Quantifying distribution shifts in single-cell data with scXMatch

Anna Möller[1,2][0000−0001−7517−8549], Miriam Schnitzerlein[5,6,7], Eric Greto[2,3,4], Vasily Zaburdaev[5,6], Stefan Uderhardt[2,3,4], and David B. Blumenthal[1][0000−0001−8651−750X]

[1] Biomedical Network Science Lab, Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
[2] Department of Medicine 3 – Rheumatology and Immunology, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Universitätsklinikum Erlangen, 91054 Erlangen, Germany
[3] Deutsches Zentrum für Immuntherapie, FAU, 91054 Erlangen, Germany
[4] Exploratory Research Unit, Optical Imaging Competence Center Erlangen, FAU, 91058 Erlangen, Germany
[5] Department of Biology, FAU, Erlangen, Germany
[6] Max-Planck-Zentrum für Physik und Medizin, Erlangen, Germany
[7] Max Planck Institute for the Science of Light, Erlangen, Germany

{anna.moeller,david.b.blumenthal}@fau.de

Abstract. A basic task that frequently arises when analyzing single-cell data is to assess if there is a global distribution shift between the data profiles of cells from two different conditions. Widely used approaches to address this task such as visual inspection of two-dimensional representations or clustering-based workflows lack a solid statistical underpinning and are notoriously unstable and prone to confirmation bias. To promote more rigorous analysis, we here present the scverse-compatible Python tool scXMatch. scXMatch is based on a non-parametric graph-based test to quantify distribution shifts in arbitrary data spaces for which a suitable distance measure is available. We evaluated scXMatch on single-cell gene expression, chromatin accessibility, and imaging-derived cell morphology data, showing that it can robustly detect distribution shifts for different types of single-cell data. scXMatch thus aims to set a new standard in the single-cell biology field, replacing easy-to-manipulate semi-manual distribution shift quantification workflows by principled statistical testing.

Keywords: Single-cell biology · Statistical testing · Distribution shifts · Graph-based analysis · Cross-matching test · Non-parametric methods · scRNA-seq · Outlier detection

1 Introduction

The following question arises in many studies in the field of single-cell biology: Is there a global distribution shift between profiles of cells of the same type in a target condition in comparison to a control condition (e.g., diseased versus healthy, treated versus untreated, knockout versus wild-type)? Depending on the type of data — single-cell RNA sequencing (scRNA-seq), single-cell ATAC sequencing (scATAC-seq), etc. — typical workflows to answer this question include:

- Visually inspect the data upon projection into a two-dimensional (2D) space, e.g., using principal component analysis (PCA) followed by popular dimensionality reduction techniques such as UMAP [13] or t-SNE [11]. Conclude that there is a global distribution shift if the 2D representations of the cells from the different conditions are visually separated.
- Carry out joint clustering of the single-cell data for the two conditions, followed by cluster-level over-representation analysis. Conclude that there is a global distribution shift if the cells from the different conditions are unevenly distributed over the clusters.
- Run statistical methods to identify individual differential features in the high-dimensional data, i.e., single variables such as differentially expressed genes (DEGs) or differentially accessible peaks that behave differently between the two conditions. Conclude that there is a global distribution shift if there are many differential features.
- Train a machine learning (ML) model that aims to separate the cells from the two conditions based on their data. Conclude that there is a global distribution shift if the ML model can accurately separate the two conditions.

Although widely used, all of these approaches have known shortcomings, which are particularly well studied for scRNA-seq data. For instance, 2D representations of scRNA-seq data have been shown to poorly preserve distances in the original high-dimensional data space, casting doubt on the reliability of visual inspection of 2D data representations [3]. Clustering-based workflows are problematic because clustering scRNA-seq and other single-cell data is notoriously sensitive to the choice of the clustering algorithm [32,9]. Similar problems apply to workflows that rely on differential feature counts or ML-based separability as also here, different subroutines (i. e., different ML models or different methods to identify differential features) often produce substantially different results [23,25]. More generally, all of these approaches are indirect, at least to some degree *ad hoc*, and therefore easily lead to confirmation bias: Many hyper-parameters and sub-routines have to be selected along the way, and it is almost always close to impossible to make principled and non-arbitrary decisions regarding all of these steps. Given enough trial and error, it is hence very often possible to achieve the "desired" results.

To address these problems, we here present scXMatch — a Python package to directly test for global distribution shifts between sets $\mathcal{X}_1 = \{\mathbf{x}_i \mid y_i = 1\}$ and $\mathcal{X}_0 = \{\mathbf{x}_i \mid y_i = 0\}$ of h-dimensional profiles $\mathbf{x}_i \in \mathbb{R}^h$ of cells i from a target condition $y_i = 1$ and control condition $y_i = 0$. scXMatch implements a runtime- and memory-efficient version of the cross-matching test proposed in [18]. The cross-matching test is a non-parametric approach to test the null hypothesis that two sets of multi-variate data points $\mathcal{X}_0$ and $\mathcal{X}_1$ were sampled from the same distribution $F_0 = F_1$. It is ideally suited to assess global distribution shifts in single-cell data, as it relies on a test statistic whose distribution under the null hypothesis can be computed exactly even when the data-generating distribution $F_0 = F_1$ is unknown (more details in Section 2 and Appendix A). The only required input besides $\mathcal{X}_0$ and $\mathcal{X}_1$ is a suitable distance measure $d : \mathbb{R}^h \times \mathbb{R}^h \to \mathbb{R}_{\geq 0}$ to quantify dissimilarity between the cell profiles.

Despite these desirable properties and although there is an R implementation of the cross-matching test [7], we are not aware of a single study which uses it to assess global distribution shifts in single-cell omics data such as scRNA-seq or scATAC-seq data. One likely reason for this is that the original version requires to compute a minimum-weight maximum-cardinality matching (MWMCM) in a complete weighted graph, where each data profile $\mathbf{x}_i$ is represented by a node. This requires quadratic memory with respect to the number of data profiles, which makes out-of-the-box usage of the cross-matching test infeasible for large-scale single-cell datasets. For scXMatch, we therefore developed a substantially more efficient version of the cross-matching test, where the complete graph is replaced by a sparse k-nearest neighbor (k-NN) graph, which reduces the test's memory requirements from quadratic to linear. A systematic evaluation on scRNA-seq, scATAC-seq, and imaging-derived morphodynamics data shows that scXMatch scales to realistically sized

