Article

Patterns

EMBEDR: Distinguishing signal from noise in single-cell omics data

Highlights

- An overview of the benefits and difficulties of dimensionality reduction
- A novel algorithm for quantifying and identifying quality within embeddings of data
- Quality can be optimized to find data scales and set algorithm parameters
- A cell-wise view of quality generates robust and interpretable representations of data

Authors

Eric M. Johnson, William Kath, Madhav Mani

Correspondence

madhav.mani@gmail.com

In brief

A novel algorithm for assessing the quality of dimensionality reduction (DR) methods is proposed and applied to several single-cell omics datasets. The method is local, quantitative, and statistical, which permits quality to be detected on a cell-wise basis in a manner comparable across parameter sets and DR methods. Optimizing DR methods per cell permits a novel embedding scheme that robustly reproduces structures in the original data.

Johnson et al., 2022, Patterns *3*, 100443
March 11, 2022
https://doi.org/10.1016/j.patter.2022.100443

CellPress

Patterns

CellPress
OPEN ACCESS

Article

EMBEDR: Distinguishing signal from noise in single-cell omics data

Eric M. Johnson,[1,2] William Kath,[1,2] and Madhav Mani[1,2,3,4,*]
[1]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208, USA
[2]NSF-Simons Center for Quantitative Biology at Northwestern University, Evanston, IL 60208, USA
[3]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA
[4]Lead contact
*Correspondence: madhav.mani@gmail.com
https://doi.org/10.1016/j.patter.2022.100443

THE BIGGER PICTURE Modern technologies have enabled biologists to construct enormous datasets containing millions of observations of thousands of measurements. These datasets push the limits of traditional analysis techniques, leaving doubts about the quality and fidelity of these methods. In this work, we present a sort of meta-algorithm, called EMBEDR, which seeks to evaluate when a certain class of methods, known as dimensionality reduction methods, are generating high-quality representations of data. We show that EMBEDR allows for visualizations of even large datasets to be interpreted with confidence. Furthermore, we show how asking about the method quality itself can lead to improved analyses of data. These improved analyses may directly impact our understanding of cellular biology, including how cells behave, grow, and respond to stimuli.

Concept: Basic principles of a new data science output observed and reported

SUMMARY

Single-cell "omics"-based measurements are often high dimensional so that dimensionality reduction (DR) algorithms are necessary for data visualization and analysis. The lack of methods for separating signal from noise in DR outputs has limited their utility in generating data-driven discoveries in single-cell data. In this work we present EMBEDR, which assesses the output of any DR algorithm to distinguish evidence of structure from algorithm-induced noise in DR outputs. We apply EMBEDR to DR-generated representations of single-cell omics data of several modalities to show where they visually show real—not spurious—structure. EMBEDR generates a "p" value for each sample, allowing for direct comparisons of DR algorithms and facilitating optimization of algorithm hyperparameters. We show that the scale of a sample's neighborhood can thus be determined and used to generate a novel "cell-wise optimal" embedding. EMBEDR is available as a Python package for immediate use.

INTRODUCTION

Advances in high-throughput measurement techniques are revolutionizing biology. The advent of single-cell omics approaches, in particular, promises to illuminate the processes of cellular differentiation, multicellular patterning, signaling, and variation at single-cell resolution.[1–13] However, omics data are high dimensional—each measured gene adds a dimension to the sample space—leading to an explosive increase in the volume occupied by the data due to the curse of dimensionality (see Figure S1 for an illustration).[14] In addition, single-cell methodologies generate significant technical noise due to the small amount of material being measured.[15–19] Thus, despite the great promise that single-cell omics approaches hold, it remains a challenge[20] to separate signal from noise in these datasets or make data-driven inferences.[14]

Faced with the challenges posed by high-dimensional datasets, a host of methods have been developed to help make quantitative inferences from the data. One such class of methods, termed dimensionality reduction (DR) methods, attempts to reduce the size (dimensionality) of the data by identifying a reduced set or combination of features (genes) on which further qualitative or quantitative analysis can be applied with more inferential power. Significant effort has been put into the

development and application of DR algorithms, such as principal-component analysis (PCA),[21] t-SNE,[22] UMAP,[23] and others.[24–38] Each of these methods attempts to find a lower-dimensional (usually 2D or 3D) representation, or *embedding*, of the data that preserves important aspects of the original data structure (for a review, see Van der Maaten et al.,[39] Gracia et al.,[40] and Espadoto et al.[41]; in application to omics data, see Fanaee-T and Thoresen[42]).

Ideally, a researcher would prefer a reduced representation of their data for gaining biological insight as it may avoid spurious conclusions caused by the curse of dimensionality. These representations can then be used to ask biologically relevant questions; for example, if cells from a tissue are sequenced, to what extent can we say that two clusters in the embedding correspond to distinct, differentiated, cell types? If clusters in such a view are connected by a bridge of cells, does this imply the existence of a path along which cells are differentiating? If cells subjected to different treatments of a drug are processed through a DR method, how is the strength of the treatment effect correlated with distance in the lower-dimensional space? Experimentally, one might be concerned with the depth of sequencing or the number of samples; how does this information get transformed into a dimensionally reduced representation of the data? Put more plainly, DR methods produce an approximate picture of the data, and we would like to know what parts display biological signal, and what parts are simply algorithmic distortions.

In traditional data analyses, statistics provides a rigorous framework with which to answer these questions, but DR methods confound the statistical distinction between signal and noise. Specifically, DR methods generically produce distortions in their representations of data, and these distortions are inhomogeneous across a representation;[30,40,43–47] are often stochastic and non-linear, meaning that the robustness and reproducibility of results is hard to assess;[41] and often require user specification of hyperparameters, where this specification is often based on heuristics rather than quantitative principles.[10,48–50] Addressing these issues provides the motivation for this work, as recovering the ability to separate signal and noise in DR output is essential for their utilization in quantitative analyses.

These difficulties with DR methods can be insidious. As an illustration, consider a sample dataset that populates the tips (vertices) of a regular tetrahedron in three dimensions. (A slightly more complicated example can be found in Figure S2.) The vertices of this tetrahedron are all equidistant in the original three dimensions of the data, but any squashing of the pyramid into two dimensions will necessarily result in the distances between some pairs of vertices being distorted. For example, flattening the pyramid onto its base will make the top vertex look artificially close to the other three. Alternatively, moving the top to a point outside the bottom triangle will make it artificially far from one of the base vertices. Real data are more complicated than a tetrahedron: cells are arranged in gene expression space in unknown geometric relationships with heterogeneous densities. But if in even simple cases one cannot match nearby regions in the original data to nearby in the DR output—or far as far—any interpretation of the dimensionally reduced representations of real single-cell data must proceed with caution.

To address the distortive effect of reducing dimensions, DR algorithms often employ stochastic or non-linear techniques, which work with remarkable success in a variety of contexts.[41] Using these techniques, however, also means that the exact outputs of a DR method will rarely look similar, whether comparing across methods, different parameter choices with the same method, or even across separate runs of the same method with identical choices of parameters. As an example, consider Figure 1, where scRNA-seq data from nearly 4,800 bone marrow cells from the *Tabula Muris* Cell Atlas[8] have been embedded in 2D using t-SNE[51] and UMAP[23] each at two different user-prescribed settings. (Throughout this work, we use k_{Eff} to parameterize t-SNE instead of perplexity. See section S3 for more information.) In each panel of this figure, the lower-dimensional representations demonstrate some apparently clustered structures, but the number, size, and shape of these clusters vary dramatically between the representations. As an example, the B cells in groups 2, 4, and 5 appear to be two to four distinct ”clusters” depending upon the panel that is considered. More extremely, the representation in (A) separates the granulocytes into two clusters, and both (A) and (B) separate the granulocytes from their progenitor cells. Without more information then, it is not obvious which of these panels best represents the high-dimensional structure of the data. An assessment of the “error” in these representations would allow for such a determination.

Together, these observations strongly motivate the need for methods to assess the size of dimensionality reduction (DR)-induced “error” associated with representing high-dimensional data in lower-dimensional spaces. We emphasize that even noiseless data will be distorted during the DR process, making error assessment a necessary component in applying these methods. However, it is also worth emphasizing that, despite these difficulties, DR for analysis and visualization is obviously useful. Our goal here is to develop a scheme that guides the user based on the data rather than merely advising the user to “be careful.” That is, an error-quantification scheme that can assess and quantify where a DR-generated representation is showing structure that is consistent with structure in the original, high-dimensional space (signal) as opposed to spurious structures that may be due to stochastic and non-linear methods (noise), would be immensely useful to the average analyst. Moreover, we specifically assert that a successful error-quantification scheme should do the following:

1. Assess quality locally: since the errors incurred in reducing the dimensionality of data are not distributed homogeneously across the lower-dimensional representation,[45,46] a quality-assessment scheme should provide *local* (per cell) estimates of DR-induced error as opposed to a single global estimate.
2. Assess variability in quality: to account for changes in quality that may be due to variation across different executions of a stochastic DR algorithm, a quality-assessment scheme should consider the *distribution* of errors across runs.
3. Assess quality statistically: a robust quality-assessment scheme should employ a *null hypothesis* to establish a “ruler” or baseline against which errors in data can be compared.

