The Specious Art of Single-Cell Genomics Supplementary Material

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

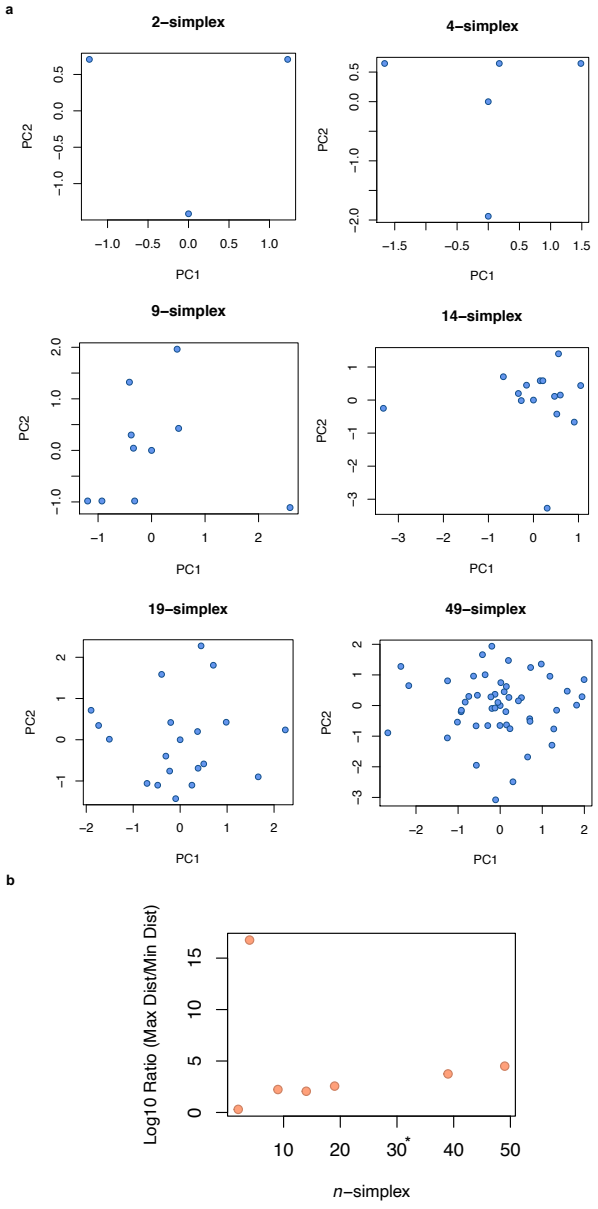
[*]Address correspondence to Lior Pachter (lpachter@caltech.edu)