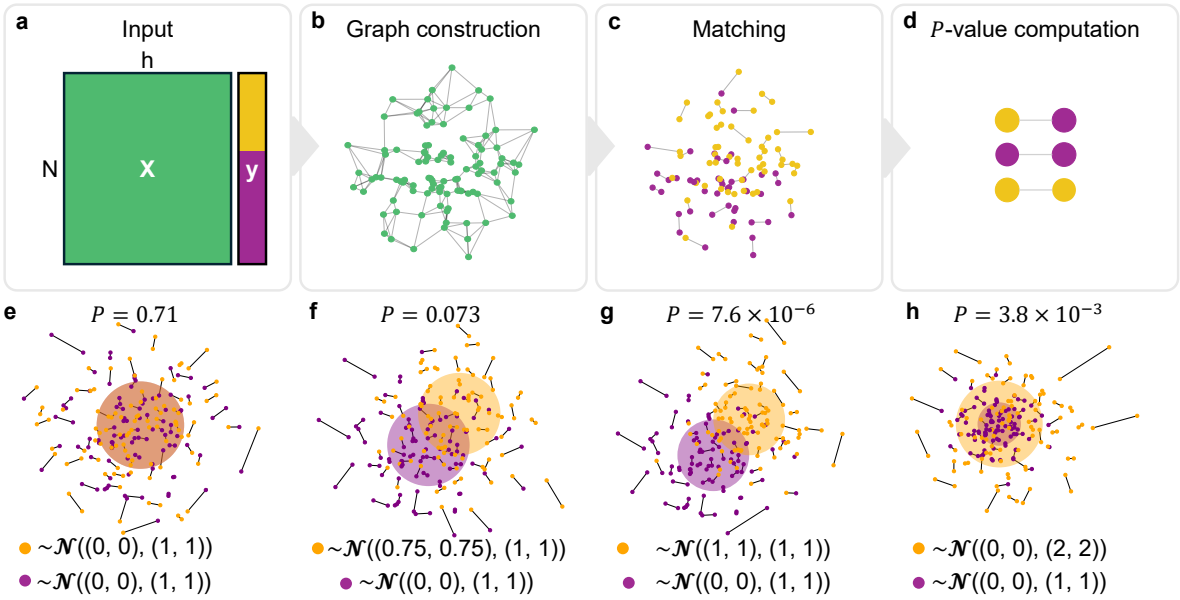

Fig. 1: Overview of scXMatch. (a) Required inputs: N single-cell profiles $\mathbf{x}_i$, binary condition labels y_i (orange/purple), a dissimilarity measure d, and the k parameter for k-NN mode. (b–c) A weighted graph G (complete or kNN) is constructed, and a minimum-weight matching M is computed. (d) The number of cross-matches between conditions yields a P-value and z-score to assess global distribution shifts. (e–h) Examples using synthetic data sampled from bivariate normal distributions show that scXMatch detects significant shifts when distributions differ in mean (f–g) or variance (h), but not when they are identical (e).

single-cell datasets and that it can reliably distinguish signal from noise, using experimental perturbation of varying strength as ground truth. A theoretical investigation shows that, for the special case of 1D data, scXMatch is strictly more sensitive than the Wilcoxon rank-sum test. To further encourage adoption within the single-cell biology research community, we implemented scXMatch as an scverse-compatible Python package and published it in the scverse ecosystem [28].

2 Method overview

The general purpose of scXMatch is to examine whether two groups $\mathcal{X}_0$ and $\mathcal{X}_1$ of single-cell data profiles are sampled from the same distribution. Figure 1 provides a high-level overview of scXMatch; see Appendix A and the original publication of the cross-matching test [18] for further details. In a first step, scXMatch constructs a weighted undirected connectivity graph G, where the node set corresponds to the set of all data profiles $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$ and edges connecting two cells i and j are weighted by the dissimilarity $d(\mathbf{x}_i, \mathbf{x}_j)$ of their profiles. Depending on available computational resources and the size of $\mathcal{X}$, the user can decide to construct G as the complete graph or as a k-NN graph. Next, scXMatch computes a MWMCM M in G, meaning that as many cells as possible are assigned to exclusive pairs, such that the sum of distances between the features of paired cells is minimized. The pairs $ij \in M$ are either iso-matches or cross-matches, depending on whether the paired cells i and j are from the same condition ($y_i = y_j$) or from different conditions ($y_i \neq y_j$). The number of cross-matches a_1 is used as a test statistic, and a P-value and a z-score are computed based on the distribution of the corresponding random variable A_1. Under the null hypothesis that $\mathcal{X}_0$ and $\mathcal{X}_1$ are sampled from the same distribution $F_0 = F_1$, the distribution of A_1 is exactly known even if $F_0 = F_1$ is unknown (as shown in Appendix A, this also holds for our efficient k-NN graph-based variant of the cross-matching test). The smaller the number of cross-matches a_1, the smaller the likelihood of observing it under the null hypothesis. That is, a small number of cross-matches is indicative of a distribution shift between $\mathcal{X}_0$ and $\mathcal{X}_1$.

A major advantage of scXMatch is that it is very flexible regarding its input. Although designed with single-cell data in mind, it can be used for any kind of multi-variate data, as long as a suitable dissimilarity

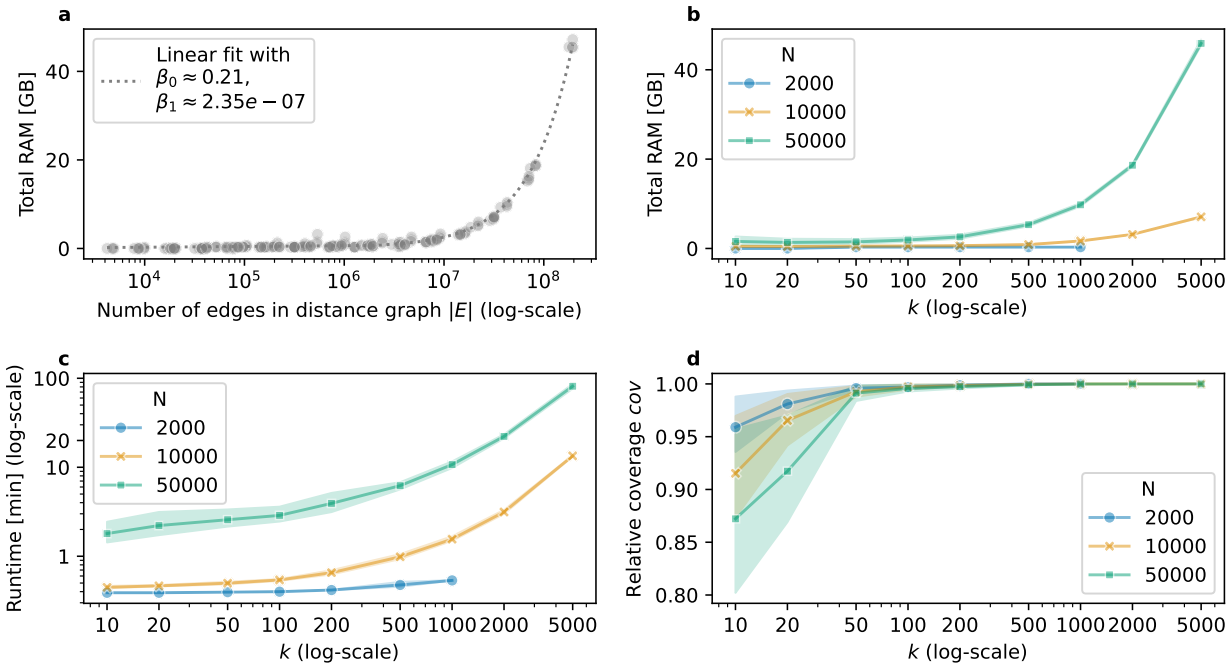

Fig. 2: Results of the scalability tests. (a) Total memory usage as a function of the number of edges $|E|$ in the k-NN graph constructed by scXMatch. (b–d) Total memory usage (b), runtime (c), and relative coverage (d) as a function of k, for datasets with different numbers of samples N. Line plots and shaded areas in (b–d) depict means and 95% confidence intervals obtained for datasets with fixed N and k and varying dimensionality $h \in \{10, 500, 1000, 2000\}$.

measure d is available. We provide out-of-the-box support for all distance functions implemented in SciPy; custom dissimilarity measures can be provided by the user as callables.

