On metrics for subpopulation detection in single-cell and spatial omics data

Siyuan Luo [1,2,*], **Pierre-Luc Germain** [2,3,4,*], **Ferdinand von Meyenn** [1], **and Mark D. Robinson** [2,3,✉]

[1]Laboratory of Nutrition and Metabolic Epigenetics, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland
[2]Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[3]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[4]Institute for Neuroscience, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland
[*]These authors contributed equally to this work.

Benchmarks are crucial to understanding the strengths and weaknesses of the growing number of tools for single-cell and spatial omics analysis. A key task is to distinguish subpopulations within complex tissues, where evaluation typically relies on *external* clustering validation metrics. Different metrics often lead to inconsistencies between rankings, highlighting the importance of understanding the behavior and biological implications of each metric. In this work, we provide a framework for systematically understanding and selecting validation metrics for single-cell data analysis, addressing tasks such as creating cell embeddings, constructing graphs, clustering, and spatial domain detection. Our discussion centers on the desirable properties of metrics, focusing on biological relevance and potential biases. Using this framework, we not only analyze existing metrics, but also develop novel ones. Delving into domain detection in spatial omics data, we develop new external metrics tailored to spatially-aware measurements. Additionally, an R package, `poem`, implements all the metrics discussed.

metric | clustering | domain detection | single-cell omics | spatial omics | benchmark | desirable properties

Correspondence: *mark.robinson@mls.uzh.ch (M.D.R.)*

Introduction

A critical feature of single-cell omics data is its ability to elucidate heterogeneous cell subpopulations within complex tissues. Identifying and distinguishing these subpopulations is a fundamental task in single-cell analysis, often involving clustering cells into meaningful groups based on the similarity of their measurements. This process is typically preceded by several preprocessing steps, including normalization, feature selection, dimensional reduction, and batch correction, and the results of clustering serve as an entry point for assessing the performance of these upstream computational tasks (Duò et al., 2018; Germain et al., 2020; Luecken et al., 2022; Luo et al., 2024; Tran et al., 2020). Thus, evaluating the effectiveness of cell subpopulation identification is crucial for informing current methodological progress in single-cell data analysis.

When assessing the performance of cell subpopulation identification, researchers usually rely on external labels (the so-called ground truth) derived either from orthogonal experimental approaches (such as fluorescence-activated cell sorting, or additional data modalities in multiomics data) or from simulation. Validation metrics are then employed to quantify the degree of concordance between the clustering outcomes and these labels (also known as external clustering metrics). These metrics define the criteria for the quality of cell subpopulation identification, and therefore shapes the field's methodological development.

However, conducting reliable and biologically meaningful performance evaluation remains challenging. Firstly, despite extensive discussions on external validation indices for general clustering tasks (de Souto et al., 2012; Amigó et al., 2009; Wu et al., 2009), there are no specific guidelines for choosing metrics in the context of single-cell biology. Researchers often follow established practices from the literature that, as we will show, may not always be optimal or well-justified. Secondly, while using multiple metrics for benchmarking is generally advisable, it can lead to inconsistent rankings of methods (Luecken et al., 2022), making it difficult to draw clear conclusions. Finally, cellular heterogeneity in biology is often organized into complex hierarchical structures, but the ground-truth labels we use typically represent only one level of this hierarchy. As we will show, evaluations that expect a match strictly at this single level, without allowing for differences in resolution, can introduce bias that misrepresents the actual biological interest.

Glossary

Partition: a grouping of a set of elements into non-empty subsets, such that every element is included in exactly one subset.
Metric: a standard measure for quantification and evaluation, not to be confused with distance metrics in mathematics and geometry.
External clustering metrics: metrics that assess how well the clustering results match a predefined (e.g. ground-truth) partition of the data.
Adjusting for chance: In clustering, some degree of concordance between two partitions is expected purely by chance. Chance adjustment removes expected random concordance, ensuring that a completely random labeling yields a constant baseline value.
Meta-cell identification: A single-cell omics analysis procedure that groups similar cells and uses their averaged profiles for downstream tasks, usually to reduce noise.
"Ambient" space: here it refers to the log-normalized count matrices filtered for HVGs in scRNA-seq data, or the fragment count matrices after genomic region selection in scATAC-seq data.
Embedding space: the low- dimensional space that cells are projected into (*e.g.* the PCA space in scRNA-seq data).
Graph-based clustering: A procedure that represents a dataset as a graph, with invididual observations as nodes connected based on their similarities, community detection algorithms (*e.g.* Louvain or Leiden) are used to divide the graph into clusters.
k-Nearest Neighbors (kNN) Graph: A graph that connect nodes to its k most similar neighbors based on a distance measure.
Shared Nearest Neighbors (SNN) Graph: A graph connecting nodes by the extent to which they share their k nearest neighbors.
Set matching indices: Metrics that first solve the optimal matching problem to match or pair each cluster to a class, then calculate some evaluation measures based on that.
Fuzzy clustering: Clustering where a data point can belong to more than one clusters with different degrees of membership.

Inspired by previous research efforts to define formal constraints, or "desirable properties", for external clustering metrics (Amigó et al., 2009; de Souto et al., 2012; Arinik et al., 2021), we structure our discussion of various metrics around the key properties most relevant for single-cell biology. We seek to provide not only guidelines for selecting metrics and interpreting benchmarking results, but also a framework for developing guidelines applicable to new metrics and related computational tasks.

In the final section, we demonstrate how desirable properties and corresponding metrics can be adapted for spatial domain detection. Moreover, we show that the insights gained from clustering tasks can be applied as heuristics, and, when combined with spatial considerations, can inform the development of new desirable properties and metrics. Overall, it highlights how our framework can generate valuable guidelines not only for choosing appropriate metrics, but also for developing novel metrics that can better serve the research questions.

Understanding partition-based metrics to evaluate clustering

To distinguish from measurements at the upstream data representation level, we use the term "partition-based metrics" for external clustering metrics (See Glossary box). For clarity, we refer to the predicted partition to be evaluated as "clusters", and the ground-truth partition as "classes". The principle for defining desirable properties of clustering metrics can be summarized as follows: Given any pair of partitions of the same set, denoted as (P_1, P_2), where P_2 is assumed to be a better clustering option than P_1 according to some intuition, a metric F should adhere to the constraint that $F(P_1) < F(P_2)$ (assuming that a higher value of F indicates a better clustering) (Amigó et al., 2009). Some of these constraints and intuitions are about features of the partition that should be reflected by the metric, while others can be about biases that should *not* influence the metric.

Rather than reviewing all desirable properties in the literature (Amigó et al., 2009; de Souto et al., 2012; Arinik et al., 2021), we focus on properties that are most relevant to clustering molecular profiles of cells: *cluster homogeneity*, *class completeness*, *class size sensitivity*, and *chance agreement neutrality*. Cluster homogeneity states that clusters should not mix objects from different classes. Class completeness requires that all elements of a single class should not be distributed across different clusters. Class size sensitivity refers to how the misassignment of an element is evaluated relative to the size of the class it belongs to. We note that whether and what exact sensitivity to class size is desired depends on the context and research question. For instance, a metric might be considered desirable if it: i) is insensitive to class size (Rezaei and Fränti, 2016), ii) penalizes errors more in larger classes than in smaller ones, or iii) penalizes errors more in smaller classes than in larger ones (de Souto et al., 2012). Lastly, chance agreement neutrality means that similarities between two partitions occuring purely by chance should not be given undue credit to inflate the final similarity score (see Glossary box or Albatineh and Niewiadomska-Bugaj, 2011).

Partition-based metrics can be categorized into three groups based on their calculation: pair-counting, information theoretic and set matching measures. Extensive literature covers what they are and how they compare to each other (Rezaei and Fränti, 2016; Wagner and Wagner, 2007; de Souto et al., 2012; Amigó et al., 2009; Wu et al., 2009; van der Hoef and Warrens, 2019; Žalik and Žalik, 2011). Instead of reviewing them all here, we focus on key aspects from each category, as listed in Table 1, and examine how they capture the desirable properties (Figure 1). Such an examination can be easily extended to additional metrics.

Category	Metric	Calculation
	Rand Index (RI)	$\frac{a+d}{n(n-1)/2)}$; the ratio of the sum of true positive and true negative pairs to the total number of object pairs (Rand, 1971).
	Wallace Homogeneity (WH)	$\frac{a}{a+c}$; the ratio of the true positive pairs to the total number of object pairs that are in the same cluster in P (Wallace, 1983).
	Wallace Completeness (WC)	$\frac{a}{a+b}$; the ratio of the true positive pairs to the total number of object pairs that are in the same class in G (Wallace, 1983).
	Adjusted Rand Index (ARI)	$\frac{\text{RI}-\text{E(RI)}}{1-\text{E(RI)}} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$; adjusting RI by accounting for the expected similarity of all pairings due to chance using the Permutation Model for clusterings. ARI is the harmonic mean of AWH and AWC (Rand, 1971).
	Normalized Class Size Rand Index (NCR)	A normalized version of RI, where each concordance quantity is divided by the maximum possible concordance for their respective class (de Souto et al., 2012).
	Mutual Information (MI)	$H(G) - H(G\|P)$; the difference between the shannon entropy of G and the conditional entropy of G given P.
	Adjusted Wallace Homogeneity (AWH), Adjusted Wallace Completeness (AWC), and Adjusted Mutual Information (AMI)	Chance adjusted version of WH, WC (Wallace, 1983) and MI (Vinh et al., 2009), respectively. For a metric M, the chance adjusted version of it is $\frac{\text{M}-\text{E(M)}}{1-\text{E(M)}}$.
	(Entropy-based) Homogeneity (EH)	$1 - \frac{H(G\|P)}{H(G)}$ if $H(G,P) \neq 0$, 1 otherwise; the ratio of MI to the individual entropy of G (Rosenberg and Hirschberg, 2007).
	(Entropy-based) Completeness (EC)	$1 - \frac{H(P\|G)}{H(P)}$ if $H(P,G) \neq 0$, 1 otherwise; the ratio of MI to the individual entropy of P (Rosenberg and Hirschberg, 2007).

Category	Metric	Calculation
	V Measure (VM)	$\frac{(1+\beta)\times \text{EH}\times \text{EC}}{\beta\times \text{EH}+\text{EC}}$; the harmonic mean between EH and EC. It is identical to normalized mutual information (NMI) when arithmetic mean is used for averaging in NMI calculation (Rosenberg and Hirschberg, 2007).
	(weighted average) F Measure (wFM)	Here we calculate weighted F1-score, where the weights are based on the sizes of classes (Rezaei and Fränti, 2016).

Table 1. Partition-based metrics. Metrics are categorized according to whether they are aggregated or track a specific desirable property, and the minimal unit of calculation (cell, class/cluster, or dataset). The notation used is common throughout the table: consider comparing the predicted partition P to the ground-truth partition G; a is the number of pairs that are in the same group both in P and G; b is the number of pairs that are in the same class in G but in different clusters in P; c is the number of pairs that are in different classes in G but in the same cluster in P; d is the number of pairs that are in different groups both in P and G; n is the total number of objects; E is the expectation operator; $H(\cdot)$ is the Shannon entropy; β is the ratio of weight attributed to homogeneity vs completeness; the expectation values of RI, WH, and WC are calculated under a generalized hypergeometric model.

partition-based metric embedding-based metric graph-based metric partition-based metric with spatial information aggregated metric property-based metric element-level evaluation class/cluster-level evaluation dataset-level evaluation

Based on how a metric reflects the properties of homogeneity and completeness, we categorize them into *property-based* metrics and *aggregated* metrics. Property-based metrics focus on a specific desirable property. Metrics that measure homogeneity include Wallace Homogeneity (WH), its adjusted version (AWH) (Wallace, 1983), Mutual Information (MI), and Entropy-based Homogeneity (EH) (Rosenberg and Hirschberg, 2007). Conversely, completeness metrics contain Wallace Completeness (WC), its adjusted version (AWC), and Entropy-based Completeness (EC). Aggregated metrics integrate both homogeneity and completeness. The Rand Index (RI) and its adjusted forms (Adjusted Rand Index, ARI, and Normalized Class Size Rand Index, NCR) are classic examples, with RI representing the harmonic mean of WH and WC, ARI being the harmonic mean of AWH and AWC, and NCR (de Souto et al., 2012) a variant of ARI using a different normalization method (Table 1). Other key aggregated metrics include the V-measure (VM), which is the harmonic mean of EH and EC, and the (weighted average) F-measure (wFM) (Rezaei and Fränti, 2016), which is the harmonic mean of precision and recall (the set-matching equivalents of homogeneity and completeness).

The use of toy examples in Figure 1a demonstrates how P_1 or P_2 should be preferred over the other according to each of the desirable properties. Figure 1b-f shows the metric values for these examples, illustrating their behavior. As expected, RI, ARI, NCR, wFM and VM effectively capture both homogeneity and completeness (Figure 1b), while WH, AWH, MI, EH reflect homogeneity independently of completeness changes. However, when MI is adjusted for chance (*i.e.* AMI), it can no longer be independent of completeness. WC, AWC and EC primarily reflect completeness, with AWC also showing a slight preference for P_2 in the homogeneity example due to its chance-adjusted nature (See Glossary box), which accounts for the difficulty of achieving the same completeness with an increased number of clusters as in P_2. Similarly, EC does not assign identical values to P_1 and P_2 in the homogeneity scenario because it lacks chance adjustment, making it sensitive to the number of clusters, particularly when both the number of samples and clusters are small (Rosenberg and Hirschberg, 2007).

Regarding class size sensitivity, Figure 1b and 1e illustrate that WH, AWH, MI, EH and wFM are insensitive to class size. NCR (de Souto et al., 2012) is unique in its tendency to penalize errors in smaller classes more heavily. The remaining metrics tend to prefer errors in smaller classes over larger ones.

When it comes to chance agreement neutrality, metrics adjusted for chance, such as ARI, NCR, AWH, AWC and AMI (Figure 1b,f), effectively neutralize the influence of random agreements. In contrast, their unadjusted counterparts—RI, WC and MI—are biased, with the exception of WH. wFM seems to have some resilience to such bias. Other unadjusted metrics including EH, EC and VM, fail to remain neutral to random agreements. Interestingly, although both EC and WC measure completeness, they differ in their preference for P_1 and P_2, indicating that they can be affected differently by random agreements.