Others have addressed the problem of DR quality assessment: work has been done to provide heuristic guidelines on

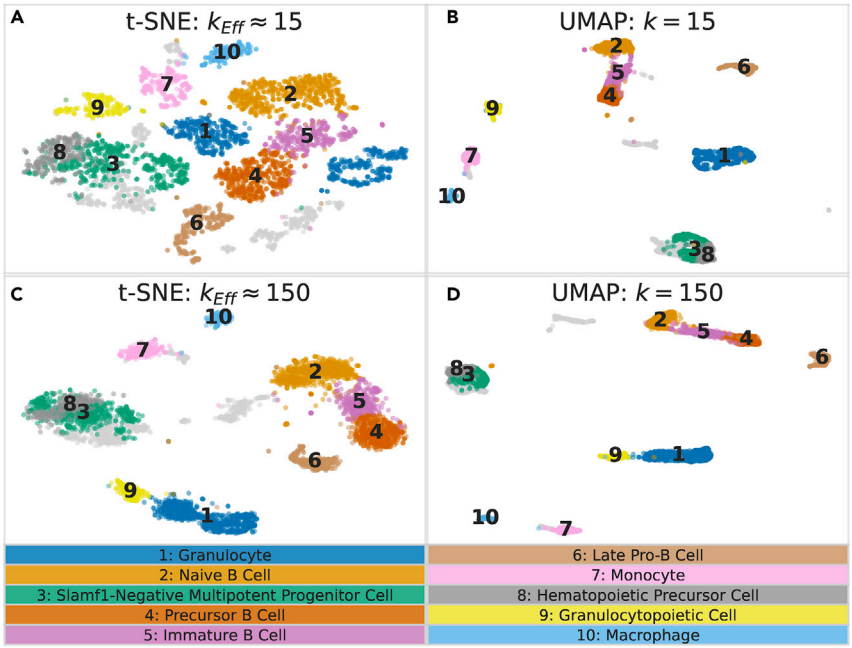

Figure 1. Features of dimensionally reduced data are sensitive to the choice of algorithm and algorithmic settings

(A and C) Four dimensionally reduced representations of RNA-seq measurements from 4,771 bone marrow cells collected by the *Tabula Muris* Consortium[8] generated by t-SNE at $k_{Eff} \approx 15$ (A) and 150 (C) (perplexity = 10 and 120, respectively; custom variation of the openTSNE implementation[51,52]) and by UMAP[23] at n_neighbors = k = 15 and 150. Ten previously annotated cell types provided by Schaum et al.[8] are colored and labeled. The same cells are colored and labeled in each panel.

(B) The number of nearest neighbors, k, is set to its default value, 15, in UMAP. Following the method in supplemental section S3, we use t-SNE with a similar number of nearest neighbors (k_{Eff} = 15) in (A).

(C and D) We visualize the data using t-SNE and UMAP, respectively, at a much larger number of nearest neighbors: $k_{Eff} \approx 150$ in (C) and k = 150 in (D).

how to appropriately use DR algorithms[10,48–50] and to make improvements to the algorithms themselves.[51–57] Several efforts to characterize the quality of DR methods have been pursued,[41,46,58] which can roughly be categorized as being global[30,58–65] or local[29,45,46,66,67] in scope, and either based on preserving distances,[65] neighborhoods,[30,46,58–60,62,68,69] or topology,[64,70,71] but in all cases they attempt to summarize the extent to which a given DR algorithm preserves some aspect of the original data's structure. In surveying this literature, and considering our basic principles, we find that what is still missing is an approach that not only assesses quality quantitatively and locally,[45,47,60,67–70,72] but also *statistically* in that it seeks to characterize the part of the natural and expected variability in quality that is due to noise.

It is with this in mind that we have developed the empirical marginal resampling better evaluates dimensionality reduction, or EMBEDR, algorithm to locally and statistically evaluate DR error. EMBEDR is a general approach that addresses the several unique concerns that arise with high-dimensional, noisy data, such as single-cell omics measurements, while also adhering to our motivating principles for a quality-assessment scheme.

RESULTS

The EMBEDR algorithm

In this section we describe the heuristic structure of the EMBEDR algorithm, as well as specific implementation details that are reflected in the figures throughout this work. Considered generally, EMBEDR is based on measuring the local, per cell, distortion of the DR method as a "quality" statistic. We then use empirical resampling methods to generate a null distribution for these statistics so that we may quantitatively assess whether a dimensionally reduced view of a cell's local neighborhood has more structure (signal) than we expect to be generated by random chance. We re-emphasize that the EMBEDR framework is agnostic to the DR method being employed and the ways in which quality is assessed. That is, while we focus on evaluating the accuracy of t-SNE and UMAP in representing single-cell omics data, EMBEDR is not specifically designed for these algorithms or datasets, but more generally to assess the quality of any DR method applied to high-dimensional data, as can be seen in Figure S12. Furthermore, to emphasize EMBEDR's most direct application, we focus on evaluating t-SNE and UMAP at the point where they are most commonly used in single-cell omics analyses: after initial data preprocessing for visualization of quality control, cell-type identification, and other results.

The EMBEDR algorithm consists of three steps: (1) the repeated *embedding* of the data (the repeated generation of low-dimensional representations of the data), (2) the construction and embedding of null datasets generated in a data-driven manner, and (3) the calculation of the quality statistics and the performance of a hypothesis test. These are illustrated in Figures 2A–2C, respectively. We elaborate on each of these three steps below. As suggested by the motivating principles, these steps focus on the calculation of a local quality statistic, the empirical embedding statistic (EES), for each sample (cell) in the dataset. We then go on to describe how our algorithm characterizes the distribution of the EES in a meaningful and useful way.

To clarify the notation throughout the rest of this paper: consider a data matrix X to be a collection of N_{cells} vectors, where each cell contains measurements for each of $N_{features}$ genes (for scATAC-seq data, this may be peaks instead of genes). Noting that, for stochastic DR algorithms, the data can be embedded multiple times to yield different lower-dimensional representations, we denote the position of the ith cell in the nth embedding by $\vec{y}_{i,n}$, where the number of embeddings is N_{embed}, and $\vec{y}_{i,n}$ is usually a 2D or 3D vector. For each cell, in each embedding, we calculate the quality statistic, which we denote $EES_{i,n}$. An asterisk (*) is used to indicate quantities that correspond to "null data" generated by resampling, so that a resampled high-dimensional data vector is $\vec{x}_i^*$ and its position in the embedded

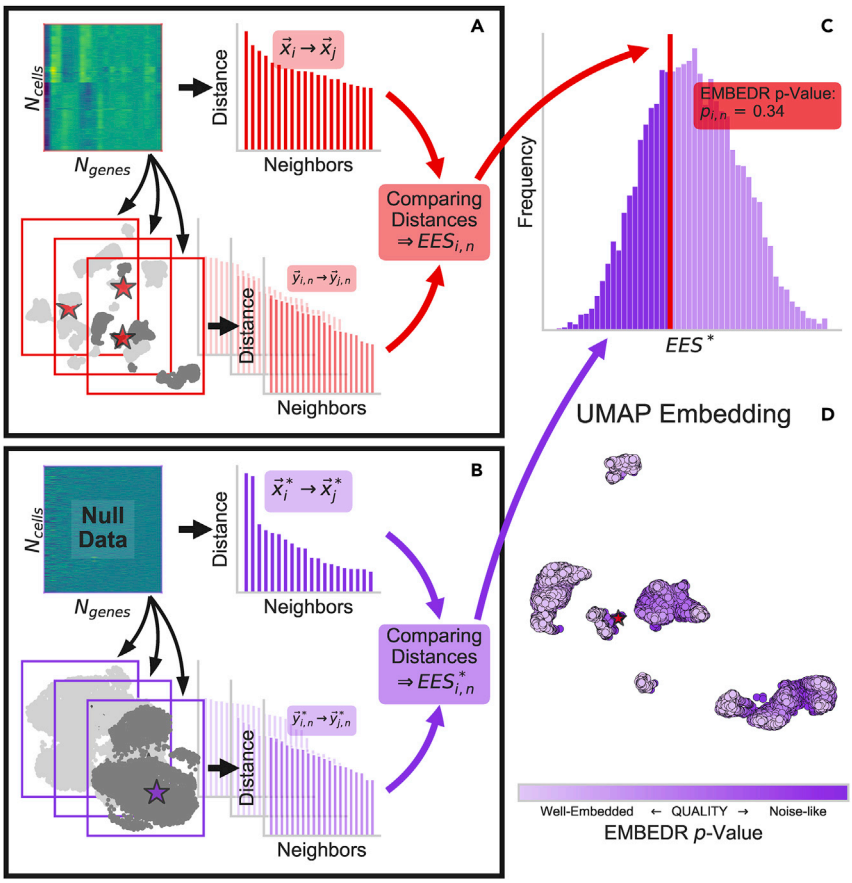

Figure 2. A schematic of the EMBEDR algorithm

(A) The data (5,037 FACS-sorted marrow cells from Schaum et al.[8] shown as a heatmap) are embedded in 2D using a DR method several times (here: UMAP with k = n_neighbors = 100). For each sample, the distances to neighboring samples are calculated in both the original data, $\vec{x}_i \rightarrow \vec{x}_j$, and the low-dimensional embedding, $\vec{y}_i \rightarrow \vec{y}_j$. An example cell is illustrated by a red star in each of the embeddings. These distance distributions are compared to calculate $EES_{i,n}$, a quality score for each cell in each embedding.

(B) The same procedure as in (A) is conducted using null datasets constructed via marginal resampling (see Figure 3). A purple star indicates a sample point in each null embedding.

(C) The individual $EES_{i,n}$ values are compared with the null distribution of EES^* to estimate a p value for each cell's embedding quality. This p value corresponds to the empirical likelihood that the null data could generate an observed or better embedding quality.

(D) The UMAP embedding of the data from (A) is shown. Cells in this embedding are colored according to the p values calculated in (C) , so that embedding quality can easily be visualized across an embedding. The light purple cells are those whose neighborhoods are better preserved than expected by random chance.

space would be $\vec{y}_i^*$. The final step of the hypothesis test process involves calculating a p value, $p_{i,n} = \text{Prob}(\text{EES}^* \leq \text{EES}_{i,n})$, using an empirically generated EES* distribution. (EES* refers to the set of $\text{EES}^*_{i,n}$ across all cells in the null data and all N^*_{embed} embeddings of the null data.)

1. Embedding the data: the first part of the EMBEDR algorithm is to use a candidate DR algorithm to embed high-dimensional data in lower dimensions. For stochastic algorithms, such as t-SNE or UMAP, this embedding may be performed multiple times with differing results as the quality of a specific sample's location can vary dramatically between embeddings (see Figures S4 and S9). The multiple embedding process is illustrated in Figure 2A using UMAP. In the final step of the algorithm, the effect of these multiple embeddings is summarized into a single quantity, so that the choice of N_{embed} is not critical to the interpretation of the output, and instead mostly impacts the resolution of the output p values (see section S4 and Figures S9 and S10) For all datasets shown, the data were embedded 25 times (except for the Allen Brain data, which were embedded 12 times), but in practice we find that ~3 embeddings is sufficient to get broad patterns in embedding quality.