3 Experimental results

3.1 scXMatch scales to large datasets

Besides the dissimilarity measure d, scXMatch requires the number of neighbors k as an additional hyperparameter when run in k-NN mode. In principle, it makes sense to select k as large as possible, since smaller k more strongly restrict the solution space for the MWMCM M (see Appendix A for details). However, setting k to larger values increases the runtime and memory requirements of scXMatch. To provide a guideline on how to select k, we therefore simulated N data points $\mathbf{x}_i \in \mathbb{R}^h$ (corresponding to data profiles of individual cells in the context of single-cell data) for all combinations $(N, h) \in \{500, 1000, 2000, 5000, 10000, 20000, 50000\} \times \{10, 500, 1000, 2000\}$. On each of the resulting 28 simulated datasets, we ran scXMatch with varying $k \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$ and recorded the overall main memory usage, the overall runtime, and the relative coverage *cov* of the computed MWMCM M, i. e., the fraction of data profiles used towards the final P-value computation. Since the purpose of the runs on the simulated data was only to benchmark scXMatch's resource requirements, we simulated data with the respective dimensions by randomly drawing from a normal distribution. Each run was executed on a single node with 16 CPU cores and sufficient memory resources.

Figure 2 shows the results. Overall, we identified main memory usage as the major limiting factor. To enable the user to select k based on the available memory, we fitted a linear regression model that estimates the total RAM usage based on the number of edges $|E|$ in the k-NN graph G, using data from all scXMatch runs on the simulated data. Since the obtained model

$$\text{estimated total RAM [GB]} = \beta_0 + \beta_1 |E|, \text{ with } \beta_0 \approx 0.21 \text{ and } \beta_1 \approx 2.35 \cdot 10^{-7} \tag{1}$$

excellently fits the data (Figure 2a) and the number of edges $|E|$ is bounded from above by $|E| \leq Nk$ (how closely $|E|$ approaches this bound depends on the symmetry of the k-NN relationships), the user can select k based on the number of samples N and the available main memory as follows:

$$k = \left\lfloor \frac{\text{available RAM [GB]} - 0.21}{2.35 \cdot 10^{-7} N} \right\rfloor \tag{2}$$

Our Python implementation of scXMatch supports this automated selection of the hyperparameter k as a function of available RAM and dataset size.

Figure 2 also contains overviews on total memory usage (Figure 2b) and runtime (Figure 2c) as a function of k for different N. The test finished in under a minute for any value of k for datasets with $N = 2000$ samples, and in under 10 minutes with $k \leq 1000$ for datasets with $N = 50000$ samples. The dimensionality h of the input data only has a very small effect on scXMatch's resource requirements, as can be seen by the tight error margins of the line plots in Figure 2b and Figure 2c. The relative coverage, which indicates how many samples were covered by the matching, exceeds 99% for all N with $k \geq 100$ (Figure 2d). For all experiments on scRNA-seq data, we thus used $k = 100$, as this setup runs in reasonable time and memory and covers a convincing amount of samples. The scATAC-seq and morphodynamics datasets were smaller, which allowed us to run scXMatch in full graph mode.

3.2 scXMatch reliably detects global distribution shifts in scRNA-seq data

A method designed to quantify global distribution shifts in single-cell data should satisfy several key criteria to be considered robust and biologically meaningful. Firstly, it should be sensitive to biologically relevant differences, with output scores that scale with the underlying perturbation strength. Secondly, it should avoid false positives, i.e., it should return weak scores when comparing subsets drawn from the same distribution. Thirdly, its output should remain stable under subsampling from the compared distributions. To systematically assess these properties, we used ten public scRNA-seq datasets from studies where cells were exposed to perturbations of increasing strengths (Figure 3a): The datasets contain scRNA-seq data for cells of different cell lines at different time points after drug treatment [12], for K562 cells upon single- and dual-guide CRISPR knockout [15], for embryonic fibroblasts at different time points after transcription factor-mediated cellular reprogramming [19], and for different neuron subtypes in the prefrontal cortex of mice that were withdrawn for 48 hours or 15 days after cocaine addiction [1]. We ran scXMatch with $k = 100$ and squared Euclidean distance as metric.

We defined three quantitative evaluation metrics (see Figure 3b–d for a visual explanation and Appendix A for formal definitions):

- $\mathrm{MI}(s)$ (Figure 3b): The monotonicity index $\mathrm{MI}(s) \in [0, 1]$ measures whether a distribution shift score s increases monotonically with increasing perturbation strength. Larger is better.
- $\mathrm{SSNR}(s)$ (Figure 3c): The signal-stronger-than-noise ratio $\mathrm{SSNR}(s) \in [0, 1]$ compares distribution shift scores derived from comparisons of condition against reference (signal) to those derived from comparisons within conditions (noise). Larger is better.
- $\mathrm{SV}(s)$ (Figure 3d): The subsample variance $\mathrm{SV}(s) \geq 0$ (Figure 3d) quantifies robustness by evaluating the variability of a distribution shift score across different subsamples of the same condition. Smaller is better.

Using these metrics, we benchmarked scXMatch against a selection of commonly applied methods for determining global distribution shifts in scRNA-seq data. As very simple baseline, we included the number of DEGs between the two compared cell populations, with DEGs determined using edgeR [17], DESeq2 [10], and the Wilcoxon rank-sum test [30], respectively. Since edgeR and DESeq2 require bulk input, we employed a pseudo-bulk strategy by randomly grouping cells within each condition in batches of size 100. Moreover, we tested against the Augur score — a more advanced competitor approach which quantifies global distribution shifts as ML-based separability [22]. While Augur is typically used to assess cell-type-specific responsiveness to changing conditions, we here applied it in a global setting by evaluating its ability to discriminate between conditions in a single cell type.