Notably, there is an inherent contradiction between the chance adjustment in metrics and a class size sensitivity that would favor the same error rate in larger classes over smaller ones. As depicted in Figure 1e, ARI gives greater advantage to P_1 than P_2 compared to RI, and similarly, AWC favors P_1 over P_2 more than WC does. Recall that both ARI and AWC are explicitly adjusted for chance. Such chance adjustment assumes that correct assignments in rare classes are more likely to happen by chance, so the adjusted score is reduced more for these

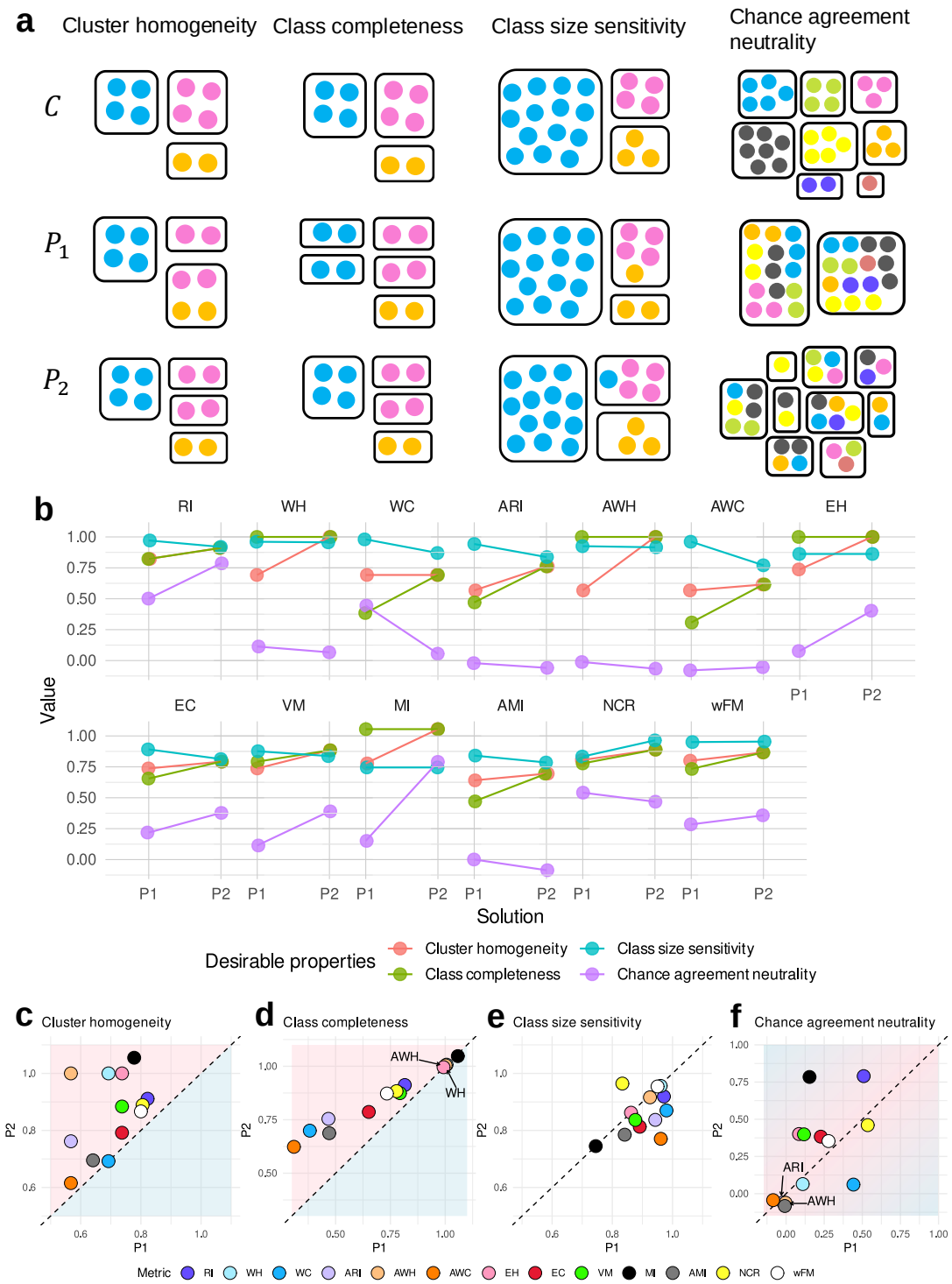

Figure 1. Desirable properties and how partition-based metrics reflect each property. Here in **a**, C represents ground truth partitions, while P_1 and P_2 are two hypothetical clustering results. In the first and second columns, P_2 should be preferred to P_1 according to the desirable properties of cluster homogeneity and class completeness, respectively. In the third column, if errors in smaller classes should be penalized more than errors in larger classes, then P_2 should be preferred over P_1, and vice versa. In the last column, P_1 and P_2 are two random partitions into different number of clusters. According to the property of chance agreement neutrality, neither P_1 nor P_2 should be preferred over the other. **b** illustrates the effectiveness of each metric in comparing P_1 to P_2 with respect to the corresponding properties. The values are calculated using the toy examples in **a**. **c**-**f** summarizes the comparisons from a property-centric perspective.

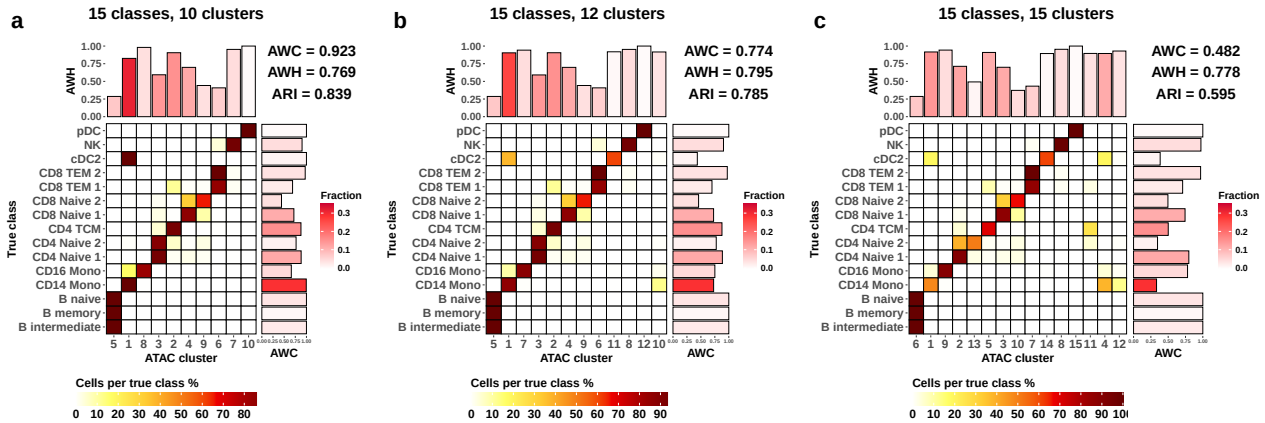

Figure 2. Real-world examples of partitions with different rankings across metrics. Here, three clustering solutions for the 10XPBMC dataset in Luo et al., 2024 are compared to the ground truth partition. The x-axis represents the predicted clusters; y-axis is the ground truth classes. The tile colors indicate the proportion of cells from the corresponding true class (each row sums up to one). A clearer diagonal structure indicates better agreement. AWH, AWC and ARI values are shown. The barplot on top shows cluster-wise AWHs (homogeneity); the barplot on the right shows class-specific AWCs (completeness). Bar color represents the proportion of cells in each cluster/class.

cases (*e.g.* P_2). However, this assumption may not always be valid in single-cell analysis, where methods that focus on major components of variability (e.g. SVD, variable feature selection, etc.) natually make it challenging to detect rare populations. While the uniform sampling assumption is often acceptable for simplicity, in studies where rare populations are highly relevant, the difficulty of characterizing smaller classes should not be ignored. Thus, although chance adjustment is generally recommended for meaningful comparisons, it can introduce bias and might not necessarily track the most desired partitioning. When the accurate identification of rare classes is crucial, metrics such as NCR should be considered.

Overall, we show here how different metrics vary in their capability to reflect various desirable properties. This sheds light on why, in many benchmark studies, rankings often vary depending on the metric (Figure 2, Luo et al., 2024; Luecken et al., 2022). The present discussion and example analyses can therefore help users understand what each metric measures, guide their usage, and lead to better benchmark interpretation.

Furthermore, we want to highlight the limitations and lack of transparency associated with using aggregated metrics including ARI (Maan et al., 2024) and VM as a universal standard for assessing clustering performance. Although ARI can provide a general sense of clustering effectiveness, it does not always correlate with the optimal clustering strategy, particularly in complex tasks. We observed that as clustering granularity increases, a noticeable trade-off emerges between homogeneity and completeness, with higher granularity favoring homogeneity at the expense of completeness (Figure S1). This phenomenon is critical in contexts such as disease-relevant cell state discovery (Figure 2) or meta-cell identification (See Glossary box). While VM provides some flexibility by allowing adjustments to the weights between EH and EC, it remains unclear how to optimally tune this weight to achieve an optimal solution.

Therefore, despite the widespread use of aggregated metrics such as ARI, we advocate for a more selective reliance on them. While such metrics are useful for broadly identifying low-performing methods, more fine-grained comparisons among top-performing methods should instead focus on property-based metrics such as MI, AWH, and AWC, as they offer clearer insights into specific aspects of clustering performance. These metrics provide an easily interpretable readout of strengths and weaknesses, allowing for a more targeted approach in method selection and evaluation.

Challenges with partition-level evaluation

Evaluating at the partition level is straightforward, but determining a partition requires selecting the number of clusters, k, or parameters that influence it, and the values of partition-based metrics vary as k varies (Wagner and Wagner, 2007; Rezaei and Fränti, 2016; see also Figure S1). Multiple strategies have been adopted to address this problem, including evaluating the metric at the "best-performing k" (Luo et al., 2024), the "true k" (based on the available ground truth labels) (Duò et al., 2018; Chen et al., 2019; Germain et al., 2020), the "detected k" (typically using default parameters) (Germain et al., 2020), or across a range of parameters that yield different k values (Luo et al., 2024). Each of these approaches, however, can introduce potential biases. For example, Figure 2a-c

show clusterings generated by the same pipeline with different k. When using the "true k" (Figure 2c), the pipeline performance is underestimated, because certain populations, such as B cell subtypes, are very similar and difficult to distinguish. Forcing the clustering to match the "true k" results in unnecessary segregation of correctly identified clusters (Figure 2c), which significantly reduces AWC (completeness) without improving AWH (homogeneity). Thus, the clusterings based on the "true k" do not always represent the most relevant solution a pipeline can generate. Selecting the "best-performing k", on the other hand, has the drawback that different metrics might not agree on the best k. This highlights the need for complementary evaluation approaches.

In addition, biological populations often exhibit a hierarchical structure (Wu and Wu, 2020), while the ground truth is typically available at a single resolution, which is not necessarily the only meaningful one. For example, within a population labeled as a single class by the ground truth, there may exist real and meaningful subpopulations. A method that identifies these subpopulations would be significantly penalized, because the partition itself does not carry any information regarding the relationships between the different classes or clusters. This issue has been discussed in Wu and Wu, 2020, where the authors proposed evaluation metrics for comparing a partition to a hierarchical ground-truth structure. As we will see, these issues can also be mitigated by evaluating at other levels of the analysis.

Evaluations on upstream data representations

To circumvent the shortcomings of partition-level evaluations or at least to complement it with other measures, a promising strategy is to evaluate *upstream* aspects of the data, such as low dimensional embeddings or graph representations. This removes the need to select a specific k or resolution to explicitly define clusters. Moreover, because these representations can preserve relationships between clusters across resolutions (i.e. two subpopulations of a class will nevertheless tend to be close to each other), evaluating graph or embedding structures helps to mitigate biases related to parameter choices and the resolution of the ground truth. As we will see, however, it is important that the assumptions underlying the representations and metrics be aligned to the relevant downstream tasks.

Below, we first extend the desirable properties of partitions to embeddings and graphs. While formulations of these properties can be found in the literature (Halkidi et al., 2015; Savic and Ivanovic, 2014), they can be understood as adaptations of the general intuitions of homogeneity and completeness. At the embedding level, these include *compactness* and *separation* of the classes; at the graph level, they are *neighborhood homogeneity* and *class compactness*. The basic intuition is that elements of a given class should be close to each other (*i.e.* compact) on the embedding or graph, and not be close to elements of other classes (*i.e.* separation or local homogeneity). Their exact meanings will vary across contexts, and we explore how different metrics, some well-established and others novel, relate to these properties.

Embedding assessment

Embedding-based metrics are often internal metrics used to evaluate clustering quality, but can also be used as external metrics using the true class labels. A cell embedding can be evaluated either directly using distance measures within the embedding space, or indirectly through secondary information derived from these distances, such as a nearest-neighbor graph. The latter will be discussed in the following section. We discuss three representative embedding metrics: the Silhouette score (Rousseeuw, 1987), the composed density between and within clusters index (CDbw) (Halkidi and Vazirgiannis, 2008), and the Density-Based Clustering Validation index (DBCV) (Moulavi et al., 2014); see Table 2 for details. They are all aggregated metrics aiming to balance compactness and separation. Notably, CDbw consists of three different components—CDbw cohesion, CDbw compactness and CDbw separation—which can be used as property-based evaluations.

Category	Metric	Calculation
	Silhouette score	$\frac{n-m}{\max(m,n)}$, where n is the mean distance between a sample and the nearest class that the sample is not a part of, and m is the mean intra-class distance (Rousseeuw, 1987).
	Composed Density between and within Clusters (CDbw)	The CDbw index consists of three components: cohesion, compactness, and separation between clusters. It uses multiple representative points selected from each cluster to calculate intra-cluster density and between-cluster distances, reflecting the geometry of the clusters and capturing changes in intra-cluster density. See Supplementary Information 1 for details (Halkidi and Vazirgiannis, 2008).
	Density Based Clustering Validation index (DBCV)	A density-based index that computes the least dense region inside a cluster and the most dense region between the clusters, to measure the within and between cluster density connectedness of clusters. See Supplementary Information 1 for details (Moulavi et al., 2014).

Table 2. Embedding-based evaluation metrics. Metrics are categorized according to whether they are aggregated or track a specific desirable property, and the minimal unit of calculation (cell, class/cluster, or dataset).

partition-based metric embedding-based metric graph-based metric partition-based metric with spatial information aggregated metric property-based metric element-level evaluation class/cluster-level evaluation dataset-level evaluation

The Silhouette score is frequently used to evaluate single cell embeddings (Luecken et al., 2022). At the dataset level, it is a ratio between the separation between classes, understood as the distance between their centroids, and the compactness of the classes, understood as variance around the centroids. As such, it implicitly assumes hyperspherical classes (Dudek, 2020; Moulavi et al., 2014). In contrast, both CDbw and DBCV indices are designed to handle arbitrarily-shaped clusters or classes. CDbw achieves this by representing each class with multiple points instead of a single one. It defines between-class separation through the minimum pairwise distance between representatives, normalized by inter-class density. Compactness, in this case, is defined based on intra-class density,

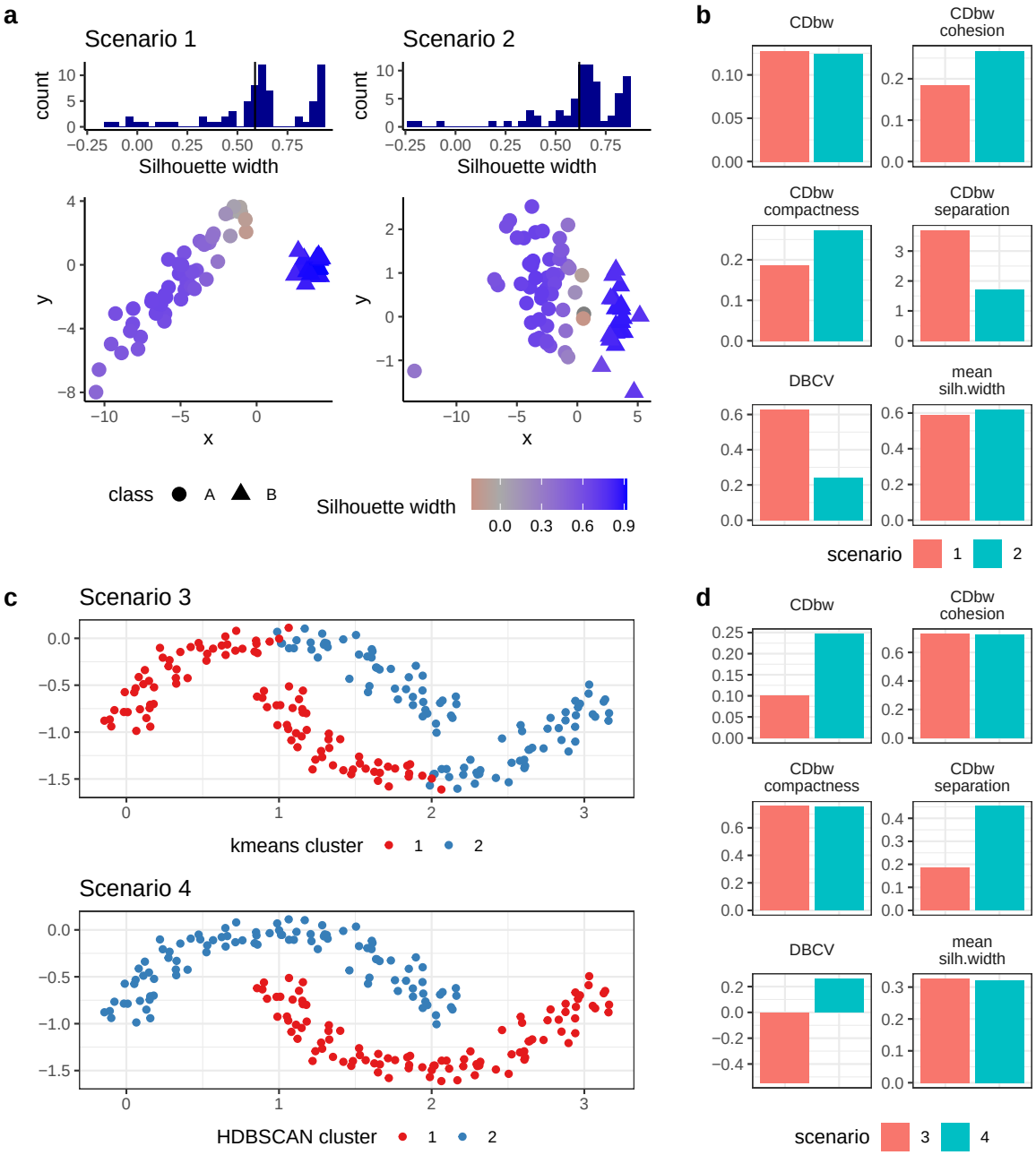

Figure 3. Examples of how class shape affects the embedding-based metrics. In **a**, both scenarios contain two classes sampled from 2D Gaussian distributions. Class A in Scenario 1 forms an elongated shape, while in Scenario 2, it is more globular. Silhouette scores for these two scenarios reveal similar distributions, despite the visually poorer separation of the two classes in Scenario 2. **b** illustrates how embedding-based metrics compare Scenario 1 to Scenario 2. In **c**, Scenario 3 and 4 highlight a moon-shaped dataset colored by clustering predictions from kmeans clustering and HDBSCAN clustering, respectively. **d** illustrates how embedding-based metrics compare Scenarios 3 and 4.

which assesses how tightly the classes' points are packed around representative points. Additionally, it introduces cohesion to measure the uniformity of intra-class density changes across representatives. DBCV also employs a notion of density, with compactness referring to the density within a class, specifically the maximum distance within the minimum spanning tree of the class, while separation corresponds to the maximum density between two classes.

