Bioinformatics Advances, 2023, **00**, vbad141
https://doi.org/10.1093/bioadv/vbad141
Advance Access Publication Date: 9 October 2023
Original Article

OXFORD

Bioimage informatics

The impact of similarity metrics on cell-type clustering in highly multiplexed in situ imaging cytometry data

Elijah Willie [1,2], **Pengyi Yang** [1,2,3,4], **Ellis Patrick** [1,2,3,5,*]

[1]Sydney Precision Data Science Centre, The University of Sydney, Camperdown, NSW 2006, Australia
[2]School of Mathematics and Statistics, The University of Sydney, Camperdown, NSW 2006, Australia
[3]Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong, China
[4]Computational Systems Biology Group, Children's Medical Research Institute, The University of Sydney, Westmead, NSW 2145, Australia
[5]Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Westmead, NSW 2145, Australia

*Corresponding author. Sydney Precision Data Science Centre, University of Sydney, Camperdown, Sydney, NSW 2006, Australia.
E-mail: ellis.patrick@sydney.edu.au

Associate Editor: Guoqiang Yu

Abstract

Motivation: The advent of highly multiplexed in situ imaging cytometry assays has revolutionized the study of cellular systems, offering unparalleled detail in observing cellular activities and characteristics. These assays provide comprehensive insights by concurrently profiling the spatial distribution and molecular features of numerous cells. In navigating this complex data landscape, unsupervised machine learning techniques, particularly clustering algorithms, have become essential tools. They enable the identification and categorization of cell types and subsets based on their molecular characteristics. Despite their widespread adoption, most clustering algorithms in use were initially developed for cell suspension technologies, leading to a potential mismatch in application. There is a critical gap in the systematic evaluation of these methods, particularly in determining the properties that make them optimal for in situ imaging assays. Addressing this gap is vital for ensuring accurate, reliable analyses and fostering advancements in cellular biology research.

Results: In our extensive investigation, we evaluated a range of similarity metrics, which are crucial in determining the relationships between cells during the clustering process. Our findings reveal substantial variations in clustering performance, contingent on the similarity metric employed. These variations underscore the importance of selecting appropriate metrics to ensure accurate cell type and subset identification. In response to these challenges, we introduce FuseSOM, a novel ensemble clustering algorithm that integrates hierarchical multiview learning of similarity metrics with self-organizing maps. Through a rigorous stratified subsampling analysis framework, we demonstrate that FuseSOM outperforms existing best-practice clustering methods specifically tailored for in situ imaging cytometry data. Our work not only provides critical insights into the performance of clustering algorithms in this novel context but also offers a robust solution, paving the way for more accurate and reliable in situ imaging cytometry data analysis.

Availability and implementation: The FuseSOM R package is available on Bioconductor and is available under the GPL-3 license. All the codes for the analysis performed can be found at Github.

1 Introduction

Technological advancements over the past decade have provided researchers the capability to simultaneously measure multiple molecular features in tissue at subcellular resolution (Lewis *et al.* 2021). Key technologies that are pioneering a new era for spatially resolved proteomics include imaging mass cytometry (IMC) (Giesen *et al.* 2014), multiplexed ion beam imaging by time of flight (MIBI-TOF) (Keren *et al.* 2019), co-Detection by indEXing (CODEX) (Black *et al.* 2021), and its successor phenocycler. These technologies can measure ~50–100 features with high throughput, enabling researchers to address complex questions about the spatial distribution and interaction of various types of cells *in situ* (Baharlou *et al.* 2019). A ubiquitous analytical step when analyzing highly multiplexed imaging data is defining functionally distinct cell groupings. While there have been recent developments in spatial analysis approaches that simultaneously phenotype cells by their cellular environment and molecular features (Lee *et al.* 2023, Liu *et al.* 2023), the most commonly used phenotyping approaches only use molecular features and are not intentionally biased by cellular interactions.

Unsupervised clustering algorithms are valuable tools for discovering both known and novel cell types in highly multiplexed data, even in cases where prior knowledge of the cell types present in an experiment is lacking (Karim *et al.* 2021). Here, we use the terminology "cell type" liberally, with clusters also potentially representing distinct known or novel cell states. In our review of the literature, over 70% of manuscripts employing highly multiplexed imaging data for analysis utilized one of three clustering algorithms. IMC and MIBI-TOF data were predominately clustered using either Phenograph (Levine *et al.* 2015), a graph-based Louvain community detection method, or FlowSOM (Van Gassen *et al.* 2015), a self-organizing map (SOM) approach, while CODEX data was predominately clustered using X-shift, a *k*-nearest neighbor (KNN) algorithm accessible through the Vortex GUI (Samusik *et al.* 2016). In the remaining manuscripts, other Louvain and Leiden graph-based

Received: 15 June 2023; **Revised:** 23 August 2023; **Editorial Decision:** 19 September 2023; **Accepted:** 7 October 2023

community detection algorithms, hierarchical clustering, and K-means clustering were used.

Despite their popularity in imaging modalities, Phenograph, FlowSOM, and X-shift were developed in 2015 and 2016 for suspension cytometry technologies, which do not share all of the same technical limitations and noise profiles with tissue-based imaging technologies. Technical artifacts present in most imaging technologies include non-specific binding (Batth *et al.* 2020) and lateral marker spillover (Bai *et al.* 2021). Additionally, in practice, these methods are often employed to generate a large set of candidate clusters, which using expert domain knowledge are then manually clustered, refined, and annotated based on biological features, such as key marker expression, and cell localization. Following this, there exist multiple avenues for further exploration of how clustering algorithms could be tailored for multiplexed imaging data.

Choosing an appropriate similarity metric is crucial for clustering algorithms as it determines how points in a dataset, in our case cells, are partitioned into clusters. Different similarity metrics, often referred to as distance metrics, can yield different clusters. While the Euclidean distance is commonly used in many clustering algorithms, recent studies have shown that correlation-based metrics, such as Pearson or Spearman correlation perform better when clustering in other multiplexed single-cell technologies (Kim *et al.* 2019, Watson *et al.* 2022). Evaluating the performance of different similarity metrics for defining cell types in multiplexed imaging data may guide the improvement or development of new clustering algorithms, which are optimal for these exciting technologies.

In this study, we systematically assess the performance of various distance- and correlation-based metrics in 15 imaging datasets, using multiple performance metrics, such as the Adjusted Rand Index (ARI), the Normalized Mutual Information (NMI), the F-Measure, and the Fowlkes–Mallows Index (FM-Index). We also compare the performance of best-practice clustering methods that currently employ different similarity metrics. Based on our assessment, which highlights the benefits of combining information from multiple similarity metrics, we introduce a new clustering algorithm called FuseSOM. FuseSOM utilizes SOMs and combines multiple similarity metrics through multi-view ensemble learning and hierarchical clustering. This algorithm aims to accurately and robustly identify cell types in multiplexed *in situ* imaging cytometry assays. Overall, our work demonstrates the impact of similarity metrics on clustering cells in multiplexed imaging cytometry data and proposes FuseSOM as a promising method for the analysis of such data.

2 Methods

2.1 Datasets

To benchmark the performance of FuseSOM on imaging datasets from various technologies, we curated a set of 15 datasets generated using different imaging technologies. We selected datasets with human intervention in manually gating cell populations or merging biologically similar clusters. The intention of selecting datasets with manual intervention in defining cell types is to reduce bias toward the original clustering method when evaluating clustering performance. Using these types of datasets also provides higher confidence in the quality of clusters since expert domain knowledge has been applied to scrutinize the clusters further. Datasets were sourced from major databases, including Zenodo, Figshare, and Mendeley. When available, we used the version data that had been processed as described in the original manuscript. We also used the same markers and the same final number of clusters for clustering as described in the manuscript. The imaging technologies used included CODEX (Black *et al.* 2021) (four datasets), IMC (Giesen *et al.* 2014) (six datasets), MIBI-TOF (Keren *et al.* 2019) (four datasets), and sequential Fluorescence *In Situ* Hybridization (seqFISH) (one dataset) (Eng *et al.* 2019). See Table 1 for a detailed description of the datasets used.

2.2 Evaluation metrics

To evaluate clustering performance for clustering solutions generated across methods, we used a set of methods; the ARI, NMI, FM-Index, and the F-Measure (Fowlkes and Mallows 1983, Steinley 2004, Hripcsak and Rothschild 2005, Kvålseth 2017). The ARI measures the similarity between two data clusterings, adjusting for chance

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}, \tag{1}$$

where n_{ij} is the number of pairs of elements that are in the same set in both clusterings, a_i is the total number of pairs in the same set for the first clustering, b_j is the total number of pairs in the same set for the second clustering, and n is the total number of elements.

NMI is a normalization of the mutual information (MI) score to scale the results between zero (no MI) and one (perfect correlation) and is defined as follows:

$$\mathrm{NMI}(X, Y) = \frac{2 \times I(X, Y)}{H(X) + H(Y)}, \tag{2}$$

where $I(X, Y)$ is the MI between clusters X and Y, and $H(X)$ $H(Y)$ are the entropies of clusters X and Y, respectively.

The FM-Index is the geometric mean of precision and recall, and it is defined as follows,

$$\mathrm{FM} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}}, \tag{3}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

The F-Measure, which is the harmonic mean of the precision and recall, is defined as follows:

$$\mathrm{F} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{4}$$

where

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

and

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{6}$$

All these metrics take values between zero and one, with zero being no similarity and one being perfect similarity.

2.3 Distance metrics

Six types of metrics across two classes that are predominantly used across machine-learning clustering literature were used in this study. The two classes include correlation-based and distance-based. The distance-based metrics were Euclidean, Manhattan, and Maximum distance, while the correlation-based metrics included Pearson correlation, Spearman correlation, and Cosine similarity. More formally, let x_{im} and x_{jm} denote the expression of a marker $m = 1, \ldots, M$ in cell $i = 1, \ldots, N$ and cell $j = 1, \ldots, N$, where G and N are the total number of markers and cells, respectively. Let $D = d_{ij}$ be a distance matrix, where d_{ij} represents the distance between cell_i and cell_j. We can then define the distance-based metrics as follows:

Euclidean distance,

$$d_{ij} = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2}, \tag{7}$$

Manhattan distance,

$$d_{ij} = \sum_{m=1}^{M} |x_{im} - x_{jm}|, \tag{8}$$

Maximum distance,

$$d_{ij} = \max_{m} |x_{im} - x_{jm}|. \tag{9}$$

Similarly, the correlation-based metrics can be defined as follows:

Pearson distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M}(x_{im} - \overline{x}_i)(x_{jm} - \overline{x}_j)}{\sqrt{\sum_{m=1}^{M}(x_{im} - \overline{x}_i)^2}\sqrt{\sum_{m=1}^{M}(x_{jm} - \overline{x}_j)^2}}\right)}, \tag{10}$$

Spearman distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M}(r_{im} - \overline{r}_i)(r_{jm} - \overline{r}_j)}{\sqrt{\sum_{m=1}^{M}(r_{im} - \overline{r}_i)^2}\sqrt{\sum_{m=1}^{M}(r_{jm} - \overline{r}_j)^2}}\right)}, \tag{11}$$

Cosine distance,

$$d_{ij} = \sqrt{2\left(1 - \frac{\sum_{m=1}^{M} x_{im}x_{jm}}{\sqrt{\sum_{m=1}^{M} x_{im}^2}\sqrt{\sum_{m=1}^{M} x_{jm}^2}}\right)}, \tag{12}$$

where r_{ij} is the rank of marker m in $cell_i$, $\overline{x_i}$ is the mean expression of cell_i, $\overline{x_j}$ is the mean expression of cell_j, $\overline{r_i}$ is the mean expression rank of cell_i, and $\overline{r_j}$ is the mean expression rank of cell_j.

2.4 Clustering algorithms

For this work, a few clustering algorithms were used for comparing the effects of distance metrics on clustering outcomes. These algorithms include hierarchical clustering, FlowSOM (Van Gassen *et al.* 2015), K-means, and Phenograph (Levine *et al.* 2015).

The "hierarchical clustering" builds a hierarchy of clusters by either merging smaller clusters into larger ones (agglomerative) or dividing a large cluster into smaller ones (divisive) using a linkage function. The process continues iteratively, resulting in a tree-like diagram called a dendrogram that represents the nested clusters. The agglomerative version was used with the average linkage function (Nielsen 2016).

The "FlowSOM" utilizes SOMs and hierarchical clustering to analyze and visualize complex datasets, particularly in flow cytometry. It groups cells into nodes on a grid based on similarity, providing insights into data structures. This technique is especially valuable in identifying and understanding cell populations. The FlowSOM algorithm (version 2.8.0) was obtained from Bioconductor.

The "K-means" clustering partitions data into "k" clusters by repeatedly assigning data points to the nearest centroid and recalculating the centroids. The process continues iteratively until the centroids stabilize. The "Base R" implementation of the K-means algorithm was used.

The "Phenograph" is a clustering method that constructs a KNN graph from data, usually applied to single-cell data analysis. Community detection is performed on this graph using the Louvain method to identify clusters or communities of similar nodes. The "phenograph" function from the *ReductionWrappers* version 2.5.4 (Smith 2023) R package was used.

2.5 FuseSOM

Here, the FuseSOM algorithm is described. The algorithm starts by taking in an m by n matrix, where m is the number of cells and n is the number of markers. Next, FuseSOM uses this matrix to generate a SOM. A SOM is a type of dimensionality reduction algorithm that maps points in a high dimensional space ($d > 2$) to a lower dimensional space ($d = 2$). The SOM architecture preserves the topological relationships between points when reduced to a lower dimension. The SOM also provides a set of points called prototypes, which are representations of points in the higher dimensional space. The SOM architecture was chosen due to its ability to preserve topological structures of the input data, and its ability to represent complex non-linear relationships in the data, allowing them to capture more intricate patterns and dependencies. Like cluster centers in the k-means algorithm, many points can be mapped to a single prototype. For a more thorough treatment of SOMs, see Miljkovic (2017). The *YASOMI* package (version 0.3) was obtained and modified to implement the SOM used in FuseSOM (Rossi 2012).

In this work, we generate a SOM, and the prototypes are used for clustering. After clustering, the clusters are projected back to the original data to classify the original data points. The SOM algorithm requires a 2-d grid(x, y) size, which determines the number of prototypes. Grid sizes of varying shapes are allowed. However, square grids are typically used. To estimate the size of the grid for a dataset, we use the method described in Patterson *et al.* (2006). This method computes the number of eigenvalues of a covariance matrix significantly different from the Tracy–Widom distribution (Tracy and Widom 1994).

Next, multiview integration combines the Pearson correlation distance, Cosine distance, Spearman correlation, and Euclidean distance between the prototypes to generate a final

distance matrix for clustering. Multiview ensemble learning is a machine-learning strategy that employs multiple diverse views or perspectives of the same data to enhance predictive modeling. In a typical multiview ensemble learning setup, each view represents a unique set of features or a unique preprocessing or transformation of the data. These different views capture various aspects of the data, which when combined, offer a more comprehensive and potentially more accurate representation. In this work, the different views are represented by the various distances computed between the cells. To combine these views, we adopted a multiview integration (Fuse) method to combine the four transformed matrices (Melssen *et al.* 2006). Formally, the multiview integration can be defined as follows:

$$D_{fused}[i,k] = \sum w_i \times D[ijk], \tag{13}$$

where $D_{fused}[i,j]$ is the combined dissimilarity between samples j and k, w_i is the weight assigned to the *ith* dissimilarity matrix, and $D[ijk]$ is the dissimilarity between j and k for the *ith* dissimilarity matrix. All distances are weighted equally. We tried a variety of weighting methods and the equal weighting (Fuse) consistently performed the best. See Supplementary Fig. S5.

The *analogue* package (version 0.17-6) is used to perform the multiview integration (Simpson and Oksanen 2021). The *psych* package (version 2.3.3) transforms the similarity matrices into distance matrices (Revelle 2022). Correlations and cosines are transformed into distances by using the formula:

$$D = \sqrt{2(1 - r_{ij})}, \tag{14}$$

where r_{ij} is the value of the correlation or cosine between feature i and j. Finally, to generate final cluster labels using the integrated distance matrices, FuseSOM takes in a parameter k, which is the number of desired clusters. Next, hierarchical clustering using the average linkage function is used to generate the final clustering solution. The *FCPS* package (version 1.3.1) was used for hierarchical clustering (Thrun and Stier 2021).

2.6 Clustering framework

For consistency when comparing distance metrics across various datasets, each dataset was sampled five times to obtain 20K cells. After this, we executed each clustering algorithm and recorded the scores based on different evaluation metrics. To compare with FuseSOM, we used a substratification framework for each dataset. This framework accepts a dataset and produces five stratified samples. The purpose of substratification is to account for potential variability in clustering outcomes. Through this framework, when a dataset is inputted, it yields five stratified datasets. Stratification involves selecting 50% of cells from every annotated class (Kim *et al.* 2019).

2.7 Cluster size estimation

For most clustering algorithms, the number of clusters k is an important hyperparameter that must be set. To this end, many methods have been developed to help practitioners choose an appropriate number for their dataset. We have included well-known methods for estimating the number of clusters as part of the FuseSOM package. These methods include the Gap statistic, the Slope statistic, the Jump statistic, the Silhouette statistic, and the within-cluster distance (WCD) (Rousseeuw 1987, Tibshirani *et al.* 2001, Sugar and James 2003, Fujita *et al.* 2014).

The "Gap statistic" compares the change in the within-cluster sum of squares (WSS) from the observed data to that of a random clustering. A large gap value indicates the observed data has a more pronounced clustering structure than expected under a random scenario.

The "Silhouette statistic" quantifies how close each data point in one cluster is to data points in neighboring clusters, with values ranging from −1 to 1; higher values indicate better-defined clusters.

The "Jump statistic" evaluates the rate of increase in the WSS as a function of the number of clusters, with large jumps indicating the possible presence of distinct groups.

The "slope statistic" identifies an "elbow" or bend in the WSS plot; the point before the stabilization or decline in the slope can suggest an optimal number of clusters.

The "WCD" measures the compactness of clusters. A smaller value indicates tighter, more well-defined clusters. Each of these statistics offers unique insights and their combined interpretation aids in selecting an appropriate number of clusters.

We also implemented a "Discriminant" method for estimating the number of clusters based on the projection pursuit of the discriminant maximum clusterability. To accomplish this, we couple hierarchical clustering with discriminant analysis and multimodality testing to estimate the number of clusters (Silverman 1981, Hartigan and Hartigan 1985, Etemad and Chellappa 1997, Samadani *et al.* 2013, Mokari *et al.* 2018, Ameijeiras-Alonso *et al.* 2019). First, we generate a dendrogram using hierarchical clustering with average linkage. Next, for each node in the resulting tree, we project the two classes onto a line such that both classes are well separated. See Supplementary Fig. S6. The dip test for multimodality testing is then applied to the distribution of the points along this line (Hartigan and Hartigan 1985). The family-wise error rate is controlled using the method described in Meinshausen (2008). Finally, the number of nodes with significant P-values is returned as the number of clusters.

3 Results

3.1 Evaluating the impact of similarity metrics

To assess the impact of similarity metrics on clustering performance, we performed hierarchical clustering on a MIBI-TOF dataset using correlation-based metrics (Pearson, Spearman, and Cosine) or distance-based metrics (Euclidean, Manhattan, and Maximum) (McCaffrey *et al.* 2022). To assess performance, we compared the hierarchical clusters with the manually curated cell-type labels identified in the manuscript (Fig. 1). On average, correlation-based metrics outperform distance-based metrics by 8.0% for ARI, 10.7% for NMI, 1.70% for FM-Index, and 8.0% for F-Measure.

To provide a comprehensive assessment of the performance of similarity metrics, we quantified the clustering performance of the metrics on 15 multiplexed *in situ* imaging cytometry datasets (Table 1). These datasets were chosen as each had some manual intervention when cell-type labels were defined. Each dataset was randomly subsampled to 20K cells five times, and each subset was clustered using hierarchical clustering with all the similarity metrics. Finally, the average was taken across the five subsets. Across the 15 datasets, correlation-based metrics consistently outperformed distance-based metrics (Fig. 2 and

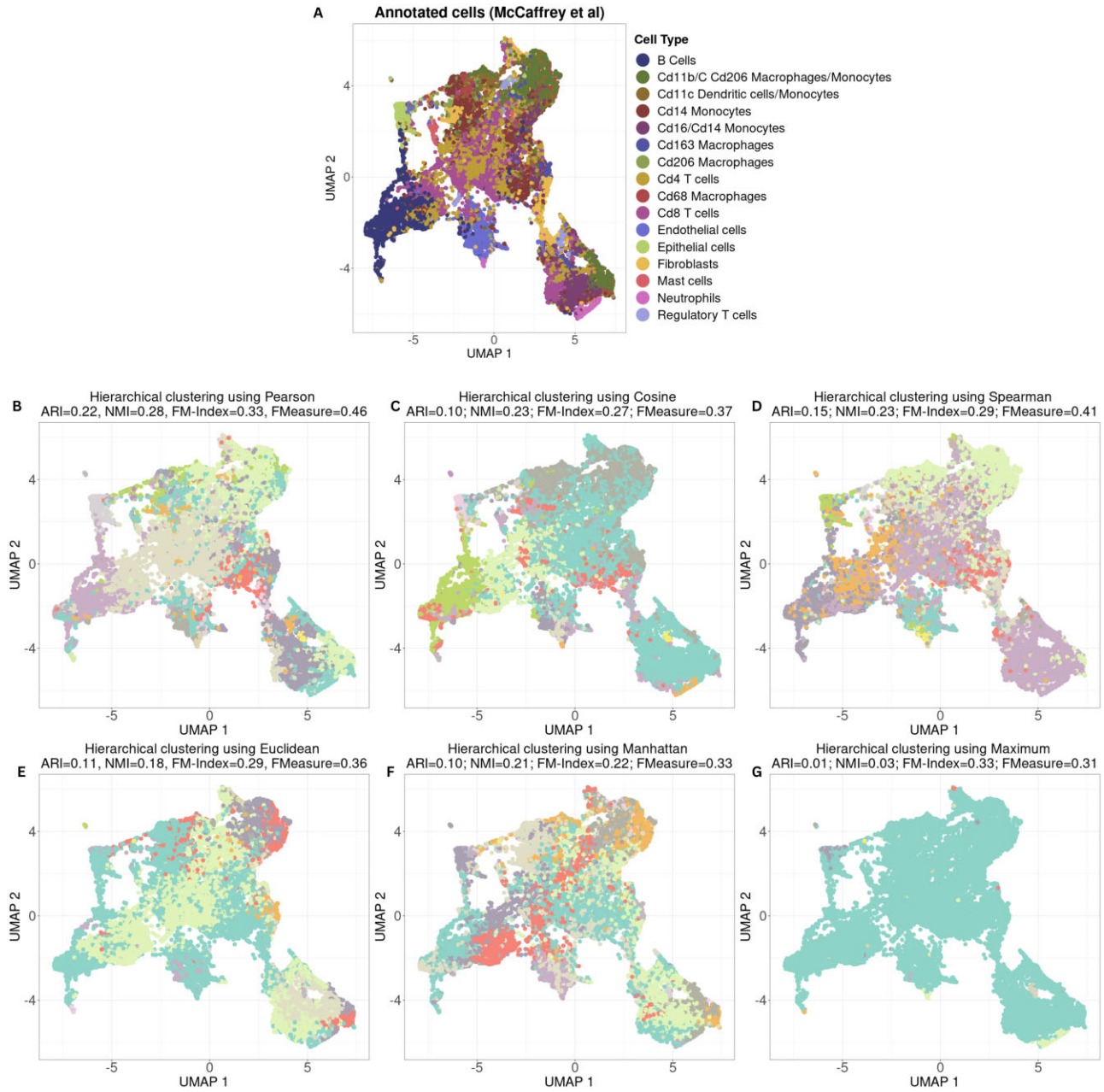

Figure 1. UMAP visualization of cells from a sample imaging dataset. (A) Cells colored by annotations from original study (McCaffrey *et al.* 2022). (B) Hierarchical clustering using Pearson's correlation and concordance quantified by ARI, NMI, FM-Index, and F-Measure. (C) Hierarchical clustering using Cosine's distance. (D) Hierarchical clustering using Spearman's correlation. (E) Hierarchical clustering using Euclidean distance. (F) Hierarchical clustering using Manhattan distance. (G) Hierarchical clustering using Maximum distance.

Supplementary Fig. S1), more accurately recapitulating the manually curated cell-type labels from their original publications. These results show the efficacy of correlation-based metrics in hierarchical clustering.

Next, we assessed if the performance differences between correlation and Euclidean distance are consistent across multiple clustering methods. We reconfigured PhenoGraph, FlowSOM, and K-means clustering to use correlation instead of Euclidean (Levine *et al.* 2015, Van Gassen *et al.* 2015). As previously, the 15 datasets were randomly subsetted to 20K cells five times and the performance scores were calculated for each subset. While there does not appear to be a benefit in clustering performance for PhenoGraph ($P > 0.05$), the overall PhenoGraph performance for both distance measures is worse when compared to the other methods using correlation-based distances (Fig. 3 and Supplementary Fig. S2). For K-means, hierarchical clustering, and FlowSOM, we observe differences in the scores between Pearson correlation compared to Euclidean distance across all evaluation metrics (Fig. 3 and Supplementary Fig. S2).

3.2 Combining similarity metrics is beneficial

Given the performance differences between the similarity metrics, we next assessed whether combining multiple metrics using strategies, such as multiview ensemble learning would further improve performance (Cao *et al.* 2020). To evaluate the efficacy of combining multiple distance metrics for clustering, we performed a comparison study combining various

Table 1. Imaging datasets used.

Dataset	Technology	Num. markers	Num. cells	Num. celltypes	Disease	Tissue	Cell annotation method
Schürch *et al.* (2020)	CODEX	49	258 385	29	Colorectal cancer	Colon	X-shift (Samusik *et al.* 2016) then supervised cluster merging
Phillips *et al.* (2021)	CODEX	52	117 170	21	Lymphoma	Skin	X-shift (Samusik *et al.* 2016) then supervised cluster merging
Brbić *et al.* (2022)	CODEX	48	248 285	21	None	Small intestine/colon	Manually annotated
Brbić *et al.* (2022)	CODEX	44	219 926	13	Normal/Bartett's esophagus	Tonsil	Manually annotated
Moldoveanu *et al.* (2022)	IMC	12	227 592	10	Melanoma	Skin	PhenoGraph (Levine *et al.* 2015) followed by K-means
Van Maldegem *et al.* (2021)	IMC	17	282 837	16	Lung cancer	Lung	PhenoGraph followed by manual splitting
Hoch *et al.* (2022)	IMC	41	864 263	10	Melanoma	Skin	Manual gating
Rendeiro *et al.* (2021)	IMC	38	515 791	17	Covid-19	Lung	Leiden followed by manual merging
Damond *et al.* (2019)	IMC	36	252 059	16	Type 1 diabetes	Pancreas	Supervised cell classifier
Bortolomeazzi *et al.* (2021)	IMC	30	218 615	9	Colorectal cancer	Colon	Seurat followed by DBSCAN
Risom *et al.* (2022)	MIBI-TOF	22	69 672	23	Breast cancer	Breast	FlowSOM (Van Gassen *et al.* 2015) followed by manual merging
Keren *et al.* (2019)	MIBI-TOF	16	201 656	6	TN breast cancer	Various	FlowSOM (Van Gassen *et al.* 2015) followed by merging by hierarchical merging
McCaffrey *et al.* (2022)	MIBI-TOF	37	30 943	16	Tuberculosis	Various	Iterative FlowSOM (Van Gassen *et al.* 2015) clustering
Liu *et al.* (2022)	MIBI-TOF	12	345 490	8	Various cancers	Various	FlowSOM (Van Gassen *et al.* 2015) followed by merging by hierarchical clustering
Lohoff *et al.* (2022)	seqFISH	50	57 536	24	None	Mouse embryos	Louvain clustering on top 50 PCs

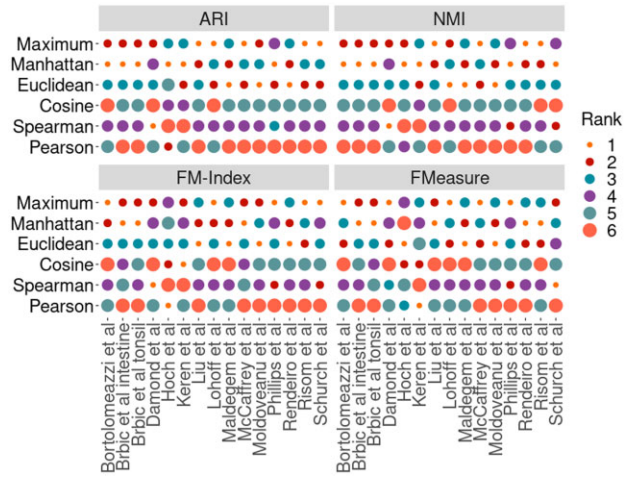

Figure 2. Benchmarking similarity metrics on agglomerative clustering of 15 multiplexed imaging datasets. Each dataset was subsetted to 20K cells five times, and the average clustering score was recorded. Results were ranked in descending order across all similarity metrics and datasets by each evaluation metric. A larger circle size indicates better performance. Correlation-based metrics are consistently ranked higher than distance-based metrics across most datasets.

combinations of Pearson, Spearman, Cosine, and the Euclidean distance. All possible combinations of metrics were used to group the prototypes generated by the SOM algorithm and then the final scores were averaged across all datasets. Of all combinations, the combination of Pearson, Spearman, Cosine, and Euclidean consistently provided the best score in all four evaluation metrics (Fig. 4). The combination of Pearson, Spearman, and Cosine also provided strong results, which implies that the addition of Euclidean does not markedly improve the overall clustering performance.

3.3 FuseSOM combines SOMs with multiview ensemble learning of similarity metrics

Here, we introduce FuseSOM for the clustering of highly multiplexed imaging data. FuseSOM leverages all the ideas already discussed by combining similarity metrics with a SOM and multiview hierarchical clustering to define cell types robustly (Fig. 5). Compared to FlowSOM, which uses Euclidean distance by default, FuseSOM has superior performance in our stratified subsampling analysis framework (Fig. 6, $P < 0.05$, and Supplementary Fig. S3), which demonstrates that a multiview ensemble of similarity metrics provides a

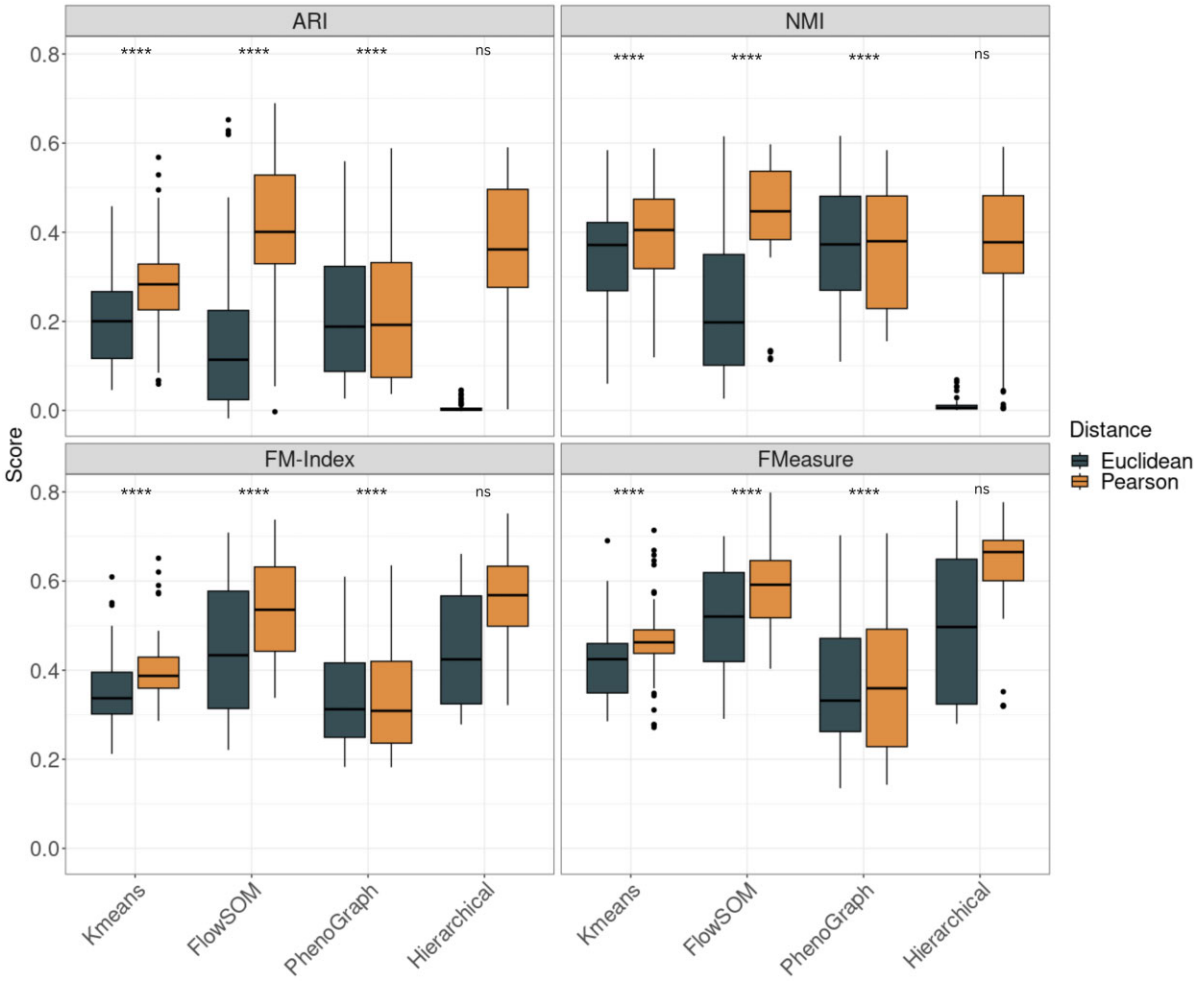

Figure 3. Boxplots of clustering performance of four clustering methods using Pearson correlation and Euclidean across four evaluation metrics (ARI, NMI, FM-index, and F-Measure). For FlowSOM, hierarchical clustering, and K-means, there is a statistically significant difference (****: $P < .0001$) in performance between Pearson and Euclidean using the Wilcoxon rank-sum test. This is not evident for PhenoGraph (ns: $P > .05$).

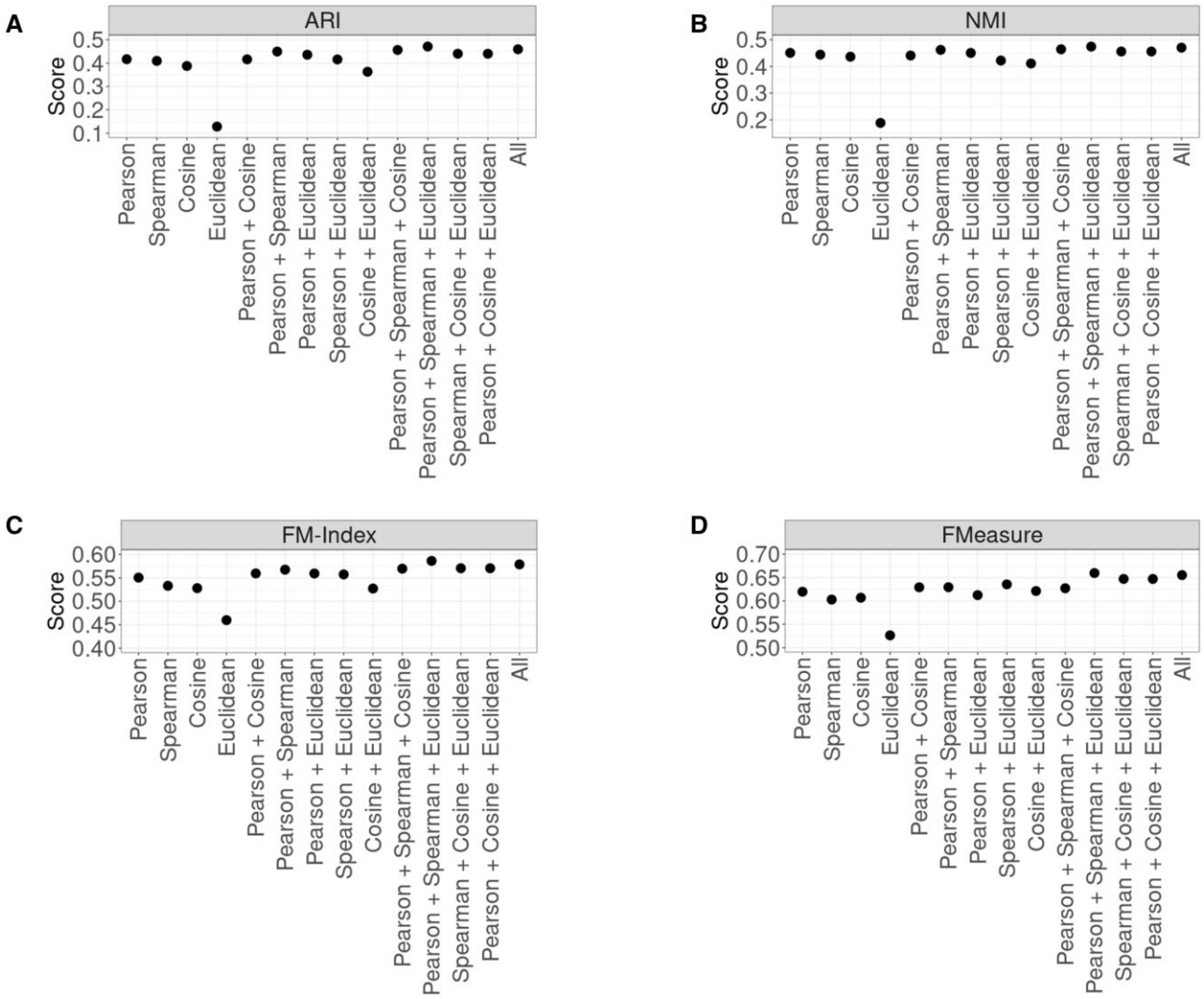

Figure 4. Scatter plots of average clustering performance of all combinations of Pearson, Cosine, Spearman, and Euclidean. (A) Scatter pot of the adjusted rand index. (B) Scatter plot of the normalized mutual information. (C) Scatter plot of the Fowlkes–Mallows index. (D) Scatter plot of the F Measure. Across all performance metrics, there is evidence that combining all distances provides a better signal for clustering.

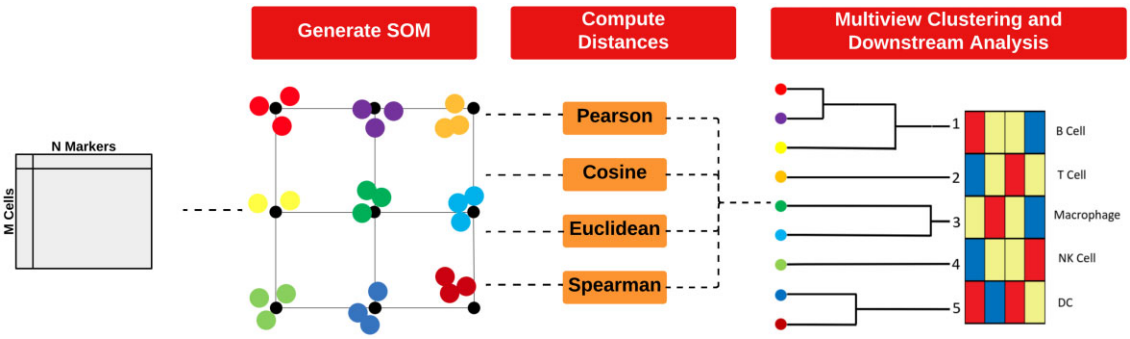

Figure 5. Overview of FuseSOM: this new scalable algorithm uses (i) a SOM to reduce the dimension of the data while preserving its topological structure; (ii) a multiview integration of various similarity metrics to capture all relevant signals; and (iii) hierarchical clustering to generate a clustering solution for further downstream analysis.

more robust clustering. The performance gain is particularly evident when looking at ARI and NMI, with average differences in scores being 32% for ARI, 27% for NMI, 10% for FM-Index, and 9.0% for F-Measure.

Several methods for estimating the number of clusters have been implemented in the FuseSOM R package. The relative error (RE) between the predicted number of clusters and the number of clusters used in the corresponding manuscripts was used to assess the accuracy of the cluster estimation methods. The Jump and Discriminant method tends to overestimate the actual number of clusters while the others tend to underestimate the actual number of clusters (Fig. 7A). Next,

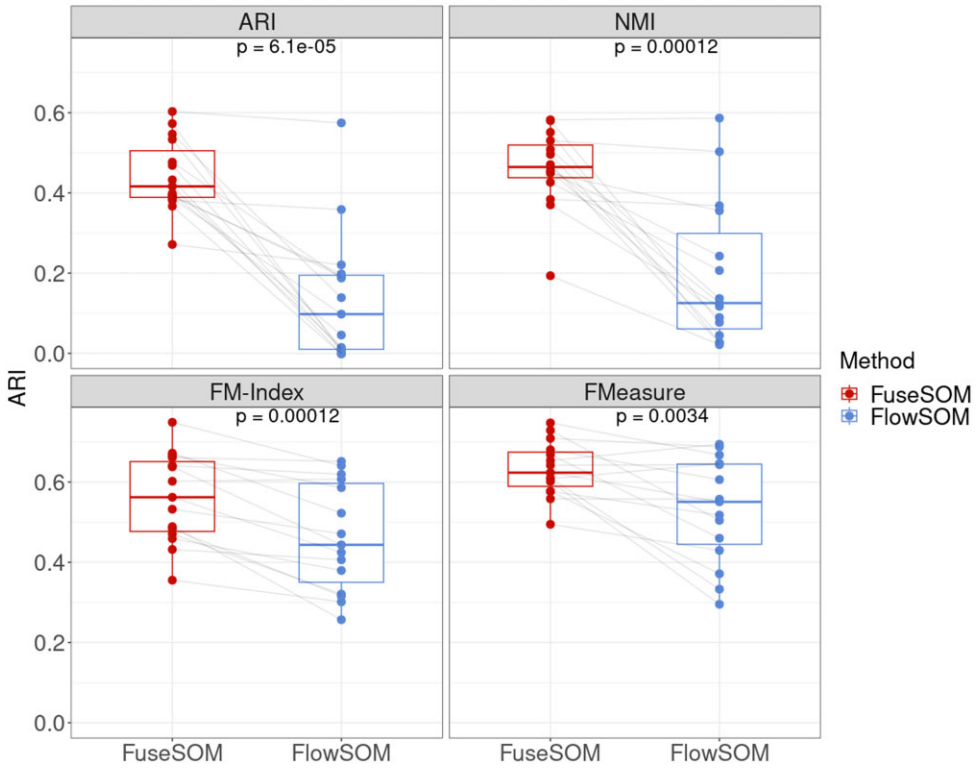

Figure 6. Paired boxplots for the average clustering performance across all datasets. For all four evaluation metrics, differences in performance are evaluated using the Wilcoxon rank-sum test.

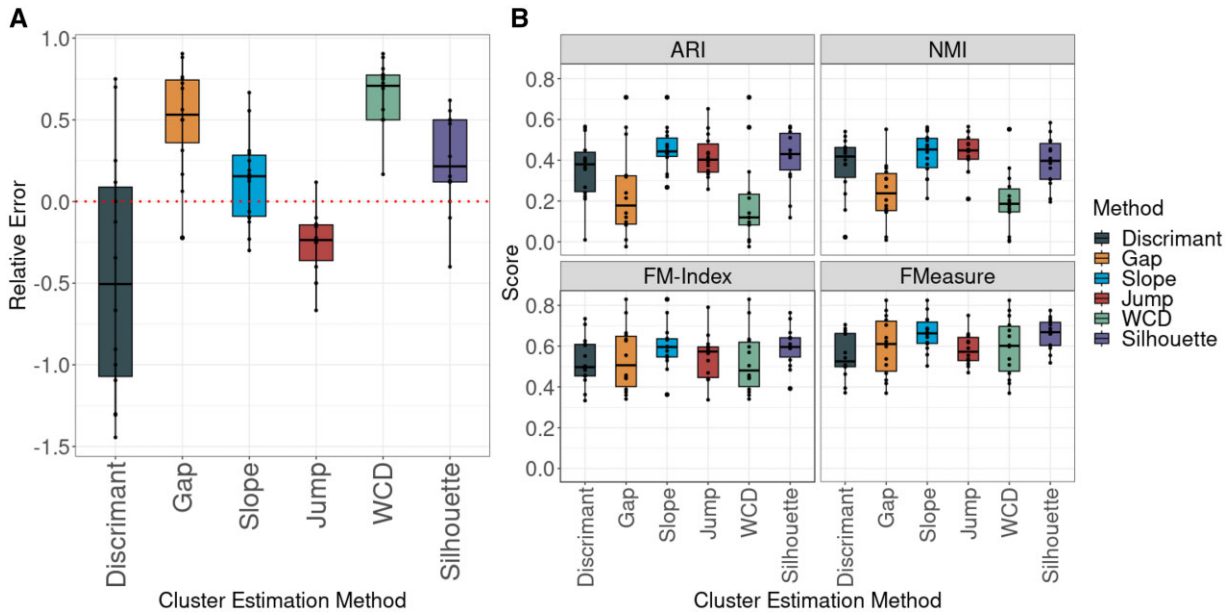

Figure 7. (A) Boxplots of cluster estimation performance for methods included in FuseSOM across all 15 datasets. The RE metric was used to gauge performance. The Jump and Discriminant methods are the top performers. The dashed line represents a RE of zero. In contrast, values above this line indicate an underestimation of the true number of clusters, and values below indicate an overestimation of the true number of clusters. (B) Boxplots of clustering performance based on the estimated number of clusters for methods included in FuseSOM. The Jump and Discriminant methods are top performers across all evaluation metrics.

FuseSOM was used to group cells in datasets using the number of groups estimated by each method. After estimating the number of clusters and then clustering, the REs and clustering scores were averaged in all datasets. When used for choosing the number of clusters, the Jump and Discriminant methods appear to have the highest average ARI and NMI, across all datasets (Fig. 7B). This demonstrates the complexity in evaluating the best metric for selecting the number of clusters.

To investigate the running time and memory usage of FuseSOM, we applied FuseSOM to the Hoch dataset (Hoch *et al.* 2022). This dataset contains 800K cells across 41 markers. Next, clustering was performed in increments of

50K cells starting from 100K cells to 400K cells (eight clustering solutions in total) to gauge how memory and running are affected by an increasing number of cells. Clustering was performed on an 11th Gen Intel® Core™ i7-1165G7 @ 2.80 GHz with four cores and 32 GB of memory. FuseSOM performance was compared against FlowSOM and PhenoGraph. In terms of running time, FuseSOM scales well and is comparable to that of FlowSOM while being faster than PhenoGraph (Supplementary Fig. S4A). For memory usage, FuseSOM is more demanding than PhenoGraph, but less demanding than FlowSOM (Supplementary Fig. S4B). Doubling the size of the data requires twice as much memory for FuseSOM, while doubling the size of the data will require more than double the amount of memory for FlowSOM. The results show that FuseSOM provides a good balance between speed and memory consumption.

4 Discussion

In this work, we performed a comparative analysis of the performance of various similarity metrics for clustering highly multiplexed *in situ* imaging cytometry assays. Using multiple clustering methods across multiple similarity metrics, we demonstrate that the choice of similarity metric affects the clustering performance of highly multiplexed cytometry *in situ* imaging data with correlation-based metrics on average outperforming distance-based metrics. We then leveraged these findings to develop a novel multiview clustering algorithm called FuseSOM and demonstrated its ability to recover semi-supervised cell-type annotations across various datasets from differing imaging technologies with reasonable accuracy. Our results comprehensively demonstrate the impact of similarity metric choice on cell-type clustering in highly multiplexed imaging cytometry data and highlight the need to develop new best-practice clustering algorithms for these technologies.

While we have demonstrated that correlation metrics are often superior to distance metrics for multiplexed imaging data, we have not shown why this is the case. We do however demonstrate that this phenomenon is consistent across the imaging platforms. Our hypothesis is that distance-based metrics, such as Euclidean and Manhattan, are sensitive to the scaling of the data and therefore are susceptible to changes in the expression of markers across images or even different regions of the tissue imaged. However, correlation-based metrics, such as Pearson and Spearman, are scale-invariant and, therefore, could be less susceptible to changes in the expression of markers driven by technical artifacts. As correlation-based metrics only consider relative expression between markers, we suspect that this makes them more robust and, therefore, more accurate in capturing cell-type-specific expression trends in highly multiplexed *in situ* imaging cytometry data.

There are many analytical decisions and data properties that can impact the phenotyping of cells. In this manuscript, we have focused solely on the choice of distance metric used for clustering. To maintain this focus in our benchmarking study, for each dataset, we used the same cell segmentation, marker quantification, cross-image marker normalization, marker selection, and number of clusters that were used in the original manuscripts. It should be expected that each of these components would impact the clustering of cells. We hope that our collection of datasets will assist in future benchmarks of each of these components.

Choosing the number of clusters to use when clustering remains as much an art as a science. As such, selecting a suitable number of clusters should always be viewed in the context of the application. For example, to identify rare cell types in biological data, one might need to deeply cluster the data to find smaller populations of cells. When using quantitative approaches to select the number, like those we have implemented in FuseSOM, our results highlight that some methods tend to identify more clusters on average and others less. Furthermore, while many clustering algorithms require the number of clusters to be chosen before executing the algorithm, there are others, such as graph-based and density-based methods, that can estimate the number of clusters as part of the algorithm. However, often other parameters, which do need to be chosen, such as the size of the neighborhood in density approaches, can inadvertently affect the number of clusters. Ultimately, there is no golden rule when selecting the number of clusters. Therefore, we encourage a user to employ their domain expertise and to use a variety of the methods we have implemented to arrive at a sensible choice for the number of clusters.

Acknowledgements

The authors thank all their colleagues, particularly at The University of Sydney, Sydney Precision Data Science Centre, and Charles Perkins Centre for their support and intellectual engagement.

Author contributions

E.P. and E.W. conceived and designed the study. E.P. and E.W. led the method development and guided the evaluation data analysis with input from P.Y. E.W. curated the imaging data, implemented all data analytics, and developed the corresponding R code. All authors wrote, read, reviewed the manuscript, and approved the final version.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

None declared.

Funding

This work was supported by the AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong to E.P. and P.Y. Australian Research Council Discovery Early Career Researcher Award [DE200100944 to E.P.] funded by the Australian Government; National Health and Medical Research Council (NHMRC) Investigator Grant [1173469 to P.Y.]; and the University of Sydney Postgraduate Excellence Award for E.W. The funding sources had no impact on the study design, in the collection, analysis, and interpretation of data, the writing of the manuscript, and in the decision to submit the manuscript for publication.

Data availability

Publicly available data were used for all evaluations. All data were downloaded as described in the originating manuscripts.

Code availability

The *FuseSOM* R package is available on Bioconductor and is available under the GPL-3 license. All the codes for the analysis performed can be found at Github.

References

Ameijeiras-Alonso J, Crujeiras RM, Rodríguez-Casal A. Mode testing, critical bandwidth and excess mass. *TEST* 2019;**28**:900–19. doi: 10.1007/s11749-018-0611-5. http://link.springer.com/10.1007/s11749-018-0611-5

Baharlou H, Canete NP, Cunningham AL *et al.* Mass cytometry imaging for the study of human diseases—applications and data analysis strategies. *Front Immunol* 2019;**10**:2657. doi: 10.3389/fimmu.2019.02657. https://www.frontiersin.org/article/10.3389/fimmu.2019.02657/full

Bai Y, Zhu B, Rovira-Clave X *et al.* Adjacent cell marker lateral spillover compensation and reinforcement for multiplexed images. *Front Immunol* 2021;**12**:652631. doi: 10.3389/fimmu.2021.652631.

Batth IS, Meng Q, Wang Q *et al.* Rare osteosarcoma cell subpopulation protein array and profiling using imaging mass cytometry and bioinformatics analysis. *BMC Cancer* 2020;**20**:715. doi: 10.1186/s12885-020-07203-7.

Black S, Phillips D, Hickey JW *et al.* CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat Protoc* 2021;**16**:3802–35. doi: 10.1038/s41596-021-00556-8. https://www.nature.com/articles/s41596-021-00556-8

Bortolomeazzi M, Keddar MR, Montorsi L *et al.* Immunogenomics of colorectal cancer response to checkpoint blockade: analysis of the KEYNOTE 177 trial and validation cohorts. *Gastroenterology* 2021;**161**:1179–93. doi: 10.1053/j.gastro.2021.06.064. https://linkinghub.elsevier.com/retrieve/pii/S0016508521031784

Brbić M, Cao K, Hickey JW *et al.* Annotation of spatially resolved single-cell data with STELLAR. *Nat Methods* 2022;**19**:1411–8.

Cao Y, Geddes TA, Yang JYH *et al.* Ensemble deep learning in bioinformatics. *Nat Mach Intell* 2020;**2**:500–8.

Damond N, Engler S, Zanotelli VR *et al.* A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metab* 2019;**29**:755–68.e5. doi: 10.1016/j.cmet.2018.11.014. https://linkinghub.elsevier.com/retrieve/pii/S1550413118306910

Eng C-HL, Lawson M, Zhu Q *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 2019;**568**:235–9. doi: 10.1038/s41586-019-1049-y. http://www.nature.com/articles/s41586-019-1049-y

Etemad K, Chellappa R. Discriminant analysis for recognition of human face images. *J Opt Soc Am A* 1997;**14**:1724. doi: 10.1364/JOSAA.14.001724. https://opg.optica.org/abstract.cfm?URI=josaa-14-8-1724

Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. *J Am Stat Assoc* 1983;**78**:553–69. doi: 10.1080/01621459.1983.10478008. http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478008

Fujita A, Takahashi DY, Patriota AG. A non-parametric method to estimate the number of clusters. *Comput Stat Data Anal* 2014;**73**:27–39. doi: 10.1016/j.csda.2013.11.012. https://linkinghub.elsevier.com/retrieve/pii/S0167947313004507

Giesen C, Wang HAO, Schapiro D *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* 2014;**11**:417–22.

Hartigan JA, Hartigan PM. The dip test of unimodality. *Ann Statist* 1985;**13**:70–84. doi: 10.1214/aos/1176346577. https://projecteuclid.org/journals/annals-of-statistics/volume-13/issue-1/The-Dip-Test-of-Unimodality/10.1214/aos/1176346577.full

Hoch T, Schulz D, Eling N *et al.* Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy. *Sci Immunol* 2022;**7**:eabk1692. doi: 10.1126/sciimmunol.abk1692. https://www.science.org/doi/10.1126/sciimmunol.abk1692

Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;**12**:296–8. doi: 10.1197/jamia.M1733. https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M1733

Karim MR, Beyan O, Zappa A *et al.* Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform* 2021;**22**:393–415. doi: 10.1093/bib/bbz170.

Keren L, Bosse M, Thompson S *et al.* MIBI-TOF: a multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci Adv* 2019;**5**:eaax5851. doi: 10.1126/sciadv.aax5851. https://www.science.org/doi/10.1126/sciadv.aax5851

Kim T, Chen IR, Lin Y *et al.* Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinform* 2019;**20**:2316–26. doi: 10.1093/bib/bby076. https://academic.oup.com/bib/article/20/6/2316/5077112

Kvålseth T. On normalized mutual information: measure derivations and properties. *Entropy* 2017;**19**:631. doi: 10.3390/e19110631. http://www.mdpi.com/1099-4300/19/11/631

Lee E, Chern K, Nissen M *et al.*; IMAXT Consortium. SpatialSort: a Bayesian model for clustering and cell population annotation of spatial proteomics data. *Bioinformatics* 2023;**39**:i131–9. doi: 10.1093/bioinformatics/btad242.

Levine JH, Simonds EF, Bendall SC *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 2015;**162**:184–97. doi: 10.1016/j.cell.2015.05.047. http://dx.doi.org/10.1016/j.cell.2015.05.047

Lewis SM, Asselin-Labat M-L, Nguyen Q *et al.* Spatial omics and multiplexed imaging to explore cancer biology. *Nat Methods* 2021;**18**:997–1012. doi: 10.1038/s41592-021-01203-6.

Liu CC, Bosse M, Kong A *et al.* Reproducible, high-dimensional imaging in archival human tissue by multiplexed ion beam imaging by time-of-flight (MIBI-TOF). *Lab Invest* 2022;**102**:762–70. doi: 10.1038/s41374-022-00778-8. https://www.nature.com/articles/s41374-022-00778-8

Liu CC, Greenwald NF, Kong A *et al.* Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *Nat Commun* 2023;**14**:4618. doi: 10.1038/s41467-023-40068-5.

Lohoff T, Ghazanfar S, Missarova A *et al.* Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat Biotechnol* 2022;**40**:74–85. doi: 10.1038/s41587-021-01006-2. https://www.nature.com/articles/s41587-021-01006-2

McCaffrey EF, Donato M, Keren L *et al.* The immunoregulatory landscape of human tuberculosis granulomas. *Nat Immunol* 2022;**23**:318–29. doi: 10.1038/s41590-021-01121-x. https://www.nature.com/articles/s41590-021-01121-x

Meinshausen N. Hierarchical testing of variable importance. *Biometrika* 2008;**95**:265–78. doi: 10.1093/biomet/asn007. https://doi.org/10.1093/biomet/asn007

Melssen W, Wehrens R, Buydens L. Supervised Kohonen networks for classification problems. *Chemometr Intell Lab Syst* 2006;**83**:99–113. doi: 10.1016/j.chemolab.2006.02.003. https://linkinghub.elsevier.com/retrieve/pii/S016974390600027X

Miljkovic D. Brief review of self-organizing maps. In: *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia. 1061–6. IEEE, 2017. doi: 10.23919/MIPRO.2017.7973581. http://ieeexplore.ieee.org/document/7973581/

Mokari M, Mohammadzade H, Ghojogh B. Recognizing involuntary actions from 3D skeleton data using body states. *Sci Iran* 2018;**27**:1424–36. doi: 10.24200/sci.2018.20446. http://scientiairanica.sharif.edu/article_20446.html

Moldoveanu D, Ramsay L, Lajoie M *et al.* Spatially mapping the immune landscape of melanoma using imaging mass cytometry. *Sci Immunol* 2022;**7**:eabi5072. doi: 10.1126/sciimmunol.abi5072. https://www.science.org/doi/10.1126/sciimmunol.abi5072

Nielsen F. Hierarchical clustering. In: *Introduction to HPC with MPI for Data Science*. Undergraduate Topics in Computer Science. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-21903-5_8

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190. doi: 10.1371/journal.pgen.0020190. https://doi.org/10.1371/journal.pgen.0020190

Phillips D, Matusiak M, Gutierrez BR *et al.* Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. *Nat Commun* 2021;**12**:6726. doi: 10.1038/s41467-021-26974-6. https://www.nature.com/articles/s41467-021-26974-6

Rendeiro AF, Ravichandran H, Bram Y *et al.* The spatial landscape of lung pathology during COVID-19 progression. *Nature* 2021;**593**:564–9. doi: 10.1038/s41586-021-03475-6. http://www.nature.com/articles/s41586-021-03475-6

Revelle W. Psych: procedures for psychological, psychometric, and personality research. Evanston, Illinois: Northwestern University, R package version 2.2.9. https://CRAN.R-project.org/package=psych. 2022.

Risom T, Glass DR, Averbukh I *et al.* Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* 2022;**185**:299–310.e18. doi: 10.1016/j.cell.2021.12.023. https://linkinghub.elsevier.com/retrieve/pii/S0092867421014860

Rossi F. yasomi: yet another self organising map implementation. R package version 0.3/r39. https://R-Forge.R-project.org/projects/yasomi/. 2012.

Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65. doi: 10.1016/0377-0427(87)90125-7. https://linkinghub.elsevier.com/retrieve/pii/0377042787901257

Samadani A-A, Kubica E, Gorbet R *et al.* Perception and generation of affective hand movements. *Int J of Soc Robotics* 2013;**5**:35–51. doi: 10.1007/s12369-012-0169-4. http://link.springer.com/10.1007/s12369-012-0169-4

Samusik N, Good Z, Spitzer MH *et al.* Automated mapping of phenotype space with single-cell data. *Nat Methods* 2016;**13**:493–6. doi: 10.1038/nmeth.3863. http://www.nature.com/articles/nmeth.3863

Schürch CM, Bhate SS, Barlow GL *et al.* Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* 2020;**182**:1341–59.e19. doi: 10.1016/j.cell.2020.07.005. https://linkinghub.elsevier.com/retrieve/pii/S0092867420308709

Silverman BW. Using kernel density estimates to investigate multimodality. *J R Stat Soc Series B Methodol* 1981;**43**:97–9. doi: 10.1111/j.2517-6161.1981.tb01155.x. https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1981.tb01155.x

Simpson GL, Oksanen J. analogue: analogue and weighted averaging methods for palaeoecology. R package version 0.17-6. https://cran.r-project.org/package=analogue. 2021.

Smith M. ReductionWrappers: wrapper exposing several Python dimensional reduction tools. R Package Version 2.5.4. 2023.

Steinley D. Properties of the Hubert-Arable adjusted rand index. *Psychol Methods* 2004;**9**:386–96. doi: 10.1037/1082-989X.9.3.386. http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.9.3.386

Sugar CA, James GM. Finding the number of clusters in a dataset: an information-theoretic approach. *J Am Stat Assoc* 2003;**98**:750–63. doi: 10.1198/016214503000000666. http://www.tandfonline.com/doi/abs/10.1198/016214503000000666

Thrun MC, Stier Q. Fundamental clustering algorithms suite. *SoftwareX* 2021;**13**:100642. doi: 10.1016/j.softx.2020.100642. https://linkinghub.elsevier.com/retrieve/pii/S2352711020303551

Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Series B Stat Methodol* 2001;**63**:411–23. doi: 10.1111/1467-9868.00293. https://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00293

Tracy CA, Widom H. Level-spacing distributions and the Airy kernel. *Commun Math Phys* 1994;**159**:151–74. doi: 10.1007/BF02100489. http://link.springer.com/10.1007/BF02100489

Van Gassen S, Callebaut B, Van Helden MJ *et al.* FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data: flowSOM. *Cytometry A* 2015;**87**:636–45. doi: 10.1002/cyto.a.22625. https://onlinelibrary.wiley.com/doi/10.1002/cyto.a.22625

Van Maldegem F, Valand K, Cole M *et al.* Characterisation of tumour microenvironment remodelling following oncogene inhibition in preclinical studies with imaging mass cytometry. *Nat Commun* 2021;**12**.

Watson ER, Mora A, Taherian Fard A *et al.* How does the structure of data impact cell–cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data. *Brief Bioinform* 2022;**23**:bbac387. doi: 10.1093/bib/bbac387. https://academic.oup.com/bib/article/doi/10.1093/bib/bbac387/6712300