The results of our benchmark are visualized in Figure 3e. On five of the ten datasets, scXMatch is Pareto-optimal with regard to the three metrics, outperforming all competitors. That is, scXMatch performs

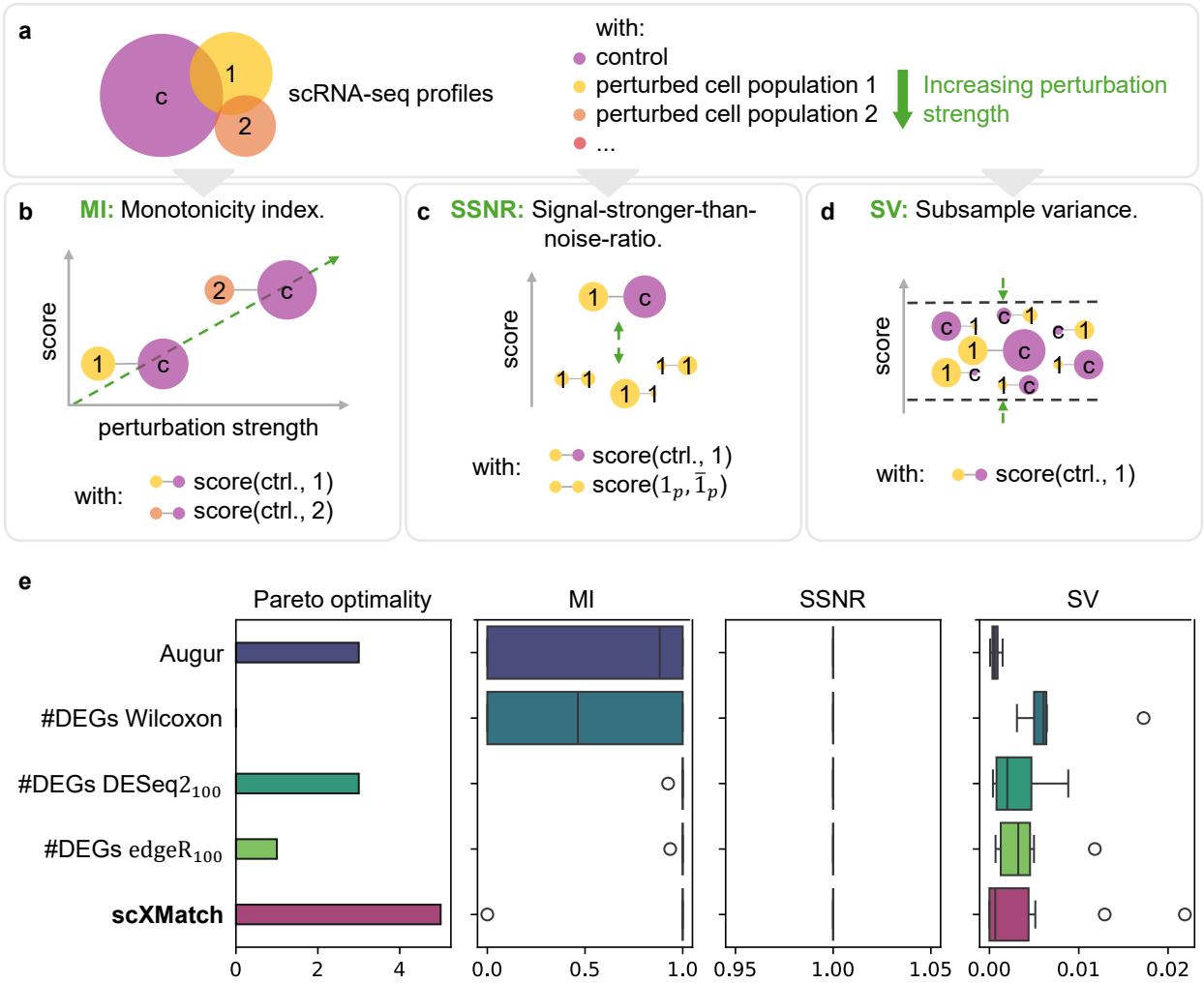

Fig. 3: Benchmarking results for scXMatch. (a) We used ten public scRNA-seq datasets from cells that were subject to perturbations of increasing strengths and defined three metrics to compare our tool scXMatch against commonly used approaches that assess distribution shifts in scRNA-seq data. The green arrows indicate the desired tendency of the respective metric. (b) MI: Comparing cell populations with increasing perturbation strength against an unperturbed reference population should lead to increasing scores. (c) SSNR: Comparing a perturbed population against the reference should lead to a higher score than comparing disjunct subsets of the perturbed population against each other. (d) SV: Comparing different subsets of the reference against different subsets of the same perturbed condition should lead to similar scores, and thus a small variance of the results. (e) Benchmarking results of the five approaches on all ten scRNA-seq datasets. For five of them, scXMatch is Pareto-optimal with regard to the three metrics (left panel).

at least as good or better than all competitors w. r. t. all three metrics. In terms of monotonicity, scXMatch achieved a perfect MI of 1.0 for nine out of ten datasets. Regarding SSNR, all tools achieved perfect scores. For SV, scXMatch produced the most stable scores five out of ten datasets (Bhattacherjee 2, McFarland 2, McFarland 3, Norman, and Schiebinger); on the others, the competitors where slightly more stable. Finally, we assessed the false positive rate (FPR) of scXMatch (for the competitor methods, we could not compute FPRs because they do not yield P-values): Out of all the 156 runs where we tested subsets of the same cell population against each other, scXMatch returned a significant raw P-value ($P < 0.05$) in 2 cases, leading to FPR $\approx 1.3\%$ for raw P-values and FPR $= 0\%$ after adjusting for multiple testing within each test group using Benjamini-Hochberg correction.

3.3 scXMatch implicitly removes outliers before assessing global distribution shifts

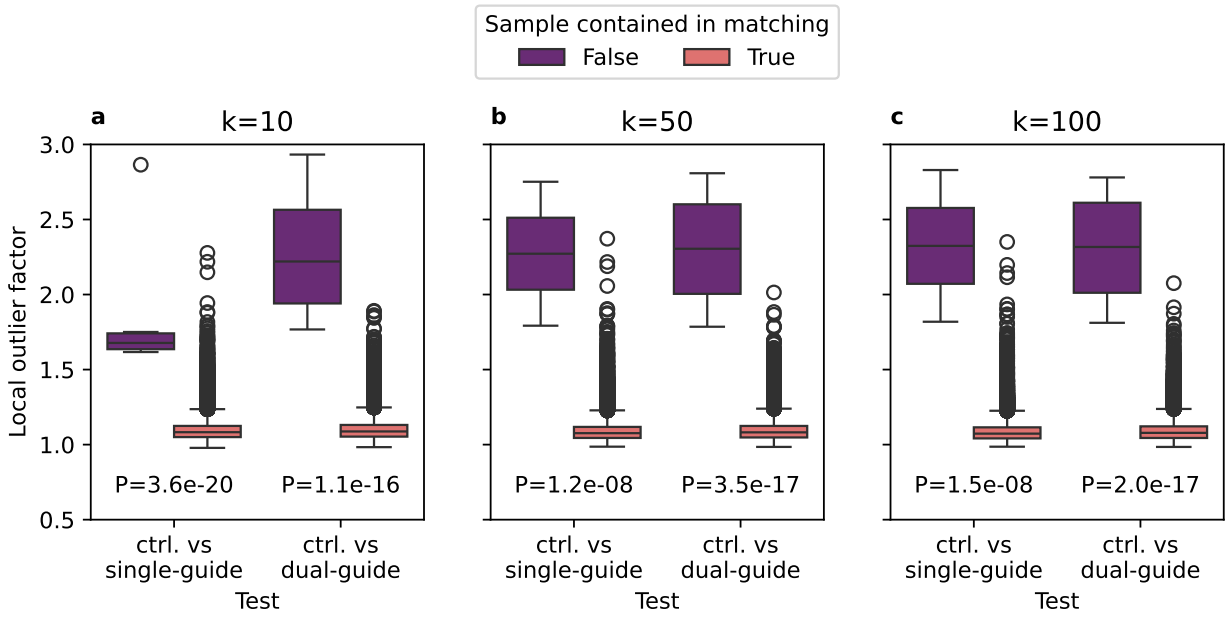

Fig. 4: Local outlier factors, colored by inclusion in the MWMCMs obtained for the Norman dataset when using scXMatch with $k \in \{10, 50, 100\}$ to compare perturbed cell populations to untreated controls. The plots are annotated with P-values of the Wilcoxon rank-sum tests (alternative hypothesis: unmatched cells have larger LOFs than matched cells).