In Figure 3, we use toy examples to explore how these three metrics behave on classes (clusters) of different shapes. In Figure 3a-b, both scenarios feature two classes sampled from 2D Gaussian distributions. By varying the covariance, the classes take on different shapes, with class A in Scenario 1 forming an elongated cluster, while in Scenario B, it is more globular. In Scenario 1, despite a clear separation between the two classes, many negative Silhouette scores appear due to the elongated shape of class A, resulting in an overall score similar to that of

Scenario 2 (Figure 3b), where the classes are visually not as well-separated. For CDbw, although the overall scores are similar between the two scenarios, its components clearly show that the class separation is better in Scenario 1, while compactness and cohesion are worse. This provides a more comprehensive comparison. From a density perspective, DBCV favors Scenario 1 (Figure 3b), likely because the density separation in Scenario 1 is better.

In Figure 3c-d, we highlight two scenarios with moon-shaped classes (*i.e.* representing non-convex clusters). Here, density-based clustering methods such as HDBSCAN can correctly separate the two classes, while kmeans clustering, which assumes globular clusters, does not perform well. This performance difference is clearly identified by DBCV and CDbw (Figure 3d), whereas the Silhouette score fails to capture it. Notably, CDbw highlights the separation differences between the two predictions in Figure 3c, but not their compactness differences (independently from the number of representative points used, see Figure S2).

We note that the assumptions about the underlying data structure captured by cell embeddings can vary depending on the downstream applications (Chari and Pachter, 2023; Lause et al., 2024). For instance, when 2D embeddings are used for visual cluster validation or exploring class relationships—where the global relationships and distances in the embedding space are key (Chari and Pachter, 2023)—the Silhouette, CDbw, and DBCV may provide valuable insights, as they all operate on the level of cluster relationships and incorporate the notion of distance. However, special caution is needed with the Silhouette score due to its sensitivity to class shapes, since shape can vary significantly between embeddings, and shape distortions introduced by non-linear dimensional reduction methods in single-cell data are common (Lause et al., 2024).

For downstream applications such as clustering, the assumptions used depend on the analysis workflow. Current best practice for clustering single-cell omics data (Luecken and Theis, 2019; Duò et al., 2018; Freytag et al., 2018) often involves constructing a kNN/SNN graph and applying graph-based clustering (see Glossary). This assumes that the embedding space preserves distances in local neighborhoods, but not necessarily global distances. Embedding-based metrics still rely on global distances, and because of this difference in assumptions, lower values from embedding-based metrics do not necessarily imply a poorer clustering outcome. Thus, we argue that metrics that focus on local neighborhood relationships, which we introduce in the next section, are better aligned with the clustering workflow. There might however be other downstream tasks for which obtaining more globular class representations would for instance be important, and in such a context the Silhouette score would still be valuable.

Given the demonstrated importance of local similarities in clustering, the most intuitive evaluation at the embedding level would be to examine how well the local neighborhood distances are preserved in the embedding space. However, there is typically no ground truth available for this. For instance, Euclidean distances in the high-dimensional "ambient" space (see Glossary) cannot be assumed to accurately represent the true relationships due to of the curse of dimensionality (Aggarwal et al., 2001; Lause et al., 2024). In addition, when single cell representations reside within a manifold, Euclidean geometry is only applicable locally (Moon et al., 2018).

Even in a low-dimensional embedding space, there are contexts in which Euclidean distances are inappropriate. Figure 4 shows a single-cell ATAC-seq dataset where a latent space was derived using Latent Semantic Indexing (LSI) (see Supplementary Information 1: Methods). In Figure 4a, Uniform Manifold Approximation and Projection (UMAP) was performed for visualization using the LSI embedding as input. Figure 4b shows Silhouette scores of true classes using either Euclidean or Cosine distances on the LSI space. Our observations revealed that for certain cell types (*e.g.*, vascular smooth muscle cells), despite UMAP's capability to effectively identify and separate these cells by relying on local connectivity, the Silhouette score using Euclidean distance in the LSI embedding space often resulted in a large proportion of negative scores, while using Cosine distance greatly alleviates this problem. In the scATAC-seq example, this is chiefly because of technical library size variation across cells: the deeper they are sequenced, the more they distinguish themselves from other classes, and hence the more spread out in the embedding (Figure 4c-e), creating Euclidean (but not Cosine) distances between cells that represent the same pattern, but differ in information content.

Graph structure assessment

Graph-based clustering has become increasingly popular across multiple fields due to its effectiveness in analyzing complex networks (Savic and Ivanovic, 2014; Almeida et al., 2011), including in single-cell omics data (Islam et al., 2011; Freytag et al., 2018; Duò et al., 2018). The general idea of graph-based metrics is to assess whether connections within communities are significantly denser than those with the rest of the network (Savic and Ivanovic, 2014). This assumes that the graph is locally homophilic, *i.e.* nodes with similar attributes are more likely to be connected (Almeida et al., 2011; Chen et al., 2022), which is guaranteed by the construction of the kNN/SNN graph. In addition, the construction of such graphs only assumes that distances are preserved locally in the embedding space, mitigating problems related to non-Euclidean spaces.

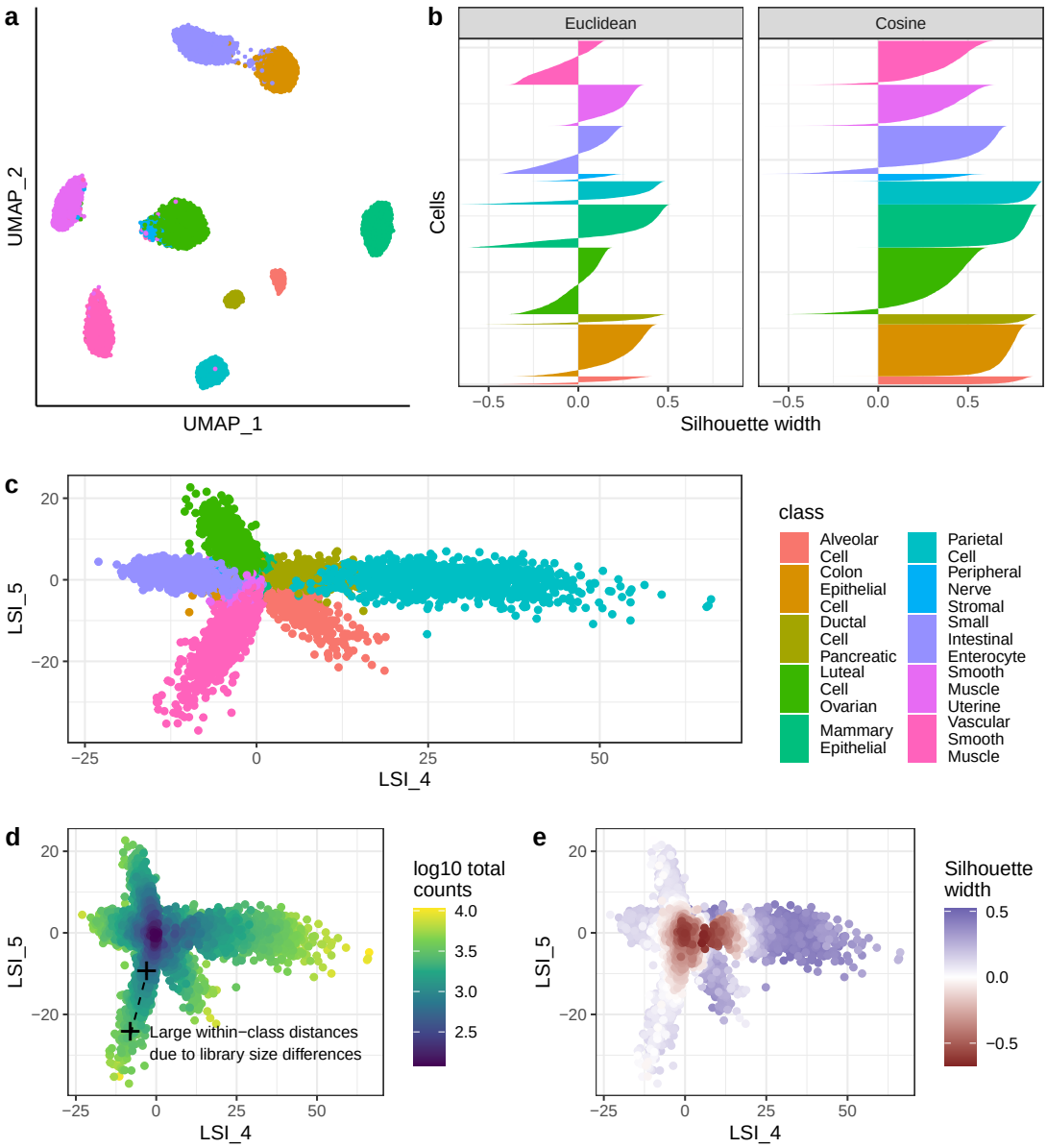

Figure 4. Real-world example of misleading Silhouette scores and Euclidean distances. Using the Atlas2 dataset from Luo et al., 2024, UMAP embeddings (**a**) show well-separated classes. However, Silhouette scores using Euclidean distance in the LSI space (left panel of **b**) show many cells with low scores; using Cosine distance strongly mitigates this effect (right panel of **b**). The LSI embedding in **c**, focusing on two LSI dimensions that best distinguish vascular smooth muscle cells—those with the highest proportion of negative Silhouettes in **a**—exhibits a star-shaped distribution (**c**). **d** shows library size for each cell, where higher values increase divergence between cell types and amplify within-class distances, collectively contributing to negative Silhouette scores (**e**).

Category	Metric	Calculation
	Modularity	For a given graph partition, it quantifies the number of edges within communities relative to what would be expected by chance. $Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$, where m is the number of edges, A is the adjacency matrix of the graph, k_i is the (weighted) degree of i, γ is the resolution parameter, and $\delta(c_i, c_j)$ is 1 if i and j are in the same community else 0 (Sueur et al., 2011).
	Local Inverse Simpson's Index (LISI)	For a given node in a weighted kNN graph, the expected number of nodes needed to be sampled before two nodes are drawn from the same class within its neighborhood (Korsunsky et al., 2019; Lütge et al., 2021).
	Neighborhood Purity (NP)	For each node in a graph, the proportion of its neighborhood that is of the same class.
	Proportion of Weakly Connected (PWC)	For a given community in a graph, the proportion of nodes that have more connections to the outside of the community than the inside of the community (Luo et al., 2024).
	Cohesion	The minimum number of nodes that must be removed to split a graph (White and Harary, 2001).
	Adhesion	The minimum number of edges that must be removed to split a graph (White and Harary, 2001).
	Adjusted Mean Shortest Path (AMSP)	For a connected graph, the average shortest paths between all node pairs, normalized by the square root of the total number of nodes. The paths are calculated in steps for either weighted or unweighted graphs. For disconnected graphs, AMSP is calculated as $\frac{\sum_i (1+m_i)}{\sqrt{N}}$, where m_i is the mean shortest path of each maximal connected subgraph, and N is the number of nodes. This metric captures the disconnectivity and spread of a graph. Note that the normalization for size is approximate, and specifically applicable for kNN graphs.
	Neighborhood Class Enrichment (NCE)	The log2 fold-enrichment of a node's class among its nearest neighbors compared to the expected proportion given its relative abundance.

Table 3. Graph-based evaluation metrics. Metrics are categorized according to whether they are aggregated or track a specific desirable property, and the minimal unit of calculation (cell, class/cluster, or dataset).

partition-based metric embedding-based metric graph-based metric partition-based metric with spatial information aggregated metric property-based metric element-level evaluation class/cluster-level evaluation dataset-level evaluation

The intuition underlying graph clustering can be linked to the desirable properties previously discussed, where *homogeneity* indicates that connections primarily occur between nodes of the same class, and *compactness* implies that any pair of nodes within a class exhibits expected similarity, reflecting the cohesiveness of the community (Table 3). Metrics for homogeneity are typically computed for the neighborhood of each node, but can be aggregated for evaluation at varying levels of granularity—per class, or across the entire graph. The simplest, most intuitive metric, which we call neighborhood purity (NP), computes the proportion of the neighborhood sharing the node's class. A diversity metric commonly used in the single-cell field, the Local Inverse Simpson's Index (LISI) (Korsunsky et al., 2019; Lütge et al., 2021), can also be inverted (i.e. the Simpson Index) to evaluate homogeneity. The Proportion of Weakly Connected (PWC) nodes (Luo et al., 2024) and modularity (Sueur et al., 2011) have also been used (See Table 3). Purity and LISI/SI operate at the node level[1], PWC at the class level, and Modularity at the graph level. Figure 5a shows the behaviour of these metrics using toy examples. While NP effectively measures the difference between graph 1 and the other two, it fails to capture the nuance in homogeneity differences between graph 2 and 3. In graph 2, pairs of classes are well distinguished from each other but completely mixed within the pair, whereas in graph 3, all four classes are partially mixed. Despite both scenarios leading to similarly low purity, the errors in graph 2 are less random than those in graph 3. This difference is captured by the LISI score. In Figure 5a, the small increase in LISI from graph 2 to 3 nicely tracks the increased class diversity (and hence decreased homogeneity) among neighborhoods that nevertheless show comparable purity. At the class level, PWC (known

[1] In the context of distinguishing populations (as opposed to measuring batch integration), the LISI/SI score of a node is less easy to interpret because it tracks dominance or diversity of the neighborhood independently of whether the neighborhood matches the node's class.

