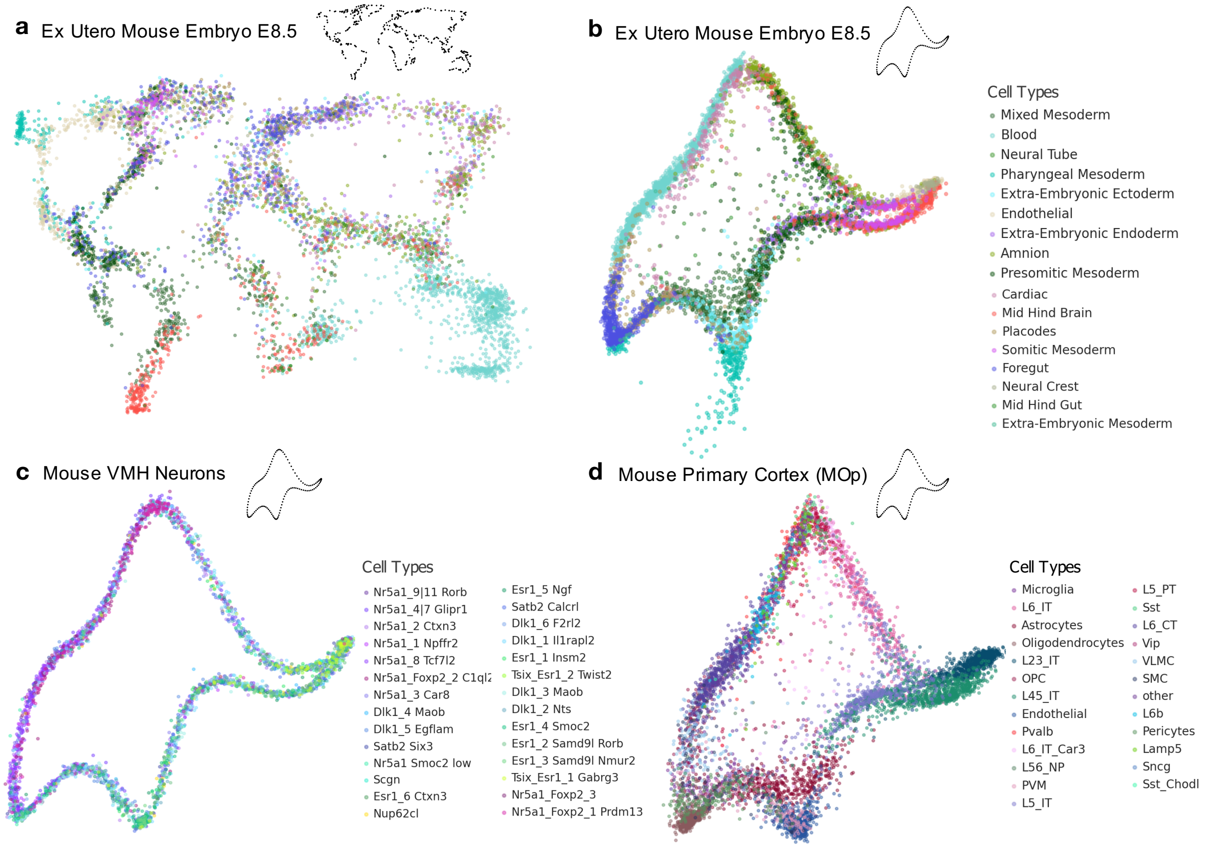

Figure 3: Picasso Embedding. a) Picasso embedding of the ex-utero mouse embryo E8.5 data fit to a world map boundary. **b)** Picasso embedding of the ex-utero mouse embryo E8.5 data fit to a 'von Neumann' elephant. **c)** Picasso embedding of the mouse VMH neuron (SMART-Seq) data fit to a 'von Neumann' elephant. **d)** Picasso embedding of the MOp data fit to a 'von Neumann' elephant. [Code a,b] [Code c] [Code d]

improving upon the t-SNE/UMAP intra-type correlations by 2-3% in both ex-utero shapes (Fig. 2a), and up to 4.6% in the MOp embedding (Fig. 2c).

Thus, Picasso can quantitatively represent these inferred relationships similarly to, or better, than the respective t-SNE/UMAP embeddings, while producing user-specified shapes. A single dataset, e.g. the ex-utero cultured mouse embryos, can be equally well represented as a world map or elephant (Fig. 3a,b), while a single shape, e.g. von Neumann's elephant, can represent two different datasets (mouse VMH neurons vs. MOp cells) (Fig. 3 c,d). This shows that two-dimensional visualizations, obtained with Picasso, t-SNE or UMAP, are not in any sense canonical and often quantitatively poor in absolute terms, casting doubt on their value for biological inference, particularly for understanding patterns of variation within, rather than between cell types (intra-distances, Fig. 2; Supplementary Fig. 8,9). Even for inter-type relationships, the rankings of cell-type neighbors as determined by pairwise distances between each type are also warped or even reversed, with separation of the cell types possibly reduced, in comparison to the ambient or input PCA space for PCA-coupled embeddings (see Methods) (Supplementary Fig. 10). These trends on limited recapitulation of inter- and intra-relationships were also present in a non-biological machine learning benchmark dataset (Supplementary Fig. 11), where high accuracy prediction is possible in higher dimensions [27].

The arbitrary nature of two-dimensional embedding was further highlighted by our ability to easily embed all the datasets as another shape: the 'flower' (Supplementary Fig. 9a,e). Many of the Picasso-generated latent spaces also displayed comparable correlations to the densVis algorithms [28] designed to improve spatial distribution of cells in t-SNE/UMAP embeddings (densSne/densMAP) (Supplementary Fig. 9c,d,f,h). Interestingly, even untrained Picasso was able to produce a two-dimensional embedding with comparable performance to existing methods. That is, the He initialization [29] of the neural network is competitive with t-SNE, UMAP and densVis on these metrics implicitly assumed to be accurately represented (Supplementary Note 4). This is not surprising as the He Gaussian initialization of weights [29] mimics the structure-preserving properties of Gaussian, random projections of data which provide the constructive proof of the Johnson-Lindenstrauss lemma [30–35] (Supplementary Note 4).

Semi-Supervision for Targeted Latent Space Construction

Given the limits on two-dimensional embedding of high-dimensional data and the resultant pitfalls of relying on such, largely arbitrary, visualizations, there is a need to rethink the role of such embeddings in single-cell expression analysis. Rather than relying on unsupervised two-dimensional embedding and subsequent visualization for validation, the goal of dimensionality reduction should be to improve results of analysis by virtue of filtering noise and extracting relevant features. In this regard, unsupervised dimensionality reduction, that does not account for the increasingly complex nature of multi-labeled genomics data including competing features in varying abundance, is likely to be suboptimal. Several publications have demonstrated advantages of supervised dimensionality reduction in constructing interpretable and separable latent space structures for marker gene extraction [36] and cell type label prediction [37], and we hypothesized that an extension of this work to multi-class, multi-label (MCML) data where each class (e.g. cell type, experimental condition, spatial location) contains a label for each cell, would improve upon unsupervised methods.