Running scXMatch in k-NN mode can lead to a matching coverage below 100% — particularly in datasets with a large number of samples. This observation raises the question if samples that are not included in the final matching share certain characteristics. Since scXMatch constructs the matching based on a k-NN graph, samples that do not reside in densely connected regions of the distance graph are more likely to be excluded from the matching. These samples may correspond to outliers.

To systematically investigate this hypothesis, we computed local outlier factors (LOFs) [2] for each cell based on the full pairwise distance matrix, using the Norman dataset as test case, where scXMatch run in k-NN mode with $k \in \{10, 50, 100\}$ resulted in matching coverages below 100%. The LOF is a widely used density-based outlier score with LOF ≈ 1 for inliers and LOF > 1 for outliers. We then compared the distribution of LOFs between matched and unmatched cells, separately for each test group matched to its respective reference. Specifically, we performed one-sided Wilcoxon rank-sum tests with alternative hypothesis that the unmatched cells have larger LOFs than those included in the matching.

Figure 4 visualizes the LOF distributions, colored by whether or not a cell was included in the matching. Across all k and test groups, the hypothesis was supported with high statistical significance. These findings suggest that scXMatch, when run in k-NN mode, effectively performs an implicit outlier removal step by excluding cells that are not embedded within well-connected regions of the data manifold.

3.4 Changes in chromatin accessibility in scATAC-seq data identified by scXMatch reflect known T cell biology

scXMatch's utility extends to other single-cell omics modalities beyond scRNA-seq data. To demonstrate this, we analyzed scATAC-seq data from [14], which maps open chromatin regions primed for gene activation. CRISPR–Cas9 perturbations of signaling components of the T cell repector (TCR) in human CD4+ T cells revealed context-dependent chromatin dynamics and accessibility. Due to the small sample sizes, we ran

Table 1: Benjamini-Hochberg corrected P-values obtained when running scXMatch on scATAC-seq data of human T cells subjected to targeted CRISPR-Cas9 knockout, sorted by decreasing significance.

Knockout, tested against control	Adjusted P-value
CD3E	1.2×10^{-25}
CD3E + CD4	1.1×10^{-20}
ZAP70	1.9×10^{-11}
NFKB2	0.44
CD4	1.0

our tool in full graph mode. Using the squared Euclidean distances for matching computation, scXMatch effectively ranked perturbation effects, with CD3E and ZAP70 knockouts yielding the most significant P-values (see Table 1). These upstream regulators are critical for structural TCR assembly (CD3E) and signal initiation/propagation (ZAP70), respectively, and drive chromatin remodeling by activating the transcription factors NFAT, AP-1, and NFκB, which prime enhancers of activation-associated genes. Their disruption caused a global loss of chromatin accessibility, reflecting their non-redundant roles in early TCR signaling [21]. In contrast, NFKB2 deficiency showed no significant effect, consistent with its role in sustaining late-phase responses rather than initiating chromatin remodeling [26]. Similarly, CD4 knockout had only a minimal effect on chromatin accessibility, since the anti-CD3/CD28 activation model used in this study bypasses the need for CD4's typical function in MHC II recognition [4]. scXMatch hence consistently prioritizes biologically coherent hits (more significant P-values for upstream vs. downstream effectors), which underscores its value for identifying global chromatin shifts in perturbation screens.

3.5 scXMatch highlights importance of dynamic imaging for differentiating resident tissue macrophage populations

To exemplify the versatility of scXMatch beyond sequencing data, we used it to re-analyze data from our previous study on the morphology of differently stimulated resident tissue macrophages (RTMs) imaged *in vivo* in the peritoneum of mice [20]. RTMs are characterized by their dynamic changes in cell shape, as they perform constant sampling of the environment [27]. To describe those morphological characteristics, our previous study introduced a collection of 31 human-interpretable morphological RTM features, including 8 features that describe static RTM morphology and 23 features that quantify dynamic activity and adaption to stimuli (Figure 5a). Overall, the resulting dataset contains data for around 350 individual RTMs in control mice and upon stimulation with 5 different stimuli: the anti-inflammatory stimulants macrophage colony stimulating factor (MCSF) and transforming growth factor β (TGF-β), the pro-inflammatory stimulants lipopolysaccharides (LPS) and interferon γ (IFN-γ), and the kinase inhibitor YM201636 (YM), which induces a halt of the sampling activity.

While our previous study identified distribution shifts in the 31-dimensional RTM morphology profiles between the different stimuli, we now sought to answer the question if the 8 static morphology quantifiers alone are sufficient to reliably distinguish RTM populations in different activation states. To do so, we used scXMatch to carry out pairwise comparisons between the different RTM populations, using all features and the static features only, respectively. The results are shown in Figure 5b. Overall, we observe that differences become less pronounced when relying on static features only, highlighting the superiority of time-lapse *in vivo* imaging in capturing morphological RTM states over conventional imaging approaches that rely on fixated tissue samples [33,8].

The strongest loss of differences can be observed for the comparison between MCSF-stimulated and control RTMs, where adjusted scXMatch P-values increase from 0.0053 when using all 31 morphological features (strongly significant) to 0.13 when using only the 8 static features (not significant). Figure 5c shows two-dimensional UMAP projections of the corresponding 31- and 8-dimensional data. While both projections suggest that the MCSF-stimulated RTMs span a subspace of the control RTMs, it is impossible to visually assess that one distribution shift is significant while the other one is not. This exemplifies the advantages of using scXMatch for assessing global distribution shifts in high-dimensional data instead of relying on visual inspection of two-dimensional projections.