   Next, an affinity between pairs of cells in the high-dimensional space is calculated by applying a Gaussian kernel with fixed entropy to the pairwise distances (as in Van Der Maaten and Hinton[22]). This is repeated in the lower-dimensional embedding except that a Student's t distribution is used to calculate affinities. The affinity distributions for each cell in high and low dimensions are compared using the Kullback-Leibler divergence, D_{KL},[73] which constitutes our quality measurement. If the D_{KL} is small, it indicates that the two distributions are similar, suggesting that the neighborhood of the embedded cell looks similar to its neighborhood in the original, high-dimensional, gene expression space. This calculation is illustrated in Figure 2A. The use of D_{KL} as a quality metric has also been used in other contexts.[30,74] For more details on how this is calculated, see section S1.
2. Null construction and embedding: the most crucial step in the EMBEDR algorithm is the data-driven construction of biologically realistic "null" datasets that can be used to generate an expectation for embedding quality from data devoid of biological signal. EMBEDR achieves this via *marginal resampling*, which is a resampling procedure where each gene's expression levels in the null data are independently drawn from the distribution for that gene in the original data. Figure 3 illustrates this process. Computationally, if X is an $N_{\text{cells}} \times N_{\text{features}}$ data matrix of single-cell omics observations, X^* can be generated by independently drawing N_{cells} samples from each column in X with replacement (the resulting X^* has the same shape as X). In this way, the null data contain biologically realistic, marginal distributions of individual features (genes, peaks, principal components, etc.)—Figure 3B

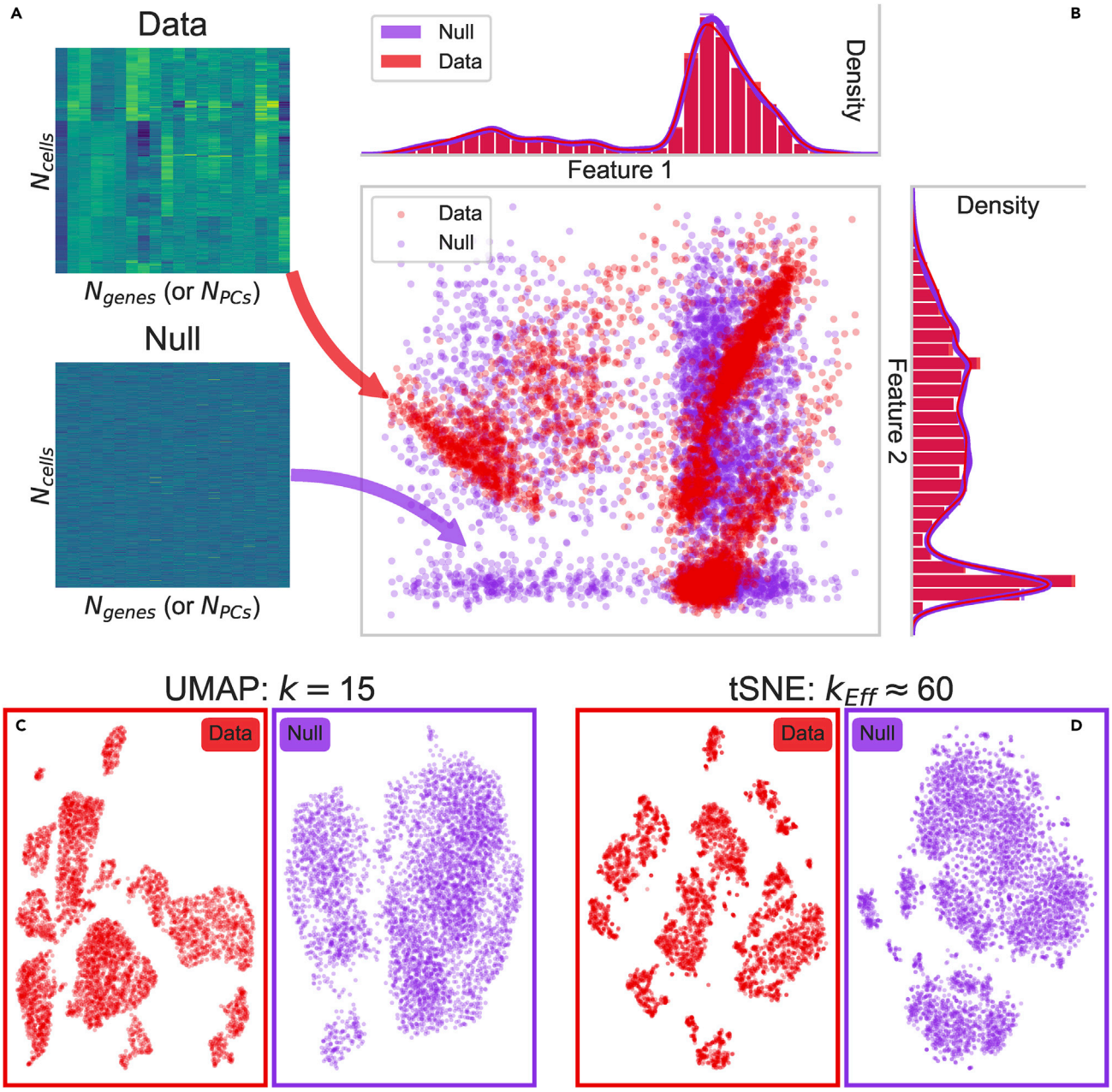

Figure 3. An overview of marginal resampling for generating null datasets

(A) Gene expression data for real and resampled scRNA-seq data (FACS-sorted marrow cells[8]) are shown as heatmaps.

(B) The first and second principal component of the data in (A) are plotted against each other, and the corresponding marginal distributions are shown to the top and right. Kernel density estimates are also plotted on the marginal distributions.

(C and D) The effect of marginal resampling to generate null distributions is shown, where the data and a null dataset are embedded using UMAP at $k = 15$ and t-SNE at $k_{Eff} \approx 60$, respectively, which correspond to the default parameters for those algorithms.

shows that genes in an scRNA-seq dataset have nearly *identical* marginal distributions in both datasets—but the *joint* distribution of genes is altered. More technically, the null dataset comprises a joint probability distribution constructed from the explicit product of the individual marginal distributions—guaranteeing statistical independence of the features in the null data. This property of independence generates a more diffuse distribution of cells relative to the real data, allowing for the assessment of whether real cells populate higher-density regions in expression space than expected. Any clustering that manifests in the null dataset is therefore a consequence of the properties of the original data's marginal distributions and the specific DR algorithm employed. In addition, in this work, the null generation takes place *after* normal data preprocessing (including normalization) so

that the cellular library size distributions are similar between the null and data samples.

As with the original data, we recommend that the null-generation and embedding process be repeated several times so that the distribution of null quality statistics, EES_i^*, is well resolved. In the examples shown in this paper, we have generated and embedded ten null datasets (three in the case of the Allen Brain data). However, we have also found that, in practice, a single null dataset is sufficient to characterize the distribution of EES_i^*, and that additional nulls mostly add improved resolution to the p value calculation outlined in the next step of the algorithm.

We note that the use of marginal resampling has been used successfully in several other contexts where the signal under examination was assumed to be a result of correlations in the data[75–77] and is similar to methods used for selecting statistically significant principal components.[78] It is reasonable to assume that correlation structures are discoverable by DR methods, as these methods leverage the covariance (PCA) or pairwise distance (t-SNE, UMAP) matrices to generate embeddings. Constructing null data via marginal resampling is also a model-free and a parameter-free process. In the context of scRNA-seq data, these resampled datasets correspond to the hypothesis that all cells are sampling a common distribution of gene expression, which is a useful and generic null hypothesis for many biologically interesting problems, such as cell-type identification, where the hypothesis would be that gene distributions depend on cell identity.

Figures 3C and 3D serve to underscore why we should generate these null data empirically: uncorrelated data are not necessarily uniform, meaning that clusters and structures can appear in DR representations of signal-less data! This is not necessarily intuitive, as one might naively expect clustering to be a consequence of cells having similar expression profiles, but clusters will be generated by many DR methods even when no such signal is present.[52] Furthermore, there are no theoretical results that describe the application of arbitrary DR methods to arbitrary data, so that marginal resampling is also a practical approach to this problem.

3. Empirical hypothesis test: the final step in the EMBEDR framework is to perform an empirical hypothesis test. Once the null data have been created and the null embedding statistics EES^* have been calculated for many samples over several embeddings, each of the sample statistics, $EES_{i,n}$, can be compared with the aggregated distribution of null statistics, as illustrated for a sample point in Figure 2C. The fraction of null statistics, EES^*, that are smaller than $EES_{i,n}$ can be used to estimate the likelihood that null data would be embedded as well by uncorrelated data. This likelihood is interpreted as an empirical p value, and can be summarized across the N_{embed} embeddings[79,80] to give a single quality metric, p_i, for each cell. For the sake of interpretability, we make an estimate of the likelihood that a cell's quality is better than that of the null across the N_{embed} embeddings by calculating $P(EES_i \leq EES^*)$, which amounts to averaging the individual embeddings' p values. See section S4 for more details.

   The EMBEDR p values can then be used, as in Figure 2D, to color each cell within an embedding indicating regions of higher or lower amounts of embedding quality. When using D_{KL} as the quality statistic, lower p values indicate that a cell's neighbors are similarly distanced in the original and low-dimensional spaces, with closer neighbors (in the original space) weighted more than those further away. The use of other quality metrics[30,45,46] would require an appropriate adjustment to this interpretation, but the interpretation of the p value as a measure of better or worse than algorithmically induced distortions does not. We demonstrate the interpretation of these p values in our results.

In practice, the EMBEDR algorithm operates in conjunction with, not as a substitute for, any DR algorithm, requiring little user input beyond what the DR method would require on its own. The algorithm has been implemented as a ready-to-use Python package on Github for t-SNE and UMAP. The rest of this section describes specific observations resulting from the application of EMBEDR to single-cell datasets.

EMBEDR reveals where DR output shows signal versus noise

Now that we have a local and statistical approach to separating signal and noise in DR output, we can start to address the difficulties introduced by DR methods in a principled way. For example, we used the tetrahedron thought experiment (Figure S2) to intuitively show how DR methods introduce heterogeneous distortions in the dimensionally reduced embeddings, but the problem here is not that these methods generate such errors, it is that they are not systematic or predictable. That is, if there were a pattern to misrepresentations in the lower-dimensional embedding, then any of its features, such as the relative separation of two clusters or a cell's similarity to its neighbors, could be inferred by taking into account that pattern. Of course, single-cell data are not as well structured as a tetrahedron, so that a heterogeneity of quality can be *expected biologically*: a single-cell dataset from a mature tissue does not always have equal numbers of distinct cell types, or the cell types might have different levels of gene expression variability. What this means practically is that the distortions in a cell's placement in the lower-dimensional representation vary in a manner that is impossible to discern "by eye." Thus, a first step toward helping researchers use DR methods confidently is to identify *where* a dimensionally reduced view of data is preserving high-dimensional structure and where it is not.