September 27, 2021

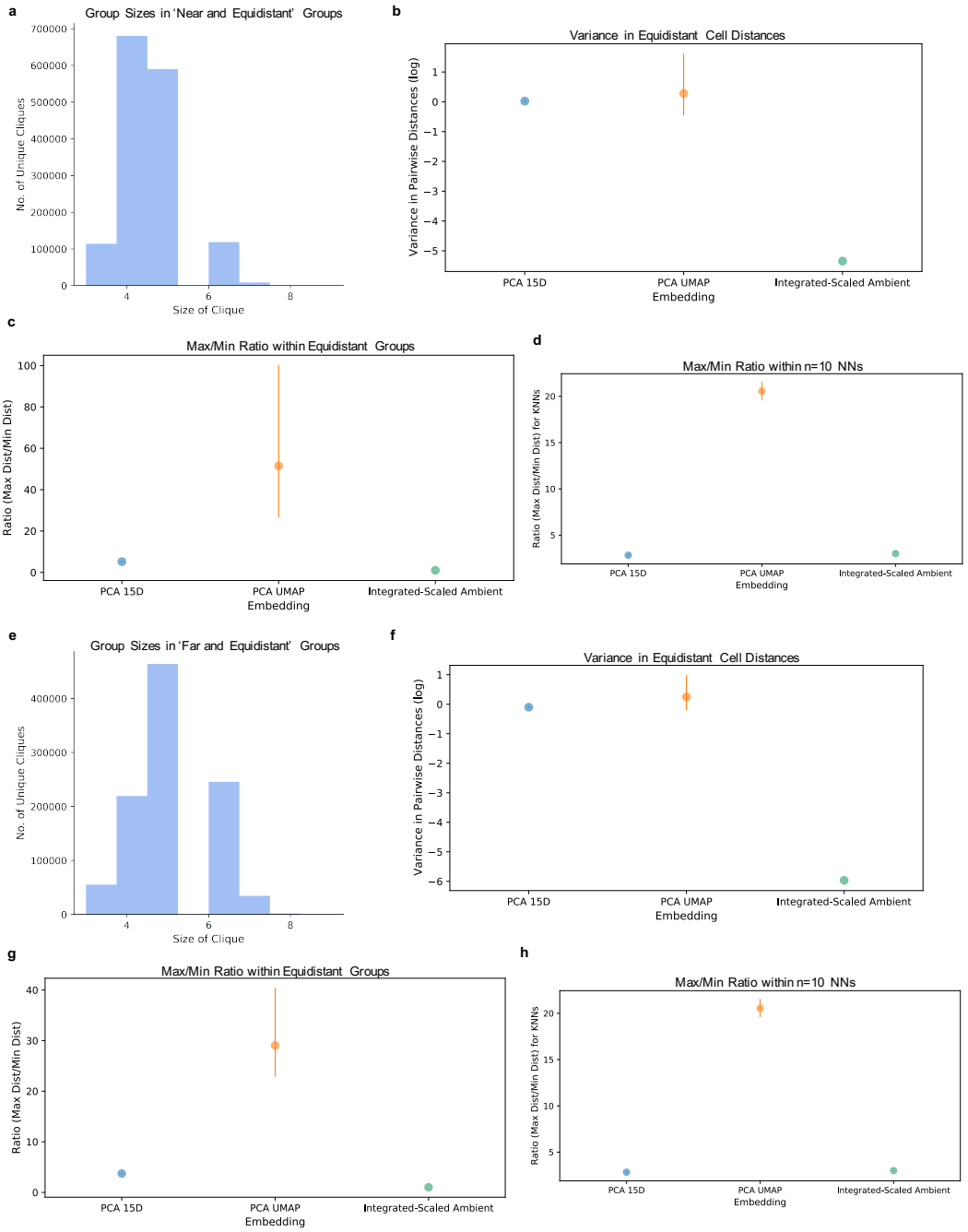
Contents

Supplementary Figures

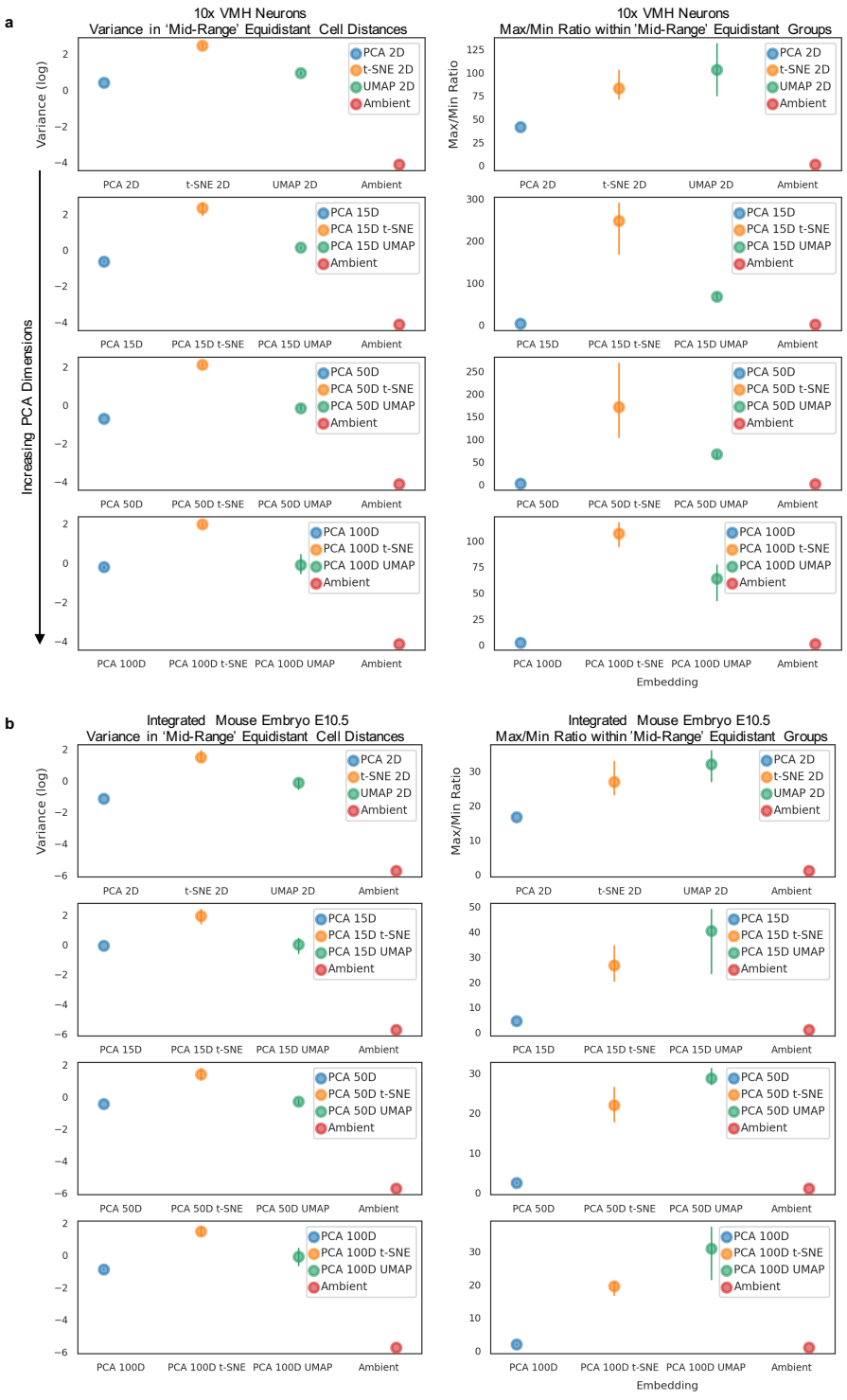
Low-Dimension Distortion



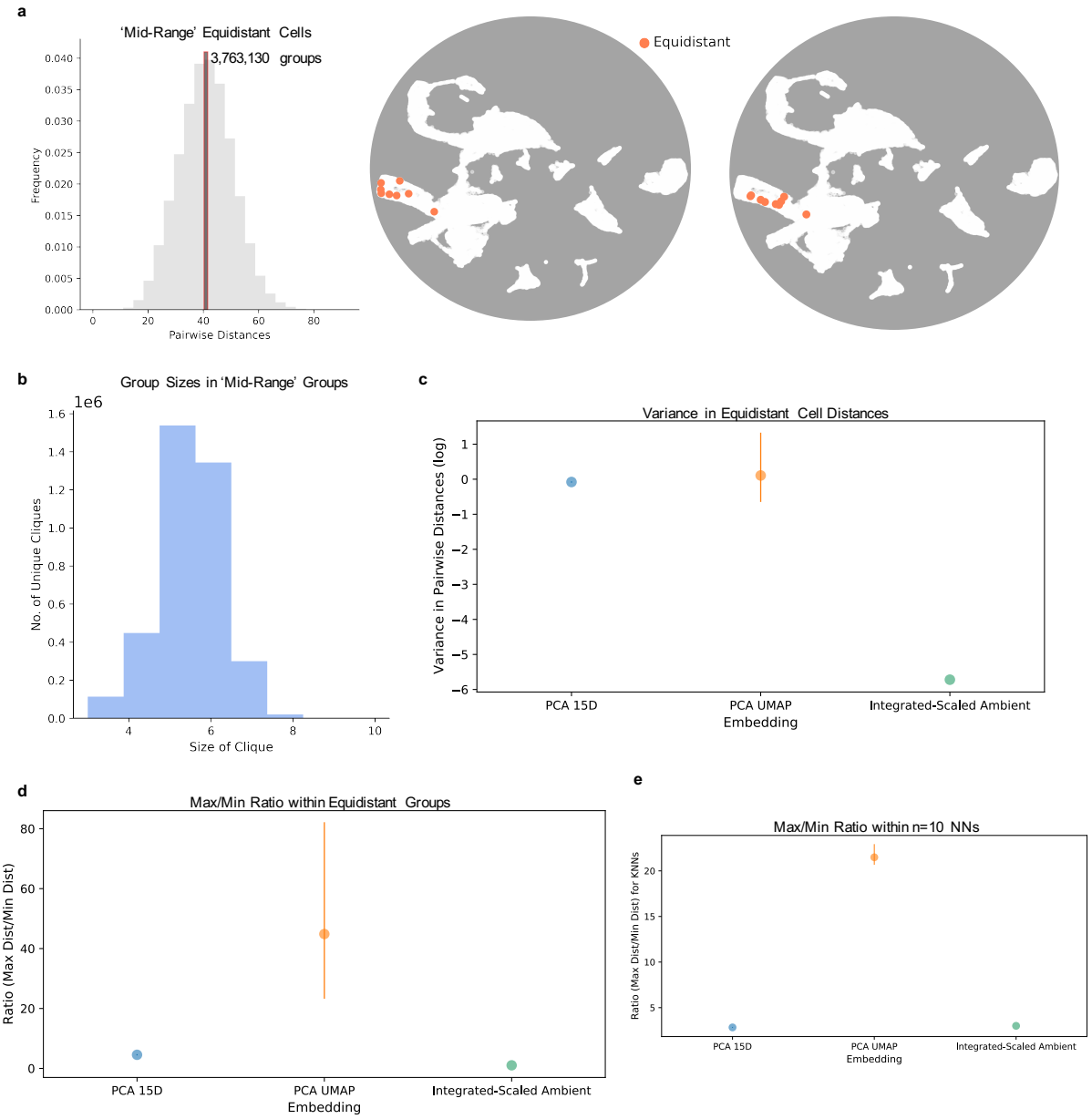
Supplementary Figure 1: Principal Components of Equidistant Points. a) First and second principal components shown for varying numbers of equidistant points, i.e. the $\mathbf{I}^n$ identity matrix in $\mathbb{R}^n$, for $n = 3, 5, 10, 15, 20$ and 50. **b)** Max/min ratios for the projections (see Methods ; Supplementary Note 2) of simplices in two-dimensions. * denotes where the minimum distance in the ratio is 0 (points are collapsed onto each other), and the max/min ratio is infinite. [Code]



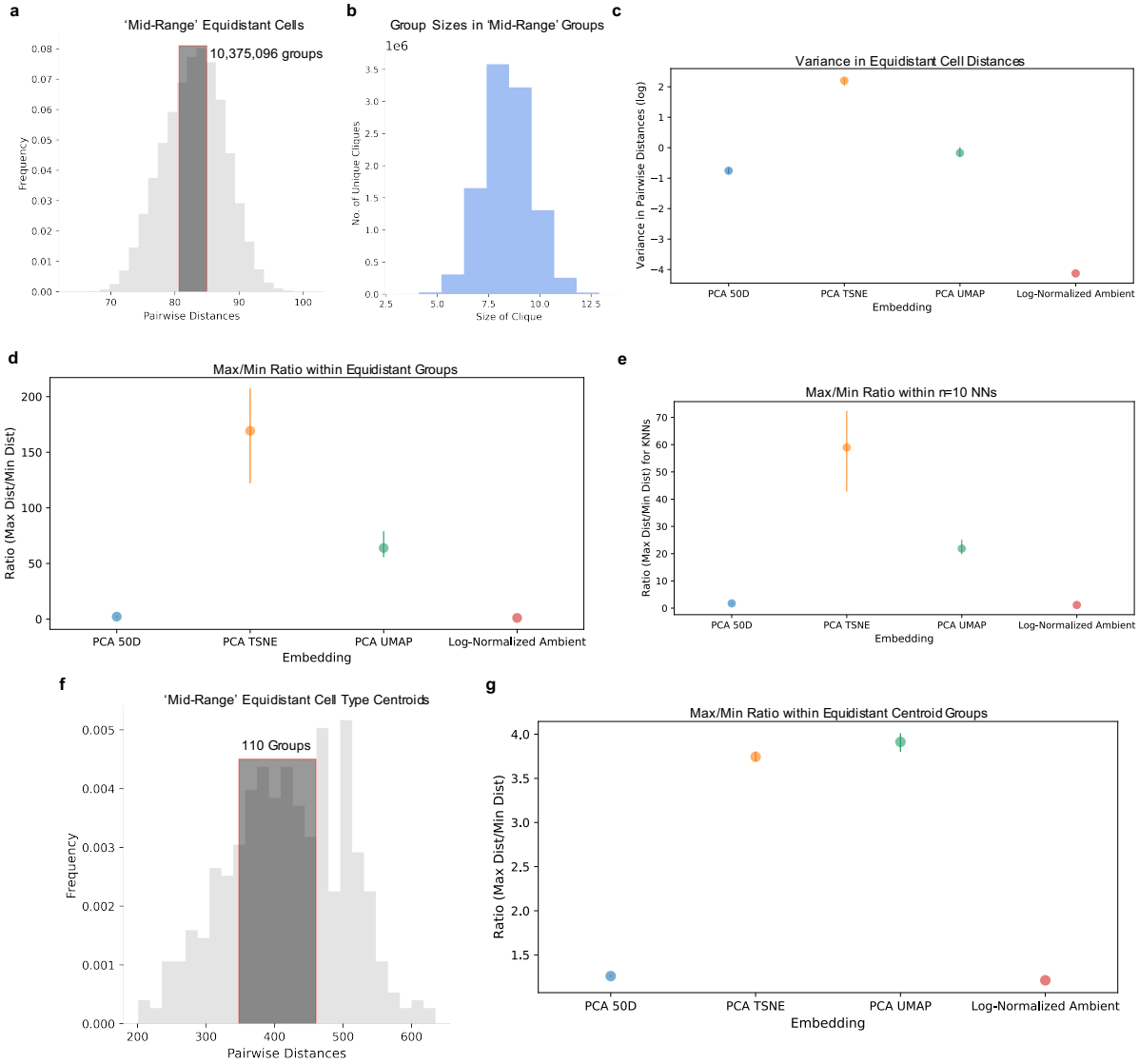
Supplementary Figure 2: Embeddings of Near and Far Equidistant Points in Integrated Ex- and In-Utero E10.5. a) Histogram of group sizes (number of equidistant cells) in the selection of 'near and equidistant' groups. **b)** Variance of pairwise distances across groups in each latent space. **c)** Ratio of the maximum to minimum pairwise distance (max/min ratio) across groups. **d)** Ratio of the maximum to minimum pairwise distance (max/min ratio) for each cell's neighborhood of 10 nearest neighbors (NNs). **e)** Histogram of group sizes (number of equidistant cells) in the selection of 'far and equidistant' groups. **f)** Variance of pairwise distances across groups in each latent space. **g)** Ratio of the maximum to minimum pairwise distance (max/min ratio) across groups. **h)** Ratio of the maximum to minimum pairwise distance (max/min ratio) for each cell's neighborhood of 10 nearest neighbors (NNs). For all plots bars denote the 95% C.I. [Code]



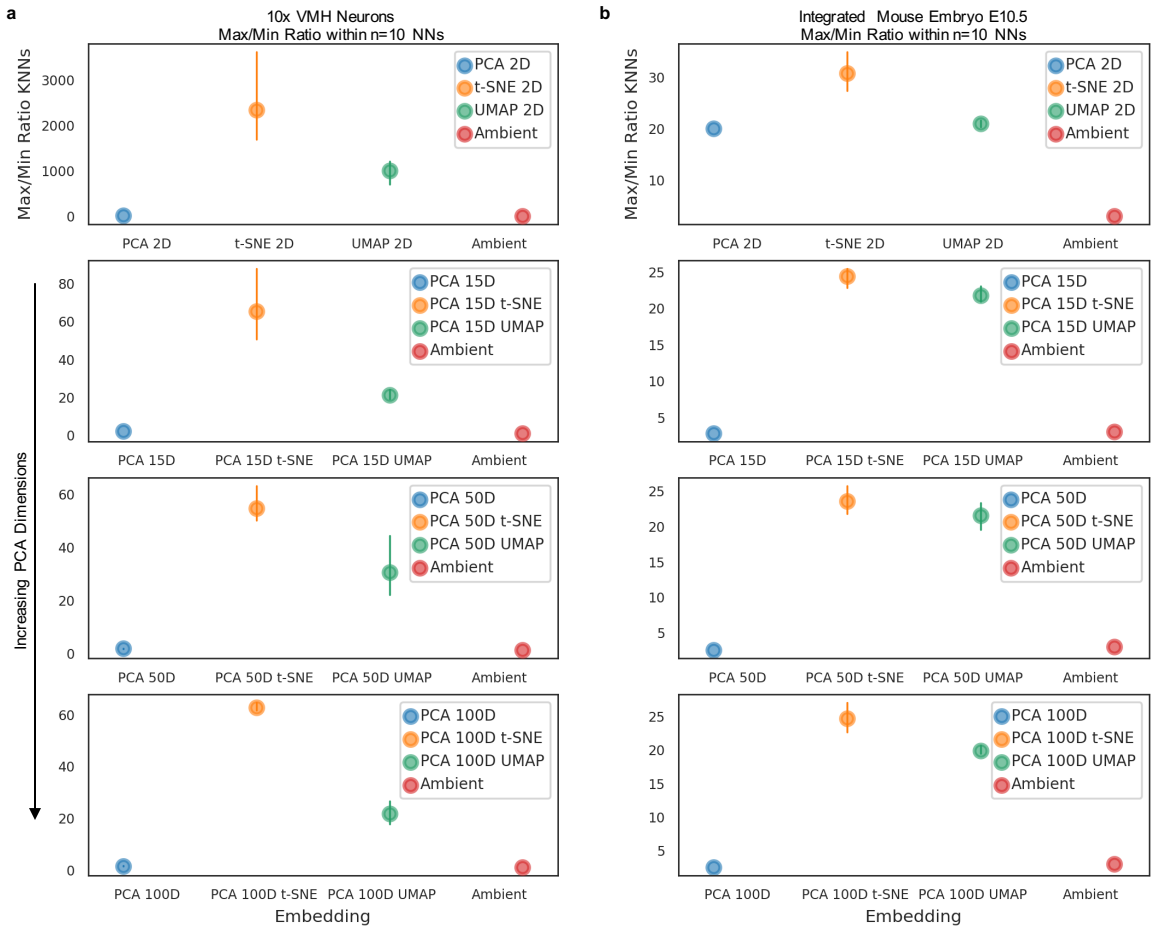
Supplementary Figure 3: Distortion of Equidistant Points Across Latent Spaces. a) Variance in pairwise distances and max/min ratios measured for the 'mid-range' equidistant groups in the 10x mouse VMH neurons dataset with and without PCA-coupled t-SNE/UMAP. '2D' latent spaces directly embed ambient space into two-dimensions, with increasing dimensions of PCA-reduced ambient data coupled with t-SNE or UMAP shown down the columns. **b)** Variance in pairwise distances and max/min ratios measured for the 'mid-range' equidistant groups in the integrated E10.5 mouse embryo dataset with and without PCA-coupled t-SNE/UMAP. For all plots bars denote the 95% C.I. [Code]