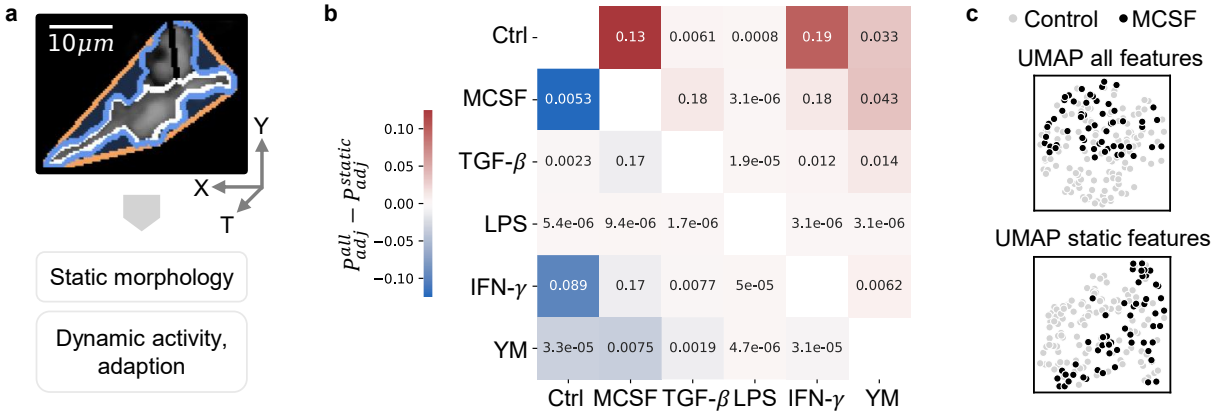

Fig. 5: Importance of dynamic imaging data for distinguishing RTM populations. (a) Snapshot of an unperturbed, sampling macrophage *in vivo* taken from [20]. Several morphological single-cell features (as schematically drawn on the image) were extracted. They characterize three categories of properties: static morphology, dynamic activity, or adaptation capabilities. While the static features can be calculated from a single image, the latter two require time-lapse images. (b) Benjamini-Hochberg-adjusted P-values returned by scXMatch for pairwise comparisons of RTM populations under different stimuli. Lower left triangle: adjusted P-values P_{adj}^{all} obtained when using all features. Upper right triangle: adjusted P-values P_{adj}^{static} obtained when using static features only. The heatmap is colored according to the difference of these two values (lower triangle minus upper triangle). Strong colors indicate large differences. Since the RTM dataset is small, we could run scXMatch in full graph mode. The standardized Euclidean distance was used as a distance measure. (c) UMAPs of the control versus MCSF-treated RTMs calculated on all and on only the static features, respectively.

4 A link between scXMatch and the two-sided Wilcoxon rank-sum test

For the special case of 1D numeric data, we establish a theoretical link between scXMatch and the Wilcoxon rank-sum test. Although we developed scXMatch for high-dimensional rather than for 1D data, we believe that this links helps readers to gain a more thorough understanding of scXMatch's mathematical foundations.

Recall that, for 1D numeric data, scXMatch's test statistic a_1 is defined as the number of cross-matches contained in a MWMCM M of a connectivity graph G defined over the elements of $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \subset \mathbb{R}$ with edge weights $w_{ij} = d(x_i, x_j)$. For our theoretical investigation, we assume the natural choice of $d(x_i, x_j)$ as the absolute difference $d(x_i, x_j) = |x_i - x_j|$. Moreover, we assume that $\mathcal{X}$ does not contain repeated elements (note that, while our proofs rely on this assumption, we conjecture that it may be possible to drop it). Under these assumptions, we show the following Proposition 1 (proof in Appendix B), which states that scXMatch's test statistic a_1 can be upper-bounded by the test statistic of the two-sided Wilcoxon rank-sum test:

Proposition 1. *Assume that $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \subset \mathbb{R}$ ($\mathcal{X}_0$ sampled from F_0, $\mathcal{X}_1$ sampled from F_1) and that $\mathcal{X}$ does not contain repeated elements. Let the distance measure d be defined as $d(x, x') = |x - x'|$, let $n = |\mathcal{X}_0|$ and $m = |\mathcal{X}_1|$ be the numbers of elements in $\mathcal{X}_0$ and $\mathcal{X}_1$, and let $U = \min\{U_0, U_1\}$ with $U_0 = |\{(x, x') \in \mathcal{X}_0 \times \mathcal{X}_1 \mid x < x'\}|$ and $U_1 = |\{(x, x') \in \mathcal{X}_0 \times \mathcal{X}_1 \mid x > x'\}| = nm - U_0$ be the test statistic of the two-sided Wilcoxon rank-sum test with alternative hypothesis $F_0 \neq F_1$. Moreover, let a_1 be the test statistic computed by scXMatch, run in full graph mode or in k-NN mode with $k \geq 2$. Then it holds that $a_1 \leq U + 2$.*

Since the P-values of both the cross-matching test implemented in scXMatch and the two-sided Wilcoxon rank-sum test are small when the corresponding test statistics a_1 and U are small, Proposition 1 implies that, for univariate numeric data without repeated elements, small Wilcoxon rank-sum test P-values force small scXMatch P-values (that is, 1D distribution shifts detectable with the Wilcoxon rank-sum test can be detected also with scXMatch).

The inverse statement, however, does not hold, as U can get large even when a_1 is small. The reason for this is that the Wilcoxon rank-sum test rejects the null hypothesis $F_0 = F_1$ only if F_0 is stochastically greater

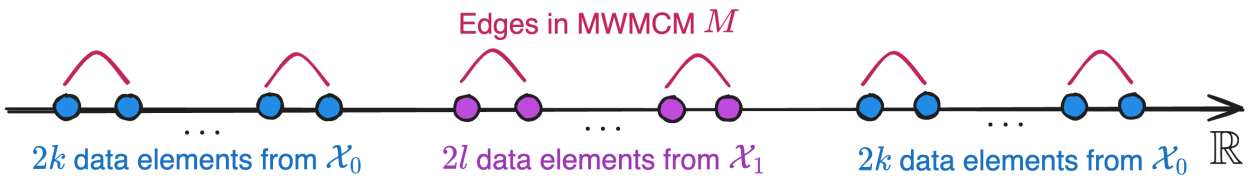

Fig. 6: 1D example where scXMatch detects a distribution shift due to differences in spread that cannot be identified with the two-sided Wilcoxon rank-sum test.

than F_1 or vice versa, whereas scXMatch can also detect distribution shifts due to differences in spread. A 2D example where scXMatch detects a distribution shift due to a spread difference is shown in Figure 1h above; Figure 6 shows a 1D example. In the arrangement visualized in Figure 6, the $4k$ numbers from $\mathcal{X}_0 \subset \mathbb{R}$ (blue dots) and the $2l$ numbers values from $\mathcal{X}_1 \subset \mathbb{R}$ (purple dots) have equal central tendencies and different spreads. The two-sided Wilcoxon rank-sum test does not detect this distribution shift but scXMatch does: Since there are no cross-matches in the MWMCM M computed by scXMatch (pink arcs), it holds that $a_1 = 0$. At the same time, the test statistic U of the Wilcoxon rank-sum test assumes the maximum possible value $U = 4kl = nm/2$: For each $x_j \in \mathcal{X}_1$, there are $2k$ elements $x_i \in \mathcal{X}_0$ with $x_i < x_j$ (implying $U_0 = 4kl$) and $2k$ elements $x_i \in \mathcal{X}_0$ with $x_i > x_j$ (implying $U_1 = 4kl$).