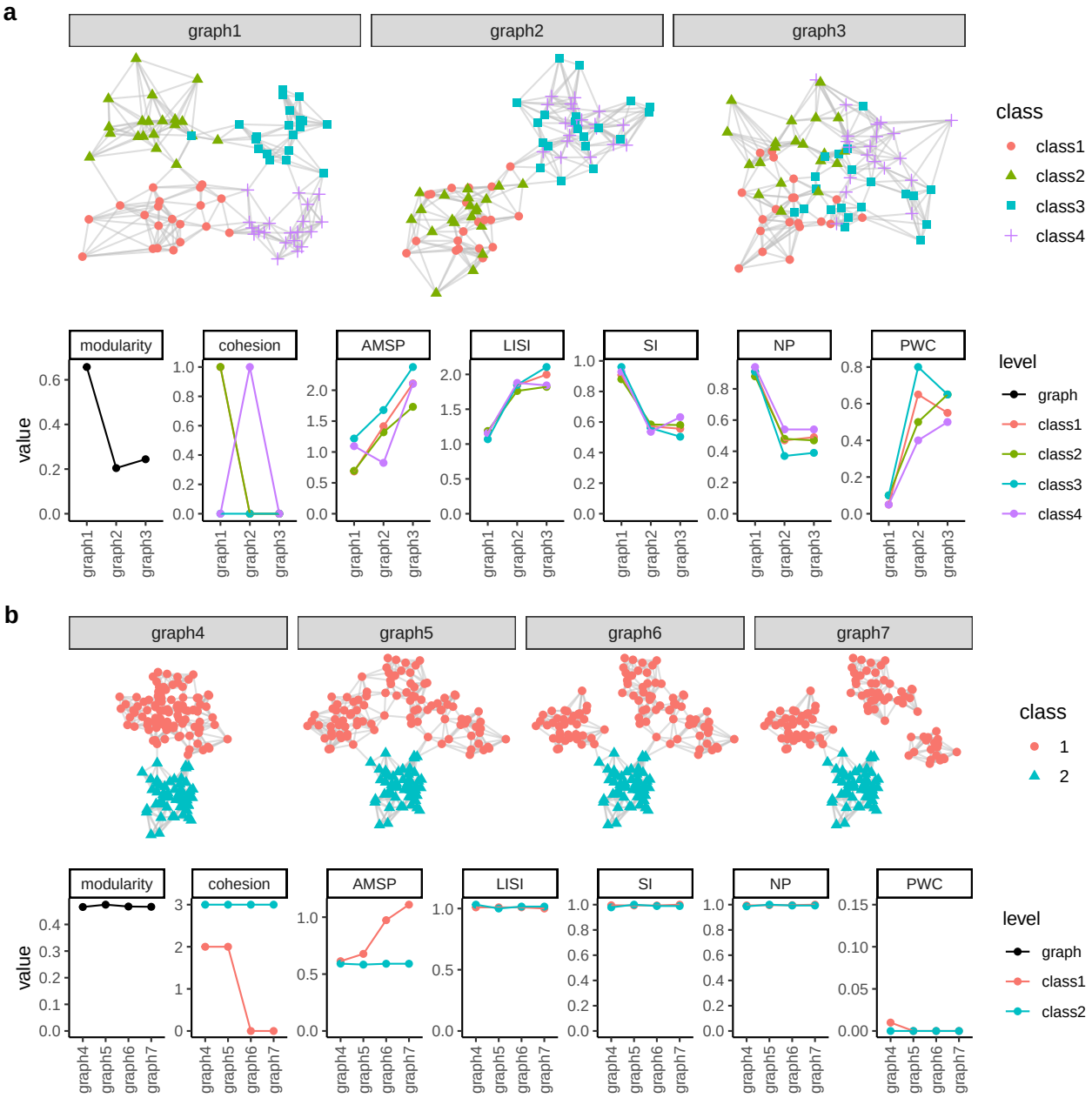

Figure 5. Example graphs and metrics computed on them. **a:** Graphs with varying levels of homogeneity and their corresponding values for different metrics. **b:** Graphs differing in class compactness of class 1, and their corresponding values for different metrics.

as Flake-ODF in Savic and Ivanovic, 2014) serves as an indicator of class separability that is especially sensitive to nodes that are at risk of being misclassified, making it a valuable measure of homogeneity (Figure S3a). In a real dataset (from Luo et al., 2024), the "B memory" and "B intermediate" subpopulations are difficult to separate during clustering (UMAP in Figure S3b and heatmap in Figure S3c), which is reflected in their large PWC score (Figure S3d). Notably, PWC remains robust across different neighborhood sizes used for constructing the SNN graph (Figure S3d). Finally, modularity is a common way to track the relationship between intra- and inter-class edges at the graph level. Of note, modularity is not limited to tracking homogeneity: because the relative weights of the edges depend on the degree of their nodes, changes in compactness can influence the modularity score. In practice, however, its focus on existing edges (as opposed to "missing" edges between elements of the same class) means that the score is mostly driven by homogeneity (see Figure 5b, "modularity").

Compactness is typically assessed at the class level, with potential to aggregate to the graph level (Table 3). Figure 5b illustrates various graphs that differ in the degree of compactness. While all graphs have a high neighborhood homogenity (as measured by LISI/SI, PWC and NP), graph 4 captures the similarity between nodes of class 1 much better than graph 7 because the subgraph of class 1 is more "compact"; class 2 is identical across all examples. Existing metrics to quantify graph compactness include graph cohesion or adhesion (White and Harary, 2001), however they are expensive to compute and therefore not scalable to large graphs. Additionally, they are limited to quantifying disconnected class subgraphs because they provide no further distinction once zero cohesion/adhesion is reached, thus making them easily influenced by a few nodes. We propose that a more robust and scalable approach is to look at the distribution of shortest paths between any two nodes for all connected components of the same class, and then aggregate over components. Here, we define a score called "adjusted mean shortest path" (ASMP), which calculates the average shortest paths between all pairs of nodes within each connected components of a class, add 1 to these values to penalize disconnectedness, and then sum across components. Since the mean shortest path in a kNN graph is linearly proportional to the number of vertices (Figure S4), we divide the value by the number of nodes to yield a metric that is less dependent on class size. This nicely captures the differences between graphs in Figure 5b.

Local homogeneity is critical to the quality of clusterings based on nearest neighbors, making the corresponding metrics highly valuable for the evaluation of (single-cell) embeddings and derived graphs. Graph compactness, which operates at the class level, affects clustering only up to a point: while discontinuities in the graph may cause improper class separation, a sufficiently connected but less compact graph is likely not to impact downstream clustering. In practice, the relevance of compactness is further complicated by the issue of ground truth resolution, since real subpopulations may exist within what is labeled as a single class. Nevertheless, class compactness may be highly relevant in tasks such as meta-cell identification (Baran et al., 2019; Persad et al., 2023), which aim at minimizing within-group variability.

Overall, graph-based metrics are particularly valuable because they avoid some of the biases that affect embedding-based metrics, including global assumptions about distances and class shape. Graph metrics instead focus on the local structure that is critical for common downstream tasks in the single-cell field. They are highly correlated to clustering homogeneity, while avoiding the problem of partition-based metrics that are highly dependent on the clustering resolution. Moreover, graph-based metrics have the advantage of being applicable, and more or less comparable, across various ways of constructing the underlying embedding and calculating distances between cells. Taken together, we believe they are the most useful to evaluate single-cell processing.

Comparison of metrics across data representations

To compare the metrics in various realistic contexts, we simulated 2240 datasets in which the number, abundance and differences between classes, as well as their spread is varied; we applied various clustering pipelines, and computed partition-, embedding- and graph-based metrics (see Supplementary Information 1: Methods). The correlation, across simulations, between pairs of metrics, and their association with specific features of the simulations, are shown in Figure S5). As expected, the major patterns of association between metrics are driven by the property they track, followed by adjustment for chance. Of note, even metrics that are adjusted for chance are significantly associated with the number of clusters called.

Domain detection in spatial (transcript)omics data

Spatially-resolved molecular measurements enhance single-cell by retaining spatial context, allowing for a more refined examination of tissue heterogeneity and cell subpopulations within physical space. One main task in the various types of spatial omics data (SOD) is the isolation of spatial homogeneous regions defined by molecular, compositional and functional similarities (Shang and Zhou, 2022; Jackson et al., 2024). So-called spatial domain detection, an analog of clustering, results in partitions of cells or spots (we refer to "spots" and array-based examples for simplicity, but the discussion also apply to other types of SOD). These are often compared to manual annotations based on immunohistochemistry and pre-defined markers, which are often taken as 'ground truth' (Yuan et al., 2024; Hu et al., 2024; Cheng et al., 2023; Liu et al., 2023a). In this context, the partition-based metrics discussed earlier can be directly applied to clustering outputs in a spatially-agnostic way. We extend this by applying the same heuristics of desirable properties, and propose new metrics that incorporate spatial information.

Category	Metric	Calculation
	Percentage of Abnormal Spots (PAS)	PAS measures the percentage of abnormal spots, which is defined as spots with a spatial domain label differing from more than half of its nearest neighbors (Shang and Zhou, 2022).
	Spatial Chaos Score (CHAOS)	CHAOS is the mean length of the graph edges in the 1-nearest neighbor (1NN) graph for each domain averaged across domains (Shang and Zhou, 2022).
	Entropy-based Local indicator of Spatial Association (ELSA)	For a site i, $E_i = E_{ai} \times E_{ci}$, where E_{ai} summarizes the dissimilarity between site i and the neighbouring sites, and E_{ci} quantifies the diversity of the categories within the neighbourhood of site i (Naimi et al., 2019) . See Supplementary Information 1 for more details.
	Spatial RI/ARI/WH/WC/ AWH/AWC	Spatial versions of the pair-sorting indices, based on fuzzy versions of the metrics. Specifically, we use the Normalized Degree of Concordance (NDC, see Hullermeier et al., 2011) and the Adjusted Concordance Index (ACI, see D'Ambrosio et al., 2021) as fuzzy versions of RI and ARI, respectively, and extend this approach to develop fuzzy versions of other metrics. In the spatial context, we first make a fuzzy representation of the true labels based on the spatial neighborhood, and then track the maximum pair concordance between the predicted labels and either the hard or fuzzy ground truth. For details, see Supplementary Information 1.
	Spot-wise Pair Concordance (SPC)	For each spot, the average concordance with all spot pairs involving that spot. This value is the same for spots sharing the same combination of cluster and class, and is especially useful for visualization. When averaged across the dataset, it corresponds to the Rand Index. A variant can be computed by excluding negative pairs (pairs discordant in both clustering and ground truth). See Figure S7a for a graphical illustration.
	Spatial SPC	Similar to the non-spatial Spot-wise Pair Concordance, but evaluates concordance in a "fuzzy" manner, following the approach used for spatial adaptations of other pair-sorting indices (see above).
	Spatially-weighted Set-matching Accuracy (SWA)	An accuracy that downweights misclassifications based on the spatial neighborhood. Instead of counting as zero in the accuracy computation, the misclassified node counts as the proportion of its spatial neighborhood that is of the node's predicted class.

Table 4. Evaluation metrics for spatial clusterings. Metrics are categorized according to whether they are aggregated or track a specific desirable property, and the minimal unit of calculation (cell, class/cluster, or dataset).

partition-based metric embedding-based metric graph-based metric partition-based metric with spatial information aggregated metric property-based metric element-level evaluation class/cluster-level evaluation dataset-level evaluation

Non-spatially-aware evaluation

Metrics such as ARI, MI, EH and EC (Table 1) have already been utilized for SOD data for domain detection (Yuan et al., 2024). Here we focus our discussion on property-based domain-level metrics, including WC, AWC, WH and AWH, highlighting their transparency and interpretability ("domain" could be either a manually-annotated class or an algorithm-derived cluster). We also introduce an element-level metric, Spot-wise Pair Concordance (SPC), which adapts the pair-counting concept to the spot-level, enhancing interpretability, particularly for visualization purposes.

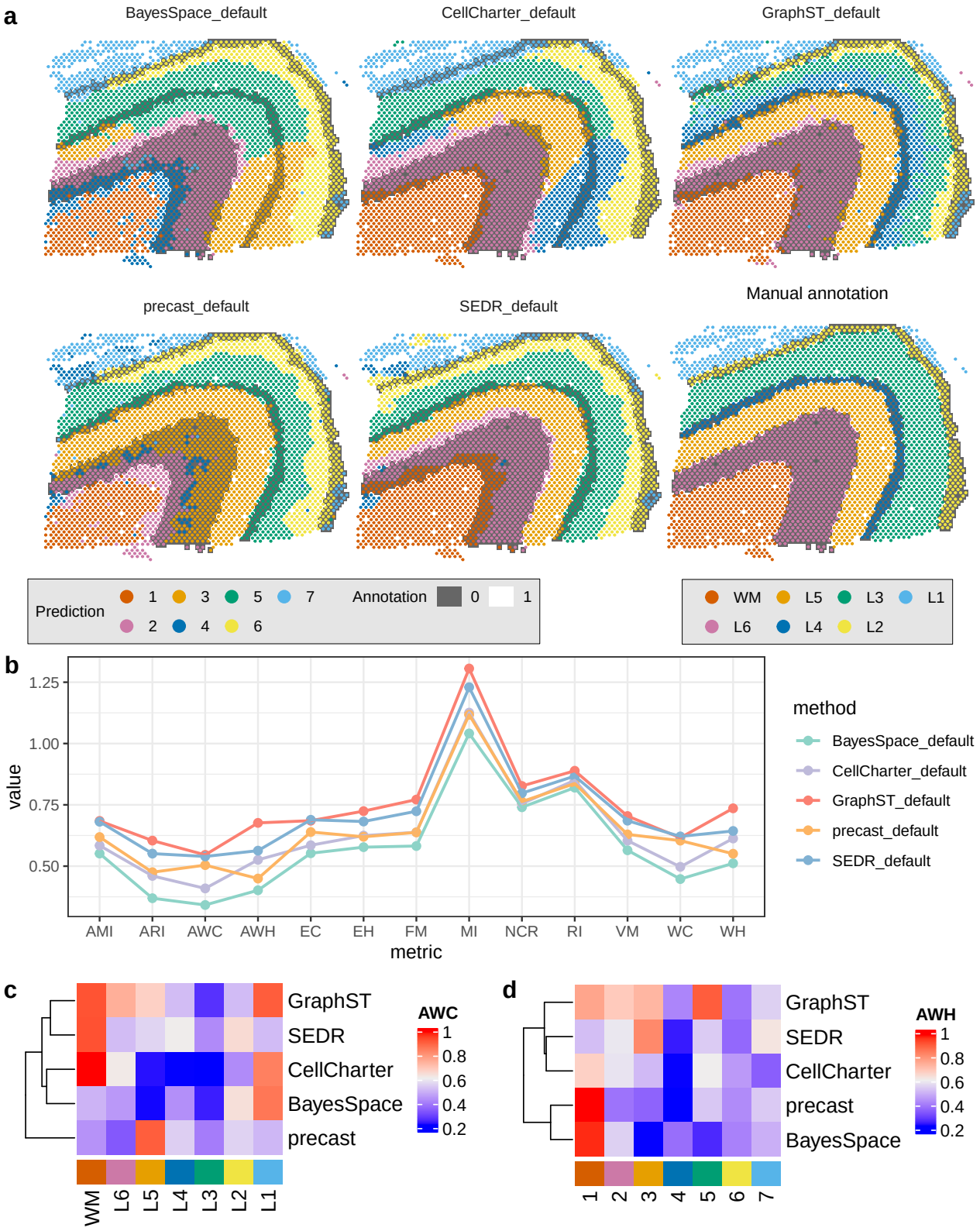

Figure 6. Partition-level metrics can be used for spatial domain detection. The LIBD 10X Visium dataset (slice 151673, Maynard et al., 2021) is used for illustration. In **a**, the last panel shows the manual annotations, while in the remaining panels, each spot is colored according to the prediction from a domain detection algorithm, with colors matched to the corresponding classes in manual annotation. The background is shaded with alternating grey and white regions, representing the annotated adjacent domains from the manual annotation. In **b**, the values of partition-based metrics for each method are calculated. **c**-**d** show the values of AWC (for each class) and AWH (for each cluster).

We apply these metrics to the LIBD human dorsolateral pre-frontal cortex spatial transcriptomics data generated with the 10x Visium platform (Maynard et al., 2021). Figure 6a shows domain detection results across five methods, along with manual annotations; all methods are set to return the "true k", and clusters are matched to classes using the Hungarian algorithm. Figure 6b presents several partition-based metrics comparing the clusters to the manual annotation. The ranking between methods are generally consistent across metrics, except for the differences between precast (Liu et al., 2023b) and CellCharter (Varrone et al., 2024). The per-domain AWC and AWH in Figure 6c provide deeper insights that are not captured by dataset-level metrics. For example, precast performs worse than CellCharter in preserving white matter (WM), while CellCharter struggles with preserving the layer-like structure in the L3, L4 and L5 classes.

Furthermore, the domain-level metrics across methods collectively reveal patterns about the manual annotations themselves. For instance, white matter, corresponding to cluster 1, consistently has high AWC and AWH scores, while class L3 shows low completeness scores. This aligns with Figure 6a, where class L3 is frequently split into multiple clusters following the shapes of different cortex layers, suggesting that this domain might exhibit heterogeneity along the axis perpendicular to the cortex layers, possibly representing a gradient layer. Lastly, cluster 4 in Figure 6d appears to be the least homogeneous, and none of the methods clearly identify class L4 of the annotation. Both precast and SEDR (Xu et al., 2024) treat cluster 4 as a miscellaneous group for spots that may be noisy or contain heterogeneous cell types, while BayesSpace (Zhao et al., 2021), CellCharter, and GraphST (Long et al., 2023) label cluster 4 as continuous areas, which are however a mixture of multiple classes based on the ground truth.