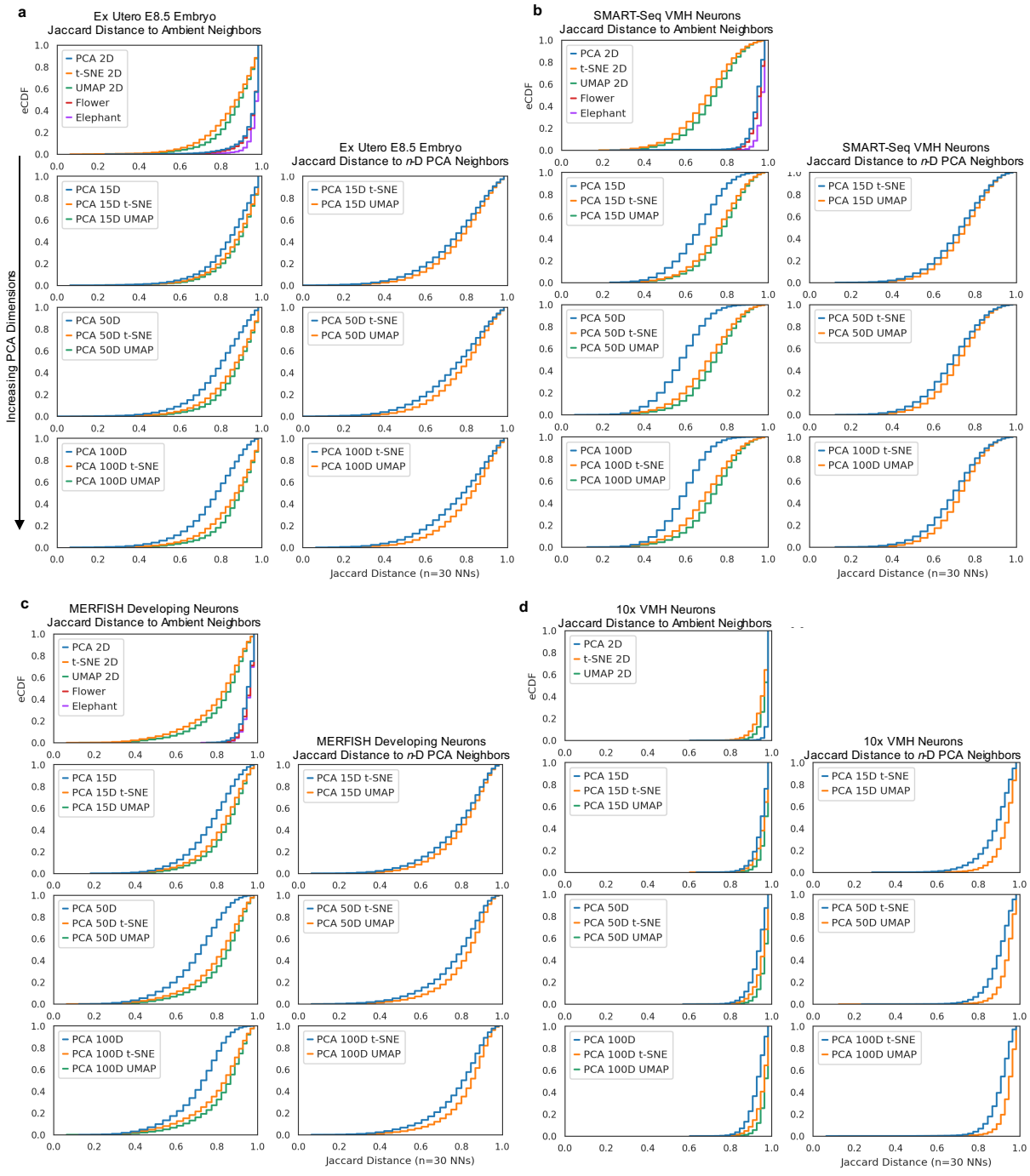
Supplementary Figure 4: Embeddings of Equidistant Points in Integrated Ex- and In-Utero E10.5. **a)** Selection of 'mid-range' sets of equidistant points, with distances close to the average pairwise distance, and their respective positions in the generated UMAP. **b)** Histogram of group sizes (number of equidistant cells) in the selection of mid-range' groups. **c)** Variance of pairwise distances across groups in each latent space. **d)** Ratio of the maximum to minimum pairwise distance (max/min ratio) across groups. **e)** Ratio of the maximum to minimum pairwise distance (max/min ratio) for each cell's neighborhood of 10 nearest neighbors (NNs). For all plots bars denote the 95% C.I. [Code]



Supplementary Figure 5: Properties of Equidistant Points in 10x VMH Neurons. a) Selection of 'mid-range' groups, with distances close to the average pairwise distance, in the 10x Mouse VMH Neurons dataset. **b)** Histogram of group sizes (number of cells in a group) in the selection of mid-range' groups. **c)** Variance of pairwise distances across groups in each latent space. **d)** Ratio of the maximum to minimum pairwise distance (max/min ratio) across groups. **e)** Ratio of the maximum to minimum pairwise distance (max/min ratio) for each cell's neighborhood of 10 nearest neighbors (NNs). **f)** Selection of 'mid-range' groups of cell type centroids, with distances close to the average pairwise distance. **g)** Ratio of the maximum to minimum pairwise distance (max/min ratio) across the groups of centroids. For all plots bars denote the 95% C.I. [Code]

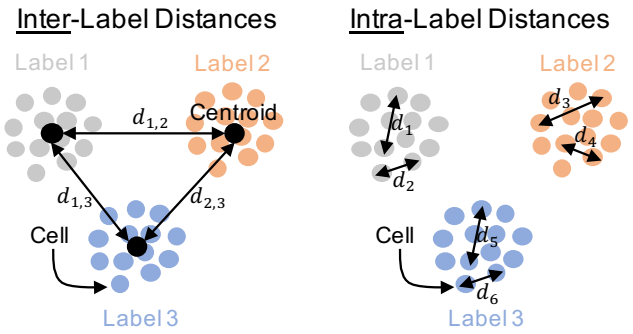


Supplementary Figure 6: Distortion of Nearest Neighbor Distances Across Latent Spaces. a) The max/min ratios measured within each cell's 10 nearest neighbors in the 10x mouse VMH neurons dataset with and without PCA-coupled t-SNE/UMAP. '2D' latent spaces directly embed ambient space into two-dimensions, with increasing dimensions of PCA-reduced ambient data coupled with t-SNE or UMAP shown down the columns. **b)** The max/min ratios measured within each cell's 10 nearest neighbors in the integrated E10.5 mouse embryo dataset with and without PCA-coupled t-SNE/UMAP. For all plots bars denote the 95% C.I. [Code]