Building on the linear-decoded autoencoder framework [38] as adapted in the Picasso algorithm, we implemented a label-based cost defined by the Neighborhood Component Analysis algorithm (NCA) [39], which optimizes the likelihood that cells of the same label are near each other ('pushing' equivalently labeled cells together) in the latent space without overfitting (Fig. 4a ; Supplementary Fig. 12) [39]. This algorithm, which we call 'MCML', combines reconstruction error with the label-aware cost to optimize nearest neighbor structure of a latent space while maintaining the higher dimensional input structure (Fig. 4a) (see Methods). Within this supervised, or semi-supervised, framework we incorporate multiple labels across the discrete and continuous spectrum, and predict labels for unlabeled cells (Fig. 4a) (see Methods).

With the embedding-based distortions of the 'mixed' nature of cells between the ex- and in-utero conditions (Fig. 1d,e), we sought to more quantitatively determine which cell types contained the greatest differences between the growth conditions in higher dimensions. We applied MCML to 'push' together cells within the same cell type and growth condition, reducing the stabilized and scaled count matrix to 15 dimensions (following the original study). From this we determined that

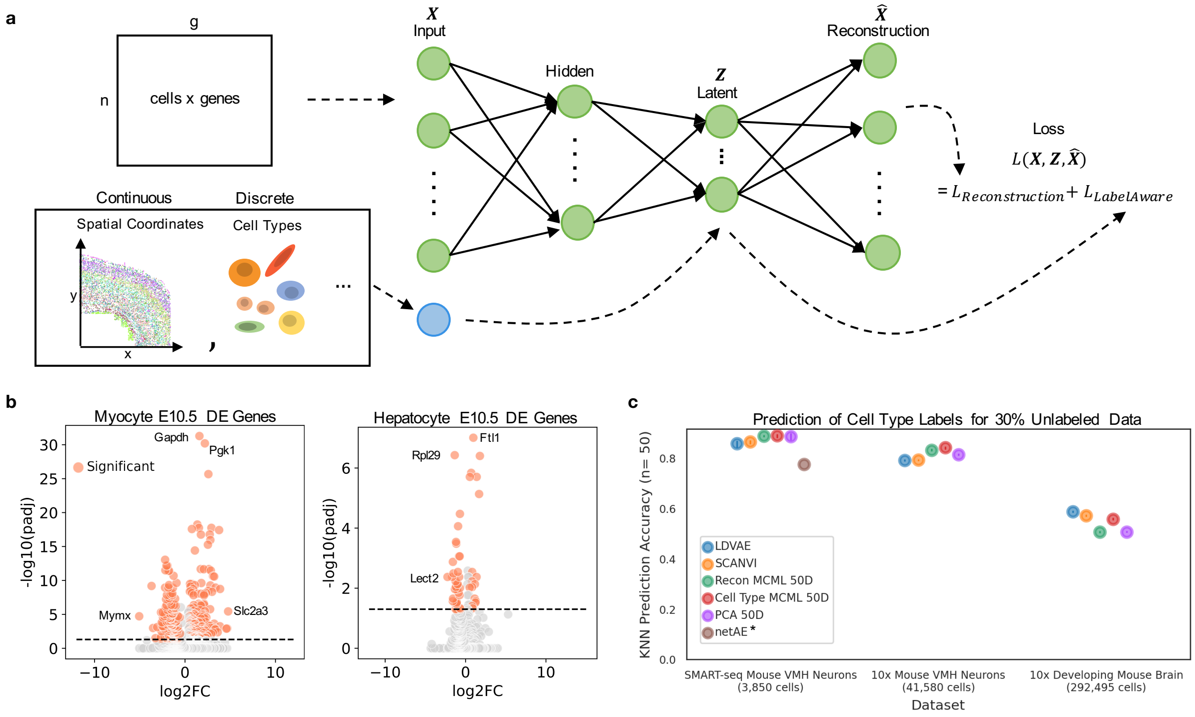

Figure 4: Integration of Discrete Metadata with MCML Reduction. a) General diagram of autoencoder structure utilized for multi-class, multi-label (MCML) tasks. **b)** Differentially expressed (DE) genes determined for the two 'internally-distant' integrated embryo E10.5 cell types using the non-parametric Wilcoxon test (see Methods). P-values adjusted for multiple testing with Benjamini-Hochberg correction. Colored dots denote significant genes, as per the original study, with padj < 0.05 and log2FC > 1. [Code] **c)** Cell type label prediction accuracy comparisons between MCML ('Cell Type MCML') with 70% randomly labeled cells and other label-aware methods (SCANVI and netAE) or only reconstruction error ('Recon MCML'), LDVAE, and other non-label based methods (see Methods). *NetAE could not be run for the two 10x datasets as it attempted to allocate over 2TB of memory, larger than our server capabilities. Bars denote the 95% C.I. [Code]

the myocytes and hepatocytes contained the largest distances between their ex- and in-utero cells (see Methods), concordant with the findings of [22] that myocytes contained greater disagreement in marker genes between the conditions (Extended Data Fig. 8 in [22]), but also highlighting hepatocytes as a cell type of interest. We then extracted differentially expressed (DE) genes between the conditions, which contribute to these internal separations, finding 398 DE genes for the Myocytes including upregulation of the GAPDH housekeeping gene in ex-utero cells, and 59 genes for the Hepatocytes including downregulation of the hepatokine LECT2 in ex-utero cells (Fig. 4b) (see Methods).

Given increasing numbers of available experimental labels and gold-standard cell type annotations across single-cell datasets, we also applied MCML to the prediction of unlabeled cells. We found comparable cell type prediction to SCANVI, a single-cell annotation package in the widely used scVI framework [40], and netAE which meshes modularity and separation of *a priori* defined clusters into a label-aware cost for latent space construction and label prediction [37] (Fig. 4c; Supplementary Fig. 13a). We were unable to run netAE for the two larger datasets as it attempted to allocate over 2TB of RAM for each (see Methods). We also measured classification accuracy for LDVAE [38], PCA, and Recon MCML (MCML with reconstruction error only) as non-label based baselines (Fig. 4c), all in 50 dimensions. Cell type labels were predicted on the 30% of unlabeled cells, for SMART-Seq and 10x sequenced mouse VMH neurons [23], and 10x sequenced mouse developing brain [41], using a KNN classifier where the prediction was made according to the majority label from a cell's 50 nearest neighbors (Fig. 4c) (see Methods). Across these datasets 'Cell Type MCML', MCML with cell type labels, demonstrated comparable performance to existing methods, with a 1.8% improvement in accuracy over SCANVI, a 2.7% improvement over LDVAE, and a 14.6% improvement on netAE for the VMH datasets (Fig. 4c). Up to 37% improvement was also found as compared to supervised, two-dimensional methods (Supplementary Fig. 14). Though LDVAE performed the best for the developing brain data, the maximum accuracy was only 0.58 (Fig. 4c; Supplementary Fig. 14).

