# SCHEMA: A general framework for integrating heterogeneous single-cell modalities

Rohit Singh[1,*,†], Ashwin Narayan[2,*], Brian Hie[1], Bonnie Berger[1,2,†]
[1]Computer Science and Artificial Intelligence Lab., MIT, Cambridge, MA 02139 USA
[2]Department of Mathematics, MIT, Cambridge, MA 02139 USA

**Abstract**

Advances in single-cell technologies now allow researchers to simultaneously measure diverse biological data modalities, including transcriptomic, proteomic, or spatial features. Sophisticated cross-modality analysis promises unprecedented and more complete insight into complex biology. Unfortunately, current methods for single-cell RNA-seq data integration focus on aligning patterns across multiple experiments and are not designed for the setting in which a single experiment simultaneously measures multiple modalities. We therefore propose SCHEMA, a general, scalable, and flexible framework for integrating diverse modalities profiled within a single experimental setting. SCHEMA has intuitive parameters that facilitate exploratory analysis and complements standard techniques like PCA and CCA (canonical correlation analysis). We first process each modality to compute a similarity (equivalently, distance) measure between cells. SCHEMA is a framework for calibrating the agreement between these modality-specific distance measures. Given a primary dataset (i.e., modality) and an arbitrary number of secondary datasets, we search for an affine transformation of the primary dataset so that distances in the transformed space have the maximum possible agreement with the secondary datasets. Crucially, the user can calibrate this by specifying a limit on how much the primary dataset can be distorted; methods like CCA lack this control. Mathematically, we quantify the concept of agreement between datasets as the correlation between their respective sets of pairwise squared distances and offer some theoretical justification for this design choice. Our optimization problem can be formulated as a quadratic program (QP), allowing for a fast algorithm with intuitive free parameters and an appealing feature-weighting interpretation. Using SCHEMA, we integrated spatial and transcriptomic modalities from the Slide-Seq experiment [1] and identified a gene-set in granule cells whose expression covaries with the spatial density of the cells. We investigated a multi-modal dataset from 10x Genomics [2], analyzing the selection pressure on residues in the CDR3 region of T-cell receptors. We also used SCHEMA to improve UMAP & tSNE visualizations by infusing metadata into RNA-seq information. As the number and diversity of multi-modal datasets grows, SCHEMA will be an essential component in the single-cell analyst's arsenal.

**Availability:** Our implementation is freely available as the Python package `schema_learn` and code is available at `http://schema.csail.mit.edu`.

*These authors contributed equally to this work.
†Corresponding authors: rsingh@csail.mit.edu, bab@mit.edu.

# 1 Introduction

The rapid emergence of biological studies performing high-throughput profiling of single cells has grown from an initial focus on transcriptomics, via single-cell RNA-sequencing (scRNA-seq), to cover many biological measurement modalities. These modalities include genomic features such as mutational differences [3] and V(D)J clonotypes [2]; epigenomic features such as chromatin accessibility [4], methylation [5], and histone marks [6]; proteomic and functional features such as surface marker expression [7] and T-cell receptor affinities [2]; and spatial features [1]. Moreover, single-cell experiments increasingly *profile more than one modality at the same time* [1, 4, 5, 7], enabling multi-faceted insight into biological processes at single-cell resolution. In one notable example, researchers simultaneously profiled gene expression, surface protein expression, T-cell V(D)J clonotypes, and T-cell receptor binding affinities against a panel of fifty ligands [2].

As these techniques mature and costs decline, we expect that multi-modal datasets will become ubiquitous, requiring sophisticated computational analysis. Unfortunately, state-of-the-art methods that integrate scRNA-seq data across multiple experiments are not well-suited to multi-modal analysis. While *multi-experiment* integration has been an active area of research, including methods like Scanorama [8], mnnCorrect [9], Harmony [10], Conos [11], and scmap [12], the problem of integration across heterogeneous modalities from a *single* experiment has not received similar attention. Although recently published methods such as Seurat v3 [13] and LIGER [14] do aim to incorporate multiple modalities, they still follow the multiple-experiment paradigm, using standard integration methods to align experiments that separately profile different modalities.

Simultaneous per-cell multi-modal measurement within a single experiment eliminates the need for computationally integrating those modalities across distinct cells from multiple experiments. Instead, the challenge that arises is to enable deeper analysis of multi-modal data. For example, how should one interpret the data if two cells look similar transcriptionally but are different epigenetically?

We therefore develop a new computational approach that generalizes to any heterogeneous multi-modal data and in particular single-cell. Although, for example, combining spatial and transcriptomic data may ostensibly seem different from a joint analysis of chromatin accessibility and proteomic data, there exist intrinsic similarities among all multi-modal analyses. First, we re-interpret each modality as describing a measure of distance between the underlying cells. The integration problem is then newly framed as reconciling the information implied by these distance measures in a researcher-specified way. Our approach is fully general: regardless of the modalities involved, we aim to support the continually-expanding breadth of single-cell profiling technologies while retaining the power, tunability and interpretability required for effective exploratory analysis.

We present Schema, the first general method for integrating multi-modal data with a specific focus on multi-modal single-cell data. It is a fast algorithm, using an efficient quadratic programming (QP) framework to learn a linear transform of the original data. Schema can accommodate multiple types of input data including numerical, categorical, or multi-dimensional. It can also be used as a feature-weighting mechanism on top of standard techniques like principal component analysis (PCA) and canonical correlation analysis (CCA). To facilitate exploration, the algorithm has intuitive parameters that allow the user to control how much to trust one dataset relative to others, thus providing flexibility. Importantly, integration need not mean agreement. Schema allows users to specify that a certain modality should be *dis*agreed with. We note that even data from single-modality single-cell experiments could benefit from Schema since the metadata associated with observations (e.g., cell type labels or batch number) can be leveraged to better interpret the core

data modality (e.g., single-cell transcriptomes.) While SCHEMA has features that are well suited to single-cell data, it is a general approach and can be used in other contexts.

To demonstrate how using SCHEMA can reveal new biological insight, we first analyzed experimental SlideSeq data [1] that places transcriptional profiles within a two-dimensional location in a tissue sample. SCHEMA identified a set of genes in granule cells in the mouse cerebellum whose expression levels are correlated with the cell's spatial density. We show that the genes that are highly expressed only in densely packed granule cells are significantly enriched for metabolism and neurotransmission, suggesting that these cells are also more active.

Moreover, to demonstrate the generality of our method to a diverse set of modalities, we investigated how variations in a T-cell receptor's CDR3 sequence influence binding specificity [2]. State-of-the-art methods assume that, even in this hypervariable region, the residue-selection pressure is consistent with a conventional BLOSUM matrix [15, 16]. To test this assumption, we used SCHEMA to rank amino acids in this region by their importance in conferring binding specificity. Our results indicate general concurrence with BLOSUM scores but also highlight key differences.

We also describe how SCHEMA can be used to make single-cell visualizations more informative. Nonlinear visualization methods such as UMAP [17] and tSNE [18] are relatively inflexible and do not allow researchers to specify which aspects of the underlying structure to highlight. We show how to incorporate metadata to make visualization more informative and flexible, which we demonstrate by highlighting the temporal structure within a dataset of aging *Drosophila melanogaster* neurons.

We discuss how our work relates to standard statistical and machine learning methods. Popular linear decomposition methods (e.g., PCA or CCA) provide limited flexibility when calibrating the agreement between datasets, which our method is designed to address. Issues involving limited flexibility, as well as lack of scalability, also arise for state-of-the-art methods in metric learning, the subfield of machine learning dedicated to learning distance functions. SCHEMA also takes a metric learning approach and, in this context, our choice of the objective function and the QP forumulation are novel contributions that enable *fast*, *tunable* and *interpretable* transformations. SCHEMA is also amenable to sketching-based acceleration [19], for which we provide theoretical guarantees and which is crucial for analyzing large-scale studies.

## 2 Intuition

To begin with, suppose we have $N$ observations; in a single-cell setting these would correspond to cells. Next, for each observation we have multiple types (i.e., modalities) of data. For example, dataset $D_1$ may record transcriptional information (represented as a $N \times k$ matrix, $k$ being the number of genes), dataset $D_2$ may have the 2-dimensional location of the cell in a tissue slice (a $N \times 2$ matrix), and dataset $D_3$ might be cell-line information (a $N \times 1$ vector of labels). If these datasets all represent views of the same underlying biology, they should be in some kind of agreement. But noise, experimental artifacts and, importantly, unknown biological factors make this hard to discern. Our approach to the heterogeneous integration task is to produce a new dataset $D^*$ that combines the information from $D_1, D_2, D_3$ and is in some agreement with each of them.

To crystallize this intuition of "agreement", we build upon an idea common to many machine learning techniques and single-cell analyses [17, 18, 20]: analyze the data exclusively in terms of distances between points in the data. Biologically, this is well justified. For example, cells with similar expression profiles typically belong to the same cell group/cluster.

We say that two datasets agree if they have a similar *neighborhood structure*. Formally, we view

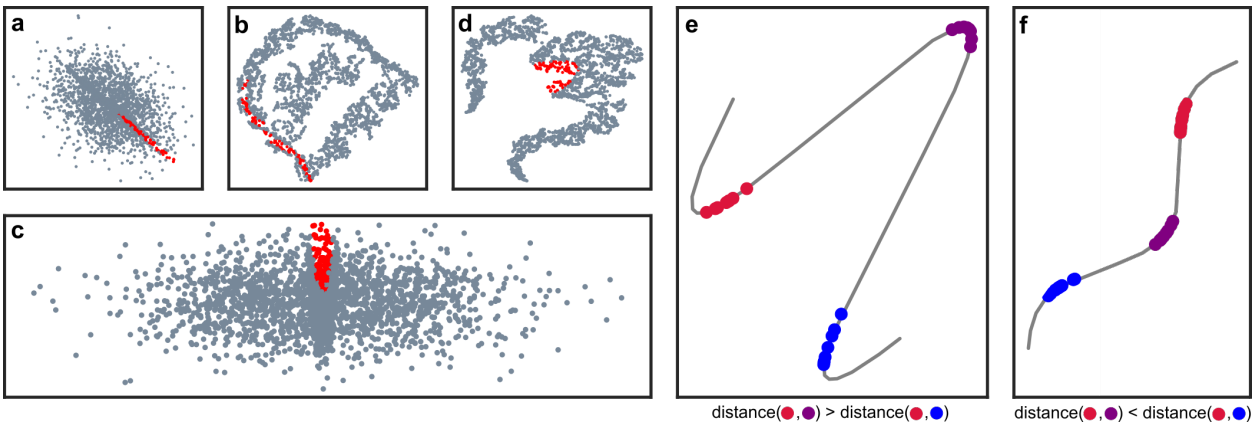


Figure 1: **Demonstration of Schema on a toy dataset.** The original dataset (**a**) has a small set of points that are differently colored (red). An analogy here would be to rare cell-types in a single-cell dataset. There is enough noise that a tSNE plot of the raw data (**b**) fails to clearly separate the red points. Running Schema produces (**c**), with the data now rotated and scaled to accentuate the red points. Note how the tSNE plot for this transformed data (**d**) clearly separates the red points from the rest. **(e,f) Manifold-reshaping interpretation of Schema.** We depict three clusters of cells $R$ (red), $P$ (purple), and $B$ (blue) in a high-dimensional space (**e**). The cells actually lie on a manifold (the gray curve) and Euclidean distance on the manifold is misleading, since it makes it appear that $R$ is closer to $B$ than to $P$, when, on the manifold, $P$ should be between $R$ and $B$. Schema can transform the data so that Euclidean distance is more reasonable. If we have some secondary data that tells us that distance$(B, P)$ and distance$(R, P)$ are less than distance$(B, R)$, we can rotate the dataset and scale the axes so that the manifold distance agrees better with Euclidean distance.