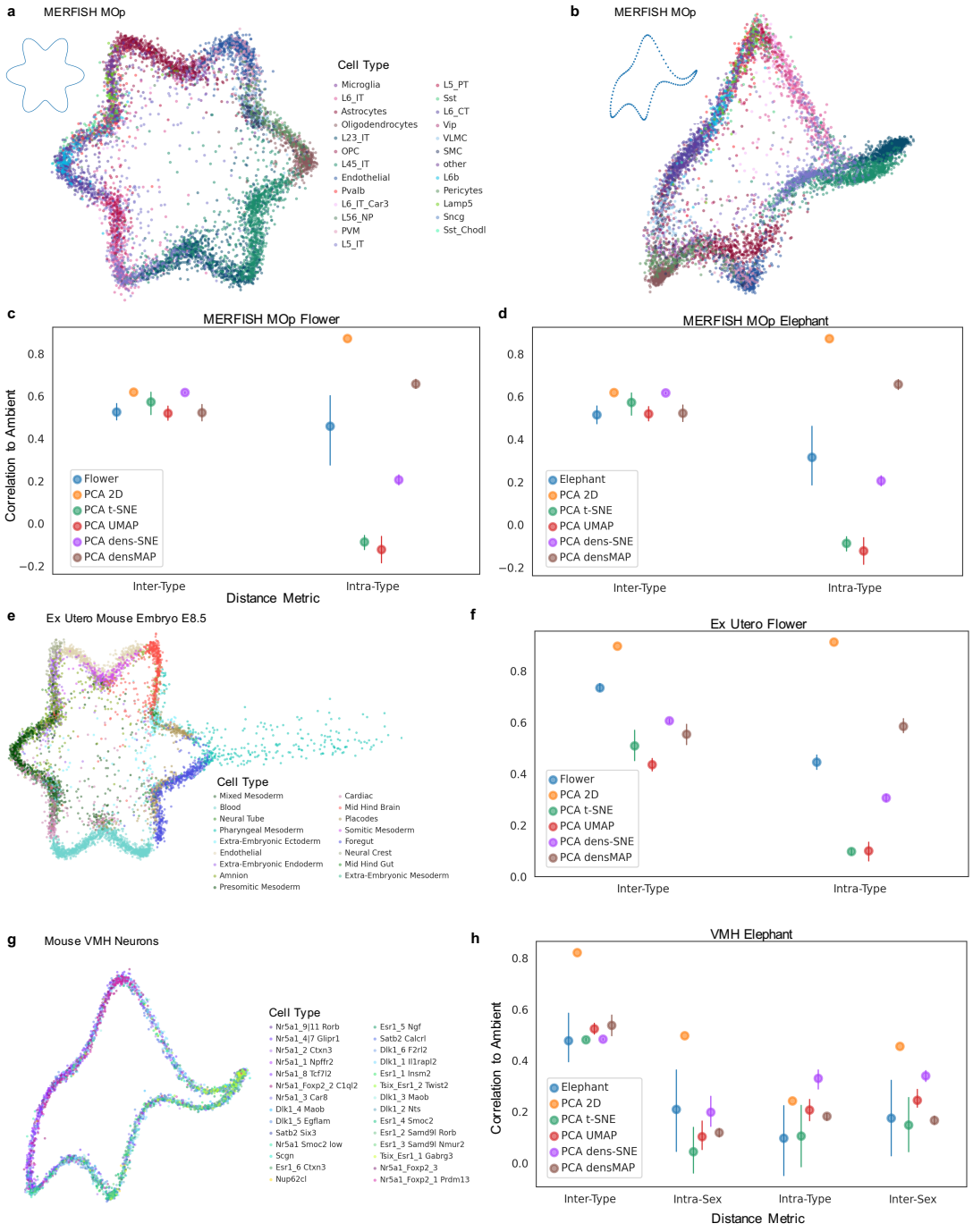


Supplementary Figure 7: Low Dimensional Dissimilarity to Higher Dimensional Neighbors. a) eCDF of Jaccard distances (dissimilarity) of 30 nearest neighbors in each designated latent space, including Picasso embeddings, as compared to the ambient/higher dimensional input space for the ex-utero E8.5 embryo data. First columns denote comparisons to ambient neighbors, and second columns denote comparisons to neighbors in the respective n-dimensional PCA space (coupled to t-SNE or UMAP). **b)** eCDF of Jaccard distances (dissimilarity) of neighbors in each designated latent space as compared to the ambient/higher dimensional input space for the SMART-Seq mouse VMH neurons dataset. **c)** eCDF of Jaccard distances (dissimilarity) of neighbors in each designated latent space as compared to the ambient/higher dimensional input space for the MERFISH developing neurons dataset. **d)** eCDF of Jaccard distances (dissimilarity) of neighbors in each designated latent space as compared to the ambient/higher dimensional input space for the 10x mouse VMH neurons dataset. [Code] [Plots]

Picasso Results

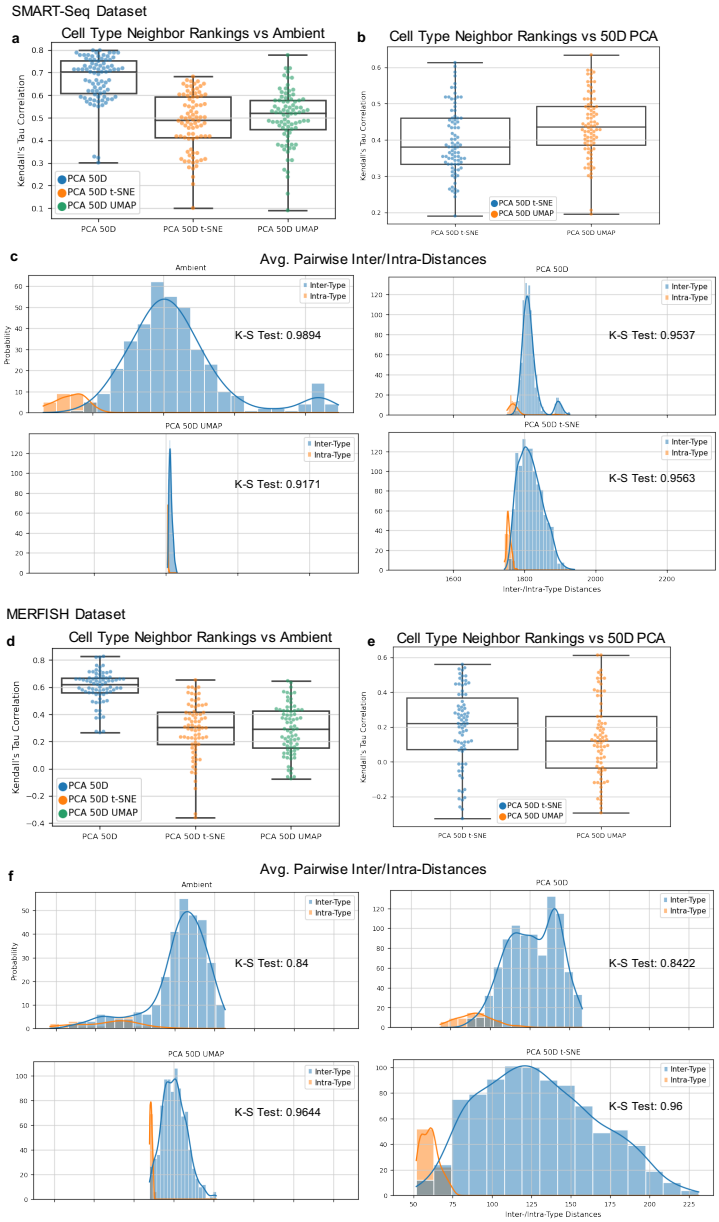


Supplementary Figure 8: Inter- and Intra-Distance Calculations. Inter-label distances, for a given class of labels, as calculated between the cells/centroids of each label. Intra-label distances as calculated between cells within each label.

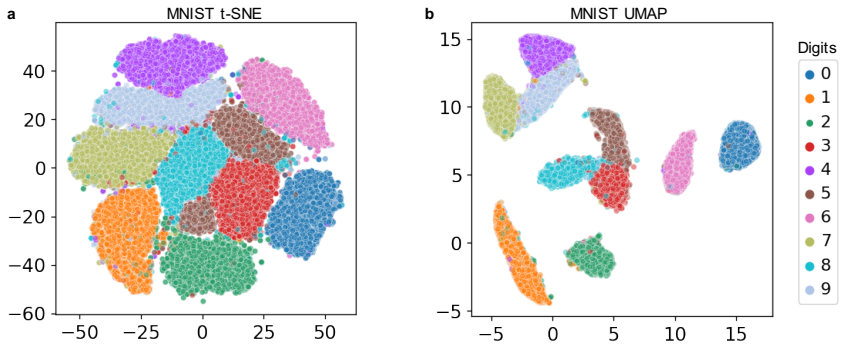


Supplementary Figure 9: Detailed Analysis of Picasso Embeddings. a) Picasso embedding of the MERFISH mouse primary motor cortex (MOp) data fit to a flower-like boundary. **b)** Picasso embedding of the MERFISH MOp data fit to a 'von Neumann' elephant. **c)** Comparison of correlation metrics between the flower Picasso embedding and the other baseline 2D embeddings, including densVis embeddings [29]. **d)** Comparison of correlation metrics between the elephant Picasso embedding and the other baseline 2D embeddings, including densVis embeddings [29]. **e)** Picasso embedding of the ex-utero mouse Embryo E8.5 data fit to a flower-like boundary. **f)** Comparison of correlation metrics between the flower Picasso embedding and the other baseline 2D embeddings, including densVis embeddings [29]. **g)** Picasso embedding of the SMART-Seq mouse VMH neurons dataset fit to a 'von Neumann' elephant. **h)** Comparison of correlation metrics between the elephant Picasso embedding and the other baseline 2D embeddings, including densVis embeddings [29]. For all plots bars denote the 95% C.I. [Code a-d] [Code e,f] [Code g,h]

Impact of Non-Linear Reduction



Supplementary Figure 10: Effect of Reduction on Separation and Neighbor Rankings. **a)** Kendall's Tau correlation (from -1 to 1) of cell type neighbor rankings for each type, to ambient rankings for SMART-Seq VMH data (see Methods). **b)** Kendall's Tau correlation of cell type neighbor rankings for each type, to 50D PCA rankings (then coupled to t-SNE or UMAP). **c)** Kolmogorov–Smirnov test statistic for measuring distance/separation between pairwise inter- or intra-type distances in the SMART-Seq VMH data. Distributions scaled to the same mean for comparison. **d)** Kendall's Tau correlation of cell type neighbor rankings for each type, to ambient rankings for MERFISH data. **e)** Kendall's Tau correlation of cell type neighbor rankings for each type, to 50D PCA rankings (then coupled to t-SNE or UMAP). **f)** Kolmogorov–Smirnov test statistic for measuring distance/separation between pairwise inter- or intra-type distances in the MERFISH data. Distributions scaled to the same mean for comparison. For box plots, whiskers denote 1.5 times the IQR. [Code]



e Fraction Correct Digits in t-SNE Clusters

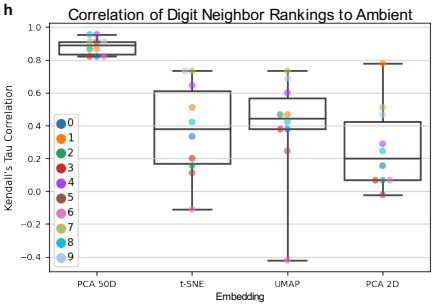
Digit	Fraction
0	0.991221
1	0.910264
2	0.945787
3	0.822215
4	0.611092
5	0.678104
6	0.983947
7	0.833839
8	0.829944
9	0.569507

f Fraction Correct Digits in UMAP Clusters

Digit	Fraction
0	0.992234
1	0.549540
2	0.955690
3	0.967705
4	0.981684
5	0.919019
6	0.989861
7	0.973025
8	0.873526
9	0.696756