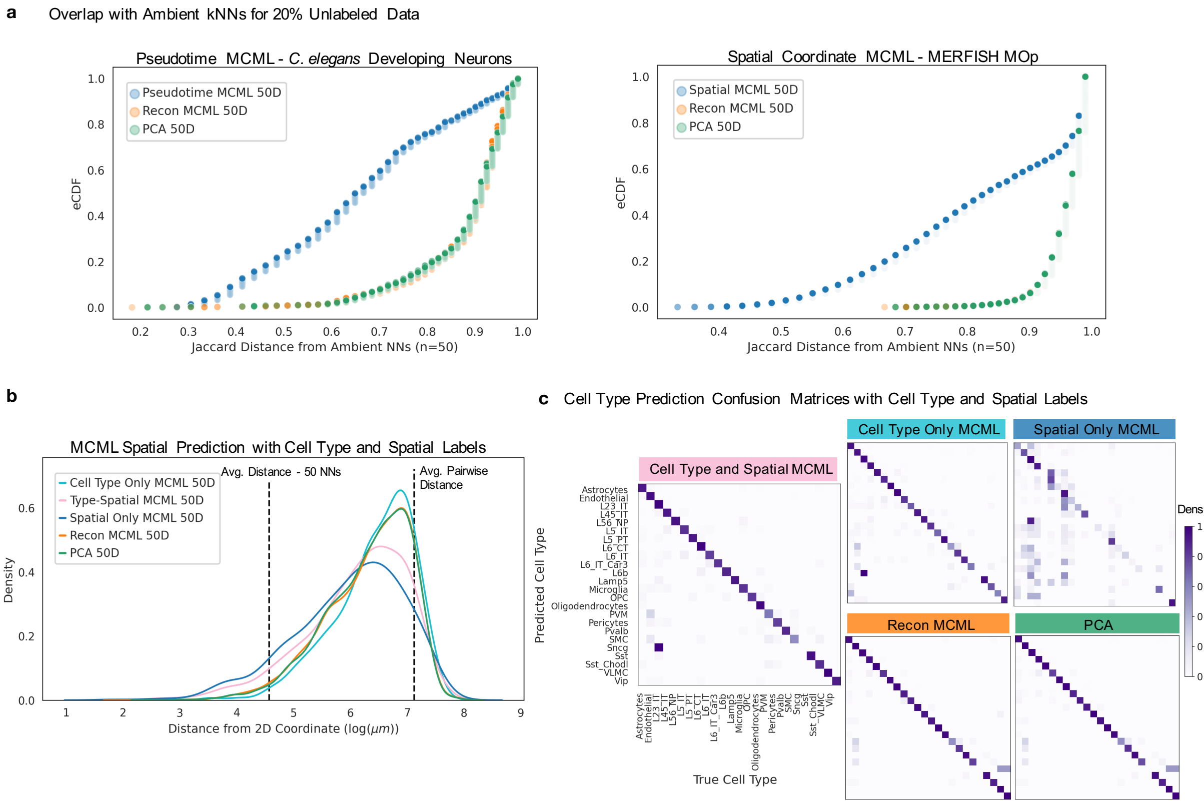

Figure 5: Extension of MCML to Continuous and Discrete Metadata. a) Jaccard distance distributions of each cell's 50 nearest neighbors in embedding space (50D) from their neighbors in the original continuous (pseudotime or spatial coordinate) space. Label-aware MCML compared to PCA and Recon (Reconstruction error only) MCML. 80% of cells were randomly labeled for MCML. **b)** Distributions of Euclidean distance of predicted spatial coordinates from the original coordinates for cell-type or spatial coordinate only MCML, spatial and cell type labeled MCML, and the respective baselines.**c)** Spatial-Type MCML confusion matrix for cell type prediction compared to confusion matrices for the single-class MCML and baseline latent spaces. [Code]

We then extended the cost function to integrate cells' continuous positions, from zero to one in pseudotime for developing *C. elegans* neurons [14], or spatial locations (physical two-dimensional coordinates) from the MERFISH MOp dataset [24], using pseudotime or spatial pairwise cell distances to weight cell-to-cell relationships in the latent space, in addition to reconstruction error (see Methods). This resulted in smaller Jaccard distances (dissimilarity) [6] of the MCML nearest neighbors to the ambient (continuous) neighbors, with average dissimilarities of 0.67 and 0.81 in MCML representations compared to averages of 0.88 and 0.96 in the baselines for the *C. elegans* and MOp embeddings (Fig. 5a) (see Methods). 1.0 denotes completely non-overlapping sets of neighbors. From here we combined discrete (cell type) and continuous (spatial location) class labels into the latent space construction, retaining the predictive properties of each of the classes individually when tested with 20% of cells unlabeled. Both the 'Type-Spatial' and 'Spatial-Only' MCML representations reduced the distance of predicted locations from the actual coordinates by 10.9% and 14.6% respectively (by 77.3 μm and 103.8 μm) compared to the baselines (Fig. 5b). The 'Type-Spatial' and 'Type-Only' representations also demonstrated comparable cell type prediction, with overall accuracy of 0.91 and 0.92 (Fig. 5c). These results show that our framework enables prediction of type, space, and time for single cells.

Additionally, we defined a label-aware loss to bias decoder reconstruction towards improving metrics of interest, such as the recapitulation of intra-label variances. We denote this method as 'bMCML' ('biased MCML') (see Methods). As a proof-of-concept we optimized latent space construction for high correlation of intra-sex (Supplementary Fig. 15a,b) or intra-type (Supplementary Fig. 15c,d) distances to the ambient space, as intra-correlations were lower than inter-correlations for most latent representations (Supplementary Fig. 15). For the SMART-Seq and 10x VMH neuron datasets, intra-sex correlations were increased by at least 126% and 26%, while intra-type correlations were increased by at least 91% and 12% as compared to the PCA and Recon MCML baselines. Although bMCML may reduce the accuracy of other metrics not in the loss, e.g. inter-distances (Supplementary Fig. 15), this demonstrates targeted preservation of desired patterns and an alternative to unsupervised reconstruction which may not capture these specific properties .