Spatially-aware evaluation

The metrics above do not use spatial information, which can be used to impose additional constraints. First, adjacent cells are more likely to share common ancestors during morphogenesis and are exposed to similar signaling environments, and therefore to be more similar (*positive spatial autocorrelation*). Second, tissue formation tends to yield continuous structures with relatively smooth boundaries (*spatial continuity*), although the extent of this continuity will vary across tissues. Third, a spot's identity is shaped not only by its own internal state, but also by its surrounding environment (*spatial contextuality*). Finally, manual annotations of spatial domains, often used as ground truth, can be uncertain, particularly near domain interfaces, where the transition between regions may be ambiguous due to spots exhibiting characteristics of multiple domains (*interface uncertainty*); this is especially the case for lower-resolution technologies.

Building on the four characteristics of SOD (above), we propose two pairs of desirable properties. The first pair is *local homogeneity* and *domain continuity*, which respectively estimate the degree of *positive spatial autocorrelation* and *spatial continuity* of SOD. They are internal evaluations, meaning they can be assessed without a ground truth. The second pair includes *neighborhood concordance*, which estimates the extent of concordance between a spot's spatial context between the predicted and reference partitions, and *interface tolerance*, which tackles *interface uncertainty*. They are external evaluations, which relate to comparisons with a ground truth.

While somewhat intuitive, the notion of domain continuity can have different meanings. On the one hand, it can simply be a higher-level implication of local homogeneity: highly homogeneous local neighborhoods will lead to smooth, continuous domains. On the other hand, there can be continuity in a strict sense with low local heterogeneity (e.g., thin striped pattern). The Spatial Chaos Score (CHAOS) (Shang and Zhou, 2022), computed at the domain level (Table 4), aims to track the latter, more specific kind of domain continuity. In contrast, two metrics are more focused on local homogeneity: the Entropy-based Local indicator of Spatial Association (ELSA) (Naimi et al., 2019) and the Percentage of Abnormal Spots (PAS) (Shang and Zhou, 2022) (Table 4). Inspired by Local Indicators of Spatial Association (LISA, Anselin, 1995), ELSA is evaluated at the spot level using categorical variables, but can be aggregated at the domain or dataset level. PAS summarizes at the domain or dataset-level the proportion of spots that are sufficiently inconsistent with their neighborhood (as such, it is analogous to the PWC graph metric). As shown in Figure 7b, such abnormal spots often highlight areas with non-smooth domain boundaries. In contrast, ELSA assigns each spot a continuous value between 0 and 1, where 0 indicates a homogeneous neighborhood (Figure 7d). ELSA highlights boundary regions, with higher values signifying higher diversity. Both PAS and ELSA can be computed across various neighborhood sizes, but are rather robust to it (Figure 7b-d, Figure S6).

We also generated toy examples in Figure 8a to investigate further the various metrics. For internal evaluation (Figure 8b), all three metrics correctly identify P3 as less smooth than P1 or P2, albeit with different dynamic ranges. While CHAOS gives the same scores to P1 and P2 due to the lack of discontinuity, PAS and ELSA consider P1 smoother, asexpected. Because these metrics are ground-truth-agnostic, the pattern is similar when comparing P4 and P5.

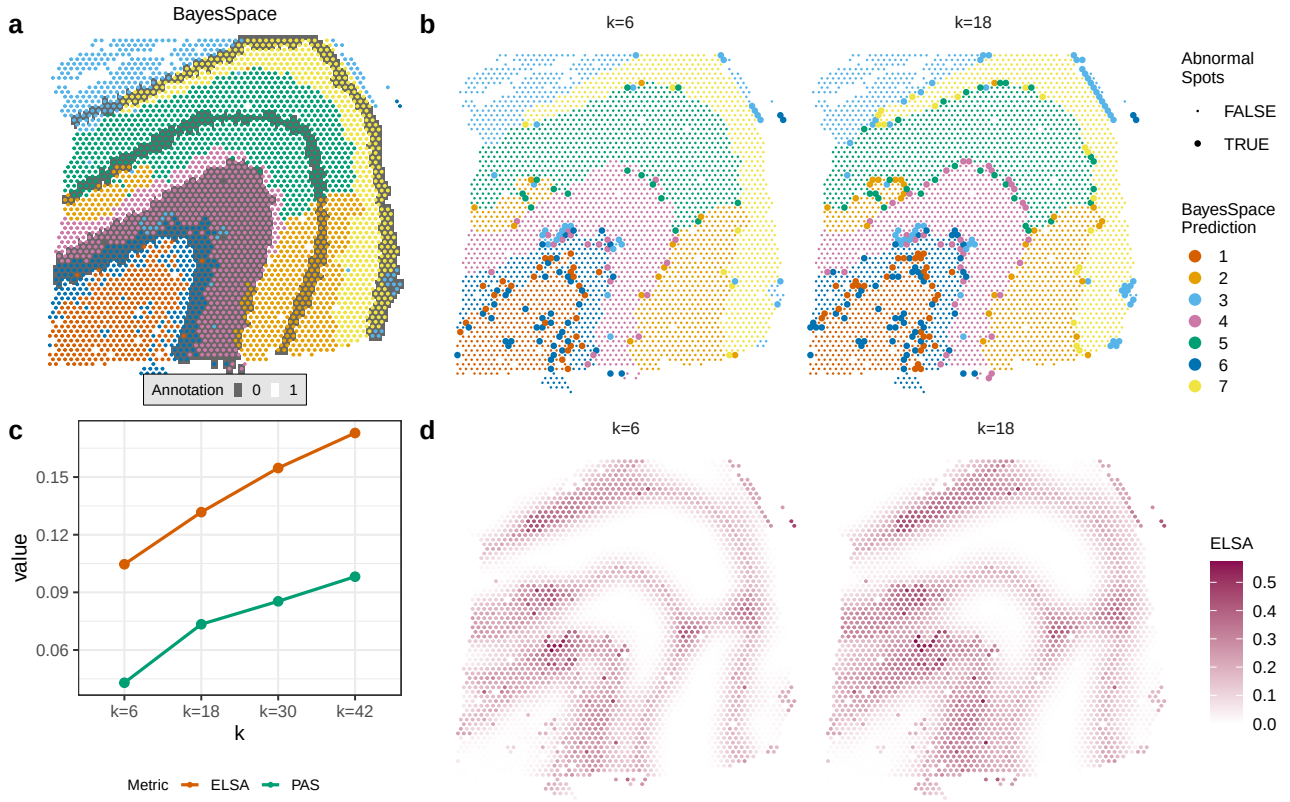

Figure 7. Metrics for spatial clustering evaluation and visualization. The same LIBD 10X Visium dataset as in Figure 6 is used. **a** is identical to the BayesSpace subpanel in Figure 6a, where the border is colored by the predicted domain, and the background is shaded using the ground truth. In **b**, spot-level metric values related to PAS are shown, indicating whether a spot is classified as abnormal or not. **c** displays how the dataset-level ELSA and PAS score change as the size of the neighborhood (k) varies. **d** shows element-level ELSA values. The same visualization for larger k are shown in Figure S6.

To the best of our knowledge, there is currently no existing external partition-based metrics that take the spatial information into account. To include this spatial dimension into partition-based evaluation, we propose two strategies.

The first approach is based on set matching (see Glossary), which treats the domain prediction as a classification task and computes a weighted accuracy score. Traditional accuracy assesses the binary agreement between each spot's predicted and true labels. In contrast, our strategy evaluates the agreement between a spot's predicted label and the true labels of both the spot and its neighborhood, capturing spatial relationships. We refer to this as the spatially-weighted set-matching accuracy (SWA, see Table 4). As shown in Figure 8c and d, SWA correctly reflects the property of interface tolerance, as misclassified spots at the domain boundary are assigned to the adjacent domain, reducing the penalties for such spots. Since P1 contains more interface spots than P2 and P3, it has higher set-matching accuracy. P2 and P3 further demonstrate that when the number of interface misclassifications is the same, SWA remains consistent, regardless of the distribution of other misclassifications within the domain. This confirms that SWA effectively captures interface tolerance in a desirable way, while not expecting more smoothness than is present in the ground truth (Figure 8c-d). In contrast, internal metrics in Figure 8b assign P4 a smoother score than P5, ignoring the non-smooth nature of GT2. Despite being simple and easy to interpret, this SWA has two major limitations: it requires matching clusters to classes, which can be ambiguous, and it is not adjusted for chance (i.e. expected agreement based on random allocations).

Another strategy we propose is to make either or both the clustering and the ground truth fuzzy (see Glossary) by averaging a spot's identity with that of its neighborhood, and then employ metrics for fuzzy clustering evaluation (Hullermeier et al., 2011; D'Ambrosio et al., 2021). Building on the permutation-based fuzzy version of the ARI (D'Ambrosio et al., 2021), we developed a fuzzy version of the (adjusted) Wallace indices (*i.e.* WH/AWH and WC/AWC (see Table 4), so that all pair-sorting partition metrics can be computed in a spatially-aware fashion (Figure 8c-d). In order to not penalize a clustering that would resemble the "hard" ground truth better than the fuzzy one, we also implemented a fuzzy-hard version that considers the maximum pair concordance with either the hard or fuzzy ground truth, which we term "spatial" versions of the pair-counting metrics. As shown in Figure 8c, the spatially-aware metrics all correctly prefer P1 to P2 and P3, and P5 to P4, thus displaying interface tolerance.

To better understand the metric behavior with fuzzy representations, let us look at domain-level metric values. When interpreting SpatialWC and SpatialAWC in Figure 8, one should consider the values for each *ground-truth* domain (*i.e.*, class), as shown in Figure S8. In Figure 8a, for spots of the red class, although P1-P3 misclassify the same number of spots into the blue cluster, the completeness of the red class in P1 is higher than in P2 and P3. Because there are more misclassified spots in P1 that are from the interface of red and blue classes that contain partial blue memberships, the error is considered less severe. On the contrary, when interpreting SpatialWH and SpatialAWH in Figure 8, one should consider the values for each *predicted* domain (*i.e.*, cluster), as shown in Figure S8. In Figure 8a, for spots of the blue cluster, although P1-P3 incorrectly mix the same number of spots from the red class into the blue cluster, these red spots in P1 are all on the interface between the blue and red classes in GT1, and thus have partial membership of the blue class, making such errors less penalized than the errors in P2 and P3. For P2 and P3, the number of misclassified spots at the class interface are identical, and thus the spatial metrics have identical values (within the range of variation of the permutations for adjusted metrics).

Finally, for element-wise external evaluation and visualization, we developed the Spot-wise Pair Concordance score (SPC) (Table 4). It can incorporate either hard-hard or fuzzy-hard representations of the truth and prediction, making it flexible for either non-spatially-aware or spatially-aware evaluation. This is useful to spatially visualize the different degrees of disagreement between a predicted domain partition and a ground truth without the need for set matching (Figure S7).

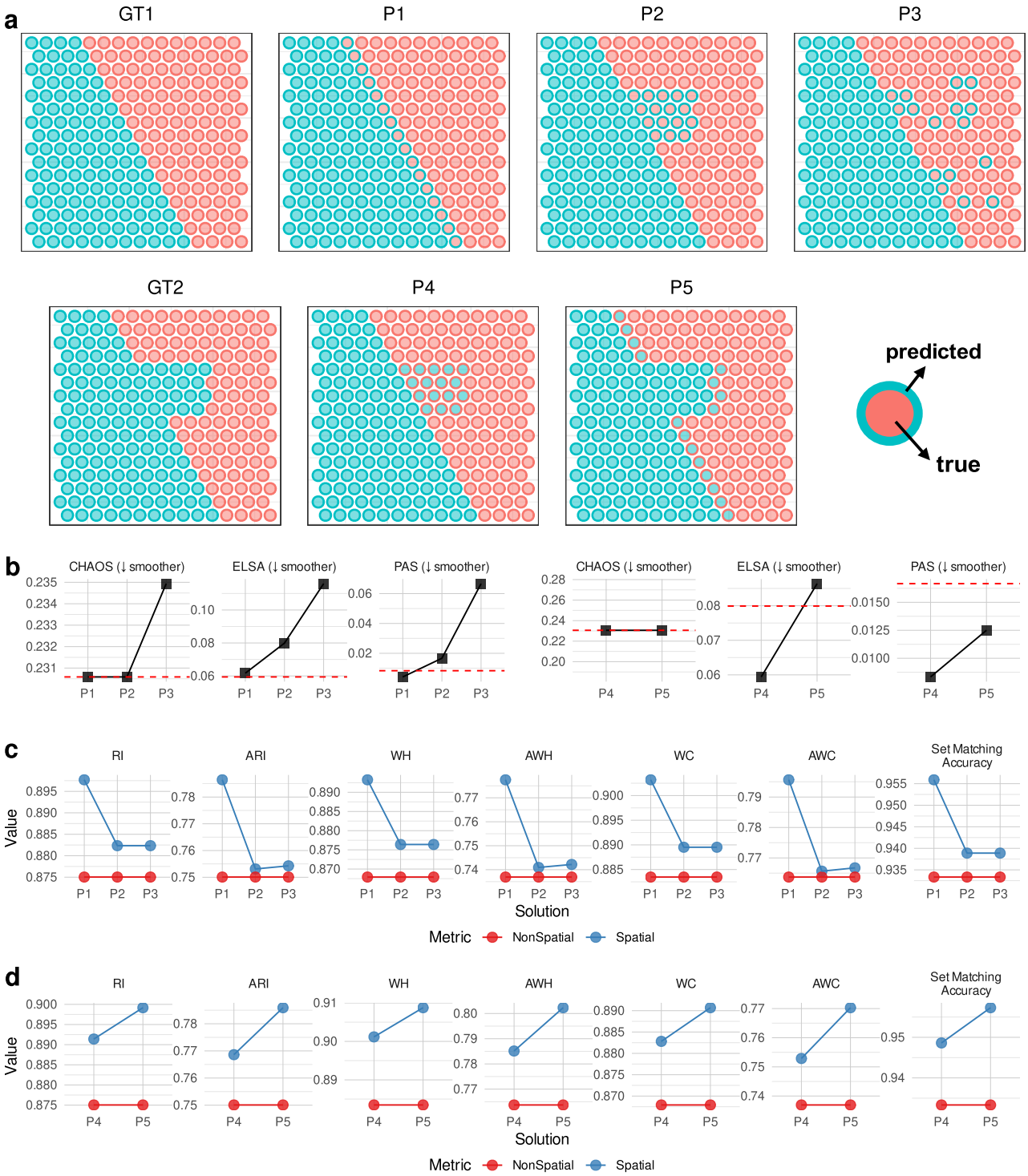

Figure 8. Examples of desirable properties for domain detection, and how different metrics reflect each property. In **a**, the colors within each spot represent the ground truth labels, while the colors around the spot borders indicate the predicted labels. GT1 and GT2 denote two datasets colored by the ground truth domain partition, with P1-P3 showing hypothetical domain detection results based on GT1, and P4 and P5 based on GT2. The number of misclassified spots remains the same across P1 to P3, and also across P4 and P5, but their spatial distribution varies. **b-d** demonstrate how each metric compares P1 with P2 and P3, and compares P4 with P5, respectively. **b** shows internal metrics (red dashed lines indicate the values for GT1 and GT2), while **c** and **d** are external metrics. For each external metric, both the spatially-aware and non-spatially-aware versions are calculated.

Discussion

Evaluation metrics play a crucial role in shaping our understanding of the strengths and weaknesses of computational methods, guiding their development, and ultimately influencing our interpretation of biology. However, the current practice of selecting evaluation metrics in the single-cell omics field often lacks a clear and rational basis. In this work, we leverage the concept of desirabe properties to gain a deeper understanding of the behavior of each metric, offering a structured guideline for rational metric selection.

To illustrate and explore these desirable properties, we make extensive use of carefully designed toy examples. These examples, while simple, offer clear insights that can be difficult to extract from real-world datasets, where multiple factors are often entangled. Despite their simplicity, these example demonstrate what does or does not influence a metric, thereby helping the interpretation in a way that extends to real-world datasets.