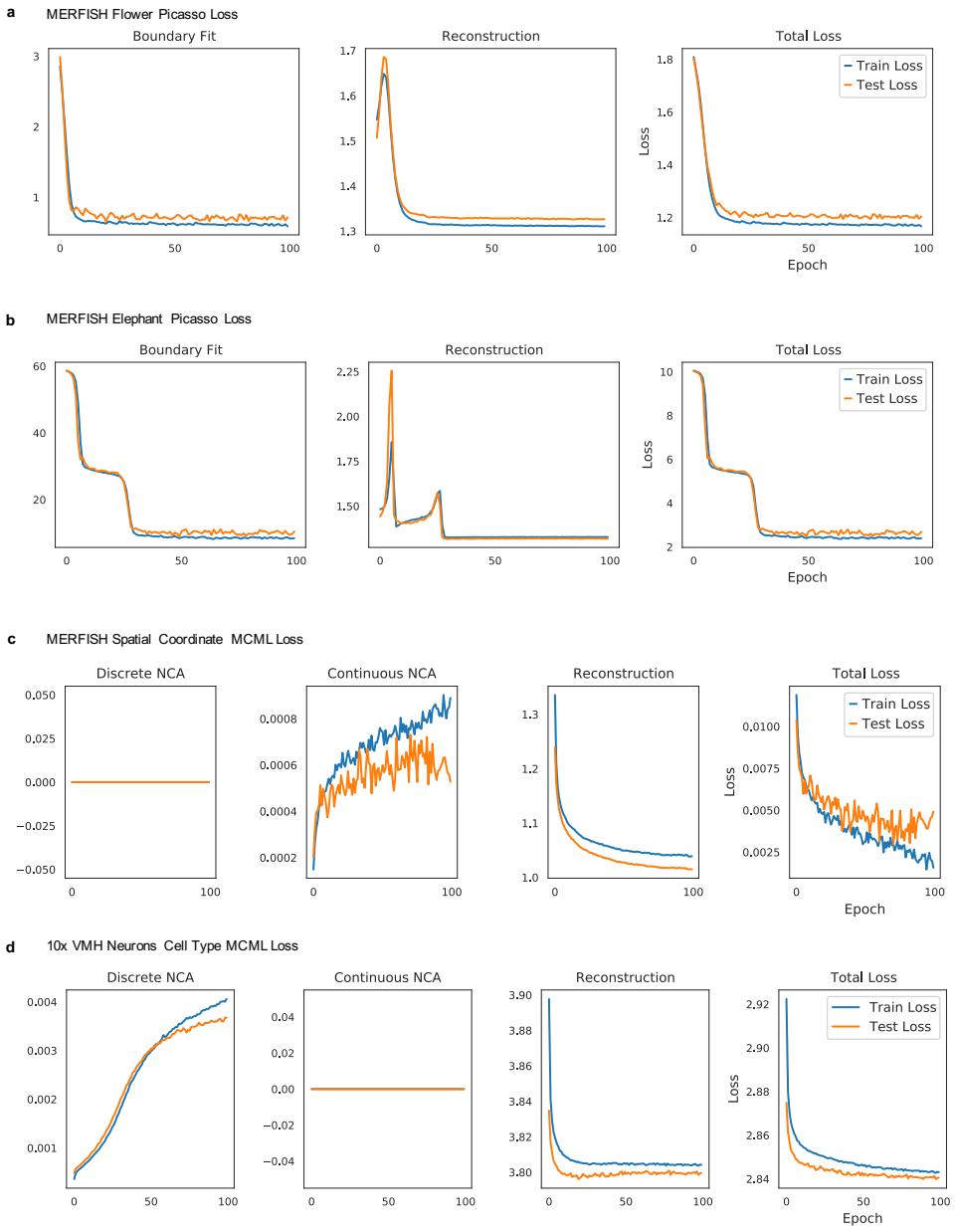
g Correlation to Ambient Intra-Distances

Latent	Pearsonr
PCA 50D	0.994942
t-SNE	0.103368
UMAP	-0.750536
PCA 2D	0.882780

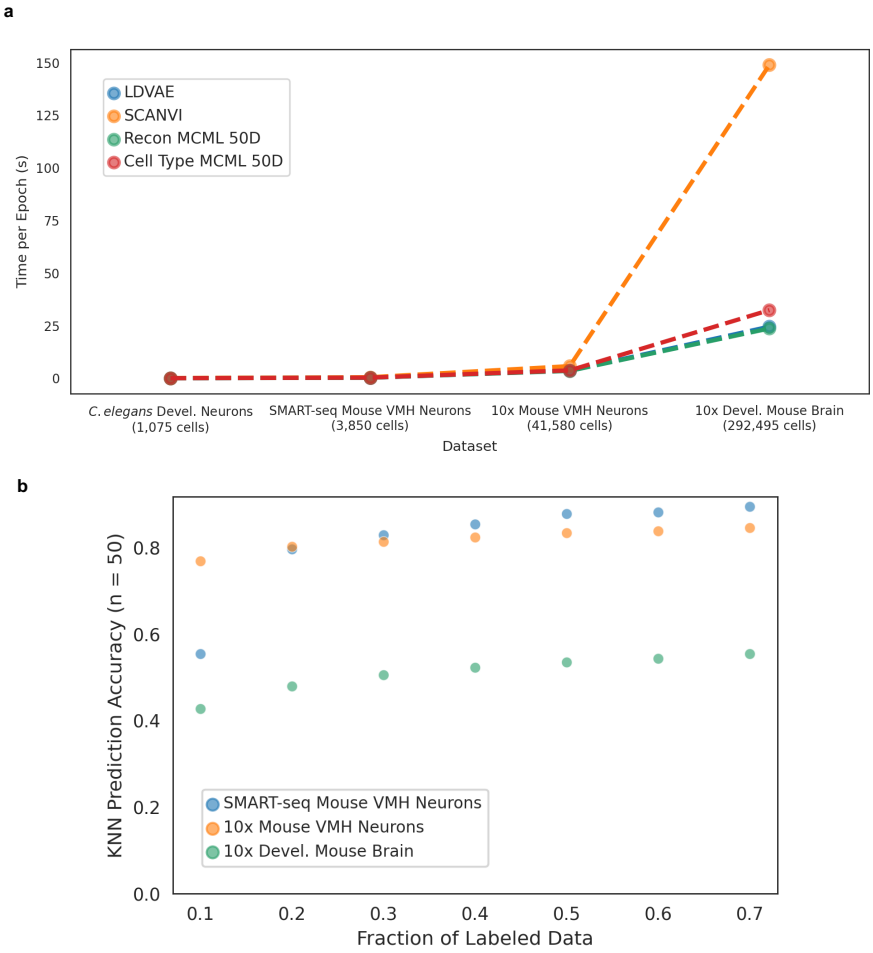


Supplementary Figure 11: MNIST Embedding Properties. **a)** Default t-SNE of the MNIST dataset. **b)** Default UMAP of the MNIST dataset. **c)** t-SNE MNIST plot with hidden points plotted in reverse order. **d)** UMAP MNIST plot with hidden points plotted in reverse order. **e)** Fraction of the correct digit in each of the ten k-means clusters from the t-SNE embedding (see Methods). **f)** Fraction of the correct digit in each of the ten k-means clusters from the UMAP embedding. **g)** Pearsonr correlation of intra-distances (internal variance) of each digit, in each embedding, to the ambient variances. **h)** Kendall's Tau correlation of each digit's neighbor rankings to ambient space. For box plots, whiskers denote 1.5 times the IQR. [Code]