In Figures 4A–4C we present lower-dimensional embeddings of the *Tabula Muris* marrow dataset at three different values of effective nearest neighbors, k_{Eff}, in t-SNE (see section S3 for a discussion on how k_{Eff} is calculated), which is a monotonic function of its perplexity parameter. We utilize k_{Eff} instead of perplexity to enable direct comparisons to UMAP and to provide a more intuitive parameterization of t-SNE. The cells in these representations are colored according to the level at which the DR method was able to preserve the high-dimensional neighborhood structure relative to noise (see the color bar in Figure 4D).

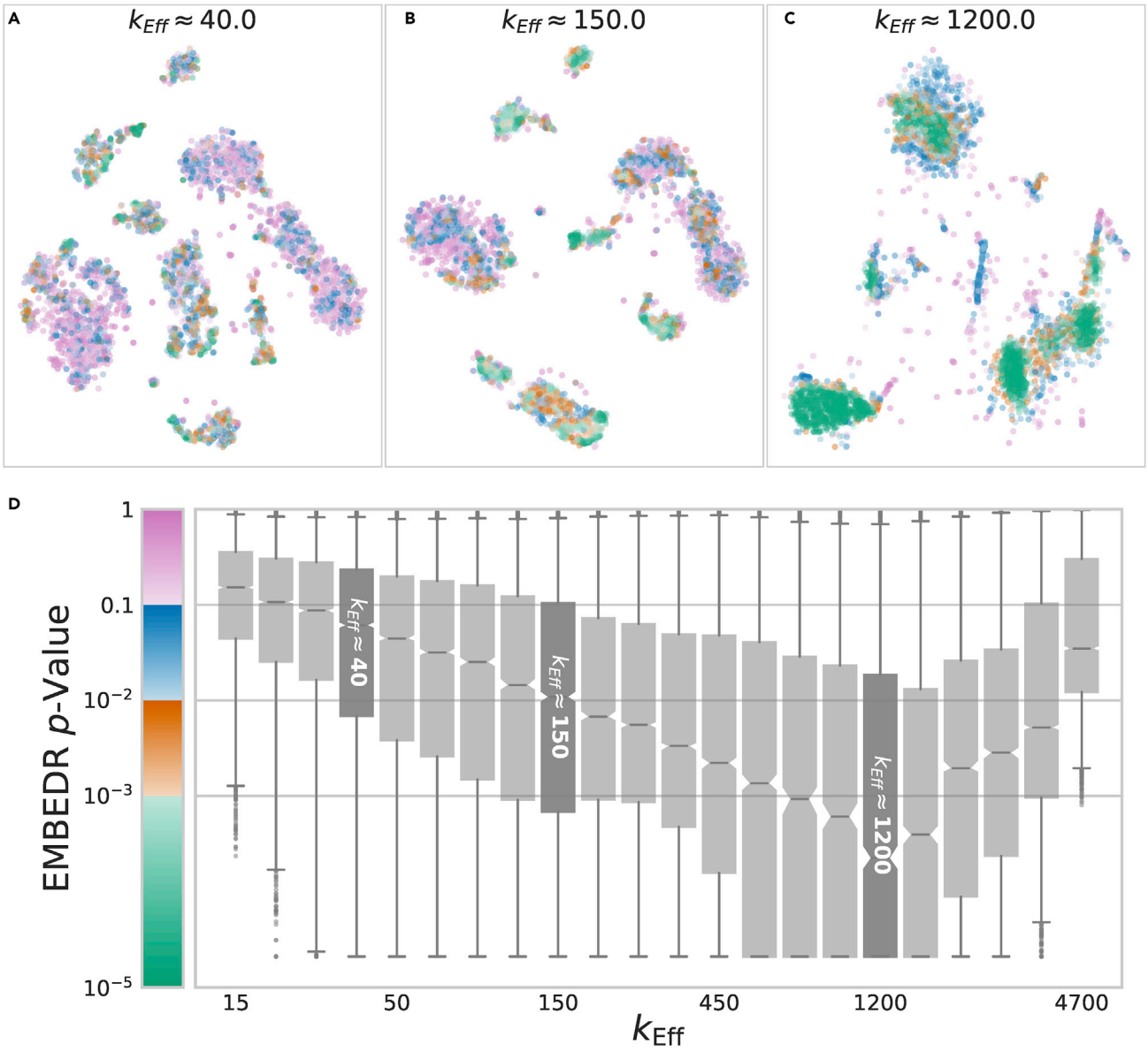

Figure 4. Optimizing DR algorithm hyperparameters generates high-quality embeddings

A total of 4,771 bone marrow cells from several mice[8] were embedded with t-SNE 5 times at several values of k_{Eff} and the EMBEDR p value was calculated using 10 null embeddings.

(A–C) Embeddings generated at three interesting values of k_{Eff}; each cell is colored by the EMBEDR p value (shown by the color bar) in (D). In (A), $k_{Eff} \approx 40$ corresponds to the default t-SNE parameter (perplexity = 30) in most implementations of t-SNE.[22,51] (C) An embedding generated using $k_{Eff} \approx 1200$ (perplexity = 1,000), which corresponds to the largest fraction of cells being well represented in the lower-dimensional embedding. Similarly, (B) shows the results at $k_{Eff} \approx 150$ (perplexity = 100), which corresponds to a second, smaller minimum in the p values.

(D) The distributions of p values are shown as box-and-whisker plots over each value of k_{Eff} and the median of the boxplot at $k_{Eff} \approx 1,200$ indicates that a substantial fraction of cells are best embedded at that hyperparameter value.

In this color map, green is used to illustrate cells whose quality is better than 99.9% of embedded cells from a null dataset. Orange then indicates cells that have a 99% chance of being better than the null, and blue indicates cells that are better represented than 90% of null cells. Pink cells are those whose neighborhoods in the lower-dimensional space are just as distorted as those generated by embedding signal-less data. As a result, this coloring allows a researcher to quantitatively understand where DR output is actually showing signal: regions of pink should not be closely interpreted since the illustrated shapes and distances are not representative of the original data. On the other hand, green regions suggest the presence of a biological signal, as the structures in those parts of the embedding are unlikely to have been generated by applying the DR method to signal-less data—i.e., they are unlikely to be due to the vagaries of the DR method. More quantitatively, a user can examine these quality

levels separately, as in Figure S11, to illuminate regions that are well embedded or poorly embedded.

Generally speaking, there are some immediate patterns worth pointing out. For example, at many values of k_{Eff}, cells that are clustered together appear to have a similar quality of embedding—there are blue (poorly embedded) clusters and green (well embedded) clusters. We will elaborate on this further in the next section. In addition, we observe that cells that are isolated from the center of mass of any cluster tend to be poorly embedded. However, we will see in the next results that such poor embedding may be largely due to the improper specification of DR method hyperparameters. More specifically, we find no correlation between the size of a cluster and its members' p values at their optimal specification, shown in Figures S16, so that both rare and common cell types are able to be assessed with EMBEDR.

It is also worth highlighting that Figure 4A employs the default parameters for t-SNE (perplexity = 30), but results in a low-quality dimensionally reduced representation of the data. Figures 4B and 4C are then a potentially surprising contrast, as large portions of the data are well represented when using hyperparameter values that are very different from common recommendations.[49] The difference in quality between these embeddings underscores the potential pitfalls of employing complex DR algorithms that require user-prescribed parameters without a quality-assessment methodology. We elaborate on this more in the next section.

In this way, EMBEDR's most immediate contribution is to provide a DR user with an intuitive map of their reduced-dimension data so that spurious structures can be separated from putative biological signals. A utility for generating plots like Figures 4A–4C is included in the Python package.

EMBEDR allows for optimization of algorithm hyperparameters

As expected, Figures 4A–4C clearly illustrate that the quality of a dimensionally reduced view of data can vary from cell to cell across the lower-dimensional space, but Figure 4D shows that quality can also depend strongly on values of DR hyperparameters. In this panel, each cell's p value is summarized as boxplots that change as we sweep across k_{Eff}, the effective number of nearest neighbors used by t-SNE to place cells in two dimensions. This figure thus allows for the detection of a "globally optimal" k_{Eff} based on where the largest fraction of cells are best embedded. For the *Tabula Muris* marrow tissue, setting $k_{Eff} \approx 1200$ corresponds to the largest fraction of minimal p values, as indicated by the shaded box in Figure 4. Interestingly, this is a far larger value for the perplexity parameter than is typically advised (perplexity = 1,000), even in some multiscale methods.[49,57] This is interesting in a practical sense, as EMBEDR provides a hyperparameter tuning scheme that differs from typical heuristics.

This result also emphasizes two important considerations. First, many DR methods—t-SNE and UMAP included—have a hyperparameter that corresponds to setting the size of "neighborhoods" in the high-dimensional data. (In section S3 we show how t-SNE's perplexity can be mapped to such a size, k_{Eff}, which we use throughout this paper.) This neighborhood size then acts like a low-pass filter in electronic circuits, in that information about cells that are further than a certain "scale" is neglected. Regardless of whether a "neighborhood" is defined as a distance or as a number of nearest neighbors, however, the scale felt by the data is always mediated by the *density* of the data in the high-dimensional space. What this means most directly is that the interpretation of the neighborhood size parameter must involve the size of the dataset. Telling a DR method to use 15 nearest neighbors will have a very different effect when applied to a dataset of 15 cells versus one with 15,000. In the former case, the effective scale is the entirety of the data, in the latter it may be the entire data or—more likely—it may be regions that differ in size for each cell depending on the data density around that cell. As a result, these DR hyperparameters must be set and interpreted uniquely for each dataset; in this paper we consider values for these parameters across their entire possible ranges. Furthermore, EMBEDR's data-sensitive statistical test means that Figure 4D can be constructed and interpreted consistently across datasets of different sizes.