Given that each metric captures different desirable properties and functions at different evaluation levels, we generally advocate the use of property-based metrics and those that operate at finer levels, such as class- or spot-level metrics, for fine-grained method comparison. As demonstrated in the clustering examples and domain detection examples, they offer greater transparency and interpretability compared to aggregated, dataset-level metrics.

We extend the most relevant desirable properties from partitions to embeddings to graphs, enabling a similar and structured evaluation across these levels, and emphasize that desirable properties should be carefully considered within the biological context of a study. By examining the relationships and trade-offs between different properties, we aim to help researchers select metrics tailored to their biological questions and the complexities of different datasets. This approach aids in developing useful metrics, as well as in addressing the challenge of interpreting benchmarking results that involve multiple metrics and datasets with inconsistent rankings of methods.

By understanding metrics from the perspective of desirable properties, researchers can more precisely assess the validity of the underlying assumptions. In our discussion, we advocate evaluating single-cell data from a connectivity standpoint at the graph level, as it requires fewer global assumptions than embedding-based evaluations, and has proven to be highly effective in capturing statistical patterns in single-cell datasets. In addition, it is preferable over partition-based evaluation because it mitigates issues caused by misaligned levels of biological granularity between the prediction and ground truth. To facilitate this practice, we introduce novel metrics–PWC and AMSP–that align with this perspective. An implementation of all of the metrics discussed here (and more), including the novel metrics and fuzzy adaptations, are available in our R package `poem` (https://github.com/RoseYuan/poem).

In addition, we extend our work to the problem of spatial domain detection in SOD. The evaluation of domain detection methods has often relied on internal metrics or external metrics that do not account for spatial relationships. To address this, we developed spatially-aware metrics for external evaluation. While we tested our external metrics primarily on array-based SOD, in principle they are applicable to non-array-based SRT datasets by defining neighborhood structures using physical distances between cell centers or boundaries.

Overall, our work emphasizes the contribution of a systematic approach to evaluation metrics grounded in desirable properties. This method prioritizes biological intuition and concerns, rather than uncritically adopting metrics from other fields. In our analysis of spatial domain detection, we demonstrate how different data modalities introduce additional biological considerations, from which we derive new desirable properties. This showcases the strength of our framework not only in providing guidelines but also in generating new principles and inspiring the development of novel metrics.

Acknowledgements

We wish to acknowledge Robinson Lab members for fruitful discussion and the core SpaceHack 2023 team (https://spatialhackathon.github.io/past.html), especially Jieran Sun, Peiying Cai, and Sarusan Kathirchelvan, for discussions and sharing partitions from methods run on the LIBD dataset.

Code availability

All the metrics we mentioned are implemented in our R package `poem` (https://github.com/RoseYuan/poem). Notebooks, R scripts, and supporting data used for reproducing the analysis and generating the visualizations in this manuscript are available at https://github.com/RoseYuan/metric_paper.

Bibliography

C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer, 2001.

A. N. Albatineh and M. Niewiadomska-Bugaj. Correcting jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification*, 5:179–200, 2011.

H. Almeida, D. Guedes, W. Meira, and M. J. Zaki. Is there a best quality metric for graph clusters? In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, pages 44–59. Springer, 2011.

E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12:461–486, 2009.

L. Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.

N. Arinik, V. Labatut, and R. Figueiredo. Characterizing and comparing external measures for the assessment of cluster analysis and community detection. *IEEE Access*, 9:20255–20276, 2021.

Y. Baran, A. Bercovich, A. Sebe-Pedros, Y. Lubling, A. Giladi, E. Chomsky, Z. Meir, M. Hoichman, A. Lifshitz, and A. Tanay. Metacell: analysis of single-cell rna-seq data using k-nn graph partitions. *Genome biology*, 20:1–19, 2019.

T. Chari and L. Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, 2023.

H. Chen, C. Lareau, T. Andreani, M. E. Vinyard, S. P. Garcia, K. Clement, M. A. Andrade-Navarro, J. D. Buenrostro, and L. Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20(1):1–25, 2019.

H. Chen, Z. Yu, Q. Yang, and J. Shao. Community detection in subspace of attribute. *Information Sciences*, 602:220–235, 2022.

A. Cheng, G. Hu, and W. V. Li. Benchmarking cell-type clustering methods for spatially resolved transcriptomics data. *Briefings in bioinformatics*, 24(1):bbac475, 2023.

M. C. de Souto, A. L. Coelho, K. Faceli, T. C. Sakata, V. Bonadia, and I. G. Costa. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *2012 Brazilian symposium on neural networks*, pages 49–54. IEEE, 2012.

A. Dudek. Silhouette index as clustering evaluation tool. In *Classification and Data Analysis: Theory and Applications 28*, pages 19–33. Springer, 2020.

A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

A. D'Ambrosio, S. Amodio, C. Iorio, G. Pandolfo, and R. Siciliano. Adjusted concordance index: an extensionl of the adjusted rand index to fuzzy partitions. *Journal of Classification*, 38:112–128, 2021.

S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo. Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7, 2018.

P.-L. Germain, A. Sonrel, and M. D. Robinson. pipecomp, a general framework for the evaluation of computational pipelines, reveals performant single cell rna-seq preprocessing tools. *Genome biology*, 21:1–28, 2020.

M. Halkidi and M. Vazirgiannis. A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29(6):773–786, 2008.

M. Halkidi, M. Vazirgiannis, and C. Hennig. Method-independent indices for cluster validation and estimating the number of clusters. *Handbook of cluster analysis*, 26:595–618, 2015.

Y. Hu, M. Xie, Y. Li, M. Rao, W. Shen, C. Luo, H. Qin, J. Baek, and X. M. Zhou. Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biology*, 25(1):212, 2024.

E. Hullermeier, M. Rifqi, S. Henzgen, and R. Senge. Comparing fuzzy partitions: A generalization of the rand index and related measures. *IEEE Transactions on Fuzzy Systems*, 20(3):546–556, 2011.

S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.

K. C. Jackson, A. S. Booeshaghi, Á. Gálvez-Merchán, L. Moses, T. Chari, A. Kim, and L. Pachter. Identification of spatial homogeneous regions in tissues with concordex. *bioRxiv: the preprint server for biology*, pages 2023–06, 2024.

I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

J. Lause, P. Berens, and D. Kobak. The art of seeing the elephant in the room: 2d embeddings of single-cell data do make sense. *bioRxiv*, 2024.

T. Liu, Z.-Y. Fang, Z. Zhang, Y. Yu, M. Li, and M.-Z. Yin. A comprehensive overview of graph neural network-based approaches to clustering for spatial transcriptomics t. liu et al. overview of spatial transcriptomics' spatial clutering. *Computational and Structural Biotechnology Journal*, 2023a.

W. Liu, X. Liao, Z. Luo, Y. Yang, M. C. Lau, Y. Jiao, X. Shi, W. Zhai, H. Ji, J. Yeong, et al. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with precast. *Nature communications*, 14(1):296, 2023b.

Y. Long, K. S. Ang, M. Li, K. L. K. Chong, R. Sethi, C. Zhong, H. Xu, Z. Ong, K. Sachaphibulkij, A. Chen, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst. *Nature Communications*, 14(1):1155, 2023.

M. D. Luecken and F. J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6): e8746, 2019.

M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

S. Luo, P.-L. Germain, M. D. Robinson, and F. von Meyenn. Benchmarking computational methods for single-cell chromatin data analysis. *Genome Biology*, 25(1):1–30, 2024.

A. Lütge, J. Zyprych-Walczak, U. B. Kunzmann, H. L. Crowell, D. Calini, D. Malhotra, C. Soneson, and M. D. Robinson. Cellmixs: quantifying and visualizing batch effects in single-cell rna-seq data. *Life science alliance*, 4(6), 2021.

H. Maan, L. Zhang, C. Yu, M. J. Geuenich, K. R. Campbell, and B. Wang. Characterizing the impacts of dataset imbalance on single-cell data integration. *Nature Biotechnology*, pages 1–10, 2024.

K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich,

M. Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.
K. R. Moon, J. S. Stanley III, D. Burkhardt, D. van Dijk, G. Wolf, and S. Krishnaswamy. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.
D. Moulavi, P. A. Jaskowiak, R. J. Campello, A. Zimek, and J. Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM, 2014.
B. Naimi, N. A. Hamm, T. A. Groen, A. K. Skidmore, A. G. Toxopeus, and S. Alibakhshi. Elsa: Entropy-based local indicator of spatial association. *Spatial statistics*, 29:66–88, 2019.
S. Persad, Z.-N. Choo, C. Dien, N. Sohail, I. Masilionis, R. Chaligné, T. Nawy, C. C. Brown, R. Sharma, I. Pe'er, et al. Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology*, 41(12):1746–1757, 2023.
W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336): 846–850, 1971.
M. Rezaei and P. Fränti. Set matching measures for external cluster validity. *IEEE transactions on knowledge and data engineering*, 28(8):2173–2186, 2016.
A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.
P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
M. Savic and M. Ivanovic. Graph clustering evaluation metrics as software metrics. In *SQAMIA*, pages 81–89, 2014.
L. Shang and X. Zhou. Spatially aware dimension reduction for spatial transcriptomics. *Nature communications*, 13(1):7203, 2022.
C. Sueur, O. Petit, A. De Marco, A. T. Jacobs, K. Watanabe, and B. Thierry. A comparative network analysis of social style in macaques. *Animal Behaviour*, 82(4):845–852, 2011.
H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.
H. van der Hoef and M. J. Warrens. Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, 46(2):353–370, 2019.
M. Varrone, D. Tavernari, A. Santamaria-Martínez, L. A. Walsh, and G. Ciriello. Cellcharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature Genetics*, 56(1):74–84, 2024.
N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080, 2009.
S. Wagner and D. Wagner. Comparing clusterings: an overview. 2007.
D. L. Wallace. A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.
D. R. White and F. Harary. The cohesiveness of blocks in social networks: Node connectivity and conditional density. *Sociological Methodology*, 31(1):305–359, 2001.
J. Wu, J. Chen, H. Xiong, and M. Xie. External validation measures for k-means clustering: A data distribution perspective. *Expert Systems with Applications*, 36(3):6050–6061, 2009.
Z. Wu and H. Wu. Accounting for cell type hierarchy in evaluating single cell rna-seq clustering. *Genome biology*, 21(1):123, 2020.
H. Xu, H. Fu, Y. Long, K. S. Ang, R. Sethi, K. Chong, M. Li, R. Uddamvathanak, H. K. Lee, J. Ling, et al. Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Medicine*, 16(1):12, 2024.
Z. Yuan, F. Zhao, S. Lin, Y. Zhao, J. Yao, Y. Cui, X.-Y. Zhang, and Y. Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, 2024.
K. R. Žalik and B. Žalik. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, 32(2):221–234, 2011.
E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. Taylor, P. Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.

Supplementary Information 1: Methods

A. Datasets

A.1. The 10XPBMC dataset used in Figure 2 and Figure S3. This dataset is the scATAC-seq part of the 10XPBMC multiomics datasets used in Luo et al. (2024), and the details about its generation, quality control (QC), and preprocessing steps can be found there. The clusterings in Figure 2 is done using the pipeline SnapATAC, and the clusterings in Figure S3 is done using the pipeline aggregation. See Luo et al. (2024) for the pipeline details.

A.2. The Atlas2 dataset used in Figure 4. See B.

A.3. The 10X Visium dataset. We use the LIBD 10X Visium dataset (slice 151673) from Maynard et al. 2021, with manual annotations from the original publication serving as the ground truth. Domain detection is performed using BayesSpace, CellCharter, GraphST, precast, and SEDR with default parameters and specifying the true number of classes. Then the Hungarian algorithm is applied to match the predicted clusters with the annotated classes.

B. Exploration of the LSI embeddings for the Atlas2 scATAC-seq dataset

This is related to Figure 4. The dataset is the scATAC-seq Atlas2 dataset used in Luo et al. (2024), and the details about its generation, quality control (QC), and preprocessing steps can be found there. After QC and filtering, the pipeline ArchR_tiles from Luo et al. (2024) is used for dimensional reduction, which generates the latent sementic indexing-based (LSI) embeddings for each cell. Following ArchR's documentation, the first LSI component is removed due to its above 0.75 Pearson's correlation with the library size. The UMAP in Figure 4a and the Silhouette width in Figure 4b are calculated based on the 2nd to 16th LSI components.

C. Composed Density between and within Clusters (CDbw)

Let C_K denotes a clustering to be evaluated. Firstly, a number of representative points are selected for each cluster in C_K using a furthest-first technique to cover the geometry of the cluster effectively. The CDbw index is defined as

$$CDbw(C_K) = Cohesion(C_K) \cdot Compactness(C_K) \cdot Sep(C_K) \tag{S1}$$

where

$$Cohesion(C_K) = \frac{Compactness(C_K)}{1 + Intra_change(C_K)} \tag{S2}$$

$$Compactness(C_K) = \frac{\sum_{i=1}^{n_s} Intra_dens(C_K, s_i)}{n_s} \tag{S3}$$

$$Sep(C_K) = \frac{\frac{1}{K}\sum_i \min_j \{Dist(C_i, C_j)\}}{1 + Inter_dens(C_K)} \tag{S4}$$

$Intra_dens(C_K, s_i)$ is the intra-cluster density, and $s_i, i \in \{1, ..., n_s\}$ is a sequence of shrinking factors $\in [0,1]$ that is used to move the representative points closer to the cluster center. For a given s_i, $Intra_dens(C_K, s_i)$ calculates the average density around all the representative points, normalized by the standard deviation of the cluster. $Intra_change(C_K)$ is the changes in the within-cluster densities over changes in s_i, which averages the difference between values of $Intra_dens(C_K, s_i)$ for adjacent values of s_i. To calculate $Sep(C_K)$, each representative of a cluster is paired with the closest representatives of any other cluster. $Dist(C_i, C_j)$, the distance between clusters C_i, C_j, is calculated as the average of distances between the pairwise closest representatives; $Dens(C_i, C_j)$, the density between clusters C_i and C_j, is an average standardized number of points in the neighborhood of the midpoint of the line between each pair of representatives; The density between clusters in the clustering, $Inter_dens(C_K)$, is the average of maximum $Dens(C_i, C_j)$ for every pair of clusters C_i and C_j.

D. Density Based Clustering Validation index (DBCV)

For a clustering solution C, the DBCV index is the weighted average of the validity indices of all clusters:

$$\text{DBCV}(C) = \sum_{i=1}^{l} \frac{|C_i|}{|O|} VC(C_i) \tag{S5}$$

where l is the number of clusters, and $|O|$ is the total number of objects. The validity index for a cluster C_i is calculated as the following:

$$VC(C_i) = \frac{\min_{j \neq i} (\text{DSPC}(C_i, C_j)) - \text{DSC}(C_i)}{\max\left(\min_{j \neq i} (\text{DSPC}(C_i, C_j)), \text{DSC}(C_i)\right)} \tag{S6}$$

DSPC(C_i, C_j) is the density separation between two clusters C_i and C_j, which is defined as the minimum mutual reachability distance between their internal nodes:

$$\text{DSPC}(C_i, C_j) = \min\left(d_{\text{mreach}}(o_i, o_j) \text{ where } o_i \in C_i \text{ and } o_j \in C_j\right) \quad \textbf{(S7)}$$

whereas DSC(C_i) is the density sparseness of a cluster C_i, being the maximum edge weight (largest distance) of the internal edges in the Minimum Spanning Tree (MST) of the cluster. The MST is constructed using the mutual reachability distance considering the objects in C_i.

$d_{\text{mreach}}(o_i, o_j)$ is the mutual reachability distance for two objects o_i and o_j, and is defined as:

$$d_{\text{mreach}}(o_i, o_j) = \max\left(\text{aptscoredist}(o_i), \text{aptscoredist}(o_j), d(o_i, o_j)\right) \quad \textbf{(S8)}$$

where $d(o_i, o_j)$ is the pairwise distance between objects o_i and o_j, and aptscoredist(o_i) is the core distance of object o_i. For an object o in a cluster C_i, its core distance with respect to all other objects in the same cluster is calculated by

$$\text{aptscoredist}(o) = \left(\frac{1}{n_i - 1}\sum_{j=2}^{n_i}\left(\frac{1}{\text{KNN}(o, j)}\right)^d\right)^{-\frac{1}{d}} \quad \textbf{(S9)}$$

where: $n_i = |C_i|$, KNN(o, j) is the distance to the j-th nearest neighbor of object o within the same cluster, and d is the dimensionality of the feature space.