5 Discussion

Statistical analysis of high-dimensional data is a common task in single-cell analysis. In this study, we introduced scXMatch — the first scalable, biologically validated, and easy-to-use implementation of a statistical test designed to quantify global distribution shifts in high-dimensional data. Instead of relying on dimensionality reduction methods such as PCA or UMAP, scXMatch bases its resulting statistics on a matching calculated in high-dimensional distance space. Unlike commonly used tools for differential feature analysis or clustering-based evaluation, scXMatch only requires a single technical hyperparameter, thereby reducing the risk of confirmation bias. In comparison to approaches like Augur scores, scXMatch assesses global distribution shifts rather than separability. This seemingly subtle difference implies that it can detect distribution shifts even if one cell population lie within a subspace of the other one (thus avoiding false negatives), and that it does not over-emphasize single features which may allow to separate the groups (thus avoiding false positives).

scXMatch implements a runtime- and memory-efficient version of Rosenbaum's cross matching test [18], making use of an efficient C++-based graph matching algorithms provided by the `graph-tool` package [16]. To further enhance memory and time efficiency, the user can specify — or let the tool automatically infer — a technical hyperparameter k, so that the matching is computed on a k-NN graph instead of the full distance graph. With the default $k = 100$, scXMatch runs in under a minute for 10000 samples, and in under 10 minutes for 50000 samples, with less than 5GB total RAM required. On datasets with a high number of samples, we showed that the matching construction via the k-NN graph implicitly removes outliers.

Aside from these technical aspects, scXMatch's results cohere with expected biological effects. On nine out of ten publicly available scRNA-seq datasets from perturbation experiments, the significance of the produced P-values was perfectly correlated with perturbation strength. We further introduced two metrics that describe specificity and robustness against subsampling, and scXMatch obtained Pareto-optimal results regarding these three criteria in five out of ten scRNA-seq datasets, outperforming all commonly used approaches. Beyond scRNA-seq data, we ran scXMatch tool on scATAC-seq data, where it returned results that perfectly align with known T cell biology, and on single-cell morphodynamic data of RTMs, where its results highlight the importance of time-lapse *in vivo* imaging to capture the functional state of RTMs.

It is important to stress that the scope of scXMatch is assessing the significance of distribution shifts between experimentally determined cell populations. Beyond this intended usage, scXMatch's P-values are less reliable. For instance, scXMatch should not be used to assess the significance of clusters that are *de novo* inferred from the single-cell data. This would amount to double dipping [24], and scXMatch will thus almost always identify significant distribution shifts. Other dedicated methods exist to assess the significance of clustering results for scRNA-seq data [5]. Caution also needs to be taken when comparing the results of

two highly significant or two non-significant scXMatch tests. When two scXMatch P-values are both very small or when they are both non-significant, we observed that the order of the P-values does not always reflect the expected biological signal.

Our study opens up at least two interesting avenues for future work. Firstly, while scXMatch tests the null hypothesis that two cell populations $\mathcal{X}_0$ and $\mathcal{X}_1$ were sampled from the same distribution, a relevant adjacent question is if $\mathcal{X}_1$ was sampled from a subspace of $\mathcal{X}_0$. By incorporating outlier detection approaches to assess if all $\mathbf{x} \in \mathcal{X}_1$ are inliers w. r. t. $\mathcal{X}_0$, scXMatch could be extended to cover also this question. However, such an extension would be *ad hoc* and thus suffer from similar weaknesses as the existing indirect distribution shift quantification approaches which motivated us to develop scXMatch. Hence, an interesting challenge for future work is to design (or rediscover) a non-parametric statistical test which can be used for single-cell data and tests the null hypothesis that $\mathcal{X}_1$ was sampled from a subspace of $\mathcal{X}_0$. Secondly, the outlier removal side effect of scXMatch's k-NN graph approach could be exploited more systematically by allowing the user to specify the expected fraction r of outliers in the dataset and then setting the parameter k such that the relative coverage of the MWMCM computed on the k-NN graph is around $1 - r$. To implement such a strategy, it would be necessary to solve the (to the best of our knowledge) currently unsolved graph-theoretical question how to estimate the size of MWMCMs in k-NN graphs.

In sum, scXMatch is the first tool specifically designed to quantify global distribution shifts in single-cell data and thus fills a gap that was previously bridged with byproducts of methods originally designed for related but different purposes. Its value to the single-cell community lies in several aspects. Firstly, the tool can be used for almost any kind of single-cell data of all realistic sizes: The test itself is non-parametric, does not expect a specific distribution, and only requires the user to select a distance function to quantify cell dissimilarity. Various distance functions are supported out-of-the-box; custom distances can be provided by the user. Secondly, the concept of matching cells and then evaluating the resulting pairs is intuitively understandable even for users without statistics background. If desired, scXMatch returns the underlying matching, which enables inspection by the user and lends interpretability to scXMatch's P-values. Finally, scXMatch is part of the scverse ecosystem and thus seamlessly interfaces with popular software for single-cell data analysis such as Scanpy [31] and anndata [29]. It is distributed via the Bioconda package manager [6] and can be installed with just one line of code: `conda install scxmatch -c conda-forge -c bioconda`.

Code and data availability

- Source code of scXMatch: https://github.com/bionetslab/scxmatch.
- Easy-to-install packaged version: https://anaconda.org/bioconda/scxmatch.
- Data processing and evaluation scripts, evaluation results, all numerical data underlying the plots, and notebooks that were used to generate all plots in this manuscript: https://github.com/bionetslab/scXMatch_paper.
- Code that was used to retrieve morphodynamic features from the macrophage imaging data: https://github.com/MiriamSchnitzerlein/MacrophageMorphodynamics.
- scRNA-seq data: pertpy's (https://github.com/scverse/pertpy) data module (`data.bhattacherjee()`, `data.mcfarland_2020()`, `data.norman_2019()`, and `data.schiebinger_2019_18day()`).
- scATAC-seq data: https://doi.org/10.5281/zenodo.7058382, file `MimitouSmibert2021.zip`.
- Macrophage morphodynamics data: https:doi.org/10.5281/zenodo.13929787.

Acknowledgments. DBB was funded by the German Research Foundation (DFG, 516188180), by the German Federal Ministry of Research, Technology and Space (BMFTR, 031L0309A), and by the Klaus Tschira Stiftung (00.003.2024). SU is supported by DFG (project-IDs 448121430, 505539112, 447268119), the Hightech Agenda Bavaria, and by an ERC starting grant (project-ID 101039438). VZ acknowledges the support by the DFG (German Research Foundation); project A7 within the Research Training Group "Immunomicrotope" (GRK 2740/447268119). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

Disclosure of Interests. All authors declare no competing interests.