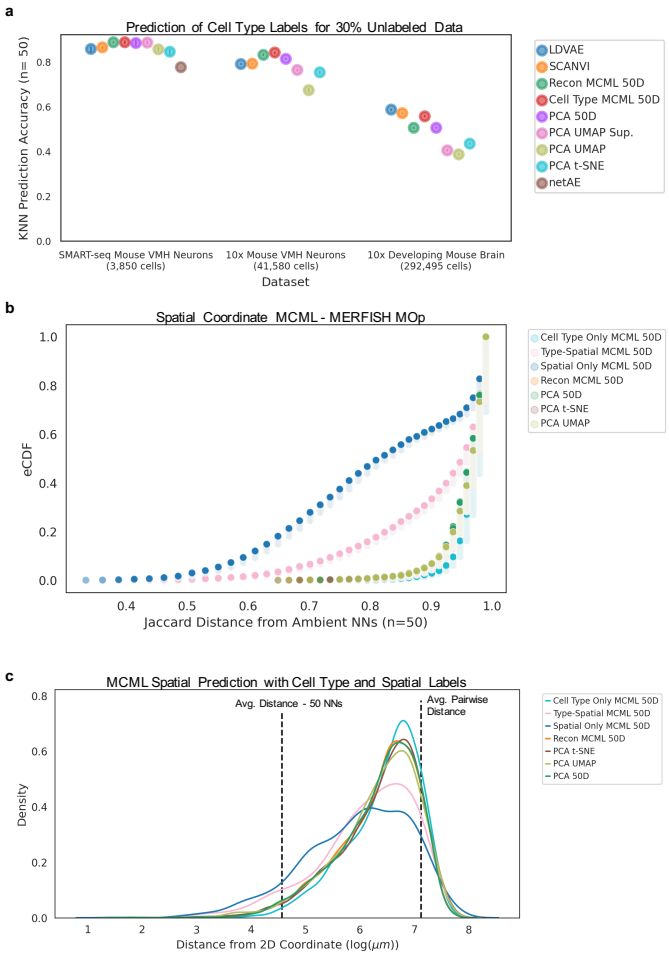
MCML Results



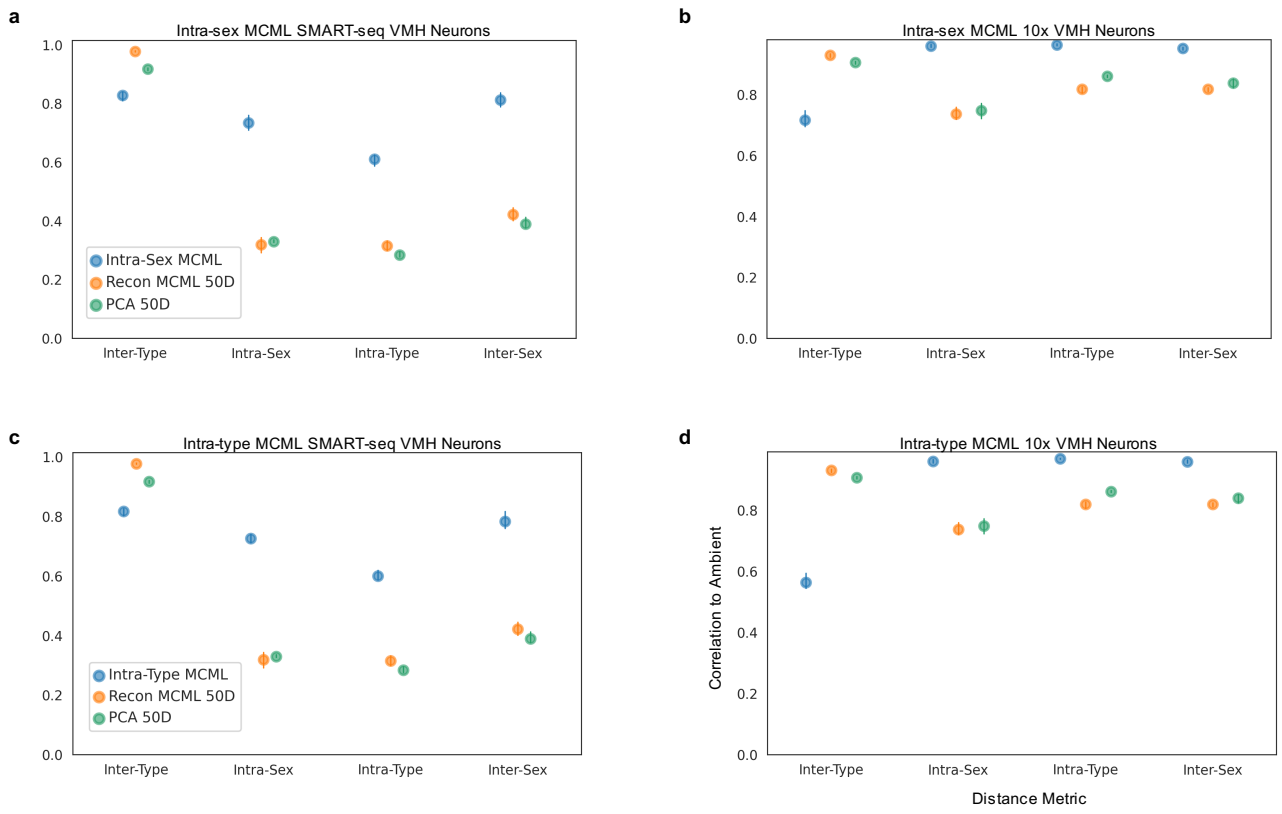
Supplementary Figure 12: Training and Testing Loss Curves for Picasso and MCML. a) Training and testing set loss plots for the MERFISH flower Picasso embedding. Loss plots include the loss for fitting to the shape's boundary ($L_{ShapeAware}$ in Methods), reconstruction loss ($L_{Reconstruction}$ in Methods), and the total combined loss. Training set represents 80% of the input data, and the testing set represents the remaining 20%. **b)** Training and testing set loss plots for the MERFISH flower Picasso embedding. **c)** Training and testing set loss plots for the MERFISH spatial coordinate MCML embedding. Loss plots include the discrete and continuous losses for the label-aware cost ($L_{LabelAware}$ in Methods), reconstruction loss ($L_{Reconstruction}$ in Methods), and the total combined loss. Only spatial coordinate labels are used (continuous values). **d)** Training and testing set loss plots for the 10x mouse VMH neuron cell type MCML embedding. Only cell type labels are used (discrete values). [Code a-c] [Code d]



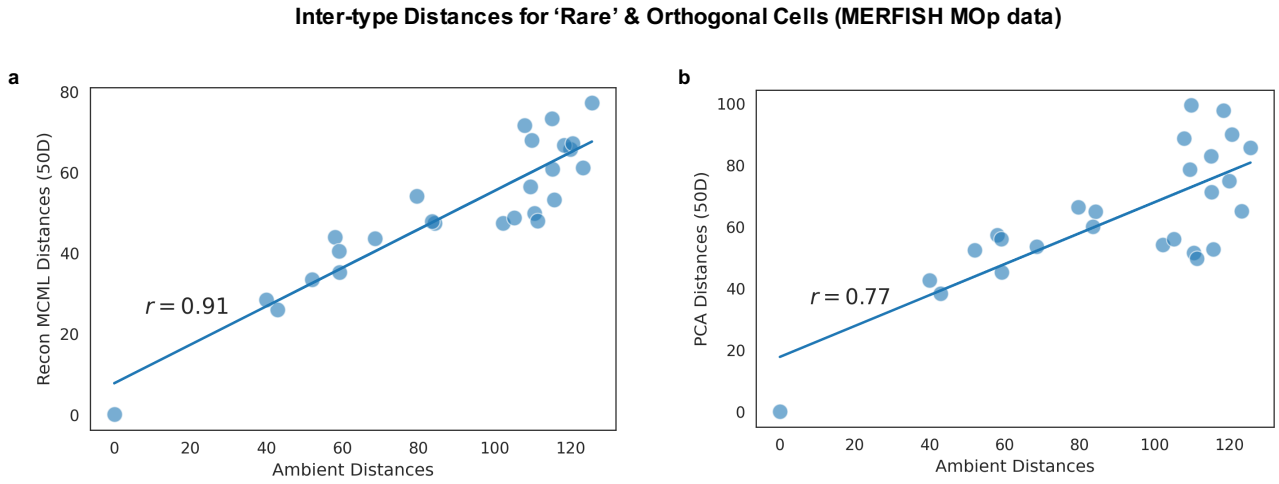
Supplementary Figure 13: Scalability of MCML for Prediction Tasks. a) Runtimes, per epoch, for each method used for cell type prediction in Fig. 4c (excluding netAE) across a range of cell numbers. **b)** Cell type label prediction accuracy for MCML on the three tested datasets, with lower fractions of labeled cells (see Methods). [Code]



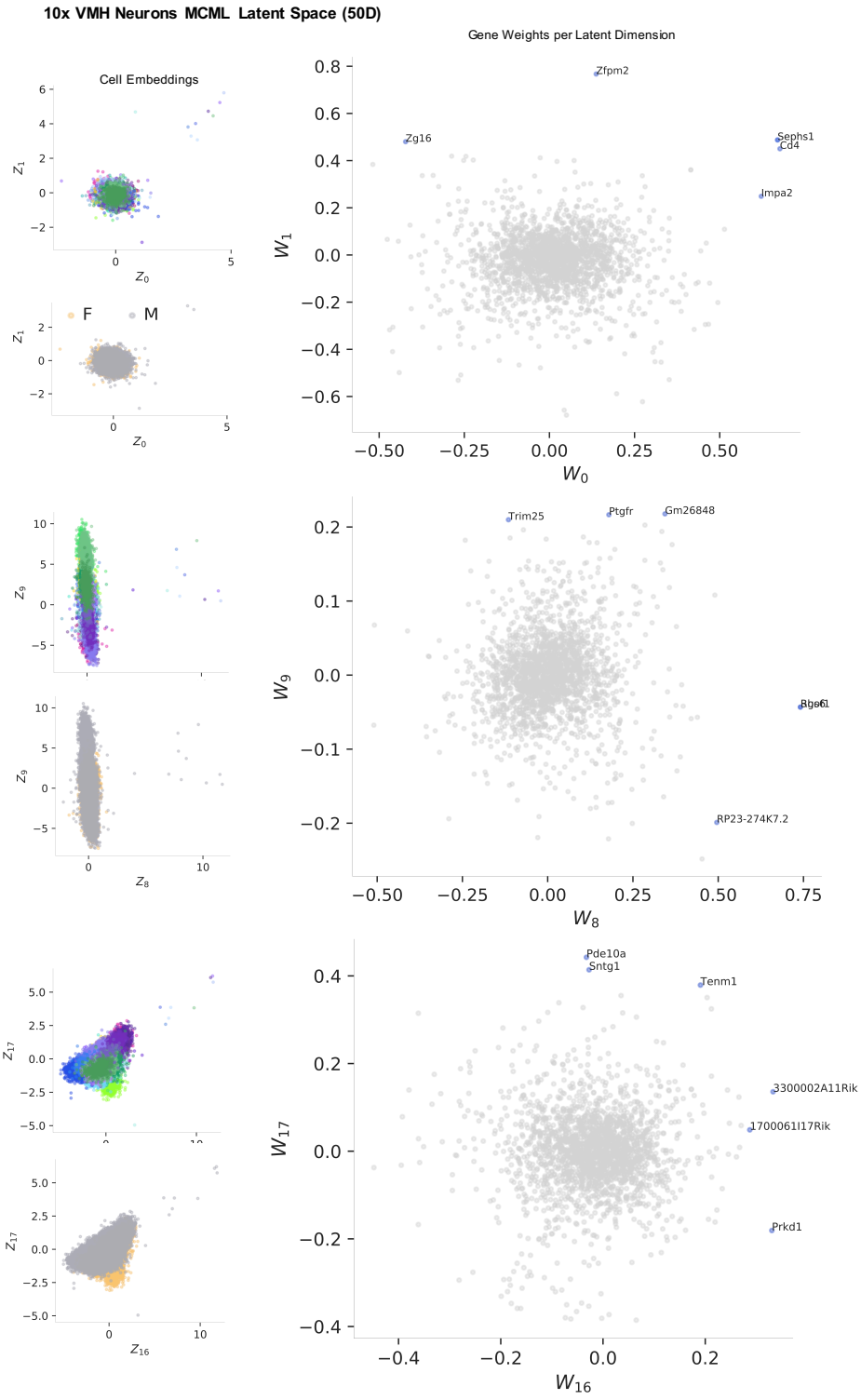
Supplementary Figure 14: Prediction Accuracy with Two-Dimensional Embeddings. a) Cell type prediction accuracy including accuracy scores for two-dimensional embeddings with semi-supervised UMAP (UMAP Sup.). For all plots bars denote the 95% C.I. **b)** eCDFs of Jaccard distance to ambient spatial neighbors including distances for two-dimensional embeddings (t-SNE and UMAP). **c)** Distributions of distance of predicted locations from actual spatial locations, including predictions from two-dimensional embeddings (t-SNE and UMAP). [Code a] [Code b,c]



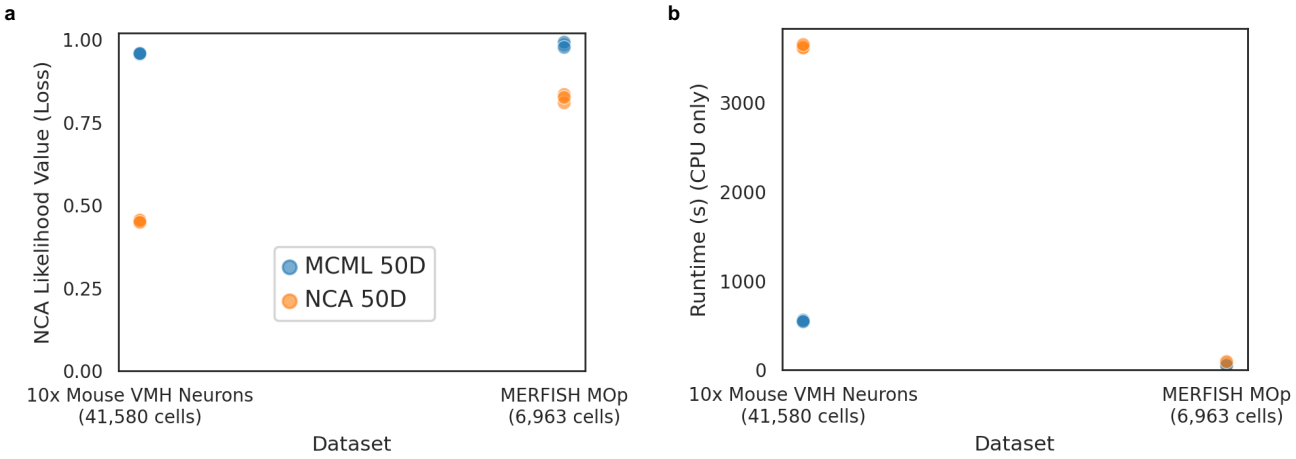
Supplementary Figure 15: Recapitulation of Ambient Properties with bMCML. a) Correlation metrics for bMCML (Biased MCML) with intra-sex distance correlation as cost function, for the SMART-Seq mouse VMH neurons. **b)** Correlation metrics for bMCML with intra-sex distance correlation as cost function, for the 10x mouse VMH neurons. **c)** Correlation metrics for bMCML with intra-type distance correlation as cost function, for the SMART-Seq mouse VMH neurons. **d)** Correlation metrics for bMCML with intra-type distance correlation as cost function, for the 10x mouse VMH neurons. (See Methods). For all plots bars denote the 95% C.I. [Code a,c] [Code b,d]