Other benefits of the MCML framework include the ability of reconstruction loss to better preserve metric correlations in rare and orthogonal cells (i.e. gene expression not shared by other cell types) as opposed to PCA (Supplementary Fig. 16) (see Methods), as PCA is designed to find directions of maximal variance which may suppress low, orthogonal expression when only top PCs are selected. The linear decoder layer also provides interpretability, analogous to the linear transformation of data with PCA, for easy extraction of genes which contribute to each of the latent dimensions (Supplementary Fig. 17) [38]. With the use of a nonlinear encoder, despite the linear decoder, and mini-batch training, MCML provides a faster and more accurate implementation of NCA itself as compared to the sklearn implementation (Supplementary Fig. 18).

Discussion

Despite claims that common dimensionality reduction techniques for single-cell genomics data preserve local [4] and/or global [5] cell relationships, our work demonstrates that blind applications of such heuristic transformations can result in significant distortions at multiple scales. Although popular two-dimensional embeddings can reflect the broader strokes of the data such as cell type inter-distances, or highlight correlations between features [42], we find that quantitative relationships between cells, nearest neighbors, and cell types are highly distorted. Researchers are therefore tasked with navigating multiple, possibly contradictory interpretations of the same data. Additionally though current methods preserve some qualitative properties of datasets, these properties can be recapitulated in an arbitrary manner, bringing into question the biological meaning of widely utilized two-dimensional representations.

We therefore believe in lessened reliance on two-dimensional artwork for the purposes of identifying biological patterns. At least if visualizations are used, they should be presented alongside the kinds of metrics we and others have proposed for quantitative assessment of 'global' and 'local' scale [6]. In particular, we urge researchers to exercise caution in assigning biological interpretations to images with no theoretical guarantees, or "canonical" properties. There is an opportunity to develop two-dimension embedding methods that, with theoretical guarantees, could provide meaningful visualizations of high-dimensional data. The lower bounds on distortion that we have derived leave room for "reasonable" embeddings, and it is an interesting open problem to achieve optimal low-distortion [43, 44]. Some promising directions include work to define more robust distortion metrics and unified embedding frameworks [45], and to preserve equidistance as possible [46] (Supplementary Note 5). For tasks such as cluster validation and trajectory inference where t-SNE and UMAP are commonly employed, quantitative/statistical metrics on marker gene specificity and strength of expression (usually employed regardless of the visual embedding), as well as expression similarity between cells, [47, 48] provide more reliable bases for analyses and targeted visuals which directly report or represent these metrics. Similarly, higher dimensional inference of differentiation trajectories [49, 50], and incorporation of probabilistic inference methods [51–53], offer meaningful and interpretable analysis approaches with possible, targeted visualization alternatives.

Beyond the goals of two-dimensional visualization, higher dimensional representations offer more flexible spaces for multiple tasks to be performed, partial to researchers' interests. To better adapt these spaces for biological investigation, we have presented a semi-supervised framework for direct incorporation of biological features into latent space structure, as opposed to unsupervised approaches unaware of the task goals. The semi-supervised MCML methodology expands the domain of latent space structure and prediction to discrete and continuous properties of cells, offers a targeted alternative to unsupervised reconstruction with bMCML, and maintains linear interpretability between the latent space and the input features [38]. Our results are based on limited parameter optimization of our multi-objective optimization schemes, however methods such as grid-search could be implemented to determine parameters e.g. the 'best' fractional weighting between label-aware and reconstruction costs (see Methods). MCML could also be extended to filter for labels which contribute to explaining variance in the data or to the accuracy of a specific task (e.g. spatial location prediction, likely dependent on multiple covariates [54]), or to parametric models of single-cell count data utilizing existing variational autoencoder models [38, 55].

Finally, we note that our work on distortion in low dimensional embeddings and our framework for semi-supervised multi-class multi-label dimensionality reduction demonstrates a step towards developing more precise answers to questions about the dimensionality of transcriptomes. Identification of groups of equidistant cells provides weak lower bounds on the dimension, which appears to be much higher than two. Our results demonstrating the advantages of semi-supervised reduction suggest such methods could be utilized to refine upper bounds on the dimension of transcriptomes, such as the dimensions necessary to separate cell type designations [56]. Moreover, these questions and results are relevant to multi-faceted datasets outside of single-cell genomics, such as in phylogenetics or population genetics, where UMAP/t-SNE are used to explore structure of genetic interactions and evolutionary relationships [57]. Our findings should also be of interest beyond the biological sciences, e.g. in chemistry [58], geology [59], astronomy [60], and the social sciences [61], where dimensionality reduction is used to find latent representations capturing key structural features of data.

Methods

Datasets and Pre-processing

All datasets used in this study are listed in Table 1, and were chosen to cover a range of sequencing platforms, experiment sizes, and experimental designs.

Dataset	Technology	Cells	Label Classes	Download Link
Ex and In Utero Mouse Embryo E10.5	10x Genomics v3	56,528	Cell Type, Growth Condition	https://ftp.ncbi.nlm.nih.gov/geo/series/GSE149nnn/GSE149372/suppl/
Ex Utero Mouse Embryo E8.5	10x Genomics v3	6,205	Cell Type, Growth Condition	https://ftp.ncbi.nlm.nih.gov/geo/series/GSE149nnn/GSE149372/suppl/
SMART-Seq Mouse VMH Neurons	SMART-Seq v4	3,850	Cell Type, Sex	https://data.mendeley.com/datasets/ypx3sw2f7c/3
10x Mouse VMH Neurons	10x Genomics v2	41,580	Cell Type, Sex	https://data.mendeley.com/datasets/ypx3sw2f7c/3
10x Developing Mouse Brain	10x Genomics v1	292,495	Cell Type	http://mousebrain.org/downloads.html
Developing *C. elegans* Embryo(Neural Lineage)	10x Genomics v2	1,075	Cell Type, Pseudotime	http://staff.washington.edu/hpliner/data/
Mouse Primary Motor Cortex (MOp)	MERFISH	6,963	Cell Type, Spatial Coordinates	https://caltech.app.box.com/folder/134209256308

Table 1: Dataset Metadata. Datasets used for the Picasso and MCML (including bMCML) analyses.

For the SMART-Seq and 10x mouse VMH datasets, cells were filtered according to the steps outlined in [23]. Unless already provided, the top 2000 highly-variable genes (HVGs) were found for all datasets using Scanpy's highly_variable_genes [48]. The top 300 genes were used for the *C. elegans* neural lineage cells as there were only 1000 cells after selecting for the ASE, ASJ, and AUA neurons [14]. Counts were log-normalized, if not already provided, with the log-count matrices representing the 'ambient' data for metric comparisons (see below). Unless otherwise indicated, 'ambient' space refers to the log-normalized count matrices filtered for HVGs. All count matrices were zero-centered and scaled before application of the Picasso, MCML methods, or PCA. All PCA analysis was performed using sklearn TruncatedSVD to 50 dimensions by default. 15 dimensions was used for the PCA of the integrated mouse embryo E10.5 dataset to facilitate direct comparison to the original study [22].