E. Entropy-based Local indicator of Spatial Association (ELSA)

For a site i, the ELSA score is the product of two terms:

$$E_i = E_{ai} \times E_{ci} \quad \textbf{(S10)}$$

where E_{ai} summarizes the dissimilarity between site i and the neighbouring sites and is calculated as

$$E_{ai} = \frac{\sum_j \omega_{ij} d_{ij}}{\max\{d\} \sum_j \omega_{ij}}, \quad j \neq i \quad \textbf{(S11)}$$

ω_{ij} is a binary weight that specifies whether the site j is within the neighbourhood of site i, and d_{ij} describes the dissimilarity between the categories at sites i and j, which is calculated as the absolute difference of the ranks assigned to the categories at sites i and j.

E_{ci} on the other hand quantifies diversity of the categories within the neighbourhood of site i:

$$E_{ci} = -\frac{\sum_{k=1}^{m_\omega} p_k \log_2(p_k)}{\log_2 m_i}, \quad j \neq i \quad \textbf{(S12)}$$

where m are the total number of categories, p_k is the probability of kth category from the m_ω categories within the neighborhood of site i, and $m_i = \begin{cases} m \text{ if } \sum_j \omega_{ij} > m \\ \sum_j \omega_{ij}, \text{ otherwise} \end{cases}$.

F. Simulations

This is related to Figure S5. To explore in a mostly unbiased way the behavior of metrics, we generate a number of simulated 2- and 3-dimensional datasets with 3 to 4 classes of varying abundances (some equal, some highly skewed), variability, and differences between classes. This is implemented in the `mockData` function of the `poem` package. Briefly, we first use multi-dimensional scaling to obtain centroids for each class that approximate the desired pairwise class differences (determined by the settings), and then generate the desired number of points around them using a normal or log-normal distribution with a given variability parameter. To create further non-gaussian spread in some classes, we concatenate points from multiple such distributions symmetrically shifted from the centroid. Each simulation was further performed with 2 seeds (see the github repository for the exact sets of parameters used), yielding 2240 simulations. For each, we then take the simulated data to be embedding-like strucutres, and calculate embedding-level metrics as well as graph metrics on a kNN graph with k=5. To compute partition-level metrics, we run k-means clustering with k ranging from 2 to 6, as well as Louvain clustering on the kNN graph with resolutions of 0.5, 1 and 2. Node- and class-specific metrics are averaged to yield global metrics. To enable a comparison of the metrics results with interpretable dimensions of the simulation parameters, we define "size imbalance" as the Simpson diversity index of the class abundances, the "signal-to-noise" as the ratio between the mean pairwise class differences and the mean class variability (i.e. standard deviation), and the "spread" as the mean class spread parameter. We then perform a linear regression of each metric on these three parameters across simulations, and for partition-level metrics we additionally include the number of clusters as a covariate.

G. Fuzzy pair-counting metrics

We use Hullermeier's Normalized Degree of Concordance (NDC, see Hullermeier et al., 2011) as a fuzzy version of the RI, and d'Ambrosio's Adjusted Concordance Index (ACI, see D'Ambrosio et al., 2021) as fuzzy ARI. The ACI adjusts the NDC for chance based on permutation of the membership probability vectors. Based on this logic, we further implemented fuzzy Wallace indices (i.e. Wallace Homogeneity and Wallace Completeness and their chance-adjusted versions). Briefly, Wallace Completeness of a class is originally defined in the non-fuzzy setting as the proportion of all the pairs of elements of that class that are also in the same cluster. In the fuzzy setting, this pair concordance is understood as the similarity in membership agreements for pairs of elements between the two fuzzy partitionings. Specifically, the agreement between a pair of elements i and j is given by $A_{ij} = \frac{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_1}{2}$, where $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ represent their respective membership vectors. The concordance between two fuzzy partitionings Q and P for that pair is then given as $C^{ij}_{PQ} = 1 - |A^P_{ij} - A^Q_{ij}|$. Since class membership is fuzzy, the extent to which this concordance matters for the computation of the completeness of a class X depends on the extent to which the pair of elements are of that class, which can be computed using different t-norm operations for set membership. While multiple such functions are implemented in our package, by default we use the product t-norm, meaning that the degree of membership of a pair of element to the class X (denoted m^{ij}_X) is given by the product of the probability that each of the two element belongs to class X. The Wallace Completeness of class X, WC_X, can then be computed as $WC_X = \frac{\sum m^{ij}_X \times C^{ij}_{PQ}}{\sum m^{ij}_X}$ across all possible pairs i and j. To obtain a chance adjusted AWC_X, we compute the randomly expected WC'_X across permutations of the predicted memberships, and then calculate $AWC_X = \frac{WC_X - WC'_X}{1 - WC'_X}$. The same is done for Wallace Homogeneity.

H. Spatial application of fuzzy metrics

To make hard labels fuzzy in a spatially-dependent fashion, we first compute the domain composition of its spatial nearest neighbors (i.e. nearest neighbors based on the Euclidean distance between spots/cells). To make the neighborhood definition more comparable across areas, in particular at the border of a field of view (where there are fewer first-degree neighbors) or in areas of low-density for non-spot-based datasets, we compute the median distance to the desired k^{th} nearest neighbor across spots/cells, and then consider for each spot/cell the neighbors that are within that range. The neighborhood domain membership ($\boldsymbol{w}_\mathsf{n}$) of a spot i, $\boldsymbol{w}^i_\mathsf{n}$, is then the number of its neighbors belonging to domain, divided by the number of neighbors. The domain membership of spot i itself, $\boldsymbol{w}^i_\mathsf{s}$, is 1 if the spot belongs to a domain, and zero otherwise. The fuzzy domain membership is then computed as $\boldsymbol{w}^i = (1-\alpha) \times \boldsymbol{w}^i_\mathsf{n} + (\boldsymbol{w}^i_\mathsf{s} \times \alpha)$, where α is a parameter controling the relative weight of the spot versus it's neighborhood (by default, we set $\alpha = 0.5$).

By comparing against a fuzzy truth, we decrease the concordance of a predicted partitioning that would be exactly like the original 'hard' truth. For this reason, we developed a fuzzy-hard version of the metrics that compares (hard) clustering labels to both a "hard" and a fuzzy ground truth. For the purpose of computing within-pair agreements, the hard labels are expressed as a weight of 1 for the given label and 0 for other labels. Otherwise the procedure is exactly like described above, except that the concordance between the prediction P and the truth Q of a pair of elements i and j is given by $C^{ij}_{PQ} = 1 - \min(|A^P_{ij} - A^Q_{ij}|, |A^P_{ij} - A^R_{ij}|)$, where Q and R are respectively the hard and fuzzy versions of the ground truth. In this context, for the purpose of computing completeness, pair membership is based on the hard ground truth labels.

Supplementary Information 2: Supplementary Figures

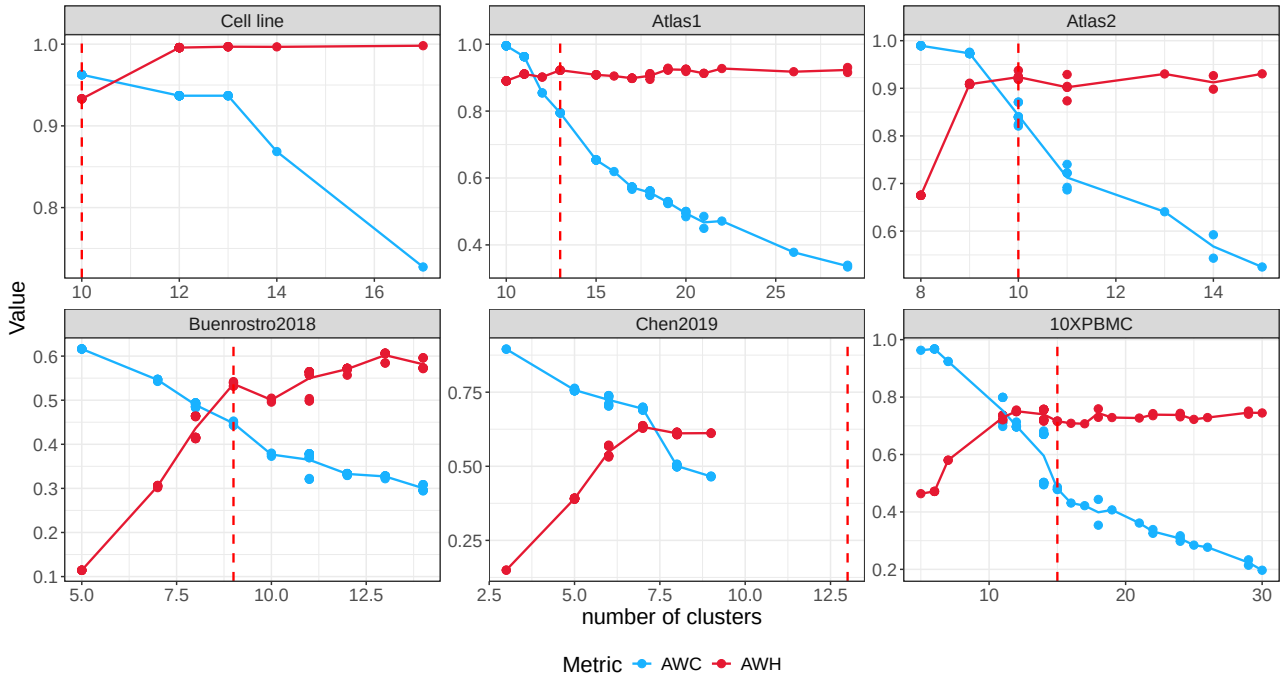

Figure S1. The trade-off between AWH and AWC as the number of clusters varies. Here six single-cell ATAC-seq datasets are used for benchmarking clustering. The red dash line is the number of classes of the ground truth. Each dot is a clustering solution. Different numbers of ATAC clusters are achieved by applying a range of resolutions during Leiden clustering. See Luo et al. (2024) for more details.

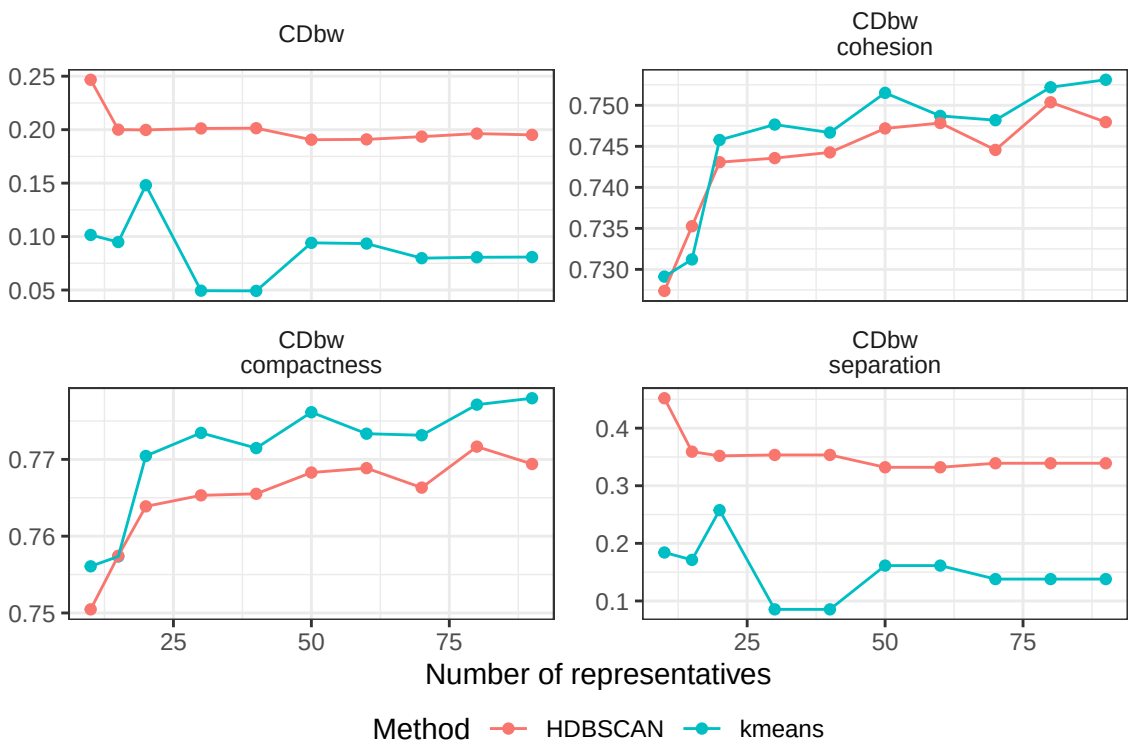

Figure S2. How the number of representative points influence CDbw metrics. In relation to Figure 3 c-d, we calculate the CDbw metrics using a variety of representative numbers. Similar to Figure 3 d where 10 representatives are used, the CDbw and CDbw separation metrics indicate that HDBSCAN produces a better clustering than kmeans for this toy example. However, the CDbw cohesion and compactness metrics fail to capture this distinction and rank them oppositely.

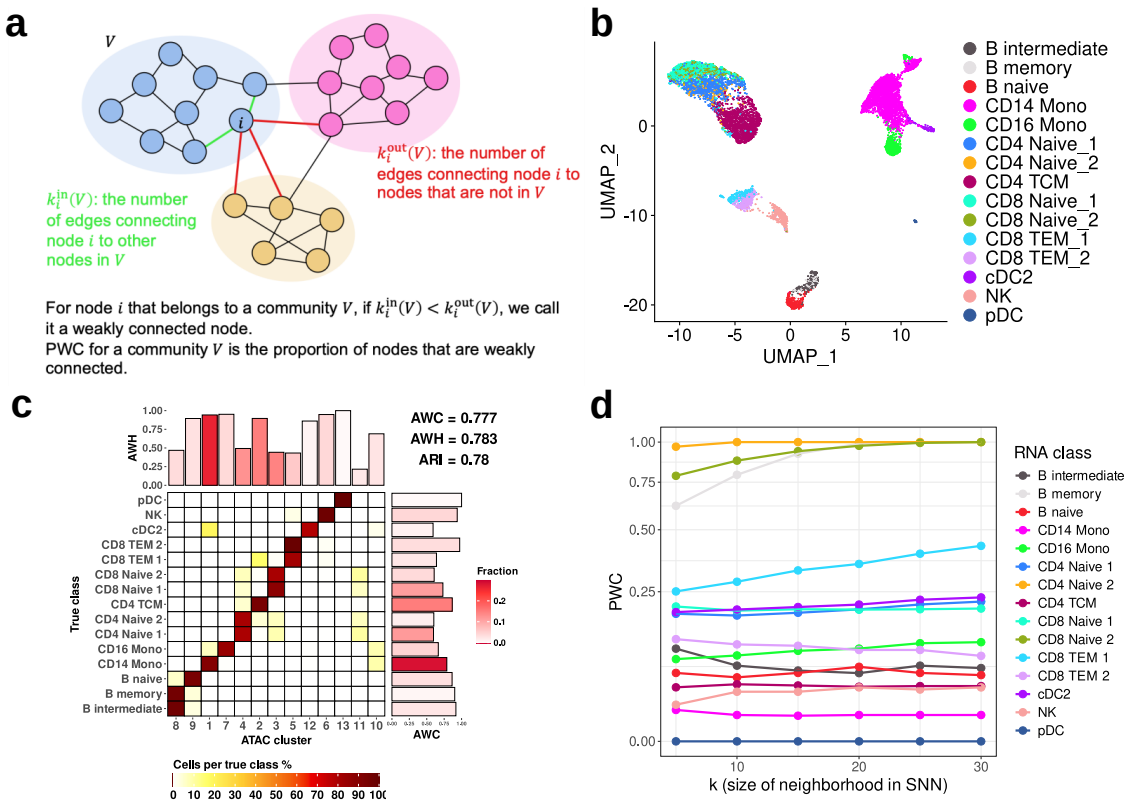

Figure S3. Illustration of the utility of PWC in single-cell datasets. a Schemetic overview of how PWC is calculated. **b** The UMAP of this dataset. **c** An example clustering results showing the classes that are failed to be seperated. **d** The value of PWC across a range of k when building the SNN graph.