Second, the fact that large sections of cells are best embedded when t-SNE considers $k_{Eff} \approx 1200$ nearest neighbors in Figure 4C means that utilizing fewer neighbors for these cells may result in spurious groupings,[52] which can be seen in the relatively poor quality of Figure 4A. As a result, the detection and interpretation of structures in low-dimensional representations need to account for whether the DR scale matches the "native" scale of the cells. The plateau in the curve in Figure 4D at $k_{Eff} \approx 150$ and the dip in the curve at $k_{Eff} \approx 1200$ means that most cells need to consider the positions of their 150 or 1,200 nearest neighbors to accurately position themselves in two dimensions, suggesting that $k_{Eff} \approx$ 150 and 1,200 correspond to native scales for these data. EMBEDR facilitates this assessment by permitting comparisons between hyperparameter choices and by assessing quality locally.

The salient features of Figure 4D in the context of the *Tabula Muris* marrow dataset are preserved across the datasets we have analyzed. For a list of datasets considered, see Table S1. A global p value sweep and a cell-optimal embedding for each dataset can be found in Figures S14 and S20–S24. In all cases, EMBEDR illustrates that (1) the quality of features in dimensionally reduced data varies in a manner that is difficult to discern "by eye," and that (2) the quality varies as a function of algorithmic hyperparameters and DR methods. Our ability to discern the local quality of dimensionally reduced data results from posing the problem statistically and the generation of data-driven null hypotheses. In addition, while it may be concerning that large portions of some DR outputs are consistent with spurious DR distortions, EMBEDR provides a quantitative tool with which to examine and improve these results.

EMBEDR allows for comparisons of DR algorithms

Novel DR algorithms are constantly being developed or adapted, so that their incorporation into single-cell analysis requires quantitative analyses of their performance. While assessments of these methods on select case studies have been performed in many studies,[10,11,13,41,48,49,81] there are no theoretical results that guarantee high-performance of any of these methods on a given dataset. Instead, our results and observations suggest that different methods will generate lower-dimensional

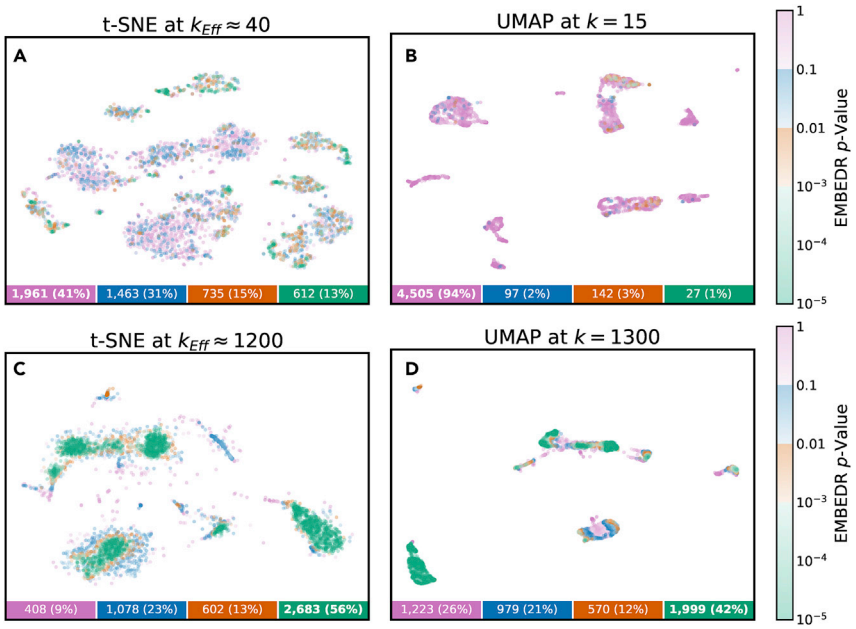

Figure 5. EMBEDR facilitates direct comparisons of DR methods

(A–D) A total of 4,711 cells from the *Tabula Muris* marrow tissue[8] are embedded by t-SNE and UMAP at default (A and B) and EMBEDR-optimized (C and D) numbers of nearest neighbors. Each cell in each embedding is colored by the EMBEDR p value according to the color bars on the right. The p values are calculated as in Figure 2 and in supplemental Section S4 using $N_{embed} = 25$ applications of t-SNE/UMAP to the data and $N^{*}_{embed} = 10$ embeddings of null data. In the boxes below each panel, the number (percentage) of cells at each p value threshold are shown (indicated by the corresponding color), with the threshold containing a plurality of cells shown in bold.

embeddings with different quality for different datasets. As a result, EMBEDR's data-driven quality assessment provides a natural tool for the comparison of DR methods applied to a common dataset.

Figure 5 illustrates this approach in action, as the quality of t-SNE and UMAP embeddings of the *Tabula Muris* marrow data are compared side-by-side. In the top row, we show embeddings generated at t-SNE's and UMAP's default parameters, while the bottom row sets k or k_{Eff} based on the optima identified in Figure 4. Below each embedding, the number of cells that meet a quality threshold are indicated, showing that at default hyperparameters neither t-SNE nor UMAP generate well-matched neighborhoods for most cells. However, the effect of optimizing t-SNE can now be seen in Figure 5C, as more than 50% of the cells have a neighborhood that is far more ordered than the random null. When UMAP uses the same number of neighbors in Figure 5D, the results are improved over the defaults (Figure 5B), but to a lesser extent than t-SNE. In Figure 5B, the null was generated by reducing the dimensionality of the resampled data using UMAP at $k = 15$. That is, the p values for each cell are determined based on how often UMAP randomly preserves structure in resampled data. In addition, the representations in Figure 5 are colored with p values generated by running t-SNE/UMAP on the data 25 times and on marginally resampled data 10 times, so that the p value indicates a *consistency* of quality as well, even though t-SNE and UMAP are stochastic and non-linear methods.

We emphasize that this should not be taken to mean that UMAP is not appropriate for the analysis of single-cell data, but only that t-SNE preserves structure better than UMAP *in this case*. We apply EMBEDR to other DR methods in Figure S12 and find similar differences in methods. Crucially, this direct, quantitative comparison of DR algorithms is an immediate consequence of our casting the quality-assessment problem as a statistical problem and by generating the null hypothesis empirically.

EMBEDR allows for a single-cell analysis of single-cell data

While our results in Figures 4 and 5 show that EMBEDR can be used to push forward global analyses of DR method quality, our earlier observations that quality is heterogeneous within a dataset suggest that we should be more careful and consider how embedding quality changes more *locally*. More directly, the existence of global optima in embedding quality at $k_{Eff} \approx 150$ and 1,200 does not imply that all cells are individually best embedded at those scales. Indeed, our expectation is that single-cell data will contain myriad densities, cell types, and expression patterns, meaning that we should expect to observe multiple scales in data generically. As a result, current methods are likely under-leveraging the information in our single-cell data by ignoring single-cell patterns.

EMBEDR provides a natural route to performing a single-cell resolution analysis of single-cell omics data as it already determines DR quality on a cell-wise basis. In Figures 6 and S13, we illustrate previously annotated cell types in the *Tabula Muris* marrow dataset to empirically demonstrate the existence of multiple scales in the data. Inspired by these observations, we propose to use a single-cell resolution analysis of single-cell data to produce a locally optimal dimensionally reduced view of data (Figure 7).

In Figure 6A, the EMBEDR p values for cells in six cell types from the *Tabula Muris* marrow dataset are shown as a function of k_{Eff}. Notice that each cell's p value "trajectory" can be followed as k_{Eff} changes, giving a cell-specific "spectrum" of quality. Considering the statistics of these spectra for each cell type shows that, indeed, some cell types are better represented at different scales than others. For example, macrophages (green) appear to be well embedded for $k_{Eff} \approx 150$, but the granulocytes and B cells are best embedded in a region around $k_{Eff} \approx 1{,}200$. In Figures 6B and 6C, two examples of embeddings at different k_{Eff} are shown to illustrate the features of these spectra. In Figure 6B, the neighborhoods of more than 80% of macrophages are better structured than noise, but in Figure 6C none of their neighborhoods are. The opposite happens for the granulocytes and B cells: using too few neighbors results in spurious clustering and over-fracturing of these cell populations; increasing to 1,200 neighbors captures that they are parts of large, diffuse regions of data space.

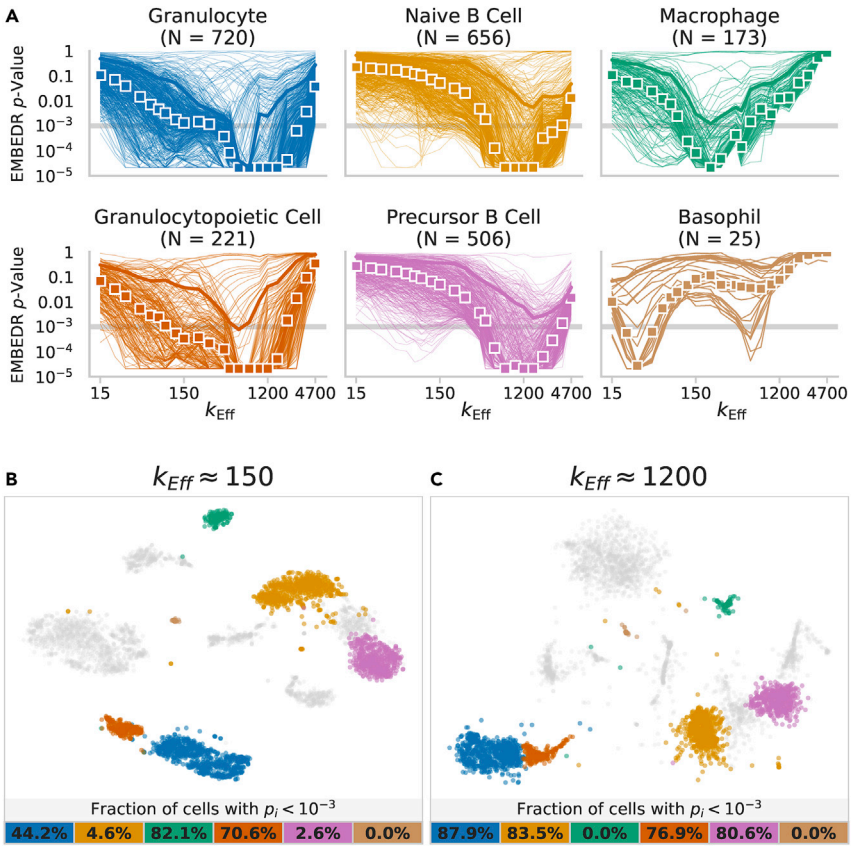

Figure 6. Different cell types are best embedded at a variety of scales

Using annotations from the *Tabula Muris* project,[8] the embedding quality of different cell types in the Marrow data can be examined individually across values of k_{Eff}.