The t-SNE and UMAP algorithms were applied to the 50 (or 15 in the case of the integrated mouse embryo E10.5 dataset) dimensional PCA embeddings with default settings. As per the discussion in [6], though slight changes in parameters can drastically impact low-dimensional embeddings, the choice of parameters for tuning is often informed by empirical observations/prior knowledge leaving open the question of which metric(s) to use for determining 'optimal' parameters. This tuning is additionally contradictory to the common use or desire of such techniques to produce 'unsupervised' representations of the data [6]. [Code]

Determining Groups of Equidistant Cells

To find equidistant cells within cell types, we selected cells from within sizeable cell types to narrow the search space, as the algorithm we used, namely clique detection in undirected graphs, is NP-complete. The cell types we investigated were 'Esr1_6' in the 10x VMH dataset and 'Chondrocytes and Osetoblasts' in the integrated embryo E10.5 dataset. We calculated all pairwise distances between the cells in the ambient space, and using those defined a graph where two cells were adjacent if the cell-cell distance was within a small fraction of the standard deviation around the mean, and the 0.1 and 0.9 quantile marks. The filtering for distances within a particular range helps to limit the size of the search space as well as produce a range of mutually equidistant cells. We used the sklearn pairwise_distances for the pairwise calculations. From the graph of ‘connected’ cells we looked for cliques, namely subsets of cells in which all cells are connected (adjacent) to each other. This same strategy was employed to determine equidistant *centroids* of cell types for the 10x VMH data. Equidistant cell type centroids were identified by constructing a graph where two nodes, associated with centroids, were adjacent if their distance was close to the average pairwise distance. To find cliques we used the find_cliques function from the networkx package, which employs a variant [62] of the Bron & Kerbosch algorithm [63], to detect cliques in undirected graphs. [Code]

Metrics for Correlation and Distortion of Ambient Space Properties

Distortion Metrics for Equidistant Cells

We used two metrics to assess distortion of equidistant cells in two dimensions. The first is the variance of the pairwise distances between cells (or centroids) in each group, as compared to the low variance in the distances in the ambient space. We also calculated the ratio of the maximum to minimum distance between cells in each group (the ‘max/min ratio’), a quantity for which we derived a lower bound (see Theorem 1 in Supplementary Note 2) :

$$\frac{D}{d} \geq \sqrt{\frac{n-2}{2}}.$$

All variance and min/max comparisons were done in the ambient space, the PCA-reduced spaced, and the UMAP/t-SNE spaces, which were generated from the PCA-space. The ambient space for the integrated embryo E10.5 data is the ‘Variance-Stabilized and Scaled’ (Fig. 1a-d) data (as opposed to solely log-normalized counts in Fig. 1e), as this was used as input for the original UMAP embedding in [22]. [Code]

These distortion metrics were also measured between every cell and its 10 nearest neighbors to demonstrate distortion outside of groups of necessarily equidistant cells. The sklearn NearestNeighbors function was used to find these 10 neighboring cells as well as for Fig. 1, to extract each cells’ 30 nearest neighbors, and compare neighbor labels, in the UMAP versus ambient space (using L_1 distance). [Code]

For comparisons of nearest neighbor overlaps in PCA-reduced and PCA-coupled t-SNE/UMAP spaces we used Jaccard distance defined as $1 - \frac{|A \cap B|}{|A \cup B|}$ where A, B represent the sets of each cell’s 30 nearest neighbors in the ambient and latent spaces respectively. [Code]

Inter- and Intra-Label Distances

To assess the relative differences and similarities within and across biological properties of interest we defined inter- and intra-label distance metrics. Inter-label distances (Supplementary Fig. 8) are calculated as pairwise L_1 (defined below) distances between the centroids of each label within a class (e.g. between centroids of each ‘cell type’ label). These represent the relative distances, or closeness, *between* labels. For two vectors $\mathbf{x}, \mathbf{y}$ the L_1 distance is defined as the absolute value of their differences:

$$d(\mathbf{x}, \mathbf{y}) = \sum_i |x_i - y_i|.$$

For this we used sklearn pairwise_distances. For more 'internal' labels such as sex, inter-label distances were calculated as the mean pairwise distance between the *cells* of each label (i.e. 'male' and 'female') within each cell type. Intra-label distances are the means of all pairwise distances within each label (Supplementary Fig. 8). These represent the relative, internal variances among the cells *within* each label. For sex labels i.e. 'male' and 'female', these means were calculated within each cell type.

We used L_1 distance as it is suitable for measuring distance between points in high dimensions, particularly in comparison to other L-norms [64, 65], and is comparable to the probabilistic Jensen-Shannon divergence in single-cell (transcriptomic) distance calculations [64]. The correlation of these metrics, in the latent space, to their values in the ambient space were then calculated by Pearson correlation. The latent spaces represent PCA, t-SNE, UMAP, Picasso, or MCML embeddings.

We used the average of all L_1 pairwise distances between cells of each type to rank the cell type neighbors for each cell type (which other types they were closest to) (Supplementary Fig. 10,11). For k-means clustering we used the default KMeans function from sklearn to determine 10 clusters for the two-dimensional embeddings of the MNIST dataset [66]. [Code]

To test recapitulation of the relative distances of 'rare' and orthogonal cells to other cell types, using PCA and the MCML framework below, we added the expression of five new 'cells' in the MERFISH MOp data with expression (single gene counts) in three gene dimensions not expressed in other cells (Supplementary Fig. 16). We then ran Recon MCML (MCML with only reconstruction error) and PCA on this new matrix, and calculated the inter-distances for this group of cells to all other cell types (Supplementary Fig. 16). [Code]

General Autoencoder Architecture

The autoencoder network used in the Picasso and MCML algorithms is outlined below. The structure of the neural network remains the same between algorithms though each has a unique set of cost functions for network optimization.

The input is a centered/scaled count matrix $\mathbf{X} \in \mathbb{R}^{n \times g}$, n cells by g genes. For MCML embeddings C is the set containing label vectors for each class k, $C : \{\mathbf{c}_1, ..., \mathbf{c}_k\}$. Classes can be discrete or continuous, and multi-dimensional in the case of continuous classes (e.g. cell type, sex, location).

The input is passed through two fully-connected layers of 128 nodes and d nodes respectively with $d = 50$ by default. Batch normalization, the ReLU activation function, and dropout regularization are applied between the layers. The second layer represents the latent representation in $\mathbb{R}^{n \times d}$ denoted as $\mathbf{Z}$. The final linear, decoder layer produces $\hat{\mathbf{X}} \in \mathbb{R}^{n \times g}$. No activation function or bias terms are used between the latent and decoder layer as the decoder output solely represents a linear transform of the latent space.

Mini-batch training was employed for all algorithms, with a default batch size of 128, though larger batch sizes were used for Picasso embeddings. Adam optimization [67] was used for network training with a default learning rate of 10^{-3} and weight-decay term of 10^{-5}.