Supplementary Figure 16: Inter-Distance Correlation for Rare, Orthogonal Cell Type. **a)** Pearson correlation of inter-type distances to ambient distances, from simulated 'rare' and orthogonal cells embedded with Recon MCML (MCML with reconstruction error only). **b)** Correlation of inter-type distances to ambient distances, from simulated 'rare' and orthogonal cells embedded with PCA. (See Methods) [Code]



Supplementary Figure 17: Extraction of Gene Weight Loadings in Latent Space. On the left, embeddings of 10x mouse VMH neuron cells in various latent dimensions of $\mathbf{Z}$ are colored by cell type and sex (Male 'M' and Female 'F'). Embedding constructed from MCML with cell type and sex labels. Right hand column shows weight loadings for genes in each of the respective latent dimensions, with highly weighted genes labeled for each dimension. [Code]



Supplementary Figure 18: Comparison of MCML and sklearn NCA Loss Values. **a)** Value of the NCA cost function (likelihood) described in [40], equivalently $L_{Discrete}$ in Methods, measured on the sklearn NCA implementation and MCML (with no/zero reconstruction error) latent output, with 10x mouse VMH neurons and MERFISH MOp datasets as input. Uses only cell type labels for MCML and sklearn NCA. **b)** Total runtime for each method, using the same processing specifications as in the prediction method runtime comparisons (see Methods), with no GPU. [Code]

Data Tables

Name	DOI Link
C. elegans Developmental Lineage	
counts.mtx	https://data.caltech.edu/records/2060
cells.csv	https://data.caltech.edu/records/2061
genes.csv	https://data.caltech.edu/records/2062
MERFISH MOp	
metadata.csv	https://data.caltech.edu/records/2063
counts.h5ad	https://data.caltech.edu/records/2064
10x VMH Neurons	
metadata.csv	https://data.caltech.edu/records/2065
tenx.mtx	https://data.caltech.edu/records/2072
var.csv	https://data.caltech.edu/records/2066
tenx_raw.mtx	https://data.caltech.edu/records/2073
SMART-Seq VMH Neurons	
metadata.csv	https://data.caltech.edu/records/2067
smartseq.mtx	https://data.caltech.edu/records/2071
smartseq_raw.mtx	https://data.caltech.edu/records/2070
gene_names.npy	https://data.caltech.edu/records/2068
smartseq.csv	https://data.caltech.edu/records/2075
Developing Mouse Brain	
gene_names.npy	https://data.caltech.edu/records/2069
dev_all_hvg.mtx	https://data.caltech.edu/records/2043
dev_all_raw.mtx	https://data.caltech.edu/records/2044
lamannometadata.csv	https://data.caltech.edu/records/2045

Supplementary Table 1: Availability of Processed Data. Links to DOI registered data for all the pre-processed data used for the MCML and Picasso analyses.

Supplementary Notes

1. Limitations of Two-Dimensional Embeddings of Equidistant Points

For completeness, we review why no more than $n+1$ points in $\mathbb{R}^n$ can be equidistant.

Let X be a set of m equidistant points $\{\mathbf{x}_1, ..., \mathbf{x}_m\}$ in $\mathbb{R}^n$, i.e. $||\mathbf{x}_i - \mathbf{x}_j|| = |\mathbf{x}_k - \mathbf{x}_l||$ for all $i \neq j, k \neq l$, and suppose, without loss of generality, that the distances between them are 1. For any given i, we note that the difference vectors $(\mathbf{x}_i - \mathbf{x}_j)$ for all $i \neq j$ are linearly independent i.e. if $\sum_{j\neq i} \lambda_j(\mathbf{x}_i - \mathbf{x}_j) = 0$ then all λ_j must be equal to 0. This is trivial for $m = 1$ and $m = 2$. For $m \geq 3$, suppose, towards contradiction, that the difference vectors are not linearly independent, i.e. there are some coefficients λ_j $(j \neq i)$ not all equal to 0 such that

$$\sum_{j\neq i} \lambda_j(\mathbf{x}_i - \mathbf{x}_j) = 0. \tag{1}$$

Choose some point $\mathbf{x}_k$ where $k \neq i$ and note that the dot product of the difference vector $(\mathbf{x}_i - \mathbf{x}_k)$ with (1) is

$$\begin{aligned}
\sum_{j\neq i} \lambda_j(\mathbf{x}_i - \mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_k) &= \lambda_k \|\mathbf{x}_i - \mathbf{x}_k\|^2 + \sum_{j\neq i,k} \lambda_j(\mathbf{x}_i - \mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_k) \\
&= \lambda_k + \frac{1}{2} \sum_{j\neq i,k} \lambda_j (\text{ by the law of cosines}) \\
&= \lambda_k + \sum_{j\neq i} \lambda_j \ = 0.
\end{aligned} \tag{2}$$

Given that (2) is true for any $k \neq i$ we can sum (1) over all $k \neq i$ to obtain

$$\begin{aligned}
\sum_{k\neq i} \left(\lambda_k + \sum_{j\neq i} \lambda_j \right) &= m \sum_{j\neq i} \lambda_j \\
\Rightarrow \quad \sum_{j\neq i} \lambda_j &= 0.
\end{aligned} \tag{3}$$

As (2) applies to any point $\mathbf{x}_k$, it follows that for any $k' \neq i$,

$$\begin{aligned}
\lambda_{k'} + \sum_{j\neq i} \lambda_j &= \lambda_k + \sum_{j\neq i} \lambda_j \\
\Rightarrow \quad \lambda_{k'} &= \lambda_k \ .
\end{aligned} \tag{4}$$

Thus all coefficients λ_j $(j \neq i)$ sum to zero (3) and are equal (4), implying they are all zero which is a contradiction. Therefore all vectors $(\mathbf{x}_i - \mathbf{x}_j)$ for $j \neq i$ are linearly independent. Since we have identified $m - 1$ linearly independent vectors, and every basis for $\mathbb{R}^n$ has size n, it follows that $m \leq n + 1$, i.e. in $\mathbb{R}^n$, there can be a maximum of $n + 1$ equidistant points.

2. Bounds on Distortion of Equidistant Points

Induced distortion has been investigated in the literature for various conformations and embedding of points, e.g. the minimum distortion bound for embedding an n-point spherical metric onto a line [44] (akin to pseudotime inference), and the number of dimensions required to embed a metric space into a low-dimension normed space (defined by some l-norm) [45]. However, investigation of the implication of these bounds in real datasets across the sciences has been limited. Here we focus on the case of equidistant points and their distortion in two-dimensions to provide a more concrete realization of such bounds in the context of single-cell gene expression.

A trivial case of the result in Supplementary Note 1 is that of $n = 2$, namely that no more than three points can be equidistant points in $\mathbb{R}^2$. This raises the question of how close to equidistant more than three points in $\mathbb{R}^2$ can be? An impossibility theorem by [20] shows that it is impossible to obtain equality among even a subset of the pairwise distances distances of more than seven points. Even near-equality is impossible; specifically, a lower bound on the ratio between the maximum and minimum pairwise distances shows that distortion, which increases with the number of points, is inevitable.

A straightforward way to see this is via the two-dimensional isodiametric inequality which states that among all shapes of a given diameter, the circle has the greatest area (for a simple proof see [21]). Formally, for any body in $\mathbb{R}^2$, the area A is bounded above by $\frac{\pi}{4}$ times the square of the diameter D (the supremum of distances between any pair of points), i.e.

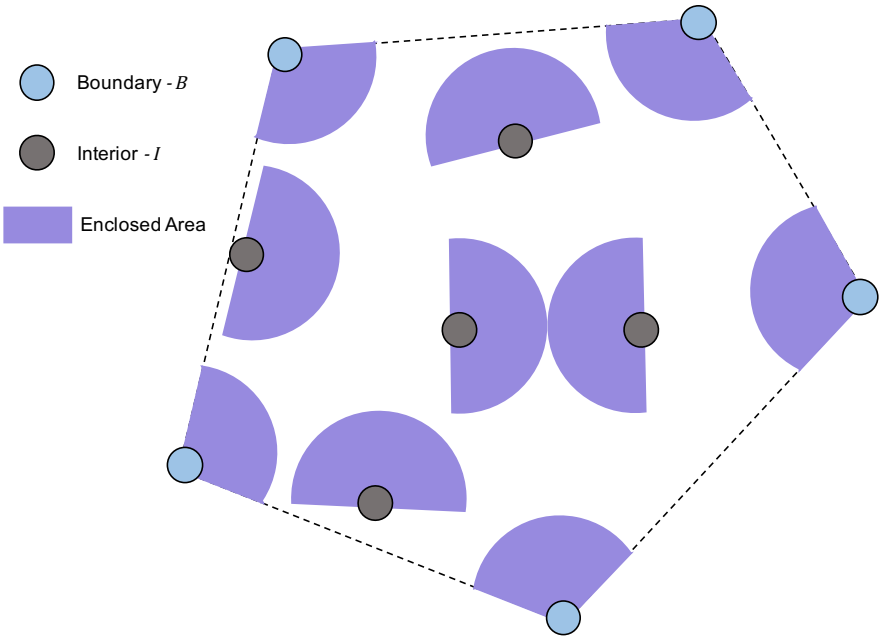
$$A \leq \frac{\pi}{4} D^2. \tag{5}$$

Theorem 1 *Given $n \geq 3$ points in $\mathbb{R}^2$, let d be the minimum distance among all pairs of points, and D the maximum distance (i.e. the diameter). The ratio of D to d satisfies*

$$\frac{D}{d} \geq \sqrt{\frac{n-2}{2}}. \tag{6}$$

Proof: Let B be the set of points consisting of the convex hull of n points in $\mathbb{R}^2$, and let I denote the remaining points, with $|B| = k$ and $|I| = n - k$. Note that for each point in I, there exists a semi-circle of radius $\frac{d}{2}$ centered at the point that does not touch any other point, or extend beyond the convex hull of the points (Supplementary Note Fig. 1). If we denote the sum of the areas of these semi-circles by A_I, we obtain

$$\begin{aligned} A_I &= \frac{1}{2}\left(\pi\left(\frac{d}{2}\right)^2\right)(n-k) \\ &= \frac{\pi d^2}{8}(n-k)\ . \end{aligned}$$



Supplementary Note Fig. 1: Bounding the Area Enclosed by Points in Two-Dimensions. Example of a set of 10 points showing the enclosed area for points in the I and B sets in the proof of Theorem 1.