(A) Six identified cell types from the bone marrow tissue are shown, where each cell with a given annotation is shown as an individual line. The colored boxes indicate the median p value across all cells with that annotation, and the solid lines indicate the 90th percentiles. Similar plots for all cell types are shown in Figure S13. Embeddings at $k_{Eff} \approx 150$ and 1,200 are shown in (B and C), respectively.

(B and C) The cells corresponding to each cell type are highlighted with the same color as in (A). Cells with an EMBEDR p value below 10^{-3} (the gray line in A) are opaque, while other cells with a highlighted annotation are lightly shaded. The fractions of such cells in an annotation are shown in the colored boxes below the embeddings. Other cell types are shown in gray for context.

More generally, in the context of datasets that may contain distinct cell types, we expect this to be reflected in these spectra, as members of the same cell type may have neighborhoods at a common scale. We observe this empirically in Figure S15, where cell annotations with more cells are best represented when t-SNE uses more neighbors. This makes sense, because if a cell is truly part of a cluster of N other cells, then incorporating spatial information from those N cells should be necessary to place that cell in an embedding. Conversely, cells from less-populous cell types may be poorly placed at high k_{Eff} because they are being positioned using cells that are not truly their neighbors. For example, the basophils are best embedded at a smaller scale ($k_{Eff} \approx 30$), which is likely because their neighborhoods are best described by only including those 25 cells.

In this way, Figure 6 demonstrates the existence of multiple scales in the data. The differences in the spectra of cells in different cell types illustrates the sizes of different neighborhoods in the data. In this figure, the cell annotations were a given, but the relationship between EMBEDR spectra and cluster sizes (Figure S15) suggests that EMBEDR may be useful for unsupervised cluster identification. The development of such a method is beyond the scope of this work and will be pursued in the future. Instead, in Figure 7 we show how adapting t-SNE to allow for scales to be set per cell results in an improved, scale-sensitive embedding that is easily interpreted biologically.

Specifically, using the spectra from Figures 4 and 6 for each cell, the value for k_{Eff} at which each cell was best embedded was determined (see section S5 for details). These values for k_{Eff} were used to generate a new similarity matrix where each cell used its own "preferred" neighborhood size to determine similarities between itself and its neighbors. This similarity matrix was then used to find a representation of the data via t-SNE. The resulting embedding is shown in Figure 7. We emphasize that this representation was determined in a completely unsupervised manner that involved no specification of t-SNE's perplexity parameter. In fact, this procedure *eliminates* the perplexity parameter from the embedding process!

Examination of this cell-wise optimized embedding using our established quantities in Figures 7B and 7C illustrates interesting patterns. In Figure 7B we see that the larger clusters are best embedded when the effective neighborhood size is large, while the smaller clusters only use $k_{Eff} \approx 100$ or fewer nearest neighbors. In this way, allowing the scales to vary locally facilitates the construction of specific and detailed structures in the embedding. These structures are robust, as reinforced by Figure 7C, where the minimal p value achieved by each cell across the parameter sweep is indicated, illustrating that all clusters were extremely well embedded at some value of k_{Eff}. In addition, Figure S19 shows that this cell-wise optimal embedding has a better average quality than default t-SNE.

In Figure 7D, we show the results of using an unsupervised clustering algorithm, DBSCAN,[82] on the cell-wise optimal embedding. That is, we took the unlabeled positions in (A), generated cluster labels, and in (D) and (E) we cross-reference these labels with the expert annotations generated by the *Tabula Muris* Consortium. Comparing these labels and annotations illustrates that the structures in this embedding are biologically relevant. Each of the seven clusters in Figure 7D clearly correspond to a class of bone marrow cell types, with almost no overlap between cell annotations except for granulocyte-monocyte progenitor cells. Similarly, the structure and arrangement of the clusters is biologically consistent: the annotated B cells (cluster 1, blue) are all

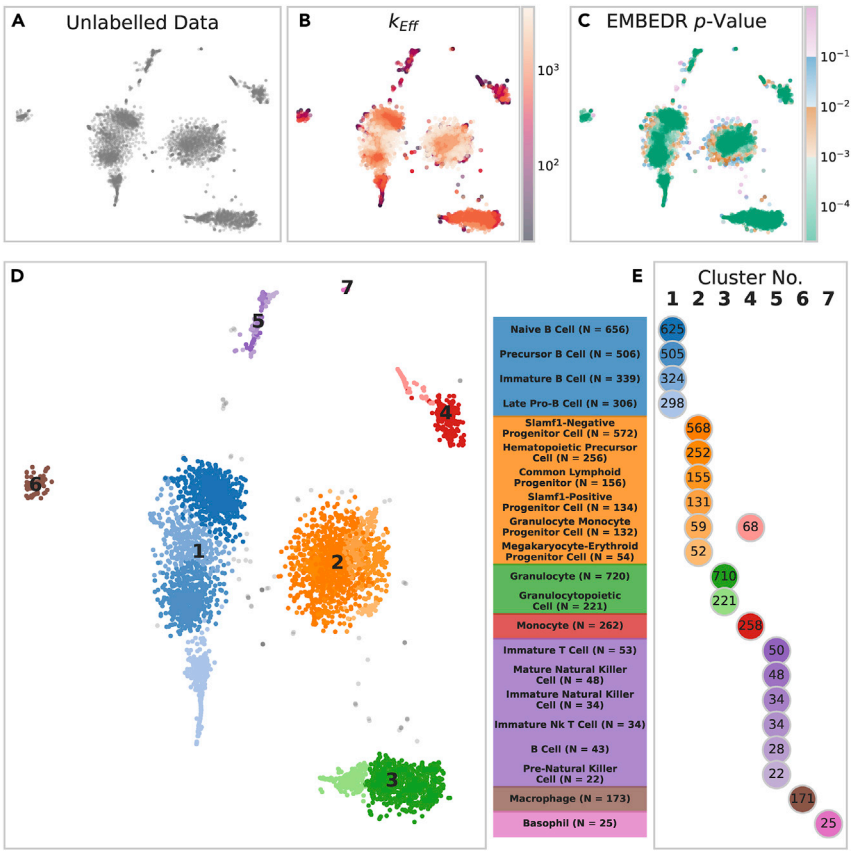

Figure 7. A cell-wise optimized embedding reveals clear biological signals

Adapting t-SNE to use a different scale for each sample in the *Tabula Muris* marrow data[8] generates a well-structured representation of the data.

(A) The unlabeled embedding is presented.

(B and C) To generate this embedding, the scale at which a cell's p value was minimized was used to set k_{Eff} for that cell. This k_{Eff} is shown in (B) and the minimal p value achieved by a cell across the sweep is shown in (C).

(D) Applying DBSCAN with eps set based on the pairwise distance (PWD) distribution of cells in the embedding (specifically, the 1.5th percentile of PWDs) detected the seven indicated clusters. Any *Tabula Muris* cell-type annotation for which more than 20 cells overlapped with a DBSCAN label was given a different shade of the cluster color.

(E) These cell annotations and colors are shown as a confusion table.

aligned according to their developmental trajectory from pro-B cells to naive B cells. At the same time, there is no differentiation pathway in the progenitor cells (cluster 2, orange), reflecting their common multipotent state. Furthermore, in Figures S17 and S18 we show that, regardless of how cluster labels are applied to this embedding, the distances between clusters in the cell-wise optimal embedding are more correlated with distances in the original data than those in a regular t-SNE representation. That is, distances between clusters in Figure 7D are actually correlated with distances between cells in gene space.

We can also see that smaller cell types, such as macrophages and basophils, are clearly separated from the larger clusters. This is another benefit of generating cell-optimal embeddings: cells that actually have small neighborhoods will be allowed to keep those small neighborhoods, even in the presence of larger clusters, which require larger scale parameters to be robustly resolved. That is, setting a DR hyperparameter large enough to resolve the structures in more populous cell types would normally squish or hide smaller, "rare" cell types, but this cell-wise optimization process protects against this. We also find that most cells have a scale at which they are well resolved, as shown in Figure S16, so that preserving these scales for each cell generates a better embedding.

We find that these results hold across other datasets. In Figures S25–S27 we recreate the process from Figure 7 for the *Tabula Muris* diaphragm and brain tissues and the MNIST digits. In each we find that structures are generated that obviously display biological (scRNA data) or visual (MNIST) meaning and that rare cell types are robustly represented in these embeddings even in the context of larger clusters.

DISCUSSION

Single-cell omics offers a path toward untold biological discovery, but its high-dimensional nature and inherent stochasticity requires the careful application of DR algorithms to make progress. The promise of DR approaches to single-cell omics data is not just to gain a visual intuition for the structure of the data, but to mitigate the curse of dimensionality and perform additional downstream quantitative analyses. As of now, the state of the art in DR currently rests on ever-changing heuristics to a degree that limits data analysis and data-driven discovery. A researcher cannot perform a comprehensive algorithm review for each new dataset, ensuring that the lack of a general approach to evaluating the quality of a DR method is preventing the community from making the most of the single-cell omics revolution. In the context of scRNA-seq, which has been the omic technology of focus in our study, cell-type classification,[8] lineage reconstruction,[48] RNA-velocity analysis,[83] and countless other approaches rely on the fidelity of dimensionally reduced data, or are limited by their inability to confidently employ DR.

The statistical approach presented in this work via the EMBEDR algorithm addresses these concerns by providing a rigorous framework for the evaluation of DR quality that can also reveal information about the data itself. The EMBEDR algorithm is relatively simple (Figure 2) and is available as a ready-to-use Python package. EMBEDR performs its quality assessment in a data-driven manner, meaning that it can be used to rigorously compare DR methods' performance (Figure 5). Perhaps more importantly, EMBEDR's local and statistical approach promises to reveal previously hidden structures in single-cell datasets while also facilitating hyperparameter optimization (Figure 4).