Picasso Cost Function for Shape Imitation

We defined two loss functions: $L_{ShapeAware}$ and $L_{Reconstruction}$, which balance the fit of the input points to the desired shape and reconstruction error in the decoder output as compared to the input. $\mathbf{S} \in \mathbb{R}^{p \times d}$ represents the coordinates comprising the desired shape, where $d = 2$ and $p \geq n$. The latent space $\mathbf{Z}$ is also limited to $d = 2$ dimensions. The pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{n \times p}$ represents Euclidean distances between the cell coordinates in $\mathbf{Z}$ and shape coordinates $\mathbf{S}$ such that

$$d_{ij} = \|z_i - s_j\|_2.$$

Using $\mathbf{D}$, we define a Boolean, $n \times p$ adjacency matrix $\mathbf{A}$, where $\sum A_i = 1$. This matrix uniquely specifies an adjacent coordinate point for every cell, in a bipartite graph mapping the n cells to the p coordinates. $\mathbf{A}$ is determined by the linear_sum_assignment scipy package, which assigns a shape coordinate to each cell by solving the minimization

$$min \sum_i \sum_j d_{ij} a_{ij}$$

where $a_{ij} = 1$ iff row i is assigned to column j. Thus,

$$L_{ShapeAware} = \sum A \odot D,$$

which we attempt to minimize i.e. map cells to their closest, unique shape coordinates. The reconstruction loss is the L_2 norm of the difference between the reconstructed and input data:

$$L_{Reconstruction} = \|\hat{\mathbf{X}} - \mathbf{X}\|_2.$$

The total loss then incorporates both loss functions, balancing their contributions with f, a user-defined fraction weighting the effect of each term on the resulting embedding:

$$L = f * L_{ShapeAware} + (1 - f) * L_{Reconstruction}. \tag{1}$$

Correlation metrics, as defined above, are measured for the output $\mathbf{Z}$, PCA to two dimensions (PCA 2D), and 2D t-SNE/UMAP (PCA t-SNE and PCA UMAP) which are run on the output of PCA to 50D by default. Picasso was tested on the SMART-Seq VMH neurons [Code], the ex-utero mouse embryo E8.5 data [Code], and the MERFISH MOp dataset [Code].

MCML Framework with Cost Function for Discrete and Continuous Properties

We use the acronym 'MCML' (multi-class multi-label) to denote the semi-supervised, label-aware methodology which directly incorporates the label-aware cost into the latent space structure (Fig. 4a). For MCML we use two loss functions: $L_{LabelAware}$ and $L_{Reconstruction}$, where $L_{Reconstruction}$ is as defined in (1). For $L_{LabelAware}$, we utilize the Neighborhood Component Analysis (NCA) algorithm from [39]. For all cells a probability matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is created where

$$p_{ij} = \frac{exp(-\|z_i - z_j\|^2)}{\sum_j exp(-\|z_i - z_j\|^2)} \;,\; \sum p_i = 1.$$

For discrete labeled data (e.g. cell type names) we define $L_{Discrete}$ for all pairs of cells i, j where

$$L_{Discrete} = \sum_k \frac{\sum_{ij} p_{ij} \mathbb{1}_{ij}}{\sum_{ij} \mathbb{1}_{ij}} \text{ where } \mathbb{1}_{ij}(c_k) := \begin{cases} 1 & \text{if } \; c_{k,i} = c_{k,j} \;, \\ 0 & \text{otherwise} \;. \end{cases}$$

Only the probabilities of cell pairs which are of the same label, for each class k, are summed and normalized to the total number of these cell pairs (which represents the maximum value of the numerator). For continuous classes of labels, such as spatial coordinates or pseudotime values, we use a separate loss function, $L_{Continuous}$. A probability weight matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ is generated for every pair of cells such that

$$w_{ij} = \frac{exp(-\|c_{k,i} - c_{k,j}\|^2)}{\sum_j exp(-\|c_{k,i} - c_{k,j}\|^2)} \;,\; \sum w_i = 1.$$

In place of the indicator function, the weights bias the masking of the original probability matrix $\mathbf{P}$ towards higher-weighted ('closer') pairs of cells. Probabilities are also normalized to the maximum of the numerator (treating the weights $\mathbf{W}$ as constants):

$$L_{Cont} = \sum_k \frac{\sum_{ij} w_{ij} p_{ij}}{\sum_i max(w_{ij})}.$$

The final loss function is

$$\begin{aligned} L_{LabelAware} &= L_{Discrete} + L_{Continuous} \\ L &= -f * L_{LabelAware} + (1-f) * L_{Reconstruction}. \end{aligned} \tag{2}$$

$L_{LabelAware}$ is negated for minimization, as opposed to maximization of the positive probabilities, and is additionally weighted by a constant factor of 10 in comparison to $L_{Reconstruction}$. For all datasets, excluding the integrated mouse embryo E10.5 dataset, the latent space $\mathbf{Z}$ is in $d = 50$ dimensions, and $d = 15$ for the Integrated data.

Measuring Distance between Cells of Differing Conditions

We applied MCML to the integrated mouse embryo E10.5 dataset, including both cell type and condition (ex- or in-utero) labels, for dimensionality reduction. We then measured the pairwise L_1 distances between the centroids of the ex- and in-utero cells within each cell type, as a measure of 'internal' distance. Within the cell types with the largest distances, we extracted differentially expressed (DE) genes between the conditions, following the metrics in the original study (genes with log2 fold-change greater than 1 and adjusted p-values greater than 0.05) [22]. Here we utilized the log-normalized data only, as it represents the counts prior to scaling and filtering for highly variable genes. We used the non-parametric Wilcoxon test to identify DE genes, with p-values adjusted with Benjamini-Hochberg correction. [Code]

Prediction Accuracy for Unlabeled Cells

To assess the ability of semi-labeled, MCML-reduced latent spaces for class-specific label prediction, we measured their accuracy in continuous and discrete label prediction for unlabeled cells. For the comparative cell type label analysis in Fig. 4c, 70% of each dataset was used as training and the remaining for testing/prediction for SCANVI, netAE, and MCML which are label-aware methods. netAE was unable to run on the 10x VMH and 10x developing brain datasets as it attempted to allocate over 2TB of RAM. [Code] All of the data was used to train LDVAE and Recon MCML (reconstruction error only, $f = 0$) as baseline prediction comparisons. Two-dimensional t-SNE, UMAP, and supervised UMAP embeddings were additionally compared for prediction accuracy, coupled to 50D PCA. For the continuous and mixed label annotation in Fig. 5, 80% of each dataset was used as training and the remaining for testing/prediction. The full count matrices were input into the MCML framework, but only the denoted percentages of cells were labeled.