each dataset as describing the same $N$ points in two different spaces and *define* that they have identical *neighborhood structures* if for any point $P$ and any $k$ ($1 \leq k \leq N-1$), the $k$-nearest neighbors of $P$ are the same in both datasets. In other words, a nearest-neighbor algorithm can not distinguish between the two datasets. Note that the distances need not be Euclidean, a feature we will exploit later. Now consider the set of $\binom{N}{2} = \frac{N(N-1)}{2}$ pairwise distances in each dataset. As we elaborate in Appendix B, there is a strong connection between neighborhood-structure similarity and the Spearman rank correlation of the two sets of pairwise distances. In particular, the correlation is 1 if and only if the neighborhood structures are identical. Drawing upon this property, we say the two datasets are in agreement if pairwise distances in one are highly correlated to the corresponding distances in the other.[1]

Our goal is to combine the information from the input datasets ($D_1, D_2, D_3$ here) into a single dataset $D^*$ that is in agreement with each of them. First, we designate one dataset (say, $D_1$ of shape $N \times k$) as the primary and the remaining datasets (here, $D_2$ and $D_3$) as secondary.[2] Specifically, we pose the following constrained optimization problem: find an embedding $D^*$ (of shape $N \times k$) that is a linear transformation of $D_1$, where the pairwise squared Euclidean distances in $D^*$ are i) maximally correlated to the corresponding distances in $D_2$ and $D_3$; and ii) correlated to the corresponding distances in the primary dataset $D_1$ above some user-specified threshold. By tuning the threshold, the user can trade off distortion from the primary dataset $D_1$ against agreement with

[1]For computational tractability, we approximate the Spearman correlation with the Pearson correlation. The difference between the two is that the former works with ranks, while the latter works with raw values.

[2]Choosing which dataset should be primary is sometimes obvious (e.g., when using meta-data). Typically, the dataset where short-range Euclidean distances would be most meaningful should be chosen (see the discussion related to manifold reshaping in Appendix B). This is not a 'reference' dataset in the sense of Stuart et al. [13]; here, the researcher is explicitly allowed to distort the primary dataset.

the secondary data in $D_2$ and $D_3$.

Introducing this constraint while keeping the optimization problem feasible is a key conceptual contribution of our paper: methods like CCA, as well as standard approaches in metric learning, place no such limits on the distortion of $D_1$. Biologically, that is deeply problematic because it de-emphasizes information in $D_1$ that does not co-vary with $D_2$ and $D_3$.

We now describe our solution to this optimization problem. Any affine transform can be broken into three components: translation, scaling (i.e., stretch or shrink the axes), and rotation. Distances are invariant with respect to translation, so only the latter two are of interest here. *The key algorithmic contribution of this paper is a Quadratic Programming (QP) formulation that computes the optimal scaling transform for the optimization problem above.*

We pair this scaling transform with a rotation produced by a change-of-basis technique, thus producing an affine transform. In particular, we use methods like principal component analysis (PCA) or non-negative matrix factorization (NMF) for this change of basis. We have found this to be surprisingly effective: doing a PCA and NMF rotates the data so that dimensions with high variance or information become axis-aligned. The QP-computed scaling transform then acts as a feature selection mechanism on top of this, identifying which axes are the most useful in maximizing agreement between the datasets.

**Runtime and features:** The QP formulation is fast, flexible and tunable: most single-cell datasets can be processed in a few minutes, arbitrarily many datasets can be handled, and there are intuitive parameters that let the user control the trade-offs involved. The user may specify non-Euclidean distance measures on $D_1, D_2, D_3$; however, distances in $D^*$ need to remain Euclidean. She may require positive correlation with some secondary datasets and negative with others. Also, relative weights for $D_2$ and $D_3$ can be specified.

## 3 Mathematical Formulation

Suppose we have $N$ observations across $r$ datasets $D_j$, $1 \leq j \leq r$, where $D_j = \{x_i^{(j)}, 1 \leq i \leq N\}$ contains data (categorical or continuous) for each observation. We will refer to $D_1$ as the *primary* dataset and the rest as secondary. Each dataset's dimensionality and domain may vary. In particular, we assume $D_1$ is $k$-dimensional. Each dataset $D_j$ should also have some notion of distance between observations attached to it, which we will denote $\rho_j$, so $\rho_j(x_n^{(j)}, x_m^{(j)})$ is the distance between observations $n$ and $m$ in $D_j$. Actually, since our entire framework below deals in *squared* distances, for notational convenience we will let $\rho_j$ be the *squared* distances between points in $D_j$; also, we drop the superscript in $x_j^{(1)}$ when referring to the primary dataset $D_1$ and its data.

The goal is to find a transformation $\Omega$ such that $\Omega(D)$ generates a dataset $D^*$ such that the Euclidean metric $\rho^*$ on $D^*$ "mediates" between the various metrics $\rho_j$, each informed by its respectively modality. Note that none of the $\rho_j$ need to be Euclidean. The above setup is quite general, and we now specify the form of the transformation $\Omega$ and the criteria for balancing information from the various metrics. Here, we limit $\Omega$ to a *scaling transform*. That is, $\Omega(D) = \{\mathrm{diag}(\omega)x \,|\, x \in D\}$ for some $\omega \in \mathbb{R}^k$ ($\mathrm{diag}(\omega)$ is a $k \times k$ diagonal matrix with $\omega$ as its diagonal entries). The scaling transform $\omega$ acts as a feature-weighting mechanism: it chooses the *features* of $D_1$ that align the datasets best (i.e. $\omega_i$ being large means that the $i$th coordinate of $D_1$ is important). We note here that a natural extension would be allowing *general linear* transformations for $\Omega$; however, in that context, the fast framework of quadratic programming would need to be substituted for the much

slower framework of semidefinite programming.

Here, our approach to integration between the metrics $\rho_j$ is to learn a metric $\rho^*$ that aligns well with all of them. Our measure of the alignment between $\rho^*$ and $\rho_j$ is given by the Pearson correlation between pairwise squared distances under two metrics. Intuitively, maximizing the correlation coefficient encourages distances under $\rho^*$ to be large when the corresponding $\rho_j$ distances are large and vice versa. This can be seen from the formula:

$$\mathrm{Corr}(\rho^*, \rho_j) = \frac{\mathrm{Cov}\left[\rho^*, \rho_j\right]}{\left(\mathrm{Var}\left[\rho^*\right] \mathrm{Var}\left[\rho_j\right]\right)^{1/2}} \tag{1}$$

To deal with multiple modalities, we try to maximize the correlation between $\rho^*$ and the distances on each of the metrics, allowing the user to specify how much each modality should be weighted. We also allow hard constraints, whereby the correlation between the transformed data and some $D_j$ has to be at least some value. Our goal is thus to find:

$$\begin{aligned} &\max_{\omega \in \mathbb{R}^k} \sum_{j=1}^{r} \gamma_j \mathrm{Corr}(\rho^*(\omega), \rho_j) \\ \text{subject to} \quad &\mathrm{Corr}(\rho^*(\omega), \rho_j) \geq \phi_j \text{ for } j \in \{1, \ldots, r\} \end{aligned} \tag{2}$$

where $\gamma_j$ and $\phi_j$ are hyperparameters that determine the importance of the various metrics. We have also highlighted that $\rho^*$ is a function of $\omega$ and is determined entirely by the solution to (2). In the rest of our discussion, we will primarily refer to $\omega$, rather than $\rho^*$.

In order to make this optimization feasible, we use the machinery of *quadratic programming*.

**Setting up the quadratic program:** Quadratic programming (QP) is a framework for constrained convex optimization problems that allows a quadratic term in the objective function and linear constraints. The general form is:

$$\begin{aligned} &\min_{v \in \mathbb{R}^s} v^T Q v + q^T v \\ \text{subject to} \quad & \\ &Gv \preceq h \\ &Av = b \end{aligned} \tag{3}$$

where $Q$ is a positive semidefinite (psd) matrix, and the notation $y \preceq z$ means the inequality is true for each coordinate (i.e., $y_i \leq z_i$ for all $i$).

To put our optimization (2) in a QP formulation, we expand the covariance and variance terms in (1), and show that the covariance is *linear* in the transformation and variance is *quadratic*:

$$\mathrm{Cov}(w, \rho_\ell) = \left(\tfrac{1}{|P|} a_\ell - \tfrac{1}{|P|^2} b_\ell\right) w \qquad \mathrm{Var}(w) = w^T \left(\tfrac{1}{|P|} S - \tfrac{1}{|P|^2} T\right) w \tag{4}$$

where $a_\ell$ and $b_\ell$ are $k$-dimensional vectors that depend only on $D_\ell$; and $S$ and $T$ are $N \times k$ matrices that depend only on $D_1$; and $P$ is the set of pairs of observations. It is also not hard to show that $(|P|^{-1} S - |P|^{-2} T)$ is psd, as required. For details of the derivation, see Appendix A.

There is one more difficulty to address. The correlation is the *quotient* of the covariance and the standard deviation, and the QP framework cannot handle quotients or square roots. However,

maximizing a quotient can be reframed as maximizing the numerator (the covariance), minimizing the denominator (the variance), or both.

