Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Variable selection for nonlinear dimensionality reduction of biological datasets through bootstrapping of correlation networks

David G. Aragones [a], Miguel Palomino-Segura [b,c,d], Jon Sicilia [b], Georgiana Crainiciuc [b], Iván Ballesteros [b], Fátima Sánchez-Cabo [e], Andrés Hidalgo [f], Gabriel F. Calvo [a,*]

[a] *Department of Mathematics & MOLAB-Mathematical Oncology Laboratory, Universidad de Castilla-La Mancha, Ciudad Real, Spain*
[b] *Area of Cell and Developmental Biology, Centro Nacional de Investigaciones Cardiovasculares Carlos III, Madrid, Spain*
[c] *Immunophysiology Research Group, Instituto Universitario de Investigación Biosanitaria de Extremadura (INUBE), Badajoz, Spain*
[d] *Department of Physiology, Faculty of Sciences, University of Extremadura, Badajoz, Spain*
[e] *Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares Carlos III, Madrid, Spain*
[f] *Vascular Biology and Therapeutics Program and Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA*

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Unsupervised learning
Feature selection
UMAP
Complex systems

ABSTRACT

Identifying the most relevant variables or features in massive datasets for dimensionality reduction can lead to improved and more informative display, faster computation times, and more explainable models of complex systems. Despite significant advances and available algorithms, this task generally remains challenging, especially in unsupervised settings. In this work, we propose a method that constructs correlation networks using all intervening variables and then selects the most informative ones based on network bootstrapping. The method can be applied in both supervised and unsupervised scenarios. We demonstrate its functionality by applying Uniform Manifold Approximation and Projection for dimensionality reduction to several high-dimensional biological datasets, derived from 4D live imaging recordings of hundreds of morpho-kinetic variables, describing the dynamics of thousands of individual leukocytes at sites of prominent inflammation. We compare our method with other standard ones in the field, such as Principal Component Analysis and Elastic Net, showing that it outperforms them. The proposed method can be employed in a wide range of applications, encompassing data analysis and machine learning.

1. Introduction

Dimensionality reduction (DR) is a fundamental task in data analysis and machine learning, aimed at extracting meaningful and concise representations from high-dimensional data. DR transforms a given dataset from an original high-dimensional space into a new low-dimensional one so that the new representation retains meaningful properties of interest. This can be done for various reasons, such as reducing noise, decreasing computational complexity, improving accuracy, enhancing interpretability, facilitating data visualization, and supporting cluster analysis [1–3]. The popularity of DR has grown in parallel with the wider availability of large datasets in multiple fields, particularly in biomedicine [4–7].

DR methods can be linear and nonlinear. Linear methods, such as Principal Components Analysis (PCA) [8], offer the advantage of lower computational complexity. However, they also possess limitations, as they can only capture linear relationships between variables. As a result, nonlinear patterns in the original data may remain hidden when applying these types of methods [3]. To circumvent this issue, a variety of nonlinear DR methods have been developed, which can handle the complex nonlinear patterns commonly found in real-world data [9]. Among these, t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) have gained significant popularity in recent years [2]. t-SNE is based on minimizing the Kullback–Leibler divergence between probability distributions [10,11], while UMAP operates by minimizing the cross-entropy between weighted graphs [9]. Although they share visual similarities, UMAP provides several advantages over t-SNE, notably, higher computational efficiency [12].

Ultimately, when comparing linear and nonlinear DR methods, two crucial factors emerge: efficiency and interpretability. Contrary to linear methods, nonlinear techniques tend to be computationally more demanding. Techniques like PCA have computational complexity that scales linearly with the number of features, whereas nonlinear methods, such as t-SNE, are more computationally expensive, with complexity

[*] Corresponding author.
E-mail address: gabriel.fernandez@uclm.es (G.F. Calvo).

https://doi.org/10.1016/j.compbiomed.2023.107827
Received 30 August 2023; Received in revised form 15 November 2023; Accepted 4 December 2023
Available online 10 December 2023

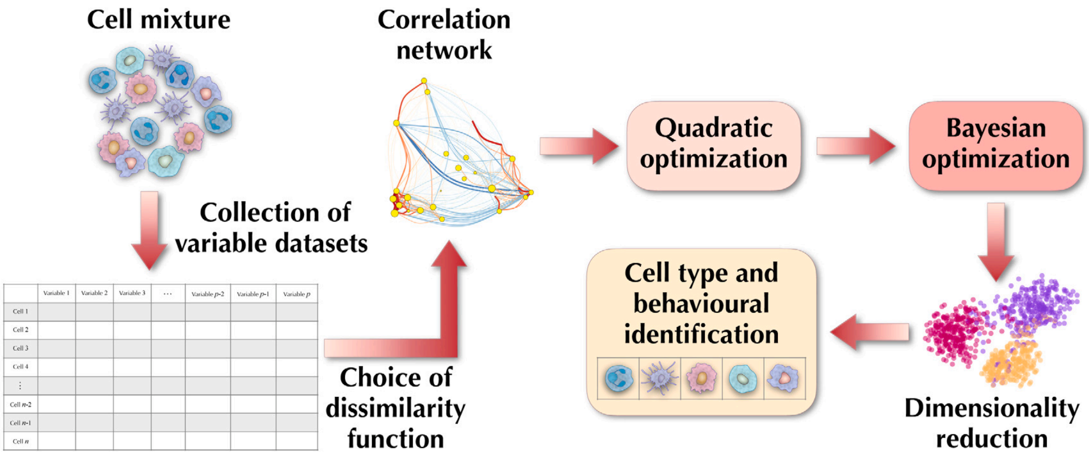

Fig. 1. Flow chart of the proposed method.

that can scale quadratically or even worse. Furthermore, nonlinear methods yield results that are less readily interpretable. The lower-dimensional representation obtained from these methods does not have a direct correspondence to the original variables. While they can still provide valuable insights into the structure of the data, understanding the precise contribution of each variable can be challenging.

Despite the usefulness of DR methods, they should be applied with caution due to two main factors: (i) the risk of overfitting, particularly for datasets with a large number of variables relative to the number of observations [13]; and (ii) the presence of hyperparameters, the values of which must be manually defined and can strongly affect the projection results [2]. One partial solution to the first limitation is to use PCA as a preliminary step, applying nonlinear DR to the first principal components of the dataset. This technique effectively reduces the number of variables, although they are no longer the original ones; instead, they become linear combinations of the original variables, which can hinder the interpretability of results.

The risk of overfitting can also be mitigated by identifying a subset of the most informative variables from the original dataset [14]. The fundamental premise is that the original data may contain irrelevant or redundant variables that can be omitted from the model [15]. However, an exhaustive search is typically computationally infeasible. Given $p \gg 1$ initial variables, a brute-force inspection of all possible combinations would involve analyzing approximately 2^p non-empty variable subsets. To address this issue, both supervised and unsupervised approaches can be utilized [16,17]. While supervised approaches are often more suitable and have been studied more extensively, they require additional knowledge that is not always available in practice [18]. This has spurred significant interest in unsupervised methods in recent years [19]. In this context, network theory approaches can be powerful tools for revealing hidden relationships among variables. Patterns rarely appear based on a single variable, but rather emerge from interactions among multiple ones [20–22]. Within this framework, variable selection is analogous to selecting the nodes of a network [15].

The challenge of identifying the most informative variables for characterizing cell states is pervasive in many high-dimensional biological datasets [23–26]. These datasets contain thousands of observations and a large number of variables, making it difficult to extract meaningful insights. Popular software packages like *Seurat* [27] are widely used for unsupervised cluster analysis but often lack the capability to select variables within the original space. This limitation poses a significant hurdle when accurately characterizing different cell states [28–31]. Addressing this challenge is crucial for advancing our understanding of complex biological systems.

In this work, we propose a method based on bootstrapping of nonlinear correlation networks and quadratic optimization, that can be applied in both supervised and unsupervised settings. Our method not only reduces the number of original variables; it does so by identifying the relevant ones to improve the performance of nonlinear DR techniques. We have tested the performance of our method and compared it with other standard procedures, particularly PCA and Elastic Net (EN), using four distinct experimental high-dimensional datasets. These datasets were collected from 4D imaging in living mice capturing over one hundred morpho-kinetic variables. These variables portray the dynamics of thousands of individual leukocytes at sites of active inflammation [32]. Our goal is to pinpoint the most informative variables that facilitate cell type/state identification, while simultaneously mitigating the risk of overfitting. While our study primarily used UMAP for DR, our approach can be easily adapted to other DR methods. Fig. 1 summarizes our framework.

2. Materials and methods

In this section, we present the experimental data and the mathematical methods used in our research. The main steps we follow are: (a) first, we estimate the importance of individual variables in an unsupervised way by bootstrapping nonlinear correlation networks; (b) we then use quadratic optimization to select a subset of variables with minimum correlations and maximum relevance; (c) subsequently, we apply Bayesian optimization to identify the hyperparameters for UMAP, the DR technique we adopted and (d) finally, we compare the results obtained from our approach with those from standard methods (PCA and EN), and also with our own procedure, using supervised information instead of bootstrapping of correlation networks to infer the relevance of each variable.

2.1. Experimental datasets

In this work, we have employed intravital imaging datasets from four different experimental settings previously published in [32,33] (see Fig. 2): (i) Influenza infection in the trachea; (ii) Inflammation of the cremaster muscle; (iii) Laser injury in the skin; and (iv) Ischemia-reperfusion in skin. In brief, we have used two imaging modalities in 4D, using either a two-photon microscopy system (TrimScope, LaVision BioTec, Bielefeld, Germany) or a spinning-disk intravital system (VIVO, from Intelligent Imaging Innovations, Denver, USA). We used timelapse capture in up to four fluorescence channels to record the morphology and movement of leukocytes (neutrophils, dendritic cells and macrophages) in their native inflamed environment in the trachea, skin, tumors and vasculature of the cremasteric vasculature. We processed and corrected the newly generated and existing datasets (trachea, skin, bone marrow and tumors), by performing drift correction and channel unmixing using custom scripts (Python 3.5) and FIJI. 4D captures were further analyzed with the Imaris software (Oxford intruments,

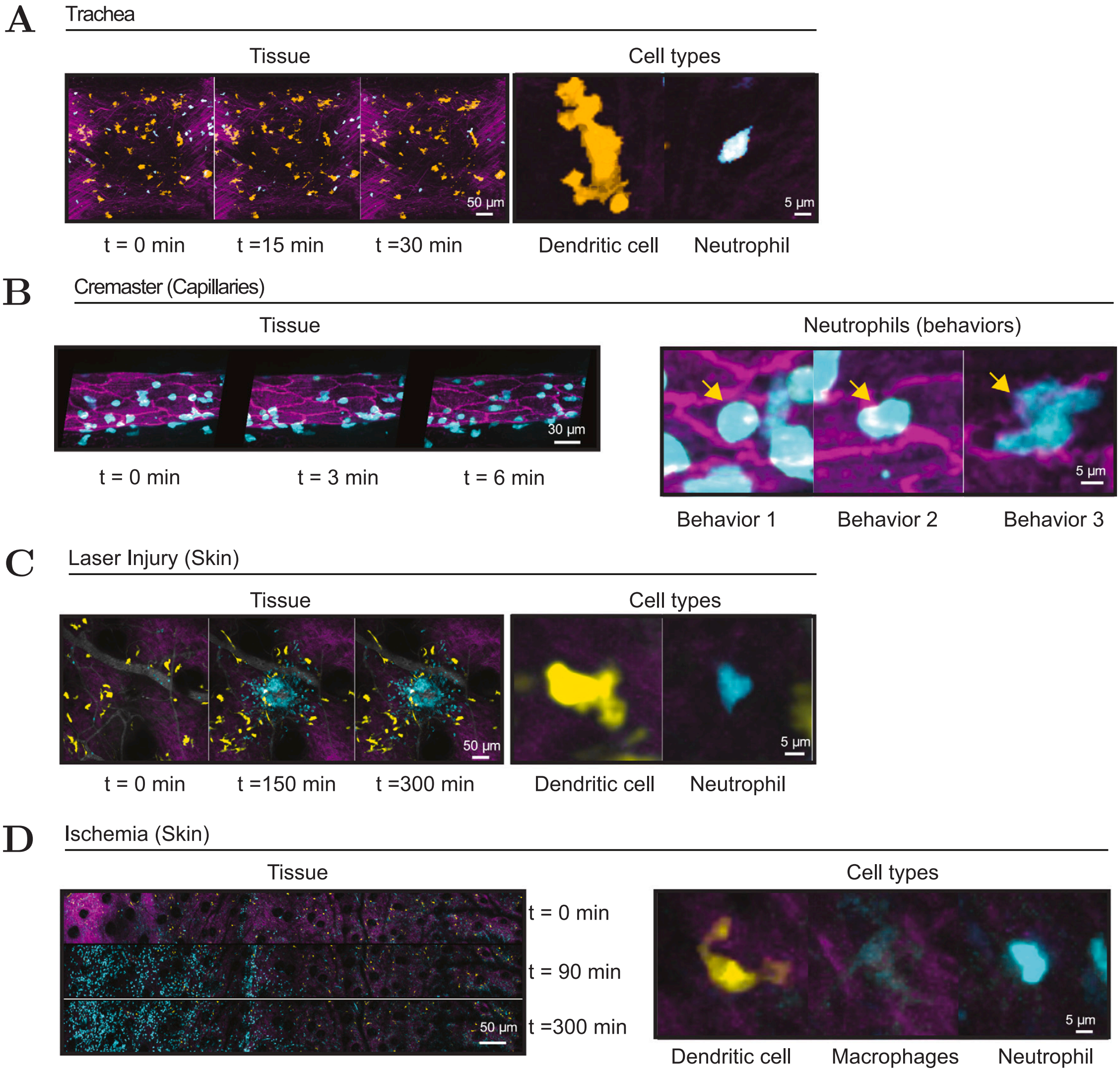

Fig. 2. Intravital imaging of immune cells (neutrophils, dendritic cells and macrophages) and tissues that generated the datasets analyzed in this study: **(A)** Influenza infection in the trachea; **(B)** Inflammation in the cremaster muscle; **(C)** Laser-induced injury in the skin; and **(D)** Ischemia-reperfusion in skin. The imaging not only facilitated dataset generation but also aided in identifying cell types and their behaviors.

9.5.1) for all tissues, except for the cremaster muscle for which we use our recently generated analytical method Automated Cell Migration Examination (ACME; available at [34]) (PMID 35066392). ACME was designed to perform automatic feature extraction for migrating cells, including automatic detection, segmentation and tracking of cells within vessels. Additional information can be found in Appendix A.

2.2. Nonlinear correlation networks

The initial stage of our procedure requires the construction of nonlinear correlation networks derived from different datasets. These networks are useful in unraveling relationships amongst variables that are not necessarily linear. Either parametric or non-parametric correlation coefficients can be used for building these networks, based on the assumptions made regarding the underlying probability distributions. In this context, we introduce a correlation matrix **P**, comprising entries denoted as ρ_{ij}, having the following structure:

$$\rho_{ij} = \mathbf{A}_i \cdot \mathbf{A}_j \,, \tag{1}$$

which involve a dot product between two unit vectors $\mathbf{A}_i$ and $\mathbf{A}_j$ (i.e., satisfying $\|\mathbf{A}\|^2 = 1$), corresponding to variables i and j, respectively. Here, $i, j = 1, 2, \dots, p$, where p is the total number of variables measured, and each unit vector has a dimension of n, denoting the total number of observations. This correlation matrix provides an approximation of the redundancy between each pair of variables, with correlation coefficients ρ_{ij} ranging from -1 to 1.

In the case of Pearson correlation coefficients, values equal to -1 or 1 indicate perfect negative or positive linear correlation, respectively, while 0 signifies no linear correlation. Alternatively, Spearman's rank correlation coefficient provides a non-parametric measure of a monotonic relationship between two datasets. Unlike the Pearson coefficient, Spearman's does not assume normal distribution for both datasets. Similar to Pearson's, Spearman's coefficient can range from -1 to 1, with extreme values indicating a perfect non-increasing or non-decreasing monotonic relationship between each variable, respectively. Values near 0 suggest a weak monotonic relationship between the variables. Other correlation coefficients, whether parametric (e.g., Point-Biserial, Phi) or non-parametric (e.g. Kendall's Tau), can similarly be expressed in the form (1), thereby offering alternatives to Pearson's or Spearman's correlation measures.

To build a correlation network, we transformed (1) into a dissimilarity $p \times p$ matrix. This allows for a more interpretable depiction of the relationships among variables. The chosen transformation is defined as (other possible forms are detailed in Appendix B):

$$d_{ij} = \sqrt{\frac{1-\rho_{ij}}{2}}\,. \tag{2}$$

This expression results in a pseudo-metric, as it fulfills the following conditions:

1. The dissimilarity between a variable and itself is zero: $d_{ii} = 0$.
2. The dissimilarity between different variables is within the interval: $d_{ij} \in [0,1]$.
3. Dissimilarities are symmetric: $d_{ij} = d_{ji}$.
4. Dissimilarities satisfy the triangle inequality: $d_{ik} \leq d_{ij} + d_{jk}$. A proof of this is presented in Appendix B.

It is important to note that property 2 leads to an interesting observation: when two different variables are perfectly redundant, their dissimilarity is zero ($d_{ij} = 0$). This feature renders d_{ij} as a pseudo-metric. Moreover, due to properties 1 and 3, d_{ij} also satisfies the definition of an undirected weighted adjacency matrix.

In order to analyze the constructed network in a 2D Euclidean space, we employed Torgerson-Gower inner-product multidimensional scaling. In this method, the variables (or nodes) of the original network are projected onto a 2D plane, striving to preserve the distances between pairs as closely as possible to those in the original space. This was accomplished by minimizing the following function:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{p\times 2}} \left(\|\boldsymbol{B}\|_F^{-1} \left\| \boldsymbol{B} - \boldsymbol{X}^T \cdot \boldsymbol{X} \right\|_F \right), \tag{3}$$

where $\boldsymbol{X}$ denotes the positions of the variables in the 2D Euclidean space, $\|*\|_F$ represents the $L_{2,2}$ norm (or the Frobenius norm), and $\boldsymbol{B}$ is a matrix obtained after centering the dissimilarity matrix given by (2). It is worth noting that this function is convex, which facilitates efficient numerical optimization. More details on the multidimensional scaling process can be found in Appendix C.

To estimate the importance of each variable in an unsupervised way, we have applied bootstrapping, which is random sampling with replacement. The goal was to analyze how stable each variable was under perturbations of the data, as an indirect measure for assessing the central tendency and dispersion of the positions of variables after multidimensional scaling. The density function of displacements for each variable after bootstrapping was estimated by means of kernel density estimation. This is a non-parametric method based on a finite sample and a kernel function:

$$\hat{f}_\nu(\Delta) = \frac{1}{m}\sum_{k=1}^{m} \phi_\nu(\Delta - \Delta_k) = \frac{1}{m\nu}\sum_{k=1}^{m} \phi\left(\frac{\Delta-\Delta_k}{\nu}\right), \tag{4}$$

where $\hat{f}_\nu$ denotes the estimation of the density function, m is the resampling size, ϕ the kernel function used, Δ the displacement from the position when employing the full dataset, Δ_k the displacement when considering resample k, and ν the corresponding bandwidth for the estimation. We employed a standard Gaussian density function as a kernel, while the bandwidth was determined by minimizing the mean integrated squared error.

2.3. Quadratic optimization

To take into account both the correlations between pairs of variables and their relevance, we formulated a quadratic optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p}} \left[\frac{1-\alpha}{2}\, \boldsymbol{\beta}\cdot(\boldsymbol{P}\odot\boldsymbol{P})\cdot\boldsymbol{\beta} - \alpha\, \boldsymbol{c}\cdot\boldsymbol{\beta} \right], \tag{5}$$

subject to $\|\boldsymbol{\beta}\|_1 = 1$ and $\boldsymbol{\beta} \geq \boldsymbol{0}$, where $\boldsymbol{\beta}$ is a vector containing the normalized weight assigned to each variable. Here, $\boldsymbol{P}$ is the chosen correlation matrix whose components are given by (1), and $\odot$ denotes the Hadamard-Schur product. In this study, we employed Spearman's rank correlation matrix, but other forms could also be used. Furthermore, $\|*\|_1$ represents the L_1 norm (also known as the Manhattan norm), and $\alpha \in [0,1]$ is a parameter that determines the relative weight assigned to the relevance of the variables compared to correlations. Lastly, $\boldsymbol{c}$ is a vector containing the estimated relevance of each variable. It is calculated as follows:

$$\boldsymbol{c} = \boldsymbol{1} - \frac{\overline{\boldsymbol{\Delta}} - \min(\overline{\boldsymbol{\Delta}})}{\max(\overline{\boldsymbol{\Delta}}) - \min(\overline{\boldsymbol{\Delta}})}\,, \tag{6}$$

where $\overline{\boldsymbol{\Delta}}$ is a vector containing the sample mean displacement for each variable. This displacement is computed based on the change in position of the variable after bootstrapping in multidimensional scaling. Notice that $\boldsymbol{c} \in [0,1]^p$, with higher values indicating a lower average displacement from the position of the variable (node) in the full network. These are precisely the variables that are considered more relevant in the unsupervised approach. Thus, our method ensures that the more stable a variable is under data perturbations, the more relevant or informative it becomes.

To solve numerically the optimization problem given in (5), we have employed an operator splitting method that does not impose strict convexity on the objective function [35]. Operator splitting methods allow for the division of the original problem into smaller, more manageable subproblems. This is particularly advantageous when handling complex, high-dimensional optimization tasks. Indeed, correlation matrices, when calculated numerically, can lead to small negative eigenvalues, violating the property that makes them positive semidefinite. This is important because a positive semidefinite matrix ensures the strict convexity of the objective function. Therefore, employing an operator splitting method that does not impose strict convexity becomes crucial. Moreover, this method is computationally less expensive than others such as the interior-point and active-set methods, proving useful when dealing with a large number of variables. We refer the reader to Appendix D for further details on first-order conditions of optimality in the quadratic optimization formulation employed here.

2.4. Bayesian optimization

One of the main limitations of nonlinear DR methods is their dependence on hyperparameters, whose values must be manually defined. In UMAP [12], the DR technique employed here, one parameter that strongly affects the outcome is the number of neighbors used in the network construction stage. Moreover, the variable selection process, based on quadratic optimization, requires an extra hyperparameter (α in (5)) to be defined. To address these issues, we employed Bayesian optimization [36], a sequential design strategy aimed at finding the global optimum of black-box functions via probabilistic surrogate models. This choice was motivated by the high computational costs associated with performing multiple DR (see Appendix E).

We initiated the process by formulating a suitable objective function that would allow us to assess the performance of DR in terms of quantifying community structure [15]. To achieve this, we resorted to the network theory concept of modularity [37,38], defined by the following expression:

$$Q(\boldsymbol{X}) = \frac{1}{2s}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\left\|\boldsymbol{X}_i - \boldsymbol{X}_j\right\|_2 - \frac{k_i\, k_j}{2s}\,\delta_{ij} \right), \tag{7}$$

where δ_{ij} is a community membership indicator function (equal to 1 if observations i and j belong to the same community and 0 otherwise), k_i is the weighted degree of node i, $\|*\|_2$ represents the L_2 norm (Euclidean norm) as a distance between observations after DR, and s is the total sum of distances. We note that (7) depends on a membership function (δ_{ij}) and thus, we need a community detection method able to extract this information in an unsupervised manner. For this, we opted for the Leiden method, which is grounded on a greedy optimization of the modularity and provides computational efficiency [39].

Upon defining the objective function, we implemented the Bayesian optimization procedure. We selected Gaussian stochastic processes as priors due to their flexibility and capability to efficiently use information from sequentially sampled hyperparameters. These hyperparameters included: (i) The number of neighbors (N) in the UMAP method, which strongly influences the DR outcomes; and (ii) the parameter α, which affects the selection of the variables subset for DR in accordance with the quadratic optimization problem given in (5). Then, the Bayesian optimization problem reads as follows:

$$\max_{N\in\mathbb{N},\,\alpha\in\mathbb{R}} Q(\boldsymbol{X})\,. \tag{8}$$

This Bayesian optimization problem is subject to $N \in [N_l, N_u]$ and $\alpha \in [0, 1]$, where the number of neighbors considered in the UMAP method was constrained with a minimum value of $N_l = 5$ and a maximum value of $N_u = n/3$.

The modularity function $Q(\boldsymbol{X})$ is computationally expensive to evaluate. For each pair of N and α, one needs to: (i) Solve the quadratic optimization problem given in (5); (ii) apply UMAP to reduce data dimensionality, which requires the minimization of the cross-entropy given by (E.1) (see Appendix E); and (iii) calculate the modularity of the network using (7). To reduce the number of evaluations, we constructed a probabilistic model for the modularity, allowing for an iterative determination of the next point to evaluate in the search space. This was accomplished by maximizing the upper confidence bound of the Gaussian process at each step:

$$\begin{aligned} f_a([N,\alpha];\,\{N_r,\alpha_r,Q_r\}\,|\,\theta) &= \mu([N,\alpha];\,\{N_r,\alpha_r,Q_r\}\,|\,\theta) \\ &\quad + \kappa\,\sigma([N,\alpha];\,\{N_r,\alpha_r,Q_r\}\,|\,\theta)\,, \end{aligned} \tag{9}$$

where f_a is the acquisition function, r the number of iterations performed in the Bayesian optimization, θ denotes the hyperparameters of the Gaussian process, κ is a constant to balance exploitation against exploration of new solutions, μ the mean of the Gaussian stochastic process, and σ the standard deviation. Further details are provided in Appendix F.

2.5. Comparison with PCA and supervised approaches

To evaluate the performance of our proposed unsupervised variable selection method for DR, we compared it to two other well-established methods. The first was PCA, which, while not strictly a variable selection method, serves to reduce the initial number of variables by generating new ones in the eigenspace of the Pearson linear correlation matrix. The second was EN, a supervised regularization method that operates based on a linear combination of penalties. In addition to these comparisons, we also adapted our method into a supervised variable selection version, utilizing classification models to assess the performance of each variable. This revised method not only provides a useful contrast, but also highlights the impact of incorporating supervised information. When such information is available, it presents a viable alternative capable of potentially enhancing the precision of the variable selection.

We started our comparison with PCA, a widely used preprocessing step in many UMAP implementations, including the popular package *Seurat* for single-cell RNA sequencing data analysis [27,40–42]. PCA returns an orthogonal basis where the new variables are uncorrelated. In our calculations, we retained the first 50 principal components and disregarded the rest. To achieve this, we computed the covariance matrix of the data using:

$$\boldsymbol{K} = \frac{1}{n-1}\,(\boldsymbol{x} - \mathbf{1}\otimes\overline{\boldsymbol{x}})^{T}\cdot(\boldsymbol{x} - \mathbf{1}\otimes\overline{\boldsymbol{x}})\,,$$

where the factor $(n-1)^{-1}$ was included for the Bessel's correction. Here, $\boldsymbol{x}$ is the matrix corresponding to the dataset analyzed (with rows being observations for each cell and columns being different variables), $\mathbf{1}$ is the all-ones vector, $\overline{\boldsymbol{x}}$ is a vector containing the mean of each variable, and $\otimes$ denotes the outer product. The eigenvalues and eigenvectors of $\boldsymbol{K}$ were calculated by employing the singular value decomposition for improved numerical accuracy.

The second method we employed for comparison was EN, an embedded method for variable selection that incorporates a regularizing term with a linear combination of L_1 (Lasso) and L_2 (Ridge) penalties [43]. It is well-suited for tackling multicollinearity problems and selection of grouped variables. The objective function for EN is given by:

$$\min_{\beta\in\mathbb{R}^{p+1}}\left[-\frac{1}{n}\sum_{i=1}^{n}\ln\mathcal{L}_i(\boldsymbol{\beta}) + \lambda_2\left(\frac{1-\lambda_1}{2}\,\|\boldsymbol{\beta}\|_2^2 + \lambda_1\,\|\boldsymbol{\beta}\|_1\right)\right], \tag{10}$$

where $\boldsymbol{\beta}$ is a vector of coefficients, λ_1 and λ_2 are parameters that control the balance between the L_1 and L_2 penalties and the overall strength of the penalties, respectively. Finally, $\mathcal{L}_i$ represents the likelihood for observation i. The optimization problem was solved using cyclical coordinate descent over each parameter with the others fixed to find an optimal combination of λ_1 and λ_2 parameters. The likelihood function employed corresponds to either the binary or multinomial logistic model, depending on the number of classes in the dataset analyzed.

The third method we employed for comparison follows the same steps as our proposed unsupervised method, with a key difference: instead of using bootstrapping of nonlinear correlation networks to obtain a proxy for the relevance of each variable, we developed classification models for each one. For binary cases, we used logistic models evaluated using the McFadden likelihood ratio index:

$$R^2_{\mathrm{McF}} = 1 - \frac{\ln\mathcal{L}_M}{\ln\mathcal{L}_0}\,, \tag{11}$$

where $\mathcal{L}_M$ is the likelihood of the model developed and $\mathcal{L}_0$ the likelihood of the null model.

For ternary cases, in contrast, we developed non-parametric classification models based on decision trees, constructed using the CART algorithm. We calculated the relevance of each variable using the adjusted Rand index (ARI):

$$ARI = \frac{\sum_{ij}\binom{c_{ij}}{2} - \left[\sum_i\binom{c_{i*}}{2}\sum_j\binom{c_{*j}}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{c_{i*}}{2} + \sum_j\binom{c_{*j}}{2}\right] - \left[\sum_i\binom{c_{i*}}{2}\sum_j\binom{c_{*j}}{2}\right]\Big/\binom{n}{2}}\,, \tag{12}$$

where c_{ij} are the elements of the confusion matrix obtained from the model developed, c_{i*} represents the sum of row i, and c_{*j} is the sum of column j of the same matrix.

3. Results and discussion

In this section, we present the main results obtained in the DR analysis, using both unsupervised and supervised approaches, of the four different datasets included in our study: (i) Influenza infection in the trachea (referred to as "Trachea" dataset); (ii) Inflammation of the cremaster muscle (the muscle of the spermatic cord, referred to as "Cremaster" dataset); (iii) Laser injury in the skin (referred to as "Laser"); and (iv) Ischemia-reperfusion in skin (referred to as "Ischemia"). These datasets were generated in [32] from imaging experiments in specific tissues and inflammatory contexts in which myeloid cells express cytoplasmatic fluorescent proteins allowing for precise spatiotemporal measurement of their morphology and movement through spinning disk or multiphoton in vivo microscopy.

3.1. Nonlinear correlation networks and bootstrapping

Firstly, in Fig. 3, we present the nonlinear correlation networks obtained by employing multidimensional scaling for each of the datasets analyzed. These projections were computed using the full datasets – that is, we took into account all available observations and variables prior to the process of bootstrapping and subset selection described in the previous section. In these networks, variables with higher positive correlations are positioned closer together, whereas those with

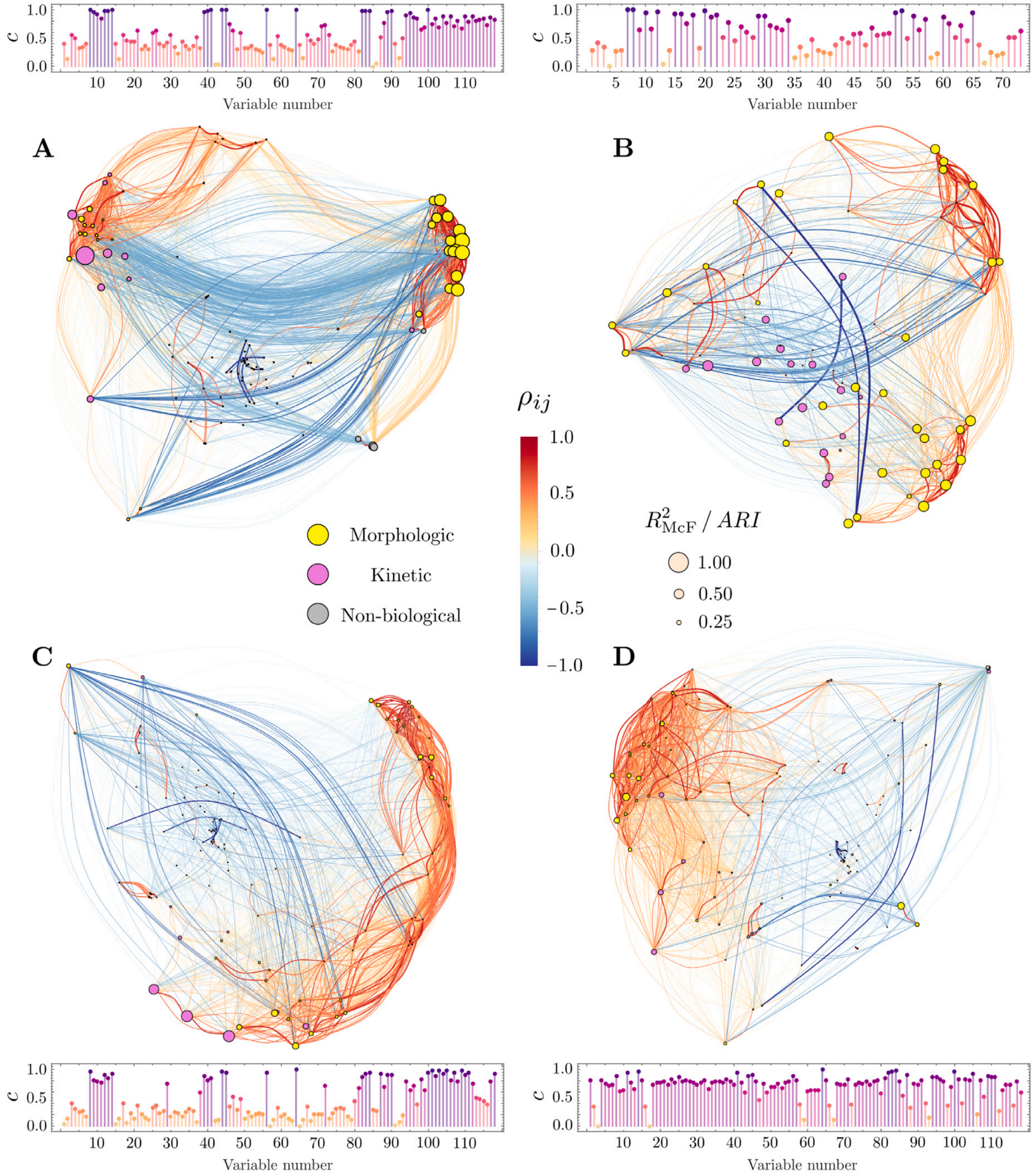

Fig. 3. Nonlinear correlation networks of different datasets analyzed from experiments with Imaris image analysis, together with the estimated relevance of each variable using (6) after bootstrapping (bar plots insets): **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Variables are shown as nodes whose diameters are proportional to their R^2_{McF} or ARI, defined in (11) and (12), respectively. Edges connecting pairs of variables have been drawn with colors (red/blue) and thicknesses proportional to the absolute value of the Spearman's nonlinear correlation coefficient for each pair. Variable types have also been highlighted, using yellow for morphologic variables and magenta for kinetic variables.

higher negative correlations are positioned further apart. Notably, in the Trachea dataset (Fig. 3.A), we observe two distinct clusters of variables. On the right, there is a cluster dominated by morphological variables (marked in yellow), which are deemed relevant according to the available supervised information (represented by larger node diameters proportional to the McFadden likelihood ratio index (11)). On the left, there is a second cluster dominated by kinetic variables (marked in magenta), which also contains some relevant variables. The remaining variables in the dataset, which are weakly correlated with these two clusters, exhibit less predictive capability between cell classes. Alongside that network, the corresponding values c, calculated from (6), yield the estimated relevance of each variable on a 0–1 scale, where values closer to 1 indicate the most informative variables.

The Cremaster dataset, depicted in Fig. 3.B, differs from the Trachea dataset in that the variables are generally less clustered. Two small clusters can be seen, one in the top-right and another in the bottom-right section, both dominated by morphologic-type variables. In contrast to the Trachea case, most variables in the Cremaster dataset exhibit significant predictive power, as measured by their ARI, which was computed from (12). Lastly, the Laser and Ischemia datasets, shown in Figs. 3.C and 3.D, display a similar pattern characterized by a large cluster of highly-correlated variables. Interestingly, irrespective of their cluster membership, most of these variables demonstrate low predictive power. This suggests that a larger subset of variables may need to be selected to distinguish meaningfully between classes upon applying DR. For additional information regarding the capability of each variable to differentiate between classes, we refer the reader to Appendix G and Appendix H. More in depth correlation analyses between variables are presented in Appendix I and Appendix J.

3.2. Variables selected

In Fig. 4, we display the variables selected for each of the analyzed datasets using both supervised and unsupervised methods. Unsupervised approaches are highlighted in light yellow, while supervised approaches are in light pink. The first row illustrates the variables selected with PCA. Even though we applied a cut-off of 50 variables, notice that they are linear combinations of all the original variables. Consequently, all variables have been used and the complete network is illustrated. In this case, we have employed a grayscale to depict the links, as correlations between variables have been removed due to the linear transformation executed by PCA.

The second row in Fig. 4 presents the variables selected using the EN method. Compared to the previous approach (PCA), several differences are apparent. Primarily, this method is supervised, therefore it requires prior knowledge of the classes present in the data (cell types and neutrophil behaviors). Unlike PCA, the selection of variables here involves selecting directly a subset of the original ones; hence, they are not transformed, which potentially improves the interpretability of results. However, the selected variables are not uncorrelated. In the Trachea dataset, we achieved the best results in terms of reducing the number of variables, from the original 118 to 27. For more details on EN results, we refer the reader to Appendix K.

The third row of Fig. 4 showcases the variables selected by the unsupervised method proposed in this study. It is evident that the number of selected variables is significantly reduced compared to PCA (which used all the original variables) and EN, across all analyzed datasets. The subgraphs corresponding to the subset of selected variables demonstrate that this method identified variables with high predictive power and low correlations. This is consistent with the logic underlying the algorithm's development, in which mean displacements from bootstrapping served as an indicator of variable relevance.

Lastly, the fourth row in Fig. 4 depicts the subgraphs corresponding to a supervised selection of variables. The process mirrored the one in the unsupervised case, except the displacements calculated from bootstrapping were substituted by the R^2_{McF} or ARI of each variable. This case is useful in order to assess the performance of the proposed method with supervised information. It is noticeable that the number of selected variables has been further reduced, having now only between 4 and 6 variables needed to describe the differences between classes across all datasets. These selected variables are among those with the highest individual predictive power, while maintaining very low correlations.

To summarize the outcomes concerning variable reduction, Table 1 has been compiled to detail the obtained results. This table illustrates the percentage reduction relative to the original variable count, with the actual number of retained variables indicated within parentheses, as determined by the L_0 norm (or sparse norm) of the vector weights ($\|\beta\|_0$). It can be seen that PCA, while involving a transformation of variables, does not reduce their original count. In contrast, EN exhibits a large reduction in the variable count for the Trachea and Laser datasets. However, its effectiveness is comparatively reduced in the Cremaster and Ischemia datasets. The unsupervised method developed in this study achieves a notable reduction in variables, averaging an 84.3% decrease across different datasets. Furthermore, the integration of supervised information, as shown in the last row, leads to a remarkable reduction of 95.1% in variable count, consistent across the analyzed datasets. For more comprehensive results, please refer to Appendix L.

Table 1
Summary of variable reduction with the different methods utilized. The percentage of reduction achieved from the original dataset is shown, along with the total number of variables retained by each method (in parentheses).

	(A) Trachea	(B) Cremaster	(C) Laser	(D) Ischemia	Mean
PCA	0.0% (118)	0.0% (73)	0.0% (118)	0.0% (118)	0.0%
EN	77.1% (27)	1.4% (72)	30.5% (82)	5.1% (112)	28.5%
UNS	86.4% (16)	89.0% (8)	74.6% (30)	87.3% (15)	84.3%
SUP	95.8% (5)	93.2% (5)	94.9% (6)	96.6% (4)	95.1%

The advantages of having a compact subset of variables are multifaceted. Firstly, this reduction substantially mitigates the risk of overfitting, thereby enabling a more accurate distinction between signal and noise. Another advantage lies in the simplicity of such a reduced subset. With fewer but crucial variables, subsequent models become simpler and more interpretable, which is of great use in many applications [44]. Furthermore, by concentrating on the most pertinent variables, cluster distinctions are sharpened, leading to more defined and discernible cluster boundaries after DR, thereby enhancing the precision of data analysis.

3.3. Evaluation of dimensionality reduction

In Fig. 5, we display the results obtained through different methods employed to improve the performance of nonlinear DR. This setup offers a visual evaluation, preceding the more rigorous quantitative performance measures. The structure of the plots mirrors that in Fig. 4. The first row presents the outcomes obtained for various datasets using PCA as a pre-processing step. The class separation is somewhat indistinct, particularly for Trachea and Cremaster, while Laser and Ischemia demonstrate better results in this respect.

The second row in Fig. 5 represents results after variable selection using EN. For Trachea and Cremaster, the outcomes are less impressive than those using PCA. This is due to the significant number of variables retained in the variable selection process. While this might be highly accurate for classification purposes (the primary objective for which EN methods were developed), it does not prove as effective for DR. However, the results for Laser and Ischemia outperform those of PCA. In these instances, the distinction between classes is visually more evident.

The third row in Fig. 5 presents the outcomes following the application of our proposed unsupervised selection method. Visually, the group distinctions outperform the two standard methods previously discussed. The results are notably good for the Trachea case, where the visual separation between clusters mirrors the true classes with impressive accuracy.

The fourth row in Fig. 5 shows the results obtained through a supervised selection of variables using our proposed method. The visual separation between groups is striking, especially for Trachea and Laser (binary cases). Results for Cremaster and Ischemia (ternary cases) are less remarkable but are still superior to any of the other previous methods employed.

We have employed various measures for a quantitative assessment of the aforementioned results, as seen in Figs. 5.E and 5.F. Fig. 5.E provides an evaluation of the selected variable subsets, using a kernel density estimation based on the R^2_{McF} or ARI of each variable within the subset. Generally, the median values for our proposed method significantly exceeded those of PCA or EN, particularly when using supervised information. The supervised variant of our method clearly demonstrated the best results in terms of selected variable subsets across all analyzed datasets.

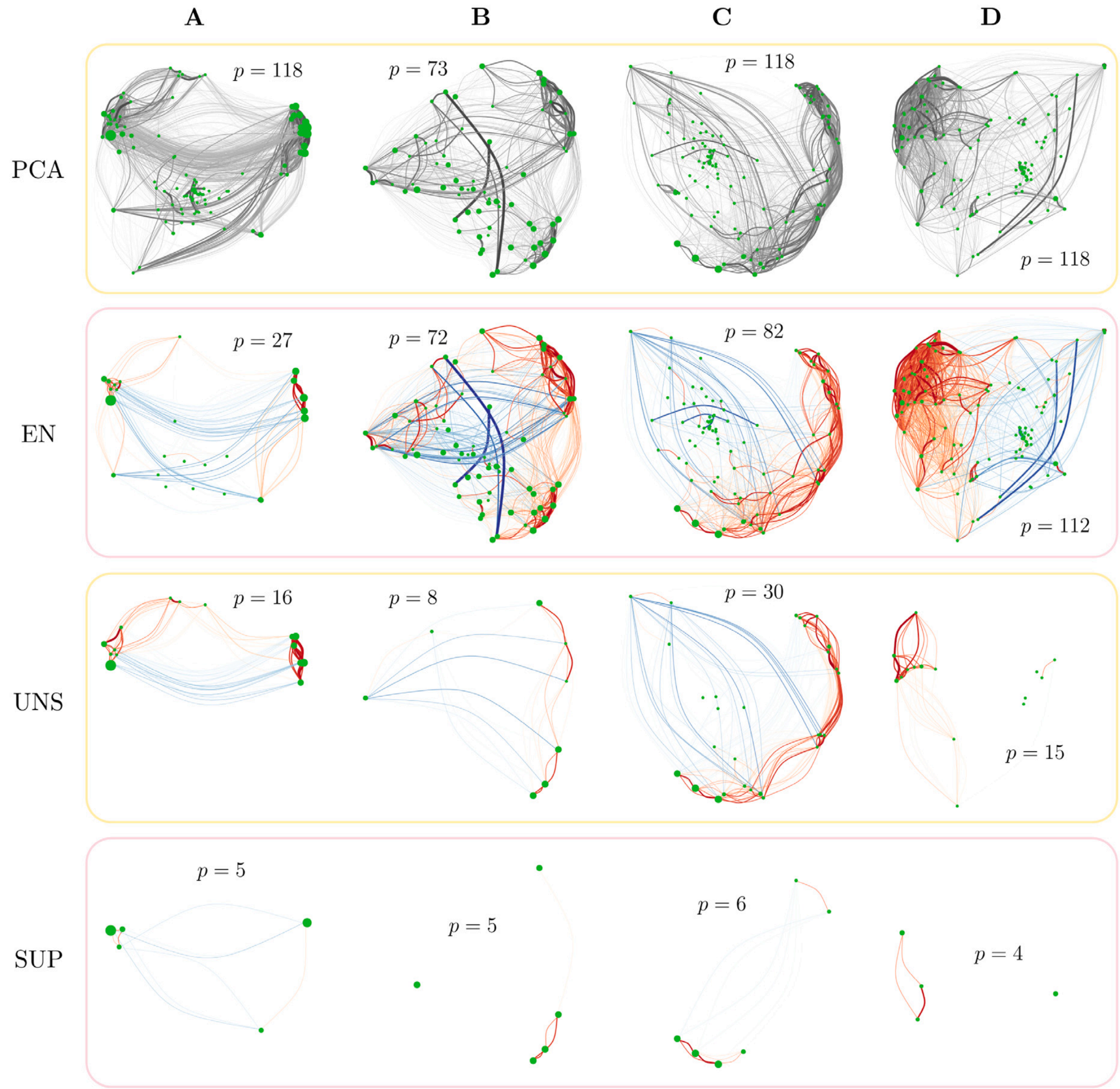

Fig. 4. Selected variables for the different datasets analyzed, organized in columns: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Rows correspond to the various methods employed: (1st) PCA; (2nd) EN; (3rd) Unsupervised selection; and (4th) Supervised selection. Methods marked in light yellow represent unsupervised approaches, while those in light pink signify supervised ones. Each plot displays the subset of variables selected for each case, indicated in green, along with the associated subgraph derived from the original correlation network.

Fig. 5.F depicts global measures, assessing overall performance rather than individual variables. We illustrate the classification capability of each selected subset for all evaluated datasets, in terms of R^2_{McF} (binary cases) or ARI (ternary ones) in blue. In all instances, we achieved measures close to 1, indicative of perfect classification capability for the subset. The silhouette score (c_s), shown in orange, evaluates the clustering capacity of the reduced dimensionality produced. Our proposed method, whether supervised or unsupervised, consistently outperformed standard methods (PCA and EN). The evaluations based on modularity (displayed in red) followed a similar pattern, with our proposed approach yielding higher values than the considered standard methods. See Table 2 for detailed results and Appendix M for an stability analysis.

Table 2
Detailed evaluation of results presented in Fig. 5.F.

		(A) Trachea	(B) Cremaster	(C) Laser	(D) Ischemia
PCA	R^2_{McF}/ARI	1.0000	0.9550	1.0000	0.9753
	c_s	0.3005	0.2067	0.2779	0.2157
	Q	0.1029	0.0732	0.0893	0.1044
EN	R^2_{McF}/ARI	1.0000	0.9382	0.9929	0.9748
	c_s	0.2537	0.1366	0.3736	0.2225
	Q	0.0988	0.0794	0.1186	0.1091
UNS	R^2_{McF}/ARI	1.0000	0.9290	0.9666	0.9465
	c_s	0.5201	0.3120	0.5007	0.4262
	Q	0.1447	0.0891	0.1315	0.1716
SUP	R^2_{McF}/ARI	1.0000	0.9329	0.8529	0.9676
	c_s	0.6521	0.2966	0.7039	0.4297
	Q	0.1830	0.0932	0.2148	0.1475

3.4. Comparison with other approaches and limitations

To conclude this section, we elaborate on the distinctive nature of our proposed method for variable selection. Our approach, rooted in the bootstrapping of correlation networks, marks a significant departure from the prevalent techniques, primarily based on supervised

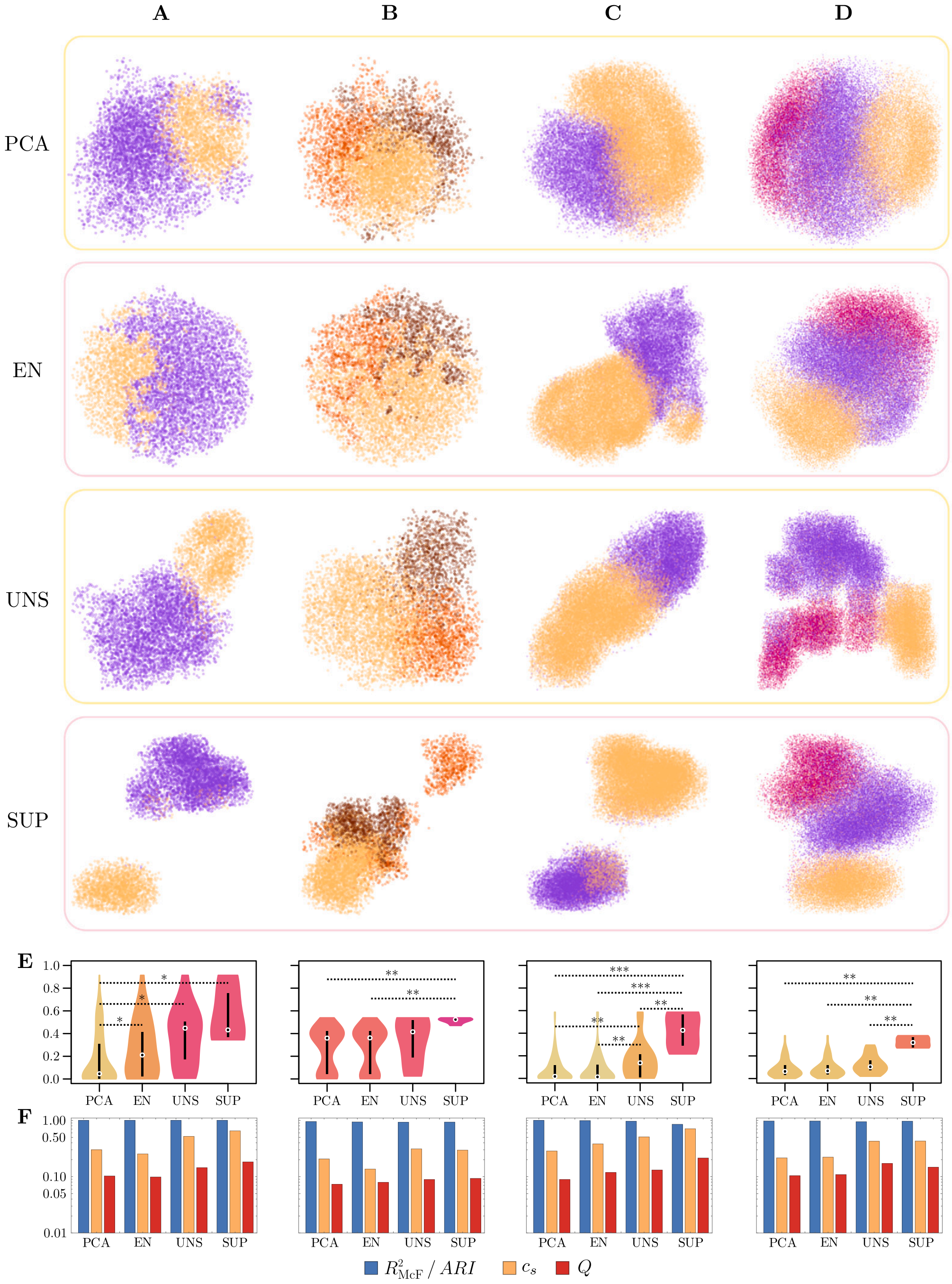

Fig. 5. UMAPs for the different datasets analyzed, arranged in columns: **(A)** Trachea, with neutrophils marked in orange and dendritic cells in violet; **(B)** Cremaster, exhibiting three different behaviors colored in shades of orange; **(C)** Laser, featuring neutrophils in orange and dendritic cells in violet; and **(D)** Ischemia, with neutrophils in orange, dentritic cells in violet, and macrophages in red. The rows correspond to the different methods employed: (1st) PCA; (2nd) EN; (3rd) Unsupervised selection; and (4th) Supervised selection. Unsupervised and supervised methods have been framed in light yellow and light pink, respectively. Row **(E)** shows the evaluation of the selected variable subsets using kernel density estimations of the R^2_{McF} (for binary cases) or ARI (for ternary cases) of the variables, with colors varying from light yellow to dark purple to indicate value differences. Medians have been compared pairwise by means of Mann–Whitney-Wilcoxon tests, employing the Benjamini–Hochberg method to account for multiple testing. Resulting p-values are indicated using the code: (*) $p < 0.05$, (**) $p < 0.01$, and (***) $p < 0.001$. Row **(F)** depicts global evaluations (in a logarithmic scale) based on the classification performance of the full subset (in blue), the silhouette score (in orange), and the network modularity (in red).

metaheuristic algorithms [45–47]. Our methodology distinguishes itself by its ability to unveil intricate relationships within datasets, thereby facilitating the discovery of intrinsic patterns. One of the key aspects is the reliance on a computationally intensive objective function (7), crucial for assessing the performance of DR. Unlike the exhaustive search paradigm employed by metaheuristic algorithms, our use of Bayesian optimization substantially reduces the need for numerous function evaluations. This optimization not only reflects computational efficiency but also brings to light the subtle interplay of variables within the data. It goes beyond the scope of metaheuristic algorithms, which typically overlook the inter-variable relationships, and instead reveals hidden associations and dependencies within correlation networks. This deepened exploration is pivotal for addressing complex biological problems with real-world applications.

Moreover, our research integrates a versatile technique applicable in both supervised and unsupervised scenarios, thus offering a broader scope compared to most existing studies. While recent research has ventured into unsupervised techniques primarily as variable filters [48], our method can identify variable subsets based on relevance and redundancy in an entirely unsupervised manner. Some studies, such as those employing network-based methods similar to ours [49], focus on variable subset identification using regularization techniques. However, these predominantly operate within supervised frameworks, whereas our approach encompasses a more comprehensive spectrum, accommodating both supervised and unsupervised settings.

Despite the promising results, our method is not without limitations. The applicability and robustness across varied datasets require further empirical validation. While our method has shown efficacy in the cases analyzed, more extensive studies across a broader range of datasets are essential to validate its robustness and generalizability. We plan to extend the application of our method to new datasets in future investigations, aiming to thoroughly assess its robustness and versatility under diverse conditions. Additionally, while we have primarily utilized modularity as our evaluation metric, future studies will explore a variety of evaluation functions to obtain a more holistic understanding of the capabilities and constraints of our method. Another important aspect is the computational cost, which is higher than PCA in terms of both runtime and memory usage (refer to Appendix N). This presents a challenge in scenarios where computational efficiency is paramount. Optimizing these computational aspects is essential to widen the practical applications of our method, making it a more viable option in a vast array of research contexts.

4. Conclusions

In this study, we introduce an approach for variable/feature selection in nonlinear Dimensionality Reduction (DR) based on the bootstrapping of correlation networks, which is applicable to both supervised and unsupervised settings. This method utilizes correlation networks constructed using a dissimilarity function that satisfies pseudometric conditions, thereby allowing for diverse function choices depending on the specific data analysis scenario. Our methodology involves a quadratic optimization algorithm that identifies the most relevant variables based on their mean displacements upon random sampling (bootstrapping) of the original datasets. This procedure places particular emphasis on distinguishing between groups or classes. To generate the nonlinear DR visualization, we then employ a Bayesian optimization strategy for hyperparameter tuning in both the quadratic optimization and the DR method.

We have applied our proposed method in both unsupervised and supervised settings to datasets of thousands of leukocytes (myeloid cells) recorded using 4D live imaging at inflammation sites. Comparisons were made with standard Principal Component Analysis (PCA) and Elastic Net (EN), enabling us to identify key features and behavioral states of these cells. Our results demonstrate the superior performance of our proposed method over the former standard approaches. Specifically, it excels in reducing the number of variables (84% of reduction on average for the unsupervised approach and 95% on average for the supervised one, on the datasets we analyzed) and extracting valuable information from DR for more effective cluster distinction in the original data, particularly, the subset of variables capable of producing the grouping into classes. Importantly, our methodology preserves the original variables without transforming them, thereby improving interpretability for subsequent analyses. Moreover, our approach, or similar analytical pipelines, could be employed to generate comparable behavioral landscapes across various tissues and physiological contexts. This enables the effective characterization of cell identities or states based on their morpho-dynamic traits. Our findings highlight the utility of measuring a wide range of variables to accurately define cell behaviors and underscore the discriminative power of specific behavioral traits in differentiating between cell types and states.

The approach we introduce in this work holds the potential to offer significant benefits not only in the current analytical context but also in future studies involving high-dimensional biological data. Our method not only addresses existing challenges but also lays a robust foundation for more accurate interpretations in the evolving realm of biological data analysis. In this field, technologies such as single-cell RNA sequencing have revolutionized our understanding of cellular heterogeneity and function [50]. Our method provides a versatile framework that could be easily extended in the context of single-cell RNA sequencing. In addition, it can be readily adapted to explore other cutting-edge biological data sources, ranging from proteomic profiles to epigenetic markers. The versatility of our approach thus paves the way for its application in diverse biological contexts, promising more refined insights into cellular processes, disease mechanisms, and potential therapeutic targets.

In conclusion, we posit that our proposed variable selection approach is a promising method with potential utility across a broad array of applications, especially within the fields of biology and biomedicine. As its strengths lie in preserving key variables and facilitating interpretability of complex data, it could be particularly beneficial in scenarios where understanding specific feature contributions is crucial. Future research might consider applying this method to an expanded range of datasets, such as genetic sequencing data, medical imaging data, or multivariate time-series data, to fully explore its versatility. Additionally, our approach can be integrated with different DR methods beyond PCA and EN, which could potentially yield even more insightful results and expand its applicability in the context of unsupervised and supervised learning. We hope that ongoing advancements in data science will lead to further refinements and novel applications of this framework.

CRediT authorship contribution statement

David G. Aragones: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Miguel Palomino-Segura:** Capture and analysis of images. **Jon Sicilia:** Computational analysis for dimensionality analysis of the original imaging datasets. **Georgiana Crainiciuc:** Capture and analysis of images. **Iván Ballesteros:** Animal generation, Sample analyses. **Fátima Sánchez-Cabo:** Computational analysis for dimensionality analysis of the original imaging datasets. **Andrés Hidalgo:** Design of experimental work, Image analyses. **Gabriel F. Calvo:** Conceptualization, Methodology, Resources, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors wish to thank Forrest C. Koch for helpful discussions about the evaluation of clustering after DR. We also thank José Antonio Palma and Diana González for additional help regarding graph styling. This research has been supported by grants awarded to G.F.C. by the Spanish Ministerio de Ciencia e Innovación and the European Union NextGenerationEU/PRTR, MCIN/AEI/10.13039/501100011033 (grant numbers TED2021-132296B-C55, PDC2022-133520-I00 and PID2022-142341OB-I00). D.G.A., Spain is supported by a research contract with reference 2023-CDT-11616 (from project with grant number TED2021-132296B-C55). A.H. was supported by RTI2018-095497-B-I00 from Ministerio de Ciencia e Innovación (MCIN), Spain, HR17_00527 from Fundación La Caixa, Spain, Transatlantic Network of Excellence, Spain

(TNE-18CVD04) from the Leducq Foundation, and FET-OPEN (no. 861878) from the European Comission. M.P.-S. is supported by a Federation of European Biochemical Societies, Spain, the EMBO ALTF (no. 1142–2020) long-term fellowship and from MICINN, Spain (RYC2021-033511-I). J.S. is supported by a fellowship (PRE2019-089130) from MICINN, Spain. The CNIC is supported by the MCIN and the Pro-CNIC Foundation, Spain.

Appendix A. Characteristics of the experimental data

In this appendix, we provide additional details about the experimental data used in the analyses. The information has been summarized in Table A.3.

Table A.3
Characteristics of the experimental data analyzed in this study.

	(A) Trachea	(B) Cremaster	(C) Laser	(D) Ischemia
Variables	118	73	118	118
Observations	7,008	7,098	32,323	49,436
Videos	2	212	1	1
Slides (Z)	18	13	10	36
Time points	60	42	154	21
Images processed	2,160	115,752	1,540	756
Microns per pixel (μm)	0.854	0.667	0.870	0.990
Step size (Z) (μm)	3	2	4	4
Total Z (μm)	54	26	44	84
Scan frequency (s)	30.0	8.8	60.0	300.0
Total time (hh:mm)	00:30	00:06	02:33	02:55

Appendix B. Conditions on the dissimilarity to be a pseudo-metric

In this appendix, we prove that the dissimilarity function given in (2) is a pseudo-metric and provide additional conditions that must be fulfilled in a general case.

To show that dissimilarity (2) satisfies the triangle inequality $d_{il} + d_{jl} \geq d_{ij}$, we first write it using (1). Thus it reads as:

$$d_{ij} = \sqrt{\frac{1 - \mathbf{A}_i \cdot \mathbf{A}_j}{2}} \,. \tag{B.1}$$

Next, notice that from (B.1) it follows that:

$$d_{ij} = \sqrt{\frac{2 - 2\mathbf{A}_i \cdot \mathbf{A}_j}{4}} = \frac{1}{2}\sqrt{\|\mathbf{A}_i\|^2 + \|\mathbf{A}_j\|^2 - 2\mathbf{A}_i \cdot \mathbf{A}_j}$$
$$= \frac{1}{2}\sqrt{\|\mathbf{A}_i - \mathbf{A}_j\|^2} = \frac{1}{2}\|\mathbf{A}_i - \mathbf{A}_j\| \,.$$

Likewise, we have $d_{il} = \frac{1}{2}\|\mathbf{A}_i - \mathbf{A}_l\|$ and $d_{jl} = \frac{1}{2}\|\mathbf{A}_j - \mathbf{A}_l\|$. We now recall the triangle inequality. For any two vectors $\mathbf{X}$ and $\mathbf{Y}$, it holds that $\|\mathbf{X}\| + \|\mathbf{Y}\| \geq \|\mathbf{X} + \mathbf{Y}\|$. Defining $\mathbf{X} = \frac{1}{2}\left(\mathbf{A}_i - \mathbf{A}_l\right)$ and $\mathbf{Y} = \frac{1}{2}\left(\mathbf{A}_l - \mathbf{A}_j\right)$ in the triangle inequality we find that:

$$\frac{1}{2}\|\mathbf{A}_i - \mathbf{A}_l\| + \frac{1}{2}\|\mathbf{A}_l - \mathbf{A}_j\| \geq \frac{1}{2}\|\mathbf{A}_i - \mathbf{A}_j\| \,,$$

which is equivalent to $d_{il} + d_{jl} \geq d_{ij}$, since $d_{lj} = d_{jl}$. Therefore, dissimilarity (2) satisfies the triangle inequality and, together with properties 1–3 indicated after (2), guarantees that it is a pseudo-metric.

As pointed out in the main text, the entries ρ_{ij} can be chosen in many ways. For example, they can be of the form:

$$\rho_{ij} = \frac{\sum_{k=1}^{n}\left[R\left(x_i^{(k)}\right) - \overline{R}\left(x_i^{(k)}\right)\right]\left[R\left(x_j^{(k)}\right) - \overline{R}\left(x_j^{(k)}\right)\right]}{\sqrt{\sum_{k=1}^{n}\left[R\left(x_i^{(k)}\right) - \overline{R}\left(x_i^{(k)}\right)\right]^2}\sqrt{\sum_{k=1}^{n}\left[R\left(x_j^{(k)}\right) - \overline{R}\left(x_j^{(k)}\right)\right]^2}} \,, \tag{B.2}$$

where ρ_{ij} is the Spearman's rank correlation coefficient between variables i and j, with $i, j = 1, 2, \ldots, p$ and p representing the total number of variables, $R\left(x_i^{(k)}\right)$ is the rank of value $x_i^{(k)}$, the kth observation of variable i, and $\overline{R}$ denotes the sample mean rank. Recall that n is the total number of observations. This correlation matrix provides a useful approximation of the redundancy between each pair of variables. Spearman's rank correlation coefficients range from -1 to 1: coefficients approaching 1 indicate a non-decreasing monotonic relationship between variables, coefficients near -1 imply a non-increasing monotonic relationship, and coefficients close to 0 suggest a weak monotonic relationship between variables.

In general, any correlation matrix involving three different variables (denoted by i, j and k) can be expressed as:

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{ij} & \rho_{ik} \\ \rho_{ij} & 1 & \rho_{jk} \\ \rho_{ik} & \rho_{jk} & 1 \end{bmatrix}. \tag{B.3}$$

Notice that this is a square matrix, symmetric, with $\mathrm{diag}(\mathbf{P}) = \mathbf{1}$, and positive semidefinite. As a result of this last property, all its leading principal minors must be non-negative. In particular, the following must hold:

$$\det(\mathbf{P}) = -\rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2 + 2\,\rho_{ij}\,\rho_{ik}\,\rho_{jk} + 1 \geq 0 \,. \tag{B.4}$$

This condition must be satisfied by any correlation measure used.

Now, we can define two regions. The first one is given by imposing condition (B.4):

$$\Omega_1 \equiv \left\{(\rho_{ij}, \rho_{ik}, \rho_{jk}) \in [-1, 1]^3 \mid \det(\mathbf{P}) \geq 0\right\} \,.$$

The second region is defined by the triangle inequality, and depends on the dissimilarity function used:

$$\Omega_2 \equiv \left\{(\rho_{ij}, \rho_{ik}, \rho_{jk}) \in [-1, 1]^3 \mid d_{ik}(\rho_{ik}) \leq d_{ij}(\rho_{ij}) + d_{jk}(\rho_{jk})\right\} \,.$$

For the dissimilarity function employed to be a pseudo-metric, the condition $\Omega_1 \subseteq \Omega_2$ must hold or, equivalently, $\Omega_1 = \Omega_1 \cap \Omega_2$. As such, for any correlation measure used, we need to verify:

$$\int_{\Omega_1 \cap \Omega_2} d\rho_{ij}\, d\rho_{ik}\, d\rho_{jk} = \int_{\Omega_1} d\rho_{ij}\, d\rho_{ik}\, d\rho_{jk} = \frac{\pi^2}{2} \,. \tag{B.5}$$

This condition is fulfilled in our case, which proves that the dissimilarity function defined in (2) is a pseudo-metric. In addition, it provides a general criterium that must be satisfied for any dissimilarity function proposed to be a pseudo-metric.

In fact, condition (B.5) can be checked with different typical dissimilarity measures used in the scientific literature [51]:

1. $d_{ij}(\rho_{ij}) = \frac{1-\rho_{ij}}{2}$.
2. $d_{ij}(\rho_{ij}) = 1 - |\rho_{ij}|$.
3. $d_{ij}(\rho_{ij}) = \sqrt{\frac{1-\rho_{ij}}{2}}$.
4. $d_{ij}(\rho_{ij}) = \sqrt{1 - |\rho_{ij}|}$.
5. $d_{ij}(\rho_{ij}) = \sqrt{1 - \rho_{ij}^2}$.
6. $d_{ij}(\rho_{ij}) = \sqrt{1 - \rho_{ij}^4}$.

Dissimilarity functions 3–6 obey the triangle inequality, while functions 1 and 2 do not. This is illustrated in Fig. B.6, where regions Ω_1 and Ω_2 have been represented. Notice that, in cases where the triangle inequality is violated, the region $\Omega_1 \not\subset \Omega_2$.

Satisfying the triangle inequality is important because it has been observed that the performance of different machine learning algorithms improves when the measure used is a pseudo-metric, as this condition can be exploited for faster and better performance [51].

Appendix C. Multidimensional scaling

In this appendix, we provide additional details about how the function given in (3) is minimized. This function can be expanded as:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{p \times 2}} \left[\left(\sum_{i=1}^{p}\sum_{j=1}^{p} b_{ij}(d_{ij})^2\right)^{-1/2}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\left[b_{ij}(d_{ij}) - \boldsymbol{X}_i \cdot \boldsymbol{X}_j\right]^2\right)^{1/2}\right] ,$$

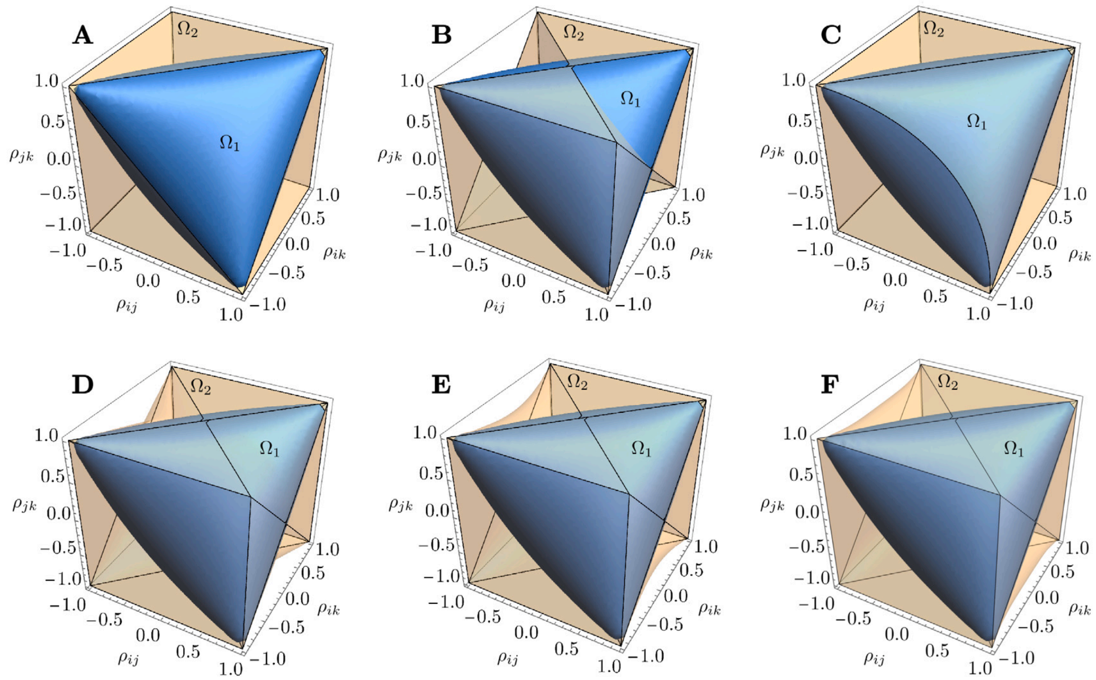

Fig. B.6. Intersection of regions Ω_1 and Ω_2 for six different dissimilarity measures: **(A)** $d_{ij}(\rho_{ij}) = (1-\rho_{ij})/2$, **(B)** $d_{ij}(\rho_{ij}) = 1-|\rho_{ij}|$, **(C)** $d_{ij}(\rho_{ij}) = \sqrt{(1-\rho_{ij})/2}$, **(D)** $d_{ij}(\rho_{ij}) = \sqrt{1-|\rho_{ij}|}$, **(E)** $d_{ij}(\rho_{ij}) = \sqrt{1-\rho_{ij}^2}$, and **(F)** $d_{ij}(\rho_{ij}) = \sqrt{1-\rho_{ij}^4}$. It can be seen that dissimilarity measures **(A)** and **(B)** do not obey condition (B.5). Notice that **(C)** is precisely the one used in this work.

where the dependence between matrix $\boldsymbol{B}$ and dissimilarity matrix $\boldsymbol{D}$ has been made explicit. In fact, matrix $\boldsymbol{B}$ has been obtained after centering $\boldsymbol{D}$ in the following form:

$$\boldsymbol{B} = -\frac{1}{2}\left(\boldsymbol{I} - p^{-1}\,\mathbf{1}\right)(\boldsymbol{D}\odot\boldsymbol{D})\left(\boldsymbol{I} - p^{-1}\,\mathbf{1}\right),$$

where $\boldsymbol{I}$ represents the identity matrix, $\mathbf{1}$ represents the all-ones matrix, and $\odot$ represents the Hadamard-Schur product.

The solution can then be found in terms of the eigendecomposition of $\boldsymbol{B}$:

$$\boldsymbol{X} = \boldsymbol{V}_{\boldsymbol{B}} \cdot \boldsymbol{\Lambda}_{\boldsymbol{B}}^{1/2},$$

with $\boldsymbol{V}_{\boldsymbol{B}}$ being a matrix containing the eigenvectors of $\boldsymbol{B}$ and $\boldsymbol{\Lambda}_{\boldsymbol{B}}$ a diagonal matrix containing the eigenvalues of $\boldsymbol{B}$. The numerical solution was found by means of the Arnoldi method.

Appendix D. Necessary conditions in the quadratic optimization problem

The operator splitting method used for solving the quadratic optimization problem is based on first-order necessary conditions [35], known as Karush–Kuhn–Tucker conditions. The Lagrangian function, remembering (5), can be formulated as:

$$L(\boldsymbol{\beta},\boldsymbol{\psi}_1,\psi_2) = \frac{1-\alpha}{2}\,\boldsymbol{\beta}\cdot(\boldsymbol{P}\odot\boldsymbol{P})\cdot\boldsymbol{\beta} - (\alpha\,\boldsymbol{c} + \boldsymbol{\psi}_1 - \psi_2\,\mathbf{1})\cdot\boldsymbol{\beta} - \psi_2,$$

where $\boldsymbol{\psi}_1$ and ψ_2 are Karush-Kuhn–Tucker multipliers. Then, the condition of stationarity can be established as follows:

$$\frac{\partial L}{\partial\boldsymbol{\beta}} = (1-\alpha)(\boldsymbol{P}\odot\boldsymbol{P})\cdot\boldsymbol{\beta} - \alpha\,\boldsymbol{c} - \boldsymbol{\psi}_1 + \psi_2\,\mathbf{1} = \mathbf{0}.$$

In addition, for primal feasibility the following must hold:

$$\|\boldsymbol{\beta}\|_1 = 1$$
$$\boldsymbol{\beta} \geq 0,$$

while for dual feasibility:

$$\boldsymbol{\psi}_1 \geq \mathbf{0},$$

and finally for complementary slackness:

$$\boldsymbol{\psi}_1 \odot \boldsymbol{\beta} = \mathbf{0}.$$

If the empirical correlation matrix $\boldsymbol{P}$ is positive definite, then these conditions are also sufficient for optimality.

Appendix E. Nonlinear dimensionality reduction

Since our main goal is to improve the performance of nonlinear DR techniques, we have employed UMAP because of its computational efficiency and strong theoretical framework. The central idea of this method is to generate a nonlinear manifold in low dimensions (in our case and more commonly in 2D) from large datasets. This is achieved by minimizing the cross-entropy function between two weighted networks, one in the original high-dimensional space and the other in the reduced new space [12]:

$$H(\boldsymbol{X}) = \sum_{i\neq j}\left[v_{ij}\ln\left(\frac{v_{ij}}{u_{ij}(\boldsymbol{X})}\right) + (1-v_{ij})\ln\left(\frac{1-v_{ij}}{1-u_{ij}(\boldsymbol{X})}\right)\right], \tag{E.1}$$

where v_{ij} is the weight corresponding to the edge between observations i and j in the original space, and u_{ij} is the weight of the edge between the same observations in the reduced space.

The first step is constructing a weighted graph in the original high-dimensional space. This graph has observations as nodes and edges based on N-nearest neighbors, with weights given by [12]:

$$v_{ij}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \bar{v}_{ij}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \bar{v}_{ij}(\boldsymbol{x}_j,\boldsymbol{x}_i) - \bar{v}_{ij}(\boldsymbol{x}_i,\boldsymbol{x}_j)\,\bar{v}_{ij}(\boldsymbol{x}_j,\boldsymbol{x}_i),$$

where $\bar{v}_{ij}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is defined as:

$$\bar{v}_{ij}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left(-\sigma_i^{-1}\max\left\{0, \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 - \tau_i\right\}\right),$$

with τ_i being the minimal distance from observation i to a neighbor, and σ_i a parameter utilized to normalize distances between neighboring observations, which can be calculated for each observation by solving the following equation:

$$\log_2 N = \sum_{j=1}^{N}\exp\left(-\sigma_i^{-1}\max\left\{0, \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 - \tau_i\right\}\right).$$

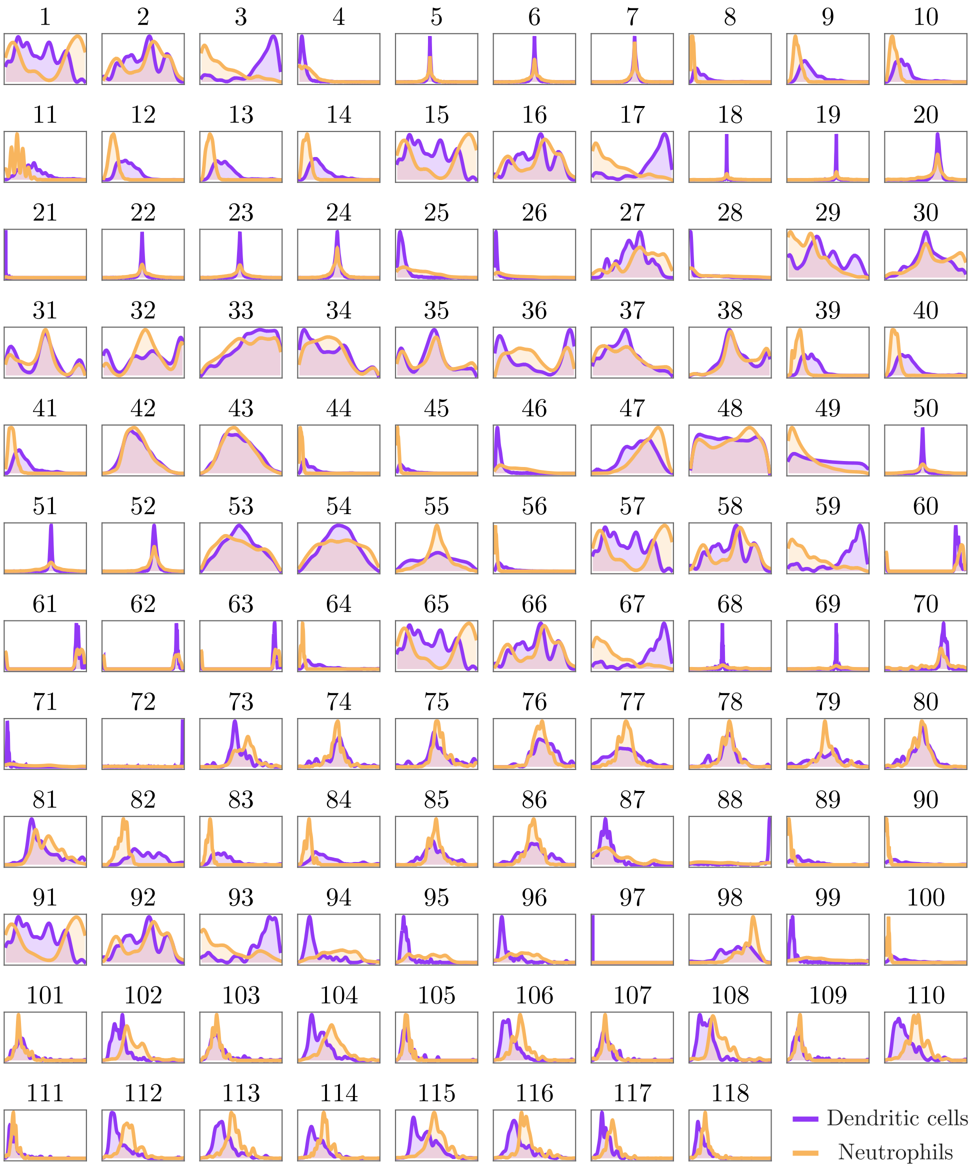

Fig. G.7. Kernel density estimations by variables and groups for the Trachea case.

The second step is the construction of a weighted graph in the reduced space. The nodes of this graph are again the observations, while the weights are given by [12]:

$$u_{ij}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left(1 + a\,\|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^{2b}\right)^{-1} ,$$

where a and b are constants. At the end, the coordinates are determined by means of numerical optimization, minimizing the cross-entropy given by (E.1). For this, we have employed the adaptive moment estimation method, a variant of the stochastic gradient descent that is less prone to getting stuck in local minima and can converge faster. The initial solution for the optimization process was set using the two first principal components to provide both faster convergence and greater stability in the optimization [52].

Appendix F. Bayesian optimization implementation

In this section, we provide more details on the Bayesian optimization implementation, which relies on the construction of a Gaussian process for the modularity, expressed as [53]:

$$Q([N,\alpha];\ \{N_r,\alpha_r,Q_r\}\,|\,\theta) = \mu([N,\alpha];\ \{N_r,\alpha_r,Q_r\}\,|\,\theta)$$
$$+\ GP([N,\alpha];\ \{N_r,\alpha_r,Q_r\}\,|\,\theta)\,,$$

where GP denotes the Gaussian process that is updated at each step. This stochastic process is defined mathematically by its mean function and its covariance function. The mean function was assumed to be zero (simple Gaussian process), while the covariance function was defined using the squared exponential kernel:

$$\mathrm{cov}(d) = \eta_1^2 \exp\left(-\frac{d^2}{\eta_2^2}\right),$$

where η_1 and η_2 are parameters to be fitted, and d represents the distance between values. The estimation was performed using simulated annealing for maximum likelihood. Finally, the Broyden–Fletcher–Goldfarb–Shanno numerical method has been employed for the maximization at each step of the acquisition function given by (9).

Appendix G. Kernel density estimations for each variable

We analyzed each variable by employing kernel density estimation. The goal was to have an initial assessment of how each variable could distinguish between classes. For the estimation, we used standard

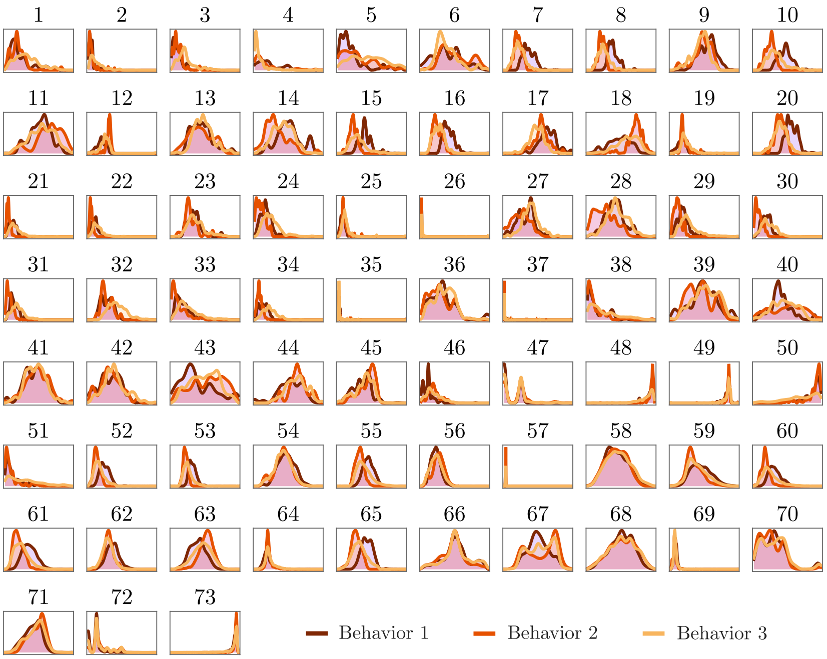

Fig. G.8. Kernel density estimations by variables and groups for the Cremaster case.