We applied sklearn's KNNClassifier with 50 nearest neighbors, weighted by their distance, for discrete label prediction in each latent space, and used the accuracy_score function from sklearn to determine the fraction of correct labels predicted (Fig. 4c). [Code 10x VMH] [Code SMART-Seq VMH] [Code Developing Brain]

The KNNRegressor from sklearn was used to predict continuous values in the same manner. We used Jaccard distance/dissimilarity [6], defined as $1 - \dfrac{|A \cap B|}{|A \cup B|}$ where A, B represent the sets of 50 nearest neighbors in the ambient and latent spaces respectively, and Euclidean distance between each cell's predicted and true value, to assess the efficacy of the continuous predictions (Fig. 5a,b). Continuous labels were either one-dimensional pseudotime values, or two dimensional coordinates for each cell's spatial location. Pseudotime values were generated using the diffusion map-based 'dpt' methods from Scanpy [48], on the PCA-reduced *C. elegans* dataset. The 'ambient' space for determining continuous-label nearest neighbors was the $n \times 1$ or $n \times 2$ matrix containing the original values for all n cells. Confusion matrices, produced by sklearn plot_confusion_matrix, were also generated to compare true and false positive cell type label predictions for MCML-generated embeddings with and without dual incorporation of discrete (cell type) and continuous (spatial coordinate) label classes. [Code]

Runtime and sklearn Comparisons

For runtime comparisons between the various cell type prediction/annotation methods in Fig. 4c (see Supplementary Fig. 13), we timed all methods on a range of datasets, processed with 1 GPU over 5 cores each with 40G of memory. [Code]

To compare the capabilities of the NCA algorithm by MCML to the standard sklearn NCA implementation, MCML was run with $f = 1$ (no reconstruction error) and sklearn's NCA with default settings, to produce 50 dimensional latent space representations incorporating cell type labels only (see Supplementary Fig. 18). The NCA loss, represented by $L_{Discrete}$, was measured for the generated latent spaces. The GPU was not utilized for these comparisons to accommodate the sklearn implementation. [Code]

Biased MCML (bMCML) with Targeted Reconstruction Cost Function

Here we denote 'bMCML' as the label-aware, biased reconstruction methodology which adapts the original MCML cost functions in (2). This targeted reconstruction loss utilizes only one term in its loss. Here L is defined by the correlation of the inter- or intra-distances (as described above) of a particular class to the ambient data $\mathbf{X}$, $\mathbf{b}$ represents the vector of the specified inter-/intra-distances in the ambient space and $\hat{\mathbf{b}}$ represents those same distances calculated for the reconstruction $\hat{\mathbf{X}}$. L is then defined in (3) as the negation of the Pearson correlation of these two vectors. Negation, again, facilitates minimization.

$$L = -\frac{\sum_i(\hat{b}_i - \bar{\hat{b}})(b_i - \bar{b})}{\sqrt{\sum_i(\hat{b}_i - \bar{\hat{b}})^2(b_i - \bar{b})^2}}. \tag{3}$$

In Supplementary Fig. 15, we demonstrate the implementation of either intra-sex (Supplementary Fig. 15a,b) or intra-type (cell type) (Supplementary Fig. 15c,d) distance correlation in the loss and the effect of these targeted losses on the resulting correlation metrics. This was tested on the SMART-Seq [Code] and 10x mouse VMH neurons [Code].

Data Availability

Download links for the original data used to generate the figures and results in the paper are listed in Table 1. Processed and normalized versions of the count matrices are available on CaltechData, with links provided in Supplementary Table 1.

Code Availability

All analysis code used to generate the figures and results in the paper is available at `https://github.com/pachterlab/CBP_2021` with Picasso and MCML analyses provided in notebooks which can be run on Google Colab. Picasso is also available at `https://github.com/pachterlab/picasso`. The MCML method as well as tools for quantitative analysis are available via a Python pip installable package from `https://github.com/pachterlab/MCML`.

Acknowledgements

Some of the computations presented here were conducted using machines in the Resnick High Performance Center, a facility supported by the Resnick Sustainability Institute at the California

Institute of Technology. We thank Gennady Gorin and Benjamin Rivière for helpful discussions regarding the MCML and Picasso analyses, Sina Booeshaghi for helpful discussions regarding NCA and dimensionality reduction, Ingileif Hallgrímsdottir for valuable feedback on the manuscript, and Páll Melsted for useful insights regarding Theorem 1. The work was supported in part by NIH grant U19MH114830 and Joeyta Banerjee was supported in part by the Caltech Summer Undergraduate Research Fellowship (SURF).

References

1. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
2. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: `1802.03426 [stat.ML]` (Feb. 2018).
3. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. en. *Nat. Commun.* **10,** 5416 (Nov. 2019).
4. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. en. *Nat. Methods* **18,** 723–732 (July 2021).
5. Heiser, C. N. & Lau, K. S. A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. en. *Cell Rep.* **31,** 107576 (May 2020).
6. Cooley, S. M., Hamilton, T., Deeds, E. J. & Ray, J. C. J. *A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data* en. July 2019.
7. Linderman, G. C. & Steinerberger, S. Clustering with t-SNE, Provably. *SIAM Journal on Mathematics of Data Science* **1,** 313–332 (Jan. 2019).
8. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. en. *Cell* (May 2021).
9. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. en. *Cell* **177,** 1888–1902.e21 (June 2019).
10. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. en. *Nat. Biotechnol.* **36,** 411–420 (June 2018).
11. La Manno, G. *et al.* RNA velocity of single cells. en. *Nature* **560,** 494–498 (Aug. 2018).
12. Ding, J. & Regev, A. Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. en. *Nat. Commun.* **12,** 2554 (May 2021).
13. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. en. *Nature* **566,** 496–502 (Feb. 2019).
14. Packer, J. S. *et al.* A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. en. *Science* **365** (Sept. 2019).
15. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. en. *Cell* **174,** 999–1014.e22 (Aug. 2018).
16. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. en. *Nat. Methods* **16,** 479–487 (June 2019).
17. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. en. *Nature* **541,** 331–338 (Jan. 2017).
18. Johnson, W. B. & Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space 26. *Contemp. Math.* **26** (1984).
19. Dasgupta, S. & Gupta, A. An elementary proof of a theorem of Johnson and Lindenstrauss. en. *Random Struct. Algorithms* **22,** 60–65 (Jan. 2003).

20. Balko, M., Pór, A., Scheucher, M., Swanepoel, K. & Valtr, P. Almost-Equidistant Sets. *Graphs Combin.* **36,** 729–754 (May 2020).

21. Littlewood, J. E. *Littlewood's Miscellany* en (Cambridge University Press, Oct. 1986).

22. Aguilera-Castrejon, A. *et al.* Ex utero mouse embryogenesis from pre-gastrulation to late organogenesis. en. *Nature* **593,** 119–124 (May 2021).

23. Kim, D.-W. *et al.* Multimodal Analysis of Cell Types in a Hypothalamic Node Controlling Social Behavior. en. *Cell* **179,** 713–728.e17 (Oct. 2019).