The EMBEDR method as proposed thus addresses the important question: "how much can I trust this dimensionally reduced view of the data?" Embedding quality is made available as a cell-wise, interpretable p value that has meaning across algorithms and datasets. This quality metric can be used to set algorithmic hyperparameters globally or locally, and can be leveraged to make inferences about the data itself. The method is robust and does not require the user to carefully specify parameters, in fact, the cell-wise optimal embedding process in Figure 7 effectively *removes* the perplexity parameter from the t-SNE algorithm.

This paper presents a broad view of the algorithm and its applications, but there are a few limitations that require further consideration. Most practically, the code as written rests on the speed of current implementations of DR algorithms that can be chained together to generate many (null) embeddings of the same data. Timings for various hyperparameter sweeps to generate figures like Figure 4 are shown in Table S2. As noted earlier, while there are benefits in principle to using many embeddings and many hyperparameter values, we find that as few as three data embeddings and a single null embedding can be sufficient. Furthermore, the recent extension of common DR methods to GPUs[84,85] or quadratic rate optimization schemes[55,86] promises drastic improvements to these runtimes, but their inclusion here was beyond the scope of this work.

The efficiency concerns also imply that there is a finite resolution to the calculated p values since the null distributions are calculated empirically. This means that the number of nulls that can be embedded determines the lower bound on the p values. Other than improved computational efficiency, remedies may include theoretical work to describe the tails of these null distributions or a principled method for parameterizing the null distribution.

Moving forward, it is clear that the nature of information that EMBEDR provides can be leveraged in a variety of ways not presented in this work. Several such directions are suggested in Figure 5, where more comprehensive efforts could be undertaken to assess the quality of DR algorithms generically, as in other studies.[40,41] Alternately, as suggested by Figure 6, a "spectral" view of embedding quality may provide an avenue for unsupervised clustering more directly. More simply, removing cells that never achieve a certain standard of quality may also be useful in improving traditional quality control processes. An extension of our null-generation process to non-normalized datasets may also permit EMBEDR to perform quality analyses of entire data-processing pipelines.

Non-computationally, Figure 4 suggests that this approach may be of widespread utility in the analysis of high-dimensional biological datasets to detect and to assess the stability of biologically relevant structures. Protecting samples with small neighborhoods from being subsumed by large-scale parameters suggests that EMBEDR's cell-wise optimal embeddings may be reliably used to detect rare cell types. In addition, the ability of the method to form model-free, non-parametric-scale spectra presents a new way to look at these datasets that may reveal heretofore unseen phenomena.

In all cases, high-dimensional and heterogeneous datasets, such as single-cell RNA-seq, require analysis techniques that account for and leverage the expected noise in the data to identify real biological signal. EMBEDR provides a robust statistical framework to achieve just that.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Madhav Mani (madhav.mani@gmail.com).

Materials availability

This study did not generate new unique materials or reagents.

Data and code availability

No new data were generated in this work. For a full list of references for the datasets considered in this work, see Table S1.

The code used to generate the results in this work is available here. Updates to the code, as well as examples and documentation, are available at our Github repository.

Single-cell data preprocessing

At this point, the EMBEDR algorithm has been designed and tested as a tool for assessing the quality of a specific DR algorithm when applied to a quality filtered and normalized dataset. DR algorithms are usually found at this point of an analysis pipeline, where they are used for visualization or confirmation of other results. In targeting EMBEDR at this stage of the process, our method allows researchers to assess the extent to which structures in their data are present or detectable in a 2D or 3D embedding. We do not, however, consider here the extent to which different data-processing steps affect a dataset's embedding quality in this work, although EMBEDR could also be utilized to evaluate this.

As a result, each of the datasets investigated in this work have been filtered and normalized following a standard protocol before we apply DR methods, such as t-SNE or UMAP. Specifically, all single-cell datasets were obtained pre-aligned from their sources. The scRNA-seq data were filtered so that each cell contained at least 500 genes and 50,000 reads. These cells were also filtered so that no cell contained more than 10% spike-ins, 10% ribosomal genes, or 40% *Rn45s*. The data were then normalized to account for each cell's library size and they were then log-transformed. The number of genes was then reduced to only the highly variable genes according to Satija et al.[87] before centering and scaling the data to have uniform variance. PCA was then applied and the first 50 components were kept (100 components were kept for the Allen Brain Institute data).

The scATAC-seq data from 10× genomics was also filtered so that each peak was found in at least 10 cells and each cell contained at least 1,000 peaks. Cells were also filtered so that their TSS enrichment score was at least 2. The cells were normalized using TF-IDF and then an SVD was applied and the first 50 components were retained.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100443.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants DMS-1547394 and DMS-1764421 and Simons Foundation grant 597491.

AUTHOR CONTRIBUTIONS

Conceptualization, E.J. and M.M.; methodology, E.J. and M.M.; software programming, E.J.; validation, E.J.; formal analysis, E.J. and M.M.; investigation, E.J.; resources, W.K. and M.M.; data curation, E.J. and W.K.; writing – original draft, E.J.; writing – review & editing, E.J., W.K., and M.M.; visualization, E.J.; funding acquisition, W.K. and M.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 17, 2021
Revised: June 25, 2021
Accepted: January 14, 2022
Published: February 8, 2022

REFERENCES

1. Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev. Cell *18*, 675–685. https://doi.org/10.1016/j.devcel.2010.02.012.

2. Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P.S., Rothenberg, M.E., Leyrat, A.A., Sim, S., Okamoto, J., Johnston, D.M., Qian, D., et al. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol. *29*, 1120–1127. https://doi.org/10.1038/nbt.2038.

3. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

4. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

5. Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science *360*, eaar3131. https://doi.org/10.1126/science.aar3131.

6. Mayer, C., Hafemeister, C., Bandler, R.C., Machold, R., Batista Brito, R., Jaglin, X., Allaway, K., Butler, A., Fishell, G., and Satija, R. (2018). Developmental diversification of cortical inhibitory interneurons. Nature *555*, 457–462. https://doi.org/10.1038/nature25999.

7. Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., and Klein, A.M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science *360*, eaar5780. https://doi.org/10.1126/science.aar5780.

8. Schaum, N., Karkanias, J., Neff, N.F., May, A.P., Quake, S.R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M.B., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. Nature *562*, 367–372. https://doi.org/10.1038/s41586-018-0590-4.

9. Kester, L., and van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. Cell Stem Cell *23*, 166–179. https://doi.org/10.1016/j.stem.2018.04.014.

10. Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 1–14. https://doi.org/10.1038/s12276-018-0071-8.

11. Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science *360*, 981–987. https://doi.org/10.1126/science.aar4362.

12. Dasgupta, S., Bader, G.D., and Goyal, S. (2018). Single-cell RNA sequencing: a new window into cell scale dynamics. Biophys. J. *115*, 429–435. https://doi.org/10.1016/j.bpj.2018.07.003.

13. Grün, D. (2018). Revealing routes of cellular differentiation by single-cell RNA-seq. Curr. Opin. Syst. Biol. *11*, 9–17. https://doi.org/10.1016/j.coisb.2018.07.006.

14. Altman, N., and Krzywinski, M. (2018). The curse(s) of dimensionality. Nat. Methods *15*, 399–400. https://doi.org/10.1038/s41592-018-0019-x.

15. Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods *14*, 565–571. https://doi.org/10.1038/nmeth.4292.

16. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.

17. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., and Garry, D.J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics *19*. https://doi.org/10.1186/s12859-018-2226-y.

18. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*. https://doi.org/10.1186/s13059-019-1874-1.

19. Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., and Zhang, N.R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. Nat. Methods *15*, 539–542. https://doi.org/10.1038/s41592-018-0033-z.

20. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*. https://doi.org/10.1186/s13059-020-1926-6.

21. Jolliffe, I.T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. *374*. https://doi.org/10.1098/rsta.2015.0202.

22. Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

23. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426.

24. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biol. Cybern. *43*, 59–69. https://doi.org/10.1007/BF00337288.

25. Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel Eigenvalue problem. Neural Comput. *10*, 1299–1319. https://doi.org/10.1162/089976698300017467.

26. Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science *290*, 2319–2323. 5500.2319. https://doi.org/10.1126/science.290.

27. Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323–2326. https://doi.org/10.1126/science.290.5500.2323.

28. Belkin, M., and Niyogi, P. (2003). Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Comput. *15*, 1373–1396. https://doi.org/10.1162/089976603321780317.

29. Chen, L., and Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. J. Am. Stat. Assoc. *104*, 209–219. https://doi.org/10.1198/jasa.2009.0111.

30. Venna, J., Kaski, S., Aidos, H., Nybo, K., and Peltonen, J. (2010). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. J. Mach. Learn. Res. *11*, 451–490.

31. Joia, P., Paulovich, F.V., Coimbra, D., Cuminato, J.A., and Nonato, L.G. (2011). Local affine multidimensional projection. IEEE Trans. Vis. Comput. Graph. *17*, 2563–2571. https://doi.org/10.1109/TVCG.2011.220.

32. Najim, S.A., and Lim, I.S. (2014). Trustworthy dimension reduction for visualization different data sets. Inf. Sci. *278*, 206–220. https://doi.org/10.1016/j.ins.2014.03.048.

33. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat. Methods *14*, 414–416. https://doi.org/10.1038/nmeth.4207.

34. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284. https://doi.org/10.1038/s41467-017-02554-5.

35. Wu, Y., Tamayo, P., and Zhang, K. (2018). Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative

embedding. Cell Syst. *7*, 656–666.e4. https://doi.org/10.1016/j.cels.2018.10.015.