Furthermore, for each of the k points in B, there is a circle sector of radius $\frac{d}{2}$ spanning the interior angle of the convex hull at that point that does not touch any other point, or extend beyond the convex hull. Since the sum of the interior angles of a k-gon is $(k-2)\pi$, we find that the sum of the areas of the circle sectors, which we denote by A_B, is given by

$$\begin{aligned} A_B &= \pi \left(\frac{d}{2}\right)^2 \left(\frac{(k-2)\pi}{2\pi}\right) \\ &= \frac{\pi d^2}{8}(k-2) \ . \end{aligned}$$

Summing A_I and A_B, we obtain a bound for the area enclosed by the n points:

$$\begin{aligned} A &\geq A_I + A_B \\ &= \frac{\pi d^2}{8}(k-2) + \frac{\pi d^2}{8}(n-k) \\ &= \frac{\pi d^2}{8}(n-2) \ . \end{aligned} \tag{7}$$

Combining the upper (5) and lower (7) bounds for the area A, we find that

$$\begin{aligned} \pi\frac{D^2}{4} &\geq \frac{\pi d^2}{8}(n-2) \\ \Rightarrow \frac{D}{d} &\geq \sqrt{\frac{n-2}{2}} \ . \end{aligned} \tag{8}$$

3. Principal Components of Equidistant Points

An implicit assumption typically relied on when applying principal components analysis (PCA) is that the linear transformation to a space that captures variation in the data is unique. While this is generically true, it fails in some pathological cases that are mostly uninteresting, but as highlighted here, relevant in understanding distortions that can result from the PCA map. We review one such case: PCA applied to equidistant points. Without loss of generality, we assume that a set of equidistant points has been scaled to the standard simplex, i.e. the points are described by the identity matrix $\mathbf{I}_n \in \mathbb{R}^{n\times n}$. Let $\mathbf{C}_n \in \mathbb{R}^{n\times n}$ be the mean-centered identity matrix where the column means are subtracted out:

$$\mathbf{C}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n;\ \mathbf{J} := \mathbf{1}\mathbf{1}^\top\ .$$

Singular Value Decomposition (SVD) of $\mathbf{C}_n$ produces the principal components of the equidistant points, and the singular values of $\mathbf{C}_n$ are found from the eigenvalues of $\mathbf{C}_n^\top\mathbf{C}_n$. These are straightforward to compute.

First, note that

$$\begin{aligned}\mathbf{C}_n^\top\mathbf{C}_n &= (\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)^\top(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)\\ &= \mathbf{I}_n - \frac{2}{n}\mathbf{J}_n + \frac{n}{n^2}\mathbf{J}_n\\ &= \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n = \mathbf{C}_n\ .\end{aligned}$$

Thus, the eigenvalues of $\mathbf{C}_n^\top\mathbf{C}_n$ are just those of $\mathbf{C}_n$. One of these is zero and the remaining $n-1$ are equal to 1. This is because

$$\begin{aligned}\mathbf{C}_n\mathbf{1} &= \mathbf{I}_n\mathbf{1} - \frac{1}{n}\mathbf{J}_n\mathbf{1}\\ &= \mathbf{1} - \mathbf{1}\\ &= 0\cdot\mathbf{1},\end{aligned}$$

and if $\mathbf{v}$ is a mean-centered vector with $\sum_i v_i = 0$ then

$$\begin{aligned}\mathbf{C}_n\mathbf{v} &= \mathbf{I}_n\mathbf{v} - \frac{1}{n}\mathbf{J}_n\mathbf{v}\\ &= \mathbf{v} - \mathbf{0}\\ &= 1\cdot\mathbf{v}.\end{aligned}$$

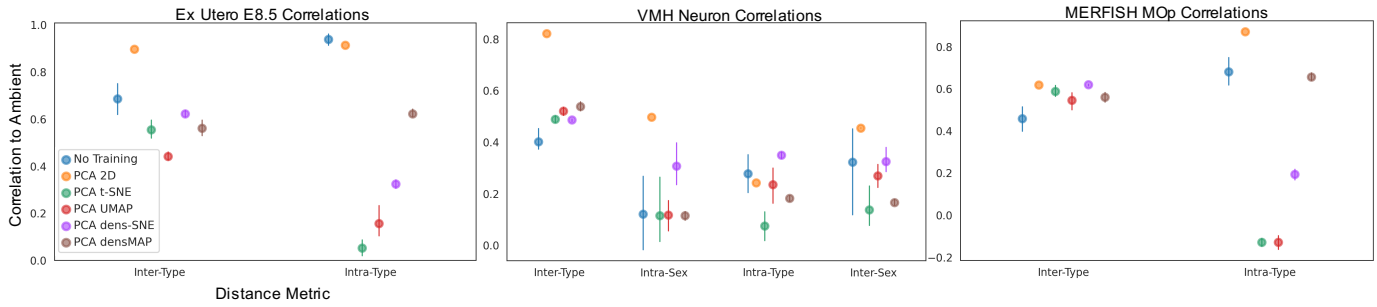
Thus, any vector $\mathbf{v}$ satisfying $\sum_i v_i = 0$ is an eigenvector with the same eigenvalue, meaning that not only are the principal components not unique, but any direction captures an equal fraction of the total variance. In practice, application of PCA to a set of equidistant points produces an arbitrary projection that will depend on software implementation details, including random number seeds and the numerical methods implemented for computing eigenvalues and eigenvectors.

4. Initialization of a Neural Network as Dimensionality Reduction

The Kaiming He initialization of neural networks, as described in [30], provides a default distribution from which weights are drawn to initialize neural networks in PyTorch. It was developed as an alternative to the practice of drawing weights from a normal distribution with fixed standard deviation [33] as a way to circumvent instability problems during backpropagation resulting from ReLU activation functions.

Instead of using a fixed standard deviation for the weights, for each layer l, He initialization draws weights from a normal distribution with mean 0 and variance $\frac{2}{n_l}$ where n_l is the number of inputs to layer l. This achieves stability in the variance across layers, i.e. the variance from the output layer will be matched to the variance at the input across layers. The derivation of the formula, which we omit, is straightforward using elementary properties of variance and expectation.

For the datasets in Figs. 2 and 3, we see that He initialization alone ('No Training') provides dimension reduction to two-dimensions that is competitive with t-SNE and UMAP on inter- and intra-distances. See (Supplementary Note Fig. 2).



Supplementary Note Fig. 2: Default Neural Network Weight Correlation Metrics. Inter- and intra-distance correlations for the ex-utero embryo E8.5 data, the (SMART-Seq) VMH neurons data, and the MERFISH MOp data. Two-dimensional embeddings with zero training epochs, i.e. a single forward pass through the autoencoder network, denoted as 'No Training'. [Code Ex Utero] [Code VMH] [Code MOp]

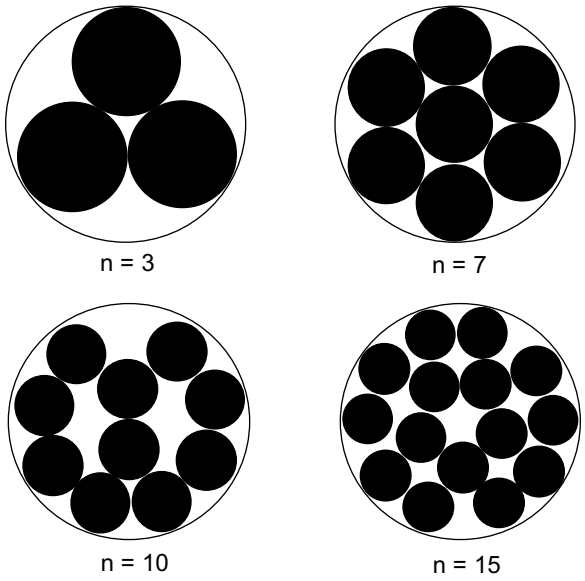
Though He initialization was designed for convolutional neural networks (CNNs) implementing ReLU activation functions, linear activation functions are also commonly employed in CNNs with other techniques for weight initialization [32,33]. Such methods include, as previously mentioned, drawing from a Gaussian distribution with fixed standard deviations [33] or a scaled uniform distribution (equivalently a Gaussian with zero mean and variance of $1/n$) [32] to initialize weights and avoid vanishing/exploding gradients over the network's many layers. It has also been noted that in particular CNN architecures, weight initialization alone can be surprisingly informative for feature learning [35]. As has been noted in the machine learning community [31], initialization methods for CNNs can be seen as analogous to the random projection of the Johnson-Lindenstrauss Lemma, whereby random $n \times k$ projection matrices selected from the standard Gaussian [19,36] or uniform distribution [34], are applied to points in $\mathbb{R}^n$, providing a linear map from $\mathbb{R}^n$ to $\mathbb{R}^k$ which preserves the pairwise distances [18]. Thus, the methods developed to minimize exponential variance over neural network layers resemble the random projections for dimensionality reduction and distortion-minimization of the Johnson-Lindenstrauss Lemma, with implications for preserving feature representation in biological datasets.

5. Minimizing Distortion

While Theorem 1 (Supplementary Note 2) shows that the distortion among points in $\mathbb{R}^2$ increases with the number of points, the bound does not resolve the question of what the best low-dimension embeddings are, in practice, for points that are equidistant in high dimension. Work on optimal packing arrangements (see, for example, [47]) has produced arrangements of points in $\mathbb{R}^2$ that maximize the minimum pairwise distance between points distributed in a unit circle. Numerical methods have been used to determine good packing arrangements, which are in some cases optimal, for up to 65 points [47] (see Supplementary Note Fig. 3 for examples). In this optimization framework, S represents a set of points placed in the unit circle, namely $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, and the minimum distance

$$d(S) = min\{\|\mathbf{s}_i - \mathbf{s}_j\| : 1 \leq i \leq j \leq n\}$$

is maximized. Such arrangements could, in principle, be utilized in low dimension embedding methods in order to minimize distortion of equidistant points at different length scales.



Supplementary Note Fig. 3: Optimal Packings. Diagrams of packed points (centers of black circles) in a unit circle which minimize the max/min ratio of the pairwise distances between the points (centers). Adapted from [47].