24. Zhang, M. *et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics* en. June 2020.

25. Mayer, J., Khairy, K. & Howard, J. Drawing an elephant with four complex parameters. *Am. J. Phys.* **78,** 648–649 (June 2010).

26. Dyson, F. A meeting with Enrico Fermi. en. *Nature* **427,** 297 (Jan. 2004).

27. Byerly, A., Kalganova, T. & Dear, I. No routing needed between capsules. *Neurocomputing* **463,** 545–553 (Nov. 2021).

28. Narayan, A., Berger, B. & Cho, H. Assessing single-cell transcriptomic variability through density-preserving data visualization. en. *Nat. Biotechnol.* **39,** 765–774 (June 2021).

29. He, K., Zhang, X., Ren, S. & Sun, J. *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification* in *Proceedings of the IEEE international conference on computer vision* (openaccess.thecvf.com, 2015), 1026–1034.

30. Wójcik, P. I. & Kurdziel, M. *Random projection initialization for deep neural networks* in *ESANN* (researchgate.net, 2017).

31. Glorot, X. & Bengio, Y. *Understanding the difficulty of training deep feedforward neural networks* in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (eds Teh, Y. W. & Titterington, M.) **9** (PMLR, Chia Laguna Resort, Sardinia, Italy, 2010), 249–256.

32. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks* in *Advances in Neural Information Processing Systems* (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) **25** (Curran Associates, Inc., 2012).

33. Achlioptas, D. *Database-friendly random projections* in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (Association for Computing Machinery, Santa Barbara, California, USA, May 2001), 274–281.

34. Saxe, A. M. *et al. On random weights and unsupervised feature learning* in *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Omnipress, Bellevue, Washington, USA, June 2011), 1089–1096.

35. Har-Peled, S., Indyk, P. & Motwani, R. Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality. *Theory of Computing* **8,** 321–350 (2012).

36. Dumitrascu, B., Villar, S., Mixon, D. G. & Engelhardt, B. E. Optimal marker gene selection for cell type discrimination in single cell analyses. en. *Nat. Commun.* **12,** 1186 (Feb. 2021).

37. Dong, Z. & Alterovitz, G. netAE: semi-supervised dimensionality reduction of single-cell RNA sequencing to facilitate cell labeling. en. *Bioinformatics* **37,** 43–49 (Apr. 2021).

38. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. en. *Bioinformatics* **36,** 3418–3421 (June 2020).

39. Goldberger, J., Roweis, S., Hinton, G. & Salakhutdinov, R. *Neighbourhood components analysis* in *Proceedings of the 17th International Conference on Neural Information Processing Systems* (MIT Press, Vancouver, British Columbia, Canada, Dec. 2004), 513–520.

40. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. en. *Mol. Syst. Biol.* **17,** e9620 (Jan. 2021).

41. La Manno, G., Siletti, K., Furlan, A., Gyllborg, D., Vinsland, E., *et al.* Molecular architecture of the developing mouse brain. *BioRxiv* (2020).

42. Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S. & Trapnell, C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. en. *Nat. Commun.* **11,** 1–6 (Mar. 2020).

43. Badoiu, M. *et al. Approximation algorithms for low-distortion embeddings into low-dimensional spaces* in *SODA* **5** (Citeseer, 2005), 119–128.

44. Matoušek, J. On the distortion required for embedding finite metric spaces into normed spaces. *Israel J. Math.* **93,** 333–344 (Dec. 1996).

45. Agrawal, A., Ali, A. & Boyd, S. Minimum-Distortion Embedding. arXiv: 2103.02559 [cs.LG] (Mar. 2021).

46. Graham, R. L., Lubachevsky, B. D., Nurmela, K. J. & Östergård, P. R. J. Dense packings of congruent circles in a circle. *Discrete Math.* **181,** 139–154 (Feb. 1998).

47. Anders, S. & Huber, W. Differential expression analysis for sequence count data. en. *Genome Biol.* **11,** R106 (Oct. 2010).

48. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. en. *Genome Biol.* **19,** 15 (Feb. 2018).

49. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. en. *Genome Biol.* **20,** 59 (Mar. 2019).

50. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. en. *Nat. Biotechnol.* **37,** 547–554 (May 2019).

51. Du, J.-H., Gao, M. & Wang, J. *Model-based Trajectory Inference for Single-Cell RNA Sequencing Using Deep Learning with a Mixture Prior* en. Dec. 2020.

52. Lin, C. & Bar-Joseph, Z. Continuous-state HMMs for modeling time-series single-cell RNA-Seq data. en. *Bioinformatics* **35,** 4707–4715 (Apr. 2019).

53. Gorin, G., Vastola, J. J., Fang, M. & Pachter, L. Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments (2021).

54. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. en. *Genome Biol.* **21,** 31 (Feb. 2020).

55. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. en. *Nat. Methods* **15,** 1053–1058 (Dec. 2018).

56. Melton, S. & Ramanathan, S. Discovering a sparse set of pairwise discriminating features in high-dimensional data. en. *Bioinformatics* **37,** 202–212 (Apr. 2021).

57. Diaz-Papkovich, A., Anderson-Trocmé, L. & Gravel, S. A review of UMAP in population genetics. en. *J. Hum. Genet.* **66,** 85–91 (Oct. 2020).

58. Andronov, M., Fedorov, M. & Sosnin, S. Exploring Chemical Reaction Space With Reaction Difference Fingerprints and Parametric t-SNE. en. *ChemRxiv* (Mar. 2021).

59. Balamurali, M. & Melkumyan, A. *t-SNE Based Visualisation and Clustering of Geological Domain* in *Neural Information Processing* (Springer International Publishing, 2016), 565–572.

60. Traven, G. *et al.* The Galah Survey: Classification and Diagnostics with t-SNE Reduction of Spectral Information. en. *ApJS* **228,** 24 (Feb. 2017).

61. Waggoner, P. D. Pandemic Policymaking: Learning the Lower Dimensional Manifold of Congressional Responsiveness. arXiv: 2011.04763 [cs.CY] (Nov. 2020).

62. Tomita, E., Tanaka, A. & Takahashi, H. The worst-case time complexity for generating all maximal cliques and computational experiments. en. *Theor. Comput. Sci.* **363,** 28–42 (Oct. 2006).

63. Bron, C. & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* **16,** 575–577 (Sept. 1973).

64. Ntranos, V., Kamath, G. M., Zhang, J. M., Pachter, L. & Tse, D. N. Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. en. *Genome Biol.* **17,** 112 (May 2016).

65. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. *On the Surprising Behavior of Distance Metrics in High Dimensional Space* 2001.

66. Deng, L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Process. Mag.* **29,** 141–142 (Nov. 2012).

67. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv: 1412.6980 [cs.LG] (Dec. 2014).