36. Tarashansky, A.J., Xue, Y., Li, P., Quake, S.R., and Wang, B. (2019). Self-assembling manifolds in single-cell RNA sequencing data. eLife *8*, 1–29. https://doi.org/10.7554/eLife.48994.
37. Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., van den Elzen, A., Hirn, M.J., Coifman, R.R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. *37*, 1482–1492. https://doi.org/10.1038/s41587-019-0336-3.
38. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nat. Commun. *10*, 390. https://doi.org/10.1038/s41467-018-07931-2.
39. Van Der Maaten, L., Postma, E., and van den Herik, J.. (2009). Dimensionality reduction: a comparative review. TiCC TR, 2009–005 https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.
40. Gracia, A., González, S., Robles, V., and Menasalvas, E. (2014). A methodology to compare dimensionality reduction algorithms in terms of loss of quality. Inf. Sci. *270*, 1–27. https://doi.org/10.1016/j.ins.2014.02.068.
41. Espadoto, M., Martins, R.M., Kerren, A., Hirata, N.S.T., and Telea, A.C. (2021). Toward a quantitative survey of dimension reduction techniques. IEEE Trans. Vis. Comput. Graph. *27*, 2153–2173. https://doi.org/10.1109/TVCG.2019.2944182.
42. Fanaee-T, H., and Thoresen, M. (2019). Performance evaluation of methods for integrative dimension reduction. Inf. Sci. *493*, 105–119. https://doi.org/10.1016/j.ins.2019.04.041.
43. Gracia, A., González, S., Robles, V., Menasalvas, E., and von Landesberger, T. (2016). New insights into the suitability of the third dimension for visualizing multivariate/multidimensional data: a study based on loss of quality quantification. Inf. Vis. *15*, 3–30. https://doi.org/10.1177/1473871614556393.
44. Lui, K., Ding, G.W., Huang, R., and McCann, R. (2018). Dimensionality reduction has quantifiable imperfections: two geometric bounds. In Advances in Neural Information Processing Systems, *31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds. *(Curran Associates, Inc.)*.
45. Aupetit, M. (2007). Visualizing distortions and recovering topology in continuous projection techniques. Neurocomputing *70*, 1304–1330. https://doi.org/10.1016/j.neucom.2006.11.018.
46. Mokbel, B., Lueks, W., Gisbrecht, A., and Hammer, B. (2013). Visualizing the quality of dimensionality reduction. Neurocomputing *112*, 109–123. https://doi.org/10.1016/j.neucom.2012.11.046.
47. Colange, B., Vuillon, L., Lespinats, S., and Dutykh, D. (2019). Interpreting distortions in dimensionality reduction by superimposing neighbourhood graphs. In 2019 IEEE Visualization Conference (VIS) (IEEE), pp. 211–215. https://doi.org/10.1109/VISUAL.2019.8933568.
48. Herring, C.A., Chen, B., McKinley, E.T., and Lau, K.S. (2018). Single-cell computational strategies for lineage reconstruction in tissue systems. Cell Mol. Gastroenterol. Hepatol. *5*, 539–548. https://doi.org/10.1016/j.jcmgh.2018.01.023.
49. Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416. https://doi.org/10.1038/s41467-019-13056-x.
50. France, S.L., and Akkucuk, U. (2021). A review, framework, and R toolkit for exploring, evaluating, and comparing visualization methods. Vis. Comput. *37*, 457–475. https://doi.org/10.1007/s00371-020-01817-5.
51. Poličar, P., Stražar, M., and Zupan, B. (2019). openTSNE: A modular Python library for t-SNE dimensionality reduction and embedding. bioRxiv, 1–2. https://doi.org/10.1101/731877.
52. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., and Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nat. Methods *16*, 243–245. https://doi.org/10.1038/s41592-018-0308-4.
53. Bodt, C.D., Mulders, D., Verleysen, M., and Lee, J.A. (2018). Perplexity-free t -SNE and twice student tt -SNE. In European Symposium on Artificial Neural Networks (Bruges).
54. Aliverti, E., Tilson, J.L., Filer, D.L., Babcock, B., Colaneri, A., Ocasio, J., Gershon, T.R., Wilhelmsen, K.C., and Dunson, D.B. (2020). Projected t-SNE for batch correction. Bioinformatics *36*, 3522–3527. https://doi.org/10.1093/bioinformatics/btaa189.
55. Häkkinen, A., Koiranen, J., Casado, J., Kaipio, K., Lehtonen, O., Petrucci, E., Hynninen, J., Hietanen, S., Carpén, O., Pasquini, L., et al. (2020). qSNE: quadratic rate t-SNE optimizer with automatic parameter tuning for large datasets. Bioinformatics *36*, 5086–5092. https://doi.org/10.1093/bioinformatics/btaa637.
56. Belkina, A.C., Ciccolella, C.O., Anno, R., Halpert, R., Spidlen, J., and Snyder-Cappione, J.E. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nat. Commun. *10*, 5415. https://doi.org/10.1038/s41467-019-13055-y.
57. Lee, J.A., Peluffo-Ordóñez, D.H., and Verleysen, M. (2015). Multi-scale similarities in stochastic neighbour embedding: reducing dimensionality while preserving both local and global structure. Neurocomputing *169*, 246–261. https://doi.org/10.1016/j.neucom.2014.12.095.
58. Lee, J.A., and Verleysen, M. (2009). Quality assessment of dimensionality reduction: rank-based criteria. Neurocomputing *72*, 1431–1443. https://doi.org/10.1016/j.neucom.2008.12.017.
59. Venna, J., and Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: an experimental study. In Lecture Notes in Computer Science (including subseries Lecture notes in artificial Intelligence and Lecture notes in Bioinformatics), G. Dorffner, H. Bischof, and K. Hornik, eds. (Springer), pp. 485–491. https://doi.org/10.1007/3-540-44668-0_68.
60. France, S., and Carroll, D. (2007). Development of an agreement metric based upon the RAND index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. In Machine Learning and Data Mining in Pattern Recognition, *4571* (Springer), pp. 499–517. https://doi.org/10.1007/978-3-540-73499-4_38.
61. Lee, J.A., and Verleysen, M. (2008). Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods. In Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008 (PMLR, Antwerp, Belgium), Proceedings of Machine Learning Research, *4*, Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y.V. de Pee, eds., *pp. 21–35*.
62. Goldberg, Y., and Ritov, Y. (2009). Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. Mach. Learn. *77*, 1–25. https://doi.org/10.1007/s10994-009-5107-9.
63. Meng, D., Leung, Y., and Xu, Z. (2011). A new quality assessment criterion for nonlinear dimensionality reduction. Neurocomputing *74*, 941–948. https://doi.org/10.1016/j.neucom.2010.10.011.
64. Paul, R., and Chalup, S.K. (2017). A study on validating non-linear dimensionality reduction using persistent homology. Pattern Recognition Lett. *100*, 160–166. https://doi.org/10.1016/j.patrec.2017.09.032.
65. Heiser, C.N., and Lau, K.S. (2020). A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. Cell Rep. *31*, 107576. https://doi.org/10.1016/j.celrep.2020.107576.
66. Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P., and Castrén, E. (2003). Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics *4*, 48. https://doi.org/10.1186/1471-2105-4-48.
67. Lespinats, S., and Aupetit, M. (2011). CheckViz: sanity check and topological clues for linear and non-linear mappings. Comput. Graph. Forum *30*, 113–125. https://doi.org/10.1111/j.1467-8659.2010.01835.x.
68. Schreck, T., von Landesberger, T., and Bremm, S. (2010). Techniques for precision-based visual analysis of projected data. In Visualization and Data Analysis 2010, J. Park, M.C. Hao, P.C. Wong, and C. Chen, eds., p. 75300E. https://doi.org/10.1117/12.838720.

69. Martins, R.M., Minghim, R., and Telea, A.C. (2015). Explaining neighborhood preservation for multidimensional projections. In Computer Graphics and Visual Computing (CGVC), R. Borgo and C. Turkay, eds. (The Eurographics Association), pp. 7–14. https://doi.org/10.2312/cgvc.20151234.

70. Rieck, B., and Leitte, H. (2015). Persistent homology for the evaluation of dimensionality reduction schemes. Comput. Graph. Forum *34*, 431–440. https://doi.org/10.1111/cgf.12655.

71. Rieck, B., and Leitte, H. (2017). Agreement analysis of quality measures for dimensionality reduction. In Topological Methods in Data Analysis and Visualization IV, H. Carr, C. Garth, and T. Weinkauf, eds. (Springer), pp. 103–117. https://doi.org/10.1007/978-3-319-44684-4_6.

72. Martins, R.M., Coimbra, D.B., Minghim, R., and Telea, A. (2014). Visual analysis of dimensionality reduction quality for parameterized projections. Comput. Graph. *41*, 26–42. https://doi.org/10.1016/j.cag.2014.01.006.

73. Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. Ann. Math. Stat. *22*, 79–86. https://doi.org/10.1214/aoms/1177729694.

74. Lee, J.A., Renard, E., Bernard, G., Dupont, P., and Verleysen, M. (2013). Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing *112*, 92–108. https://doi.org/10.1016/j.neucom.2012.12.036.

75. Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774–786. https://doi.org/10.1016/j.cell.2009.07.038.

76. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., and Stanley, H.E. (2002). Random matrix approach to cross correlations in financial data. Phys. Rev. E *65*, 066126. https://doi.org/10.1103/PhysRevE.65.066126.

77. Aparicio, L., Bordyuh, M., Blumberg, A.J., and Rabadan, R. (2020). A random matrix theory approach to denoise single-cell data. Patterns *1*, 100035. https://doi.org/10.1016/j.patter.2020.100035.

78. Dobriban, E. (2020). Permutation methods for factor analysis and PCA. Ann. Stat. *48*, 2824–2847. https://doi.org/10.1214/19-AOS1907.

79. Loughin, T.M. (2004). A systematic comparison of methods for combining p-values from independent tests. Comput. Stat. Data Anal. *47*, 467–485. https://doi.org/10.1016/j.csda.2003.11.020.

80. Heard, N.A., and Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. Biometrika *105*, 239–246. https://doi.org/10.1093/biomet/asx076.

81. Gisbrecht, A., and Hammer, B. (2015). Data visualization by nonlinear dimensionality reduction. Wiley Interdiscip. Rev. Data Mining Knowl. Discov. *5*, 51–73. https://doi.org/10.1002/widm.1147.

82. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., and Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans. Database Syst. *42*, 1–21. https://doi.org/10.1145/3068335.

83. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.

84. Chan, D.M., Rao, R., Huang, F., and Canny, J.F. (2018). t-SNE-CUDA: GPU-accelerated t-SNE and its applications to modern data. In 2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) (IEEE), pp. 330–338. https://doi.org/10.1109/CAHPC.2018.8645912.

85. Agrawal, A., Ali, A., and Boyd, S. (2021). Minimum-distortion embedding. Found. Trends® Mach. Learn. *14*, 211–378. https://doi.org/10.1561/2200000090.

86. de Bodt, C., Mulders, D., Verleysen, M., and Lee, J.A. (2020). Fast multiscale neighbor embedding. IEEE Trans. Neural Netw. Learn. Syst. 1–15. https://doi.org/10.1109/TNNLS.2020.3042807.

87. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.