References

1. Bhattacherjee, A., Djekidel, M.N., Chen, R., Chen, W., Tuesta, L.M., Zhang, Y.: Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. Nat. Commun. **10**(1), 4169 (2019). https://doi.org/10.1038/s41467-019-12054-3
2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (2000). https://doi.org/10.1145/335191.335388
3. Chari, T., Pachter, L.: The specious art of single-cell genomics. PLoS Comput. Biol. **19**(8), e1011288 (2023). https://doi.org/10.1371/journal.pcbi.1011288
4. Damle, N.K., Doyle, L.V.: Stimulation via the CD3 and CD28 molecules induces responsiveness to IL-4 in CD4+CD29+CD45R-memory T lymphocytes. J. Immunol. **143**(6), 1761–1767 (1989). https://doi.org/10.4049/jimmunol.143.6.1761
5. Grabski, I.N., Street, K., Irizarry, R.A.: Significance analysis for clustering with single-cell RNA-sequencing data. Nat. Methods **20**(8), 1196–1202 (2023). https://doi.org/10.1038/s41592-023-01933-9
6. Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., Köster, J., Bioconda Team: Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat. Methods **15**(7), 475–476 (2018). https://doi.org/10.1038/s41592-018-0046-7
7. Heller, R., Small, D., Rosenbaum, P.: crossmatch: The Cross-Match Test (2024), https://CRAN.R-project.org/package=crossmatch, r package version 1.4-0
8. Kim, J., Pavlidis, P., Ciernia, A.V.: Development of a high-throughput pipeline to characterize microglia morphological states at a single-cell resolution. eNeuro **11**(7), ENEURO.0014–24.2024 (2024). https://doi.org/10.1523/ENEURO.0014-24.2024
9. Liang, X., Cao, L., Chen, H., Wang, L., Wang, Y., Fu, L., Tan, X., Chen, E., Ding, Y., Tang, J.: A critical assessment of clustering algorithms to improve cell clustering and identification in single-cell transcriptome study. Brief. Bioinform. **25**(1), bbad497 (2024). https://doi.org/10.1093/bib/bbad497
10. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
11. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
12. McFarland, J.M., Paolella, B.R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W.N., Jones, A., Chambers, E., Dionne, D., Bender, S., Wolpin, B.M., Ghandi, M., Tirosh, I., Rozenblatt-Rosen, O., Roth, J.A., Golub, T.R., Regev, A., Aguirre, A.J., Vazquez, F., Tsherniak, A.: Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat. Commun. **11**(1), 4296 (2020). https://doi.org/10.1038/s41467-020-17440-w
13. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [stat.ML] (2018). https://doi.org/10.48550/arXiv.1802.03426
14. Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.S., Yeung, B.Z., Papalexi, E., Thakore, P.I., Kibayashi, T., Wing, J.B., Hata, M., Satija, R., Nazor, K.L., Sakaguchi, S., Ludwig, L.S., Sankaran, V.G., Regev, A., Smibert, P.: Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat. Biotechnol. **39**(10), 1246–1258 (2021). https://doi.org/10.1038/s41587-021-00927-2
15. Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., Weissman, J.S.: Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science **365**(6455), 786–793 (2019). https://doi.org/10.1126/science.aax4438
16. Peixoto, T.P.: The graph-tool python library. figshare (2014). https://doi.org/10.6084/m9.figshare.1164194, http://figshare.com/articles/graph_tool/1164194
17. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010). https://doi.org/10.1093/bioinformatics/btp616
18. Rosenbaum, P.R.: An exact distribution-free test comparing two multivariate distributions based on adjacency. J. R. Stat. Soc. Series B Stat. Methodol. **67**(4), 515–530 (2005). https://doi.org/10.1111/j.1467-9868.2005.00513.x
19. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., Lander, E.S.: Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell **176**(4), 928–943.e22 (2019). https://doi.org/10.1016/j.cell.2019.01.006
20. Schnitzerlein, M., Greto, E., Wegner, A., Möller, A., Aust, O., Ben Brahim, O., Blumenthal, D.B., Zaburdaev, V., Uderhardt, S.: Cellular morphodynamics as quantifiers for functional states of resident tissue macrophages in vivo. PLoS Comput. Biol. **21**(5), e1011859 (2025). https://doi.org/10.1371/journal.pcbi.1011859
21. Shah, K., Al-Haidari, A., Sun, J., Kazi, J.U.: T cell receptor (TCR) signaling in health and disease. Signal Transduct. Target. Ther. **6**(1), 412 (2021). https://doi.org/10.1038/s41392-021-00823-w

22. Skinnider, M.A., Squair, J.W., Kathe, C., Anderson, M.A., Gautier, M., Matson, K.J.E., Milano, M., Hutson, T.H., Barraud, Q., Phillips, A.A., Foster, L.J., La Manno, G., Levine, A.J., Courtine, G.: Cell type prioritization in single-cell data. Nat. Biotechnol. **39**(1), 30–34 (2021). https://doi.org/10.1038/s41587-020-0605-1
23. Soneson, C., Robinson, M.D.: Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods **15**(4), 255–261 (2018). https://doi.org/10.1038/nmeth.4612
24. Song, D., Chen, S., Lee, C., Li, K., Ge, X., Li, J.J.: Synthetic control removes spurious discoveries from double dipping in single-cell and spatial transcriptomics data analyses. bioRxiv (2024). https://doi.org/10.1101/2023.07.21.550107
25. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., Levine, A.J., La Manno, G., Skinnider, M.A., Courtine, G.: Confronting false discoveries in single-cell differential expression. Nat. Commun. **12**(1), 5692 (2021). https://doi.org/10.1038/s41467-021-25960-2
26. Sun, S.C.: Non-canonical NF-κB signaling pathway. Cell Res. **21**(1), 71–85 (2011). https://doi.org/10.1038/cr.2010.177
27. Uderhardt, S., Martins, A.J., Tsang, J.S., Lämmermann, T., Germain, R.N.: Resident macrophages cloak tissue microlesions to prevent neutrophil-driven inflammatory damage. Cell **177**(3), 541–555.e17 (2019). https://doi.org/10.1016/j.cell.2019.02.028
28. Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Scverse Community, Berger, B., Pe'er, D., Regev, A., Teichmann, S.A., Finotello, F., Wolf, F.A., Yosef, N., Stegle, O., Theis, F.J.: The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat. Biotechnol. **41**(5), 604–606 (2023). https://doi.org/10.1038/s41587-023-01733-8
29. Virshup, I., Rybakov, S., Theis, F.J., Angerer, P., Wolf, F.A.: anndata: Access and store annotated data matrices. J. Open Source Softw. **9**(101), 4371 (2024). https://doi.org/10.21105/joss.04371
30. Wilcoxon, F.: Individual comparisons by ranking methods. In: Springer Series in Statistics, pp. 196–202. Springer New York, New York, NY (1992). https://doi.org/10.1007/978-1-4612-4380-9_16
31. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. **19**(1), 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
32. Yu, L., Cao, Y., Yang, J.Y.H., Yang, P.: Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. **23**(1), 49 (2022). https://doi.org/10.1186/s13059-022-02622-0
33. Zanier, E.R., Fumagalli, S., Perego, C., Pischiutta, F., De Simoni, M.G.: Shape descriptors of the "never resting" microglia in three different acute brain injury models in mice. Intensive Care Med. Exp. **3**(1), 39 (2015). https://doi.org/10.1186/s40635-015-0039-0