We now have the ingredients for the QP and can frame the optimization problem as:

$$\max_{w \in \mathbb{R}^k} \sum_{j=1}^{r} \gamma_r \mathrm{Cov}(w, \rho_j^2) - \alpha \mathrm{Var}(\rho^*) - \lambda ||w - \mathbf{1}||^2 \tag{5}$$

subject to

$$\mathrm{Cov}(w, \rho_j) \geq \beta_j \text{ for } 1 \leq j \leq r$$
$$w \succeq \mathbf{0}$$

where $\mathbf{0}$ and $\mathbf{1}$ are the all-zeros and all-ones vectors (of the appropriate length) respectively. Here, $\lambda$ is the hyperparameter for regularization of $w$, which we want to penalize for being too far away from the all-ones vector (i.e. equal weighting of all the features). One could also regularize the $L_2$ norm of $w$ alone (i.e. incorporate $-\lambda||w||^2$) which would encourage $w$ to be small; we have found that empirically the choices yield similar results.

This program can be solved by standard QP solvers (see Appendix A for the full details of how to put the above program in canonical form for a solver), and the solution $w^*$ can be used to transform unseen input data, using $\omega^* \in \mathbb{R}^k$, where $\omega_i^* = \sqrt{w_i^*}$.

**Hyperparameters:** A well-known challenge for machine learning algorithms is interpretability of hyperparameters. Here, the QP solver needs values for $\lambda$, $\alpha$, and $\beta$, and specifying these in a principled way is a challenge for users. Our approach is thus to allow the user to specify more natural parameters. Specifically, we allow the user to specify minimum correlations between the pairwise distances in $D^*$ and each of the $D_i$; and also the ratio of the largest value of $w$ to its average value, Formally, the user can specify $s_i$ and $\bar{w}$ such that:

$$\mathrm{Corr}(\rho^*, \rho_i) \geq s_i \text{ for } 1 \leq i \leq r$$
$$\frac{||w||_\infty}{||w||_1} \leq \frac{\bar{w}}{k} \tag{6}$$

While these quantities are not directly optimizable in our QP formulation (5), we can access them by varying the hyperparameters $\alpha, \beta, \lambda$. We note that SCHEMA currently supports the constraint $s_i$ only for $i = 1$, i.e., the primary dataset. In future, we intend to support all $s_i$.

Intuitively, we note that the choice of $\lambda$ controls whether $w$ satisfies $\bar{w}$; and $\alpha$ and $\beta$ control whether the correlation constraints $s_i$ are satisfied. To satisfy these constraints, we simply *grid search* across feasible values of $\{\alpha, \beta, \lambda\}$: we solve the QP for fixed values of $\alpha, \beta, \lambda$, keeping only the solutions for which the $\{s_i, \bar{w}\}$ constraints are satisfied. Of these, we choose the most optimal. The efficiency of quadratic programming means that such a grid search is feasible, which gives users the benefit of easily interpretable and natural hyperparameters.

**Preprocessing transforms:** Standard linear decompositions, like PCA or NMF are useful as preprocessing steps for SCHEMA. PCA is a good choice in this regard because it decomposes along directions of high variance; NMF is slower, but has the advantage that it is designed for data that is non-negative (e.g., transcript counts). The transform $\omega$ that we generate can be interpreted as a *feature-weighting* mechanism, identifying the directions (in PCA) or factors (in NMF) most relevant to aligning the datasets. The user can also employ a feature-set that is a union of features from two methods (e.g., PCA & CCA).

**Motivating the choice of correlation as an objective:** As a measure of the alignment between our transformation and a dataset, correlation of pairwise distances is a flexible and robust measure. An expanded version of arguments in this paragraph is available in the appendices. Given a pair of dataset, the connection between their pairwise-distance Spearman rank correlation and the neighborhood-structure similarity is deep: if the correlation is greater than $1 - \epsilon$, the fraction of misaligned neighborhood-relationships will be less than $O(\sqrt{\epsilon})$. There is a *manifold* interpretation that is also compelling: assuming the high-dimensional data lie on a low-dimensional manifold, small distances are more accurate than large distances, so the *local* neighborhood structure is worth preserving. We can show intuitively that optimizing the correlation aims to preserve local neighborhood structure (Appendix B). Using correlation in the objective also affords the flexibility to broaden $\mathrm{Corr}(w, \rho_j)$ in (2) to any function $f_j$ of the metric: i.e. $\mathrm{Corr}(w, f_j \circ \rho_j)$; this allows us to invert the direction of alignment or more heavily weigh local distances. As scRNA-seq dataset sizes reach millions of cells, even calculating the $O(N^2)$ pairwise distances becomes infeasible. In this case, we sample a subset of the pairwise distances. As an estimator, sample-correlation is a robust measure. This allows SCHEMA to perform well even with relatively small subsets; in fact, we only need a sample-size *logarithmic* in our desired confidence level to generate high-confidence results (Appendix C). This enables SCHEMA to continue scaling to more massive scRNA-seq datasets.

## 4 Connections to Existing Approaches

**Linear Decomposition Techniques**: Like popular linear decomposition techniques (PCA, SVD, LDA, PLS, and CCA)[3], SCHEMA runs quickly and performs only linear transformations of the data. However, there is a gap between methods like PCA and SVD, which explain the variance in a *single* dataset, and those like LDA, PLS, and CCA, which focus on explaining the covariance *between* datasets. The former cannot integrate information across datasets; the latter de-emphasize structure within a dataset. The assumption underlying SCHEMA is that both the variance *within* a dataset and covariance *across* datasets carry important biological information. Specifically, our focus on the importance of preserving *neighborhood structure* in the primary dataset explicitly emphasizes the within-dataset relationships we care about.

SCHEMA effectively complements these existing decomposition techniques. As discussed earlier, one can use a change-of-basis transform as a preprocessing step, so that the QP-derived scaling transform SCHEMA generates puts weights on the transformed features in a way that preserves neighborhood structure. This allows, for example, principled dimensionality reduction by choosing the heaviest weighted features instead of just an arbitrary number of principal components.

**Metric Learning:** Our approach falls under the framework of *metric learning*, where the goal is to learn an appropriate distance measure on data. In essence, the transform $\omega$ we produce induces a metric on the primary dataset that effectively integrates information across the heterogeneous modalities. Most downstream machine learning techniques yield better results with an accurate distance measure on the underlying data.

We pursue the same intuition that existing work in the metric learning [21–24] leverages: that additional information can be useful to transform our primary dataset. As with these methods, we formulate an optimization problem that searches for an affine transformation of the input data.

[3]PCA: principal component analysis, SVD: singular value decomposition, CCA: canonical correlation analysis, PLS: partial least squares, LDA: linear discriminant analysis

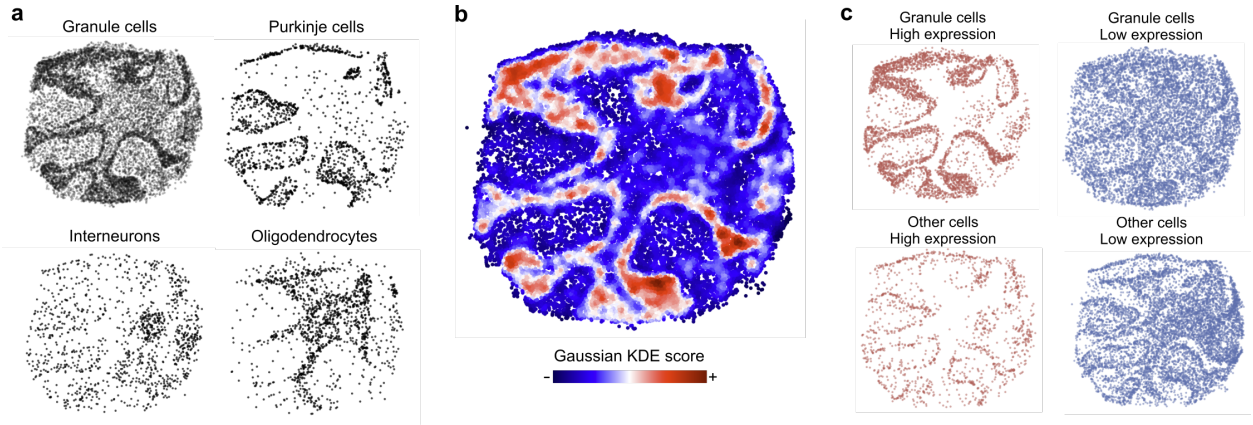


Figure 2: **A subset of genes in granule cells has expression that covaries with spatial cellular density.** All figures correspond to a mouse cerebellum tissue sample from Rodriques et al. [1]; in each, a dot corresponds to a cell in the corresponding $(x, y)$ position on the puck (glass slide with tissue slice.) (**a**) Spatial distribution of the most common cell-types in the tissue section: granule cells, Purkinje cells, interneurons, and oligodendrocytes. Note the variation in spatial density, especially noticeable for granule cells. (**b**) We quantified this spatial density variation by computing a two-dimensional Gaussian kernel density estimate over the granule cells. Locations more densely surrounded by granule cells are assigned a higher score. (**c**) Schema is able to identify a set of genes that are highly expressed only in densely-packed granule cells. The four figures here show mutually disjoint sets of cells: granule cells with high expression of the gene set, granule cells with low expression of the gene set, other cells with high expression, and other cells with low expression. Here, a cell is said to have high expression of the gene set if the cell's loading on this gene-set ranks in the top quartile.

However, our approach is novel in its speed and generality. By restricting to a scaling transform, we obtain a fast QP formulation with intuitive hyperparameters, which can easily handle arbitrarily many heterogeneous modalities and large single-cell datasets. In contrast, two popular metric learning methods, LMNN [23] & ITML [24], failed to converge within 4 hours on a medium-sized dataset (23,822 cells $\times$ 3,000 genes). Additionally, previous methods provide no constraints on the distortion of the primary dataset. Since the primary dataset itself contains important information, we contend that these distortion constraints are crucial and biologically motivated.

Our choice of correlation as our objective function is also novel. Most of the above methods instead use some form of classification error or Kullback-Leibler divergence. In addition to being easier to generalize to multiple heterogeneous modalities, requiring high correlation helps preserve neighborhood structure in the modalities, which is useful especially in a biological context.

# 5 Results

We show the results of applying Schema in three distinct use-cases, spanning a diverse set of modalities that include spatial, transcriptomic, proteomic, functional, and temporal information.

**Spatial and transcriptomic modalities.** We investigated Slide-Seq data from Rodriques et al. [1], in which the researchers augmented a scRNA-seq experimental setup with spatial barcodes, allowing for a cell's transcriptome to also be spatially located within a tissue of interest. A key contribution of the paper was using transcriptomic information to characterize the spatial layout of cell types within mouse brain tissue. Using this spatial information, we noted that granule cells seem to have regions of both low and high spatial density (Fig 2 in this paper; also Fig 2B of [1]).

To further probe the biological significance of cellular density, we applied Schema to relate

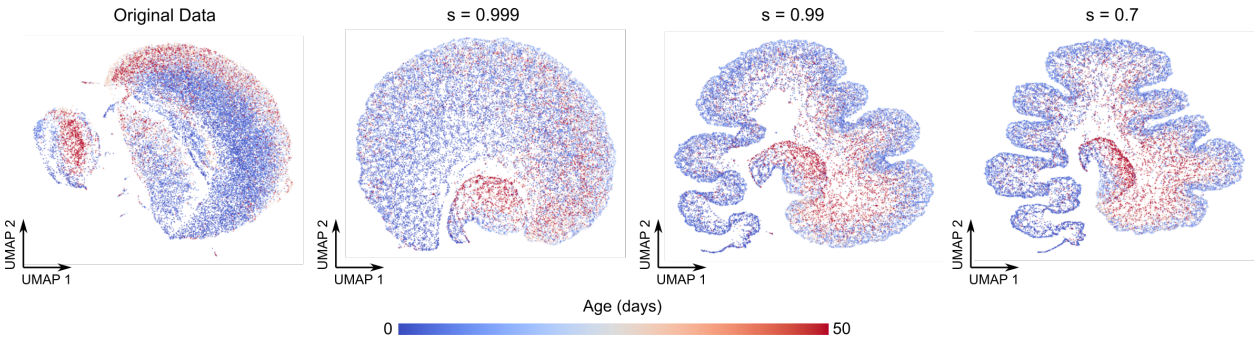


Figure 3: **Incorporating temporal meta-data into UMAP visualizations of aging neurons.** UMAP visualization of scRNA-seq profiles of *D. melanogaster* neurons at 0, 1, 3, 6, 9, 15, 30, and 50 days after birth, representing the full range of a typical *D. melanogaster* lifespan. The transcriptomic data was transformed using SCHEMA to reflect the temporal metadata associated with each cell and then visualized using UMAP. Decreasing the minimum correlation constraint $s$ results in a UMAP structure that increasingly reflects an age-related trajectory.

spatial density variation to differential gene expression. Specifically, we aimed to identify genes that are highly expressed primarily in densely-packed granule cells. To do so, we learned a two-dimensional Gaussian-kernel density function on granule cell location that assigns higher scores to regions with denser cell packing. We then ran SCHEMA using the gene expression matrix as the *primary dataset*, on which our *change-of-basis transform* was based on top-100 NMF factors of the primary dataset. Our *secondary datasets* consisted of the numeric kernel-density scores, as well as the categorical labels corresponding to the four most common non-granule cell types. We required positive correlation with cell density, negative correlation with non-granule cell type labels, and high correlation with granule cell type labels in the optimal transformation. SCHEMA then identified a gene-set that is highly expressed primarily in densely-packed granule cells (Figure 2).

The density-related genes that SCHEMA identifies are strongly enriched for processes involving metabolism and neurotransmitter signaling (Figure S1), including gene ontology (GO) terms related to cellular respiration-related electron transport (`GO:0022904`, FDR $q = 1.07 \times 10^{-21}$) and neurotransmitter secretion (`GO:0007269`, FDR $q = 6.23 \times 10^{-7}$). Likewise, enriched GO cellular components involve the mitochondrion and the synapse (Figure S2). This suggests that granule cells located in denser regions of the brain have higher levels of metabolic and signaling activity, perhaps reflecting greater levels of connectivity and communication with adjacent neurons.

**Temporal and transcriptomic modalities.** We next wanted to use SCHEMA to improve the informativity of single-cell visualizations. Widely used visualization methods such as UMAP do not allow a researcher to specify aspects of the underlying data that should be highlighted in the visualization. We used SCHEMA to highlight the age-related structure in a scRNA-seq dataset of *D. melanogaster* neurons profiled across the full *D. melanogaster* lifespan. We ran SCHEMA on a 20-component NMF change of basis of the gene expression matrix, using temporal metadata as our secondary dataset and visualizing the transformed result in two-dimensions using UMAP. While some age-related structure exists in the original dataset, SCHEMA-based transformation of the data clearly captured a cellular trajectory consistent with biological age (Figure 3). This structure becomes visually apparent even at relatively high values of the minimum correlation constraint (e.g., $s = 0.99$ in Figure 3). SCHEMA enables visualizations that capture biological patterns of particular interest to researchers, while preserving much of the distance-related correlation structure of the original primary dataset. Our results also apply to other kinds of metadata; for example, we can also infuse cluster hierarchy information into a tSNE visualization (Figure S3).

**Protein sequence and binding specificity modalities.** The V(D)J recombination mechanism in T-cells leads to high sequence diversity in the hypervariable CDR3 segments of T-cell receptors (TCRs). Understanding how this sequence variability relates to the binding specificities of TCRs has been a long-standing research goal [25]. Recent computational approaches [15, 16] have computed the sequence similarity score between two CDR3 regions using a BLOSUM62 matrix and posited that the computed score correlates with similarity in binding affinity. We wanted to see if the use of BLOSUM scores is appropriate, especially given that V(D)J recombination introduces a highly specific set of selection pressures on CDR3 sequence. We therefore used SCHEMA to estimate which amino acids in CDR3 segments confer binding specificity. We analyzed a single-cell dataset that recorded clonotype data for 62,858 T-cells and their binding specificities against a panel of 50 ligands [2]. For each T-cell, we represented fixed-length sub-sequences of CDR3 $\alpha$ and $\beta$ chains as 20-count vectors (one for each amino acid). We maximized the correlation of this primary dataset with the binding specificity (a 50-bit vector), our secondary dataset. We interpreted the 20 feature weights produced by SCHEMA as amino acid importance scores and compared them to the diagonal entries of the BLOSUM62 matrix (Figure S4). The two sets of scores are roughly consistent: His and Cys are difficult to substitute, while Ala, Lys, Thr, Ile, and Val can be more easily replaced. Interestingly, unlike BLOSUM, SCHEMA indicates Pro can be replaced relatively easily but Phe and Glx can not. Further investigation is needed to resolve if this is an artifact of the dataset or if modifying the BLOSUM matrix makes it better suited to TCR sequences.

## 6 Discussion

Potential algorithmic developments promise to make SCHEMA an even more powerful integration tool. SCHEMA pairs a change-of-basis with a scaling transform for reasons of speed: exploring the space of *all* linear transforms could be more powerful but converts the problem to the generally much-slower framework of *semidefinite programming*. However, since the constraints of the program are well-structured, and the size of the psd matrix constraint is small, we are developing a fast, bespoke optimization framework that can leverage these advantages to optimize over all linear transforms.

The "kernel trick" for distances [26] is a powerful machine learning tool that allows distances to be implicitly computed in some (potentially nonlinear) feature space and is commonly used to analyze scRNA-seq data. SCHEMA can be paired with the kernel trick for *nonlinear* feature selection and can thus be used to determine the best kernels for representing the primary dataset.

We have motivated correlation as an intuitively useful objective function for preserving and aligning neighborhood structure of datasets and demonstrated its utility empirically. We aim to further develop the theory underlying the interaction between the correlation coefficient and the low metric entropy of biological data, which will hopefully lead to related objective functions that allow more explicit control of the level of neighborhood structure that is preserved.

Lastly, we note that SCHEMA is highly general and has potential applications beyond the examples in this paper, including to other modalities, experimental strategies, and multi-modal biological datasets beyond single cell biology (e.g., large population datasets like the UK Biobank). We make Schema available at `http://schema.csail.mit.edu`.

## 7 Acknowledgements

We thank Benjamin DeMeo for most helpful discussions.

## References


[1] Samuel G Rodriques et al. "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". In: *Science* 363 (2019), pp. 1463–1467.

[2] *A New Way of Exploring Immunity - Linking Highly Multiplexed Antigen Recognition to Immune Repertoire and Phenotype - 10x Genomics*. URL: https://www.10xgenomics.com/resources/application-notes/a-new-way-of-exploring-immunity-linking-highly-multiplexed-antigen-recognition-to-immune-repertoire-and-phenotype/.

[3] Xiao Dong et al. "Accurate identification of single-nucleotide variants in whole-genome-amplified single cells". In: *Nature Methods* 14 (2017), pp. 491–493. ISSN: 15487105. DOI: 10.1038/nmeth.4227.

[4] Junyue Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells." In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. DOI: 10.1126/science.aau0730. URL: http://www.ncbi.nlm.nih.gov/pubmed/30166440.

[5] Ino D. Karemaker and Michiel Vermeulen. "Single-Cell DNA Methylation Profiling: Technologies and Biological Applications". In: *Trends in Biotechnology* 36.9 (2018), pp. 952–965. ISSN: 18793096. DOI: 10.1016/j.tibtech.2018.04.002.

[6] Assaf Rotem et al. "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state". In: *Nature Biotechnology* 33 (2015), pp. 1165–1172. ISSN: 15461696. DOI: 10.1038/nbt.3383.

[7] Marlon Stoeckius et al. "Simultaneous epitope and transcriptome measurement in single cells". In: *Nature Methods* 14 (2017), pp. 865–868. ISSN: 15487105. DOI: 10.1038/nmeth.4380.

[8] Brian Hie, Bryan Bryson, and Bonnie Berger. "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama". In: *Nature Biotechnology* 37.6 (June 2019), pp. 685–691. ISSN: 15461696. DOI: 10.1038/s41587-019-0113-3.

[9] Laleh Haghverdi et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature Biotechnology* 36.5 (2018). ISSN: 15461696. DOI: 10.1038/nbt.4091.

[10] Ilya Korsunsky et al. "Fast, sensitive, and accurate integration of single cell data with Harmony". In: *bioRxiv* (2018). DOI: 10.1101/461954. URL: http://dx.doi.org/10.1101/461954.

[11] Nikolas Barkas et al. "Joint analysis of heterogeneous single-cell RNA-seq dataset collections". In: *Nature Methods* 16 (2019), pp. 695–698. DOI: 10.1038/s41592-019-0466-z.

[12] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. "Scmap: Projection of single-cell RNA-seq data across data sets". In: *Nature Methods* 15.5 (Apr. 2018), pp. 359–362. ISSN: 15487105. DOI: 10.1038/nmeth.4644.

[13] Tim Stuart et al. "Comprehensive Integration of Single-Cell Data". In: *Cell* (2019). ISSN: 10974172. DOI: 10.1016/j.cell.2019.05.031.

[14] Joshua D. Welch et al. "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity". In: *Cell* 177.7 (June 2019), pp. 1873–1887. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2019.05.006. URL: https://www.sciencedirect.com/science/article/pii/S0092867419305045?via%3Dihub.

[15] Pradyot Dash et al. "Quantifiable predictive features define epitope-specific T cell receptor repertoires." In: *Nature* 547.7661 (2017), pp. 89–93. DOI: 10.1038/nature22383. URL: http://www.ncbi.nlm.nih.gov/pubmed/28636592.

[16] Neerja Thakkar and Chris Bailey-Kellogg. "Balancing sensitivity and specificity in distinguishing TCR groups by CDR sequence similarity". In: *BMC Bioinformatics* 20.1 (Dec. 2019), p. 241. ISSN: 1471-2105. DOI: 10.1186/s12859-019-2864-8. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2864-8.

[17] Leland McInnes and John Healy. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv* (2018). ISSN: 2475-9066. DOI: 10.21105/joss.00861.

[18] Laurens van der Maaten. "Accelerating t-SNE using Tree-Based Algorithms". In: *Journal of Machine Learning Research* 15 (2014), pp. 1–21. URL: http://homepage.tudelft.nl/19j49/tsne;.

[19] Brian Hie et al. "Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape". In: *Cell Systems* 8.6 (2019), pp. 483–493.

[20] Debajyoti Sinha et al. "Dropclust: Efficient clustering of ultra-large scRNA-seq data". In: *Nucleic Acids Research* 46.6 (2018). ISSN: 13624962. DOI: 10.1093/nar/gky007.

[21] Jacob Goldberger et al. "Neighbourhood Components Analysis". In: *Advances in Neural Information Processing Systems*. 2005, pp. 513–520.

[22] Eric P Xing et al. "Distance Metric Learning, with Application to Clustering with Side-Information". In: *Advances in Neural Information Processing Systems*. 2002. URL: http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.

[23] Kilian Q Weinberger and Lawrence K Saul. "Distance Metric Learning for Large Margin Nearest Neighbor Classification". In: *Journal of Machine Learning Research* 10 (2009), pp. 207–244.

[24] Jason V. Davis et al. "Information-theoretic metric learning". In: *Proceedings of the 24th international conference on Machine learning - ICML '07*. New York, New York, USA: ACM Press, 2007, pp. 209–216. ISBN: 9781595937933. DOI: 10.1145/1273496.1273523. URL: http://portal.acm.org/citation.cfm?doid=1273496.1273523.

[25] Nishant K. Singh et al. "Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes". In: *The Journal of Immunology* 199.7 (2017), pp. 2203–2213. ISSN: 0022-1767. DOI: 10.4049/jimmunol.1700744.

[26] Bernhard Schölkopf. "The Kernel Trick for Distances". In: *Advances in Neural Information Processing Systems*. 2001, pp. 301–307. URL: http://papers.nips.cc/paper/1862-the-kernel-trick-for-distances.

[27] J. Durbin and A. Stuart. "Inversions and Rank Correlation Coefficients". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2 (July 1951), pp. 303–309. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1951.tb00095.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1951.tb00095.x.

[28] Y. William Yu et al. "Entropy-Scaling Search of Massive Biological Data". In: *Cell Systems* 1.2 (Aug. 2015), pp. 130–140. ISSN: 2405-4712. DOI: 10.1016/J.CELS.2015.08.004. URL: https://www.sciencedirect.com/science/article/pii/S2405471215000587.

[29] Bosiljka Tasic et al. "Shared and distinct transcriptomic cell types across neocortical areas". In: *Nature* 563.7729 (Nov. 2018), pp. 72–78. DOI: 10.1038/s41586-018-0654-5. URL: http://www.ncbi.nlm.nih.gov/pubmed/30382198.

[30] Dmitry Kobak and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics". In: *bioRxiv* (May 2019), p. 453449. DOI: 10.1101/453449. URL: https://www.biorxiv.org/content/10.1101/453449v2.

# A Details of the Quadratic Program

We introduce some notation to condense the expressions. Define $w \in \mathbb{R}^k$ where $w_i = \omega_i^2$, $\delta_{ij} \in \mathbb{R}^k$ with $(\delta_{ij})_s = ((x_i)_s - (x_j)_s)^2$ (i.e. squared elements of $x_i - x_j$) and, for convenience, let $P$ be the set of pairs of observations $P = \{\{i,j\} : 1 \leq i \leq j \leq N\}$. Using the fact that the covariance between variables $X$ and $Y$ is given by $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, and the variance as $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X]$, we can expand:

$$\begin{aligned}
\mathrm{Cov}(w, \rho_\ell) &= \frac{1}{|P|} \sum_{\{i,j\}\in P} \rho_\ell(x_i^{(\ell)}, x_j^{(\ell)}) \delta_{ij}^T w - \frac{1}{|P|^2} \left( \sum_{\{i,j\}\in P} \delta_{ij}^T w \right) \left( \sum_{\{i,j\}\in P} \rho_\ell(x_i^{(\ell)}, x_j^{(\ell)}) \right) \\
&= \left( \frac{1}{|P|} a_\ell - \frac{1}{|P|^2} b_\ell \right)^T w \\
\mathrm{Var}(w) &= \frac{1}{|P|} \sum_{\{i,j\}\in P} w^T \delta_{ij} \delta_{ij}^T w - \frac{1}{|P|^2} \left( \sum_{\{i,j\}\in P} \delta_{ij}^T w \right)^2 \\
&= w^T \left( \frac{1}{|P|} S - \frac{1}{|P|^2} T \right) w
\end{aligned}$$

where $a_\ell$ and $b_\ell$ are $k$-dimensional vectors that depend only on $D_\ell$; and $S$ and $T$ are $N \times k$ matrices that depend only on $D_1$.

Explicitly, we derive:

$$\begin{aligned}
a_\ell &= \sum_{\{i,j\}\in P} \rho_\ell(x_i^{(\ell)}, x_j^{(\ell)}) \delta_{ij} \\
b_\ell &= \left( \sum_{\{i,j\}\in P} \rho_\ell(x_i^{(\ell)}, x_j^{(\ell)}) \right) \sum_{\{i,j\}\in P} \delta_{ij} \\
S &= \sum_{\{i,j\}\in P} \delta_{ij} \delta_{ij}^T \\
T &= \left( \sum_{\{i,j\}\in P} \delta_{ij} \right) \left( \sum_{\{i,j\}\in P} \delta_{ij}^T \right)
\end{aligned}$$

We recall the general optimization problem (5) that needs to be mapped to this framework:

$$\max_{w \in \mathbb{R}^k} \sum_{j=1}^{r} \gamma_r \mathrm{Cov}(w, \rho_j) - \alpha \mathrm{Var}(\rho^*) - \lambda ||w - \mathbf{1}||^2 \tag{7}$$

subject to

$$\mathrm{Cov}(w, \rho_j) \geq \beta_j \text{ for } 1 \leq j \leq r$$
$$w \succeq \mathbf{0}$$

and the framework for quadratic programming in (3) that this needs to be mapped to:

$$\min_{v \in \mathbb{R}^s} v^T Q v + q^T v \tag{8}$$

subject to

$$Gv \preceq h$$
$$Av = b$$

The mapping is straightforward:

$$v = w$$
$$Q = \frac{1}{|P|} S - \frac{1}{|P|^2} T + \lambda I_k$$
$$q = -2\lambda \mathbf{1} - \sum_{j=1}^{r} \gamma_\ell \left( \frac{1}{|P|} a_\ell + \frac{1}{|P|^2} b_\ell \right)$$

We also require that $Q$ be positive semidefinite (psd). This is also straightforward to show. We can write:

$$Q = \lambda I_k + \frac{1}{|P|} \sum_{\{i,j\} \in P} (\delta_{ij} - \mu)(\delta_{ij} - \mu)^T$$

where $\mu = \frac{1}{|P|} \sum_{\{i,j\} \in P} \delta_{ij}$, so it is a sum of psd matrices.

For the linear constraint, we express $G$ as a block matrix:

$$G = \begin{pmatrix} H & \mathbf{0} \\ \mathbf{0} & -I_k \end{pmatrix}$$

where each row in $H$ is given by:

$$H_j = -\frac{1}{|P|} a_j - \frac{1}{|P|^2} b_j \text{ for } 1 \leq j \leq r$$

and $h$ is an $r + k$-dimensional vector, where:

$$h_j = \begin{cases} -\beta_j & \text{for } 1 \leq j \leq r \\ 0 & \text{for } r + 1 \leq j \leq r + k \end{cases}$$

If $\beta_j = 0$ for some $j$ (i.e. no correlation constraint) the corresponding rows can be deleted from $H$ and $h$. We note here that while the framework can handle as many correlation constraints as the user desires, our current implementation only works with a correlation constraint on the primary dataset (i.e. $\beta_j = 0$ for all $j > 1$); we believe for most use cases, this will be sufficient.

We have no equality constraints in our optimization, so $A$ and $b$ from (8) are not needed.

# B Motivation for Correlation

## B.1 Spearman Correlation of Pairwise Distances & Neighborhood Structure

We first introduce some notation. $D_1$ and $D_2$ are the two datasets under consideration, each covering the same set of points $\mathcal{S}$, with $|\mathcal{S}| = N$. Each dataset has associated metrics $\rho_1$ and $\rho_2$, respectively. We denote the Spearman rank correlation between the $\binom{N}{2}$ pairwise distances within each of $D_1$ and $D_2$ as $pwcs(\rho_1, \rho_2)$. Also, we will often refer to point-triplets $A, B, C \in \mathcal{S}$. We recapitulate that the Spearman correlation is analogous to the more popular Pearson correlation with the distinction that the former works with ranks while the latter works with actual values.

Below, we assume *w.l.o.g.* that the $\binom{N}{2}$ pairwise distances in a dataset are distinct. If they are not, it is easy to break the ties in a way that preserves the ordering of all non-distinct distances. For example, let $\delta$ be the smallest *non-zero* difference between two pairwise distances. For each point in $\mathcal{S}$, we pick a random direction and move it $\frac{\delta}{100}$ units along that direction.

**Lemma 1.** *$pwcs(\rho_1, \rho_2) = 1 \iff$ the neighborhood structures of $D_1$ and $D_2$ are identical.*

*Proof.* If $pwcs(\rho_1, \rho_2) = 1$, then the pairwise distances in $D_1$ and $D_2$ are ranked identically; this follows from the definition of Spearman correlation. In other words, $\rho_1(A, B) < \rho_1(A, C) \iff \rho_2(A, B) < \rho_2(A, C)$. This, in turn, implies that, for each $A$, its distance to the remaining $N - 1$ points is in the same order in both datasets. Thus, the neighborhood-structures are identical.

The other direction also follows directly from our definition of identical neighborhood structures. If $D_1$ and $D_2$ have identical neighborhood structures, then for any point $A \in \mathcal{S}$ and any $k$, the set of $k$-nearest neighbors of $A$ are the same in $D_1$ and $D_2$. This is equivalent to stating that the $j$-th nearest point to $A$ is the same in $D_1$ and $D_2$, for all $j$. Repeating this for all points establishes that the pairwise distances are ranked identically in $D_1$ and $D_2$. □

We now describe a more general connection between $pwcs(\rho_1, \rho_2)$ and neighborhood similarity in the case when there *are* some mis-alignments between $D_1$ and $D_2$.

**Lemma 2.** *$pwcs(\rho_1, \rho_2) > 1 - \epsilon \implies mismatch(D_1, D_2) < \sqrt{2\epsilon}$*
*where $mismatch(D_1, D_2)$ is the fraction of point-triplets that are misaligned, i.e., $\rho_1(A, B) < \rho_1(A, C)$ but $\rho_2(A, B) > \rho_2(A, C)$.*

*Proof.* We first convert pairwise distances in $D_1$ and $D_2$ to ranks (i.e., $1, \ldots, \frac{N(N-1)}{2}$). The rank-ordering $\mathcal{R}_2$ for $D_2$ is a permutation of the rank-ordering $\mathcal{R}_1$ for $D_1$. The lemma above is essentially relating two measures of distance between these permutations: the Spearman distance and the Kendall Tau distance. The latter counts the number of inversions, i.e., the number of pairs $p_{AB}, p_{AC} \in \mathcal{R}_1$ whose order is inverted in $\mathcal{R}_2$. We appeal to a well-known result from Durbin and Stuart [27] that states

$$r_S \leq 1 - \frac{1 - r_K}{2(M + 1)}[(M - 1)(1 - r_K) + 4]$$

where $r_K$ is the Kendall Tau correlation and the $r_S$ is the Spearman correlation between the permutations, and $M$ is their length (here, $M = \frac{N(N-1)}{2}$). For large N, this simplifies to

$$\frac{(1 - r_K)^2}{2} \geq 1 - r_S$$

The inequality stated in the lemma directly follows from this. □

The results above provide support to our intuition that Spearman rank correlation of pairwise distances is a direct way of measuring distortions of neighborhoods. Furthermore, we can show intuitively that maximizing the Spearman correlation encourages global and not local movement, so as to minimally distort neighborhoods. However, rank correlation is hard to compute (and certainly not doable in a QP framework). While the Pearson correlation can be quite different than the rank correlation in general, we find that empirically it captures the right intuition on the large, low-metric-entropy datasets we encounter in biological settings.

### B.2 Manifold Reshaping

Recent work by Yu et al. [28] has provided evidence that even very high-dimensional biological data often actually lie on some implicit low-dimensional manifold. Thus, the proper way to measure the similarity between points is using *manifold's* metric and not the Euclidean metric of the high-dimensional ambient space. For machine learning applications that operate on distances between points, it is therefore crucial that we can access the *correct* distances, even if we do not know the manifold.

Assuming the data manifold is "smooth" (as practitioners usually do), the Euclidean distances between *nearby* points closely approximate the distances on the manifold. Therefore, we can assume that the small distances in the primary dataset are accurate and should be preserved, and as importantly, the ranking of the distances should be preserved. Concretely, if $\rho(x_i, x_j) > \rho(x_i, x_\ell)$, then we should encourage $\rho^*(x_i, x_j) > \rho^*(x_i, x_\ell)$; in other words, the local neighborhoods of points should not be distorted.

Suppose our dataset consists of well-separated clusters (common in scRNA-seq datasets, where the clusters could potentially represent different cell types/lines) and lies on a manifold as in Figure 1. The within-cluster distances are well represented by the Euclidean distance in the ambient space, but the between-cluster distances are not. The goal for our algorithm is to encourage *global* movement of the *entire* clusters, rather than distorting the neighborhoods within each cluster.

We consider a simplified dataset and show that optimizing the Spearman rank coefficient encourages global and not local moves. Suppose we have two clusters, one centered at the origin (call it $O$), and the other (call it $M$) drawn from a Gaussian with mean $\mu$ and variance $\sigma^2 I_k$ (the analysis does not require an isotropic Gaussian but it is cleaner).

We define here our notion of *local* and *global* movements. In the local scenario, we perturb the points in $M$ randomly to generate $T$:

$$T = \{x + z \,|\, x \in M, z \sim \text{Normal}(\mathbf{0}, \tau^2 I_k)\}$$

In other words, the points in the cluster get some isotropic Gaussian noise with variance $\tau^2$.

In the global scenario, we want all the points in the cluster to move in generally the same direction. The simplest way to achieve this is to choose a Gaussian with variance only in one direction. To that end, we fix a random vector $\nu$ such that $\|\nu\| = 1$. Then, we generate another perturbation $S$ analogously:

$$S = \{x + z \,|\, x \in M, z \sim \text{Normal}(\mathbf{0}, k\tau^2 \nu\nu^T)\}$$

Thus, $S$ is the points in $M$ perturbed in the direction of $\nu$. We multiply the variance by $k$ to make sure that norms of variances of the Gaussians for both $S$ and $T$ are the same.

We consider the *inter*-cluster distances between $S$ or $T$ and $O$ and compare them to the distances between $M$ and $O$. The goal is to show that the global perturbation has a higher Spearman rank correlation with the original distances than the local perturbation.

First, we simplify the model: assume that $\mu$ is large with respect to the variances of the clusters, so we can approximate distances between *any* point in $O$ and some $x$ in any of $M, S, T$ as $\|x\|$. We can also consider the clusters as random variables:

$$\begin{aligned} M &\sim \text{Normal}(\mu, \sigma^2 I_k) \\ S &= M + Z_S \\ Z_S &\sim \text{Normal}(\mathbf{0}, k\tau^2 \nu\nu^T) \\ T &= M + Z_T \\ Z_T &\sim \text{Normal}(\mathbf{0}, \tau^2 I_k) \end{aligned}$$

Note that $S$ and $T$ are both Gaussian as well. To evaluate which perturbation weakens the Spearman rank correlation the most, we determine the probability that the perturbation *swaps* the norms of two points in the cluster. Probabilistically, we compare:

$$\begin{aligned} &\Pr(\|M_1 + (Z_T)_1\|^2 > \|M_2 + (Z_T)_2\|^2 \text{ given } \|M_1\|^2 < \|M_2\|^2) \\ &\Pr(\|M_1 + (Z_S)_1\|^2 > \|M_2 + (Z_S)_2\|^2 \text{ given } \|M_1\|^2 < \|M_2\|^2) \end{aligned}$$

Note that the norm of a multivariate Gaussian follows a $\chi^2$-distribution, so the above probabilities can be simulated using that distribution. Figure S5 illustrates the outcome of this simulation. Across a wide range of paramter choices, the *global* perturbation causes fewer swaps or rank and therefore a higher Spearman rank correlation. In future work, we will aim to prove this fact analytically.

# C Concentration Bounds

Our approach is to show that, given a $\hat{\omega}$ that has been calculated based on a random sample, the correlation coefficient between *all* pairwise distances cannot be too different than the correlation coefficient computed on the sample. To do this, we use Chernoff bounds, which bound how far away a random variable can be from its expectation, on the covariance and variance terms of correlation coefficient given by (1). This gives us a bound on how far away the correlation coefficient on the whole population can be from the one calculated on the sample.

Let $P$ be a random subset of all possible interactions. For now, we assume that interactions are chosen uniformly at random. Solving the optimization problem (2) with our sample $P$ yields $\hat{\omega}$, an estimator for the true optimal transform $\omega$. We show that $\hat{\omega}$ approximates $\omega$ well by showing that the pairwise distances of $\hat{\omega}(D)$ have high correlations with the secondary datasets as long as $\hat{\omega}$ has high correlations on the subsample.

Formally, we will guarantee, for any $\alpha, \delta > 0$ and sample size at least $|P| = O\left(\frac{\log(1/\alpha)}{\delta^2}\right)$,

$$\left|\text{Corr}(\hat{w}, \rho_j) - \widehat{\text{Corr}}(\hat{w}, \rho_j)\right| < \delta \text{ with probability at least } 1 - \alpha, \tag{9}$$

where $\widehat{\text{Corr}}(\,\cdot\,,\,\cdot\,)$ is the sample correlation coefficient.

This is a powerful result, made possible by our restriction to scaling transforms, which are easy to analyze. First of all, note that we only need a sample-size *logarithmic* in our desired confidence level in order to get strong concentration, allowing analysis of massive scRNA-seq datasets.

To begin our analysis, let $W \succeq 0$ be a $k \times k$ psd matrix (in our specific case it will be diagonal, but this analysis will generalize to any psd matrix, which motivates the generalization to all psd matrices in Section 6). We also assume randomly draw pairwise distances $\delta$ uniformly from the set of pairs of points in our primary dataset. Here, we focus on the correlation between the transformed dataset and the primary dataset (i.e. the one that appears in the constraint in all of our examples). Analyses for correlations between the transformed data and the secondary datasets will be similar.

Consider the form of the (population) correlation:

$$\mathrm{Corr}(W, \rho_1) = \frac{\overbrace{\mathbb{E}[\delta^T W \delta \delta^T \delta]}^{A} - \overbrace{\mathbb{E}[\delta^T W \delta]}^{B} \overbrace{\mathbb{E}[\delta^T \delta]}^{C}}{\underbrace{\mathrm{Var}^{1/2}(W)}_{D} \underbrace{\mathrm{Var}^{1/2}(\rho_1)}_{E}} \tag{10}$$

If, for our samples, we can determine confidence intervals of size $2\epsilon$ for each of the terms $A, B, C, D, E$, then we can bound the distance away from the correlation on the *entire* set of pairwise distances. This distance is maximized when $A$ is as small as possible, and $B, C, D$, and $E$ is as large as possible. So:

$$\begin{aligned}
\widehat{\mathrm{Corr}}(W, \rho_1) &\geq \frac{(A-\epsilon) - (B+\epsilon)(C+\epsilon)}{(D+\epsilon)(E+\epsilon)} \\
&\approx \frac{A - BC - (1+B+C)\epsilon}{DE(1+\epsilon/D)(1+\epsilon/E)} \\
&\approx \left(\frac{A-BC}{DE} - \frac{B+C+1}{DE}\epsilon\right)(1-\epsilon/D)(1-\epsilon/E) \\
&\approx \left(\frac{A-BC}{DE} - \frac{B+C+1}{DE}\epsilon\right)(1-\epsilon/D-\epsilon/E) \\
&\approx \left(\frac{A-BC}{DE}\right)\left(1 + \frac{D+E}{DE}\epsilon\right) - \frac{B+C+1}{DE}\epsilon \\
&= \mathrm{Corr}(W, \rho_1)\left(1 - \frac{\mathrm{Var}^{1/2}(W) + \mathrm{Var}^{1/2}(\rho_1)}{\mathrm{Var}^{1/2}(W)\mathrm{Var}^{1/2}(\rho_1)}\epsilon\right) - \frac{1 + \mathbb{E}[\delta^T W \delta] + \mathbb{E}[\delta^T \delta]}{\mathrm{Var}^{1/2}(W)\mathrm{Var}^{1/2}(\rho_1)}\epsilon
\end{aligned}$$

Thus, for a desired overall confidence level $\eta$, the relationship between $\epsilon$ and $\eta$ is given by:

$$\epsilon = \left(\frac{\mathrm{Var}^{1/2}(W)\mathrm{Var}^{1/2}(\rho_1)}{\max\left\{\mathrm{Var}^{1/2}(W) + \mathrm{Var}^{1/2}(\rho_1), 1 + \mathbb{E}[\delta^T W \delta] + \mathbb{E}[\delta^T \delta]\right\}}\right)\eta$$

To show that we can bound each of the terms $A, B, C, D, E$ we use *Hoeffding's inequality* to limit how far away the terms can be from their expectations. Let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from bounded range $[a, b]$, and set $s = b - a$, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then Hoeffding's inequality states:

$$\Pr\left[\bar{X} - \mathbb{E}X \geq t\right] \leq \exp\left(-\frac{nt^2}{s^2}\right)$$

This can be converted into giving a (one-sided) confidence interval of length $t$ by substituting the probability on the left with a desired confidence level $\alpha$, and solving for $n$, which gives a statement:

$$\mathbb{E}X \geq \bar{X} - t \text{ with confidence } 1-\alpha \text{ for } n \geq \frac{s^2 \log(1/\alpha)}{t^2} \tag{11}$$

We begin by applying the inequality on term $A = \mathbb{E}[\delta^T W \delta \delta^T \delta]$ by bounding $\delta^T W \delta \delta^T \delta$. It is clear that $|\delta^T W \delta \delta^T \delta| \leq |\delta^T W \delta||\delta^T \delta|$, so we can bound each individually. Note that we can assume without loss of generality that $W$ is diagonal here, because otherwise (since it is psd), we could write $W = UDU^T$, where $D$ is diagonal and $U$ is unitary; setting $y = U\delta$ yields $\left|\delta^T W \delta\right| = \left|y^T D y\right|$, and, by unitarity, $\|\delta\| = \|y\|$.

Then, by Cauchy-Schwarz:

$$\left|\delta^T W \delta\right| \leq \left|\sum \delta_i W_{ii} \delta_i\right| \leq \|W\| \|\delta\|^2 \tag{12}$$

where $\|W\|$ is the matrix-norm, i.e. $\|W\| = \sqrt{\mathrm{Tr}(W^T W)}$. So for a diagonal matrix, $\|W\|^2 = \sum W_{ii}^2$. We can bound $||\delta|| \leq \max_{x_i, x_j \in D}\{\|x_i - x_j\|\} \equiv \mathrm{diam}(D)$.

Thus, $\left|\delta^T W \delta \delta^T \delta\right| \leq \|W\| \mathrm{diam}^4(D)$.

To get a confidence interval of size $\epsilon$, we plug into (11), so we require:

$$N \geq \frac{\|W\| \mathrm{diam}^8(D) \log(1/\alpha)}{\epsilon^2}$$

Note that the diameter is an *extremely* coarse bound for the above bound. Morally, one can replace "diameter" with "variance", and the user has control over $\|W\|$ by choice of hyperparameters. We also note that the sample complexity improves drastically if we focus only or *local* distances, as discussed in Section 6.

The same analysis can be used for terms $B$ and $C$ in (10), but the dependency on the diameter is not as bad for those terms, so term $A$ is the worst case.

Now, we consider the variance terms $D$ and $E$. For term $E$, note:

$$\mathrm{Var}(\delta^T \delta) = \mathbb{E}[(\delta^T \delta - \mathbb{E}[\delta^T \delta])^2]$$

Again, $\left|\delta^T \delta - \mathbb{E}[\delta^T \delta]\right|$ is bounded by the maximum squared distance in the dataset $\mathrm{diam}^2(D)$, so we can use the Hoeffding inequality from above in the same way.

And term $D$ takes the same form as above, but with $\delta^T W \delta$ instead of $\delta^T \delta$. As noted in (12), this is a bounded random variable as well, but here with bound $\|W\|^2 \mathrm{diam}^4(D)$.

Thus, in order to get a uniform confidence interval across all the terms, we require:

$$N \geq \frac{\|W\|^2 \mathrm{diam}^8(D) \log(1/\alpha)}{\epsilon^2} \tag{13}$$

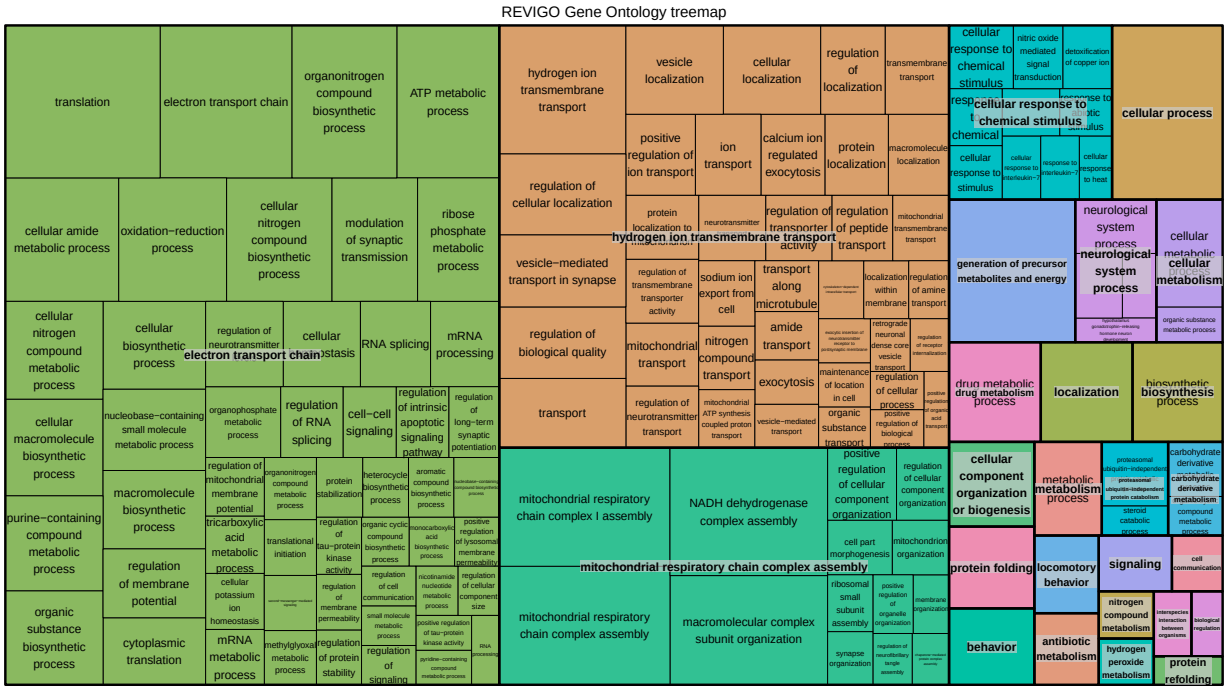


Figure S1: **Summary of density-dependent GO process enrichments.** Genes were ranked according to association with granule cell density (Figure 2) according to the SCHEMA transformation. Significant GO process enrichments on the ranked list of genes was obtained using the GOrilla web tool (`http://cbl-gorilla.cs.technion.ac.il`) and terms with a Bonferroni-Hochberg false discovery rate (FDR) $q$-value of less that 0.001 were summarized and visualized as a treemap using the REVIGO web tool (`http://revigo.irb.hr`). Box area corresponds to $-\log_{10}(q)$ of the respective term.

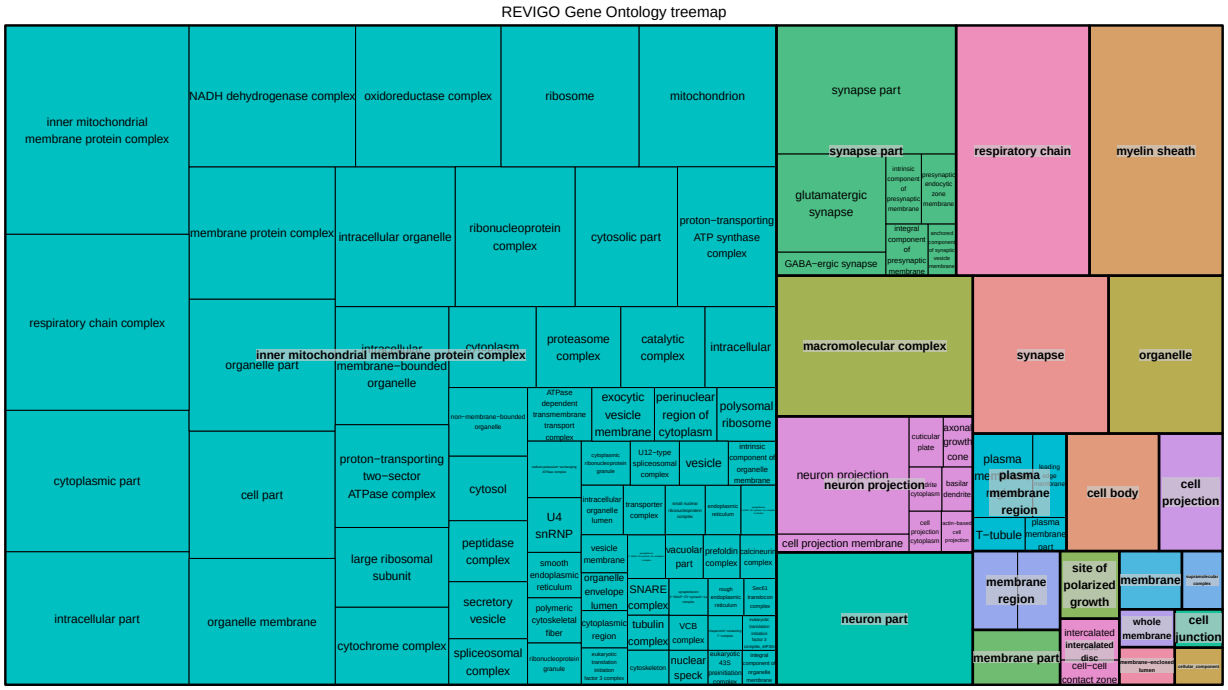


Figure S2: **Summary of density-dependent GO component enrichments.** Genes were ranked according to association with granule cell density (Figure 2) according to the SCHEMA transformation. Significant GO component enrichments on the ranked list of genes was obtained using the GOrilla web tool (`http://cbl-gorilla.cs.technion.ac.il`) and terms with a Bonferroni-Hochberg false discovery rate (FDR) $q$-value of less that 0.001 were summarized and visualized as a treemap using the REVIGO web tool (`http://revigo.irb.hr`). Box area corresponds to $-\log_{10}(q)$ of the respective term.

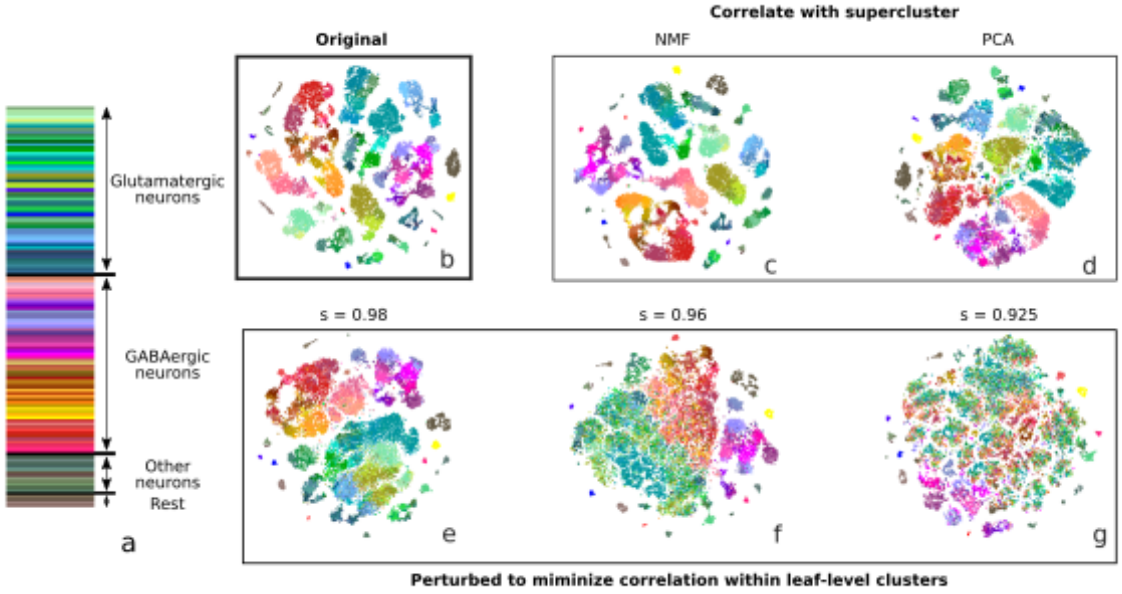


Figure S3: **Incorporating cluster information as meta-data into tSNE visualizations.** Here we analyzed mouse brain cell data from [29] using the clustering hierarchy described there. We followed the tSNE recipe in [30] for our preliminary processing. **(a):** We follow the color scheme from [29] where the colors were chosen based on cluster similarity. We have also outlined the higher-level supercluster that each leaf-level cluster is part of. **(b):** The original tSNE plot loses some of the global structure and puts some sibling clusters far apart (e.g., in the GABAnergic neuron supercluster). **(c,d):** We used Schema to maximize the correlation between gene expression (the primary dataset) and the cluster and supercluster labels (secondary datasets). We experimented with both NMF and PCA as change-of-basis transforms. In both cases, tSNE's visualization results in a more globally consistent placement of leaf-level clusters. **(e,f,g)**: As an alternative exploration, we used leaf-level cluster labels as the secondary dataset and used Schema to *minimize* the correlation with the primary dataset (gene expression). The intuition is that points in biologically-close clusters will be the first to be merged. As we increase the minimum correlation threshold $s$, it is clear that sibling clusters within a supercluster are the first to be merged. Note that we did *not* specify the supercluster data in these figures but are still able to rediscover that structure.

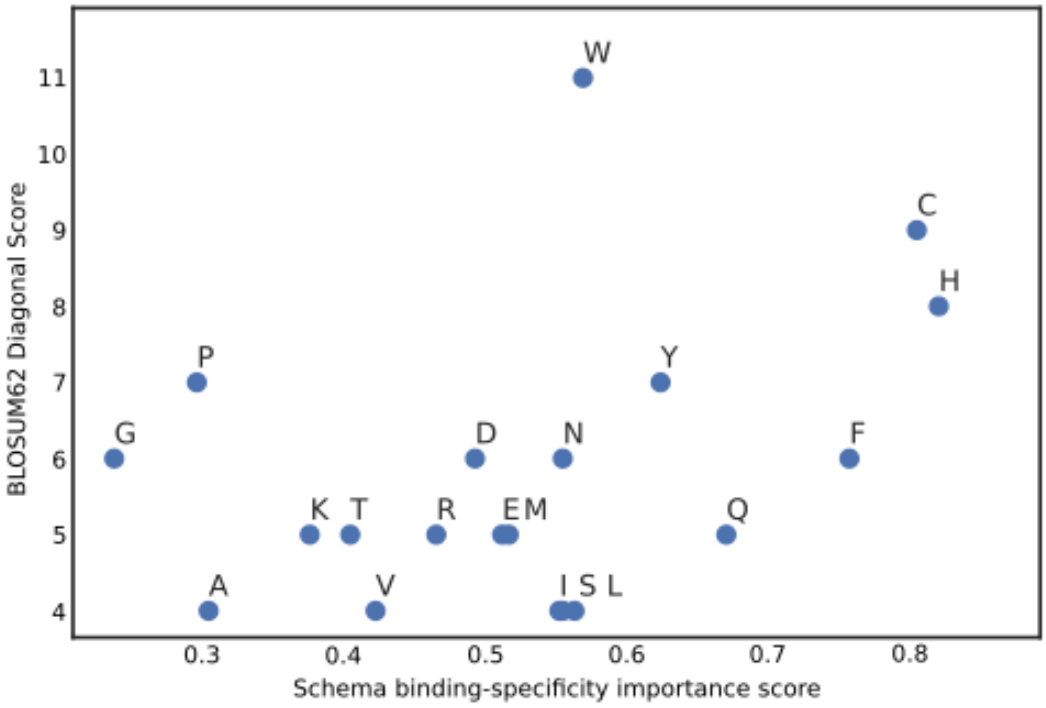


Figure S4: **Comparison of BLOSUM62 scores of residue conservation with SCHEMA-computed scores of residue importance in binding affinity.**. The point labels are the single-letter amino acid codes. The SCHEMA scores are feature ranks averaged over an ensemble of SCHEMA runs, covering various $\bar{w}$ and $s$ choices.

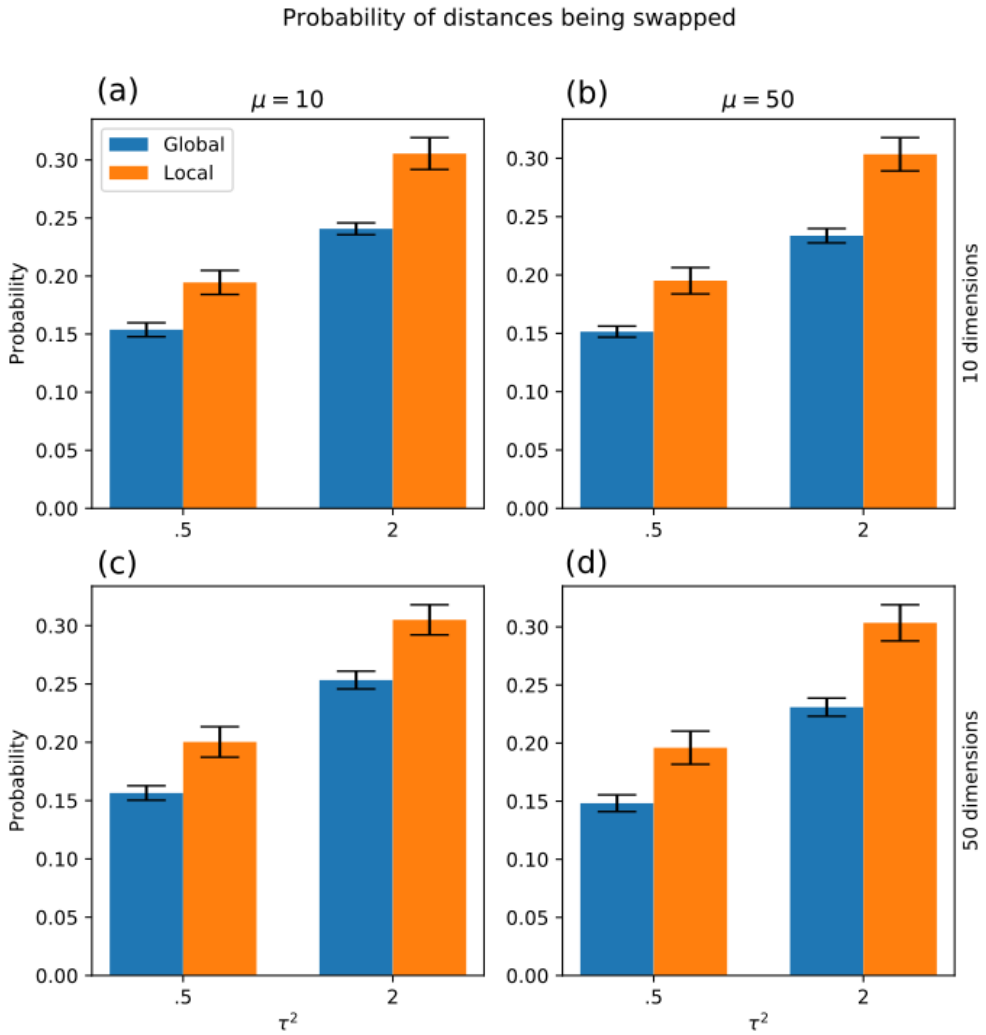


Figure S5: **The correlation objective emphasizes global movement:** Here, we demonstrate empirically that correlation as an objective function encourages *global* movement of clusters, not local movements for different sets of parameters, as discussed in Appendix B. In each plot, the left two bars have low perturbation variance and the right two bars are high perturbation variance; the top row uses 10-dimensional and the bottom uses 50-dimensional vectors; and the left column places the perturbed cluster 10 units from the origin, and the right column places it 50 units away. The height of the bar represents the proportion of ranks that get swapped because of the perturbation. From this it is clear that across all these parameter choices, *global* movement is encouraged because fewer swaps happen.