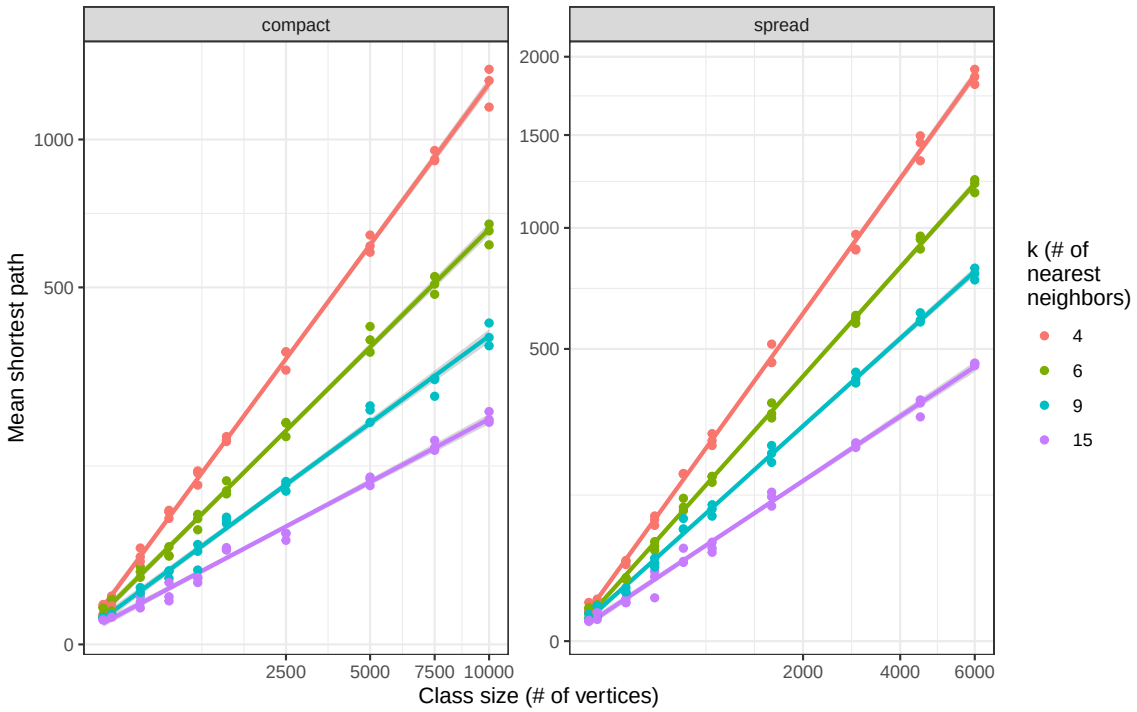

Figure S4. The mean shortest path (in steps) is linearly proportional to the number of vertices. Shown are simulated two-class graphs containing a more compact and more spread class, simulated with the same parameters except for a size factors. Each size is simulated with 3 random seeds.

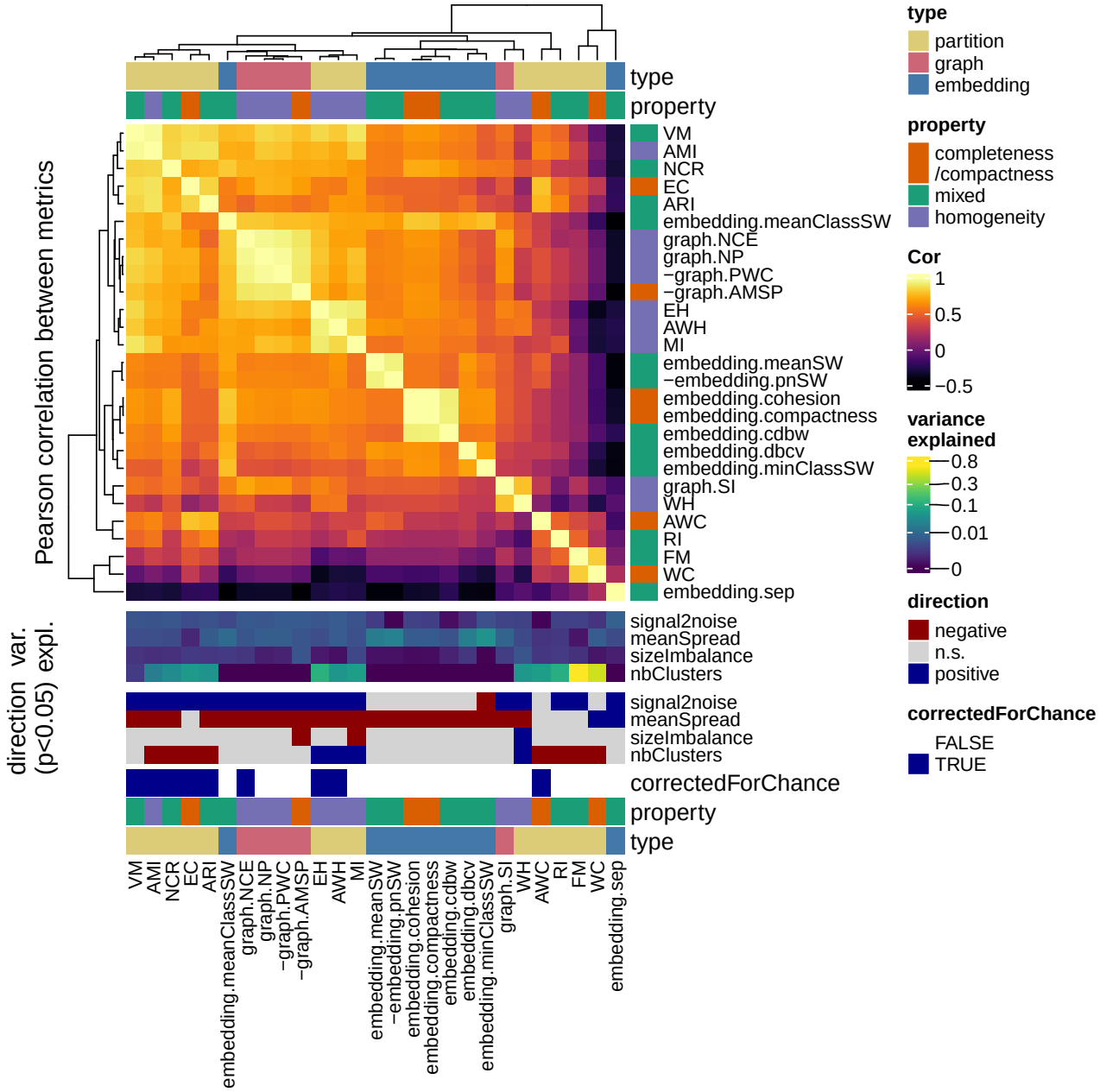

Figure S5. Comparison of metrics across data representations. As explained in Supplementary Information 1, 2240 simulated datasets of 3 or 4 classes with varying abundance, variability, and inter-class distances are used. The main heatmap shows the Pearson's correlation coefficients between metrics. The bottom heatmaps are the percentage of variance explained by each simulation factor using a linear regression model, the effect directions, and the metric categories, respectively.

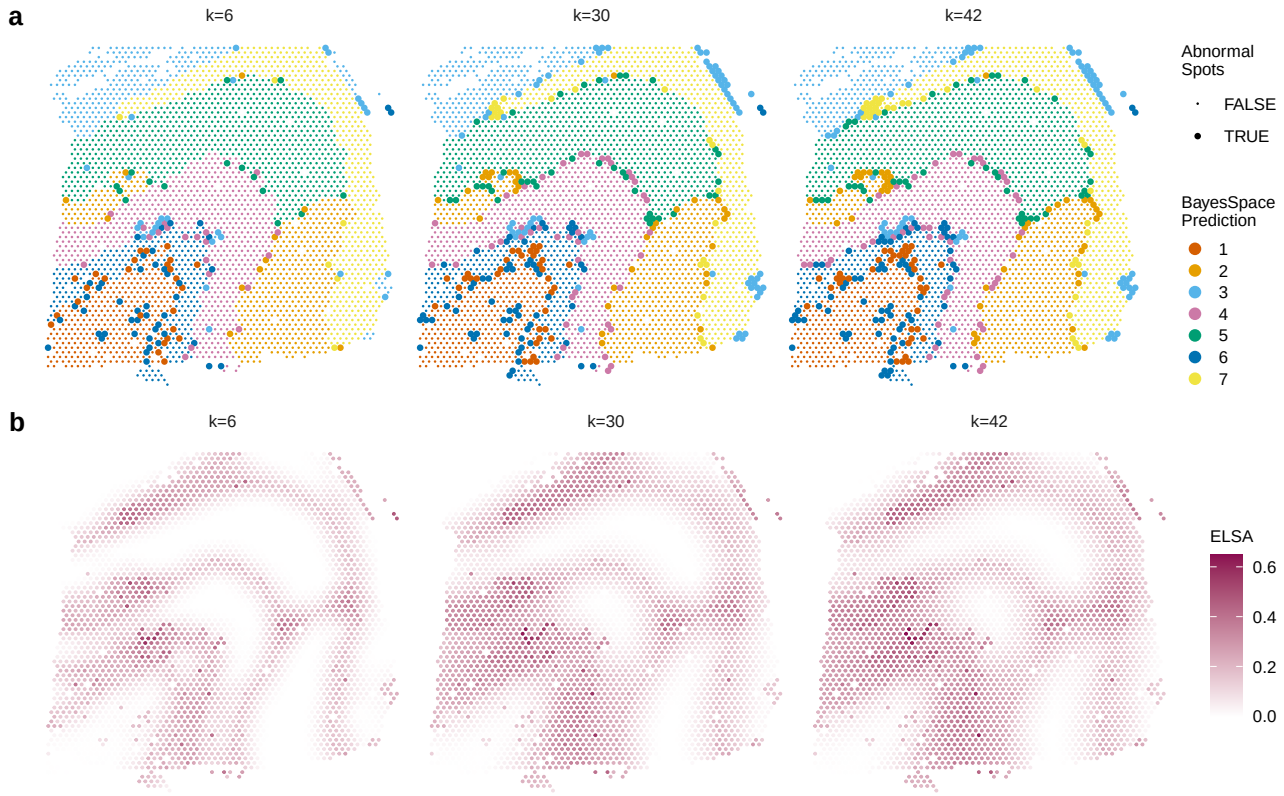

Figure S6. Element-level internal metrics as the size of neighborhood changes. This is related to Figure 7. k is the size of the neighborhood used to calculate the element-level metric values. Score in **a** indicates whether a spot is classified as abnormal or not; Score in **b** is the ELSA values.

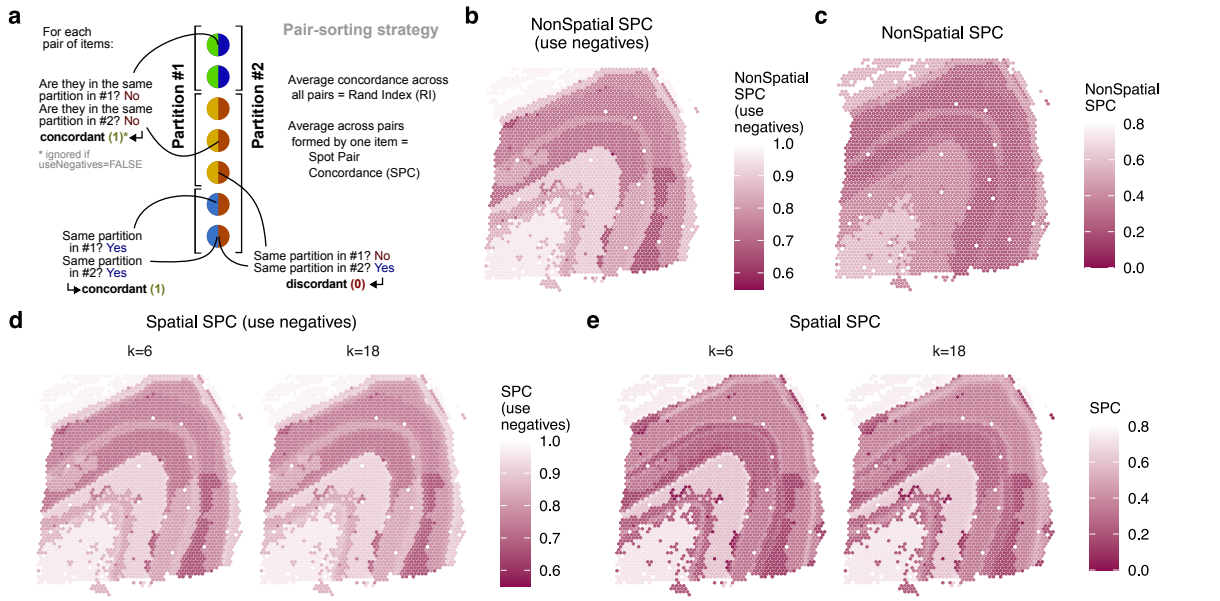

Figure S7. Variants of the Spot-wise Concordance (SPC) score. a Illustration of the pair-counting strategy. **b**-**e** show the values of SPC variants for BayesSpace's prediction as in Figure 7. **b**-**c**: SPC not accounting for spatial relationships, and using (**b**) or not using (**c**) negative pairs. **d**-**e**: Spatially-aware SPC using (**d**) or not using (**e**) negative pairs, and across different neighborhood sizes (k).

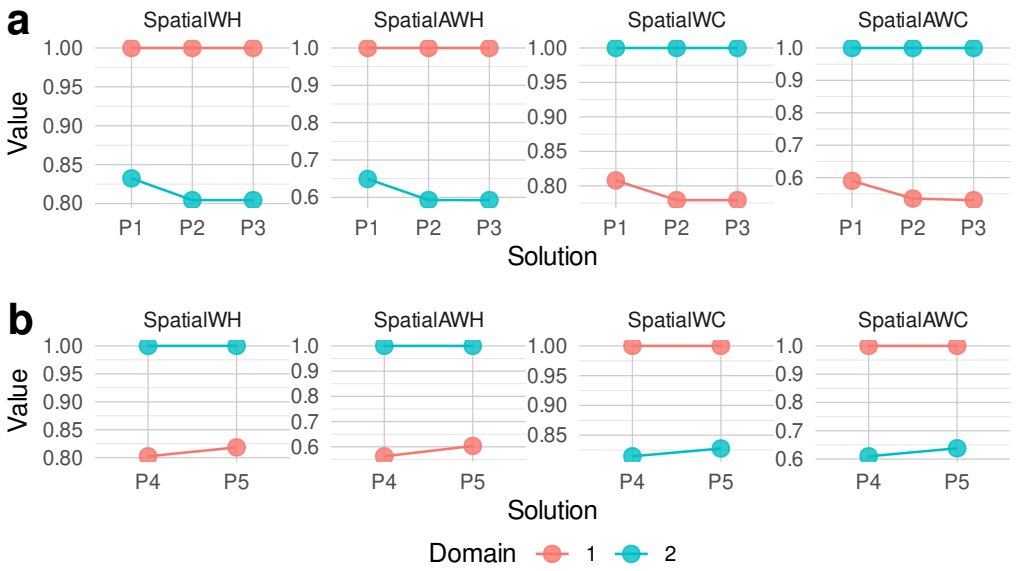

Figure S8. The values of domain-level partition metrics. Related to Figure 8. The cluster-specific (fuzzy-hard version of) spatial WH and AWH, as well as class-specific (fuzzy-hard version of) spatial WC and AWC are calculated for P1-P3 in **a**, and for P4-P5 in **b**, respectively.

Bibliography

A. D'Ambrosio, S. Amodio, C. Iorio, G. Pandolfo, and R. Siciliano. Adjusted concordance index: an extensionl of the adjusted rand index to fuzzy partitions. *Journal of Classification*, 38:112–128, 2021.

E. Hullermeier, M. Rifqi, S. Henzgen, and R. Senge. Comparing fuzzy partitions: A generalization of the rand index and related measures. *IEEE Transactions on Fuzzy Systems*, 20(3):546–556, 2011.

S. Luo, P.-L. Germain, M. D. Robinson, and F. von Meyenn. Benchmarking computational methods for single-cell chromatin data analysis. *Genome Biology*, 25(1):1–30, 2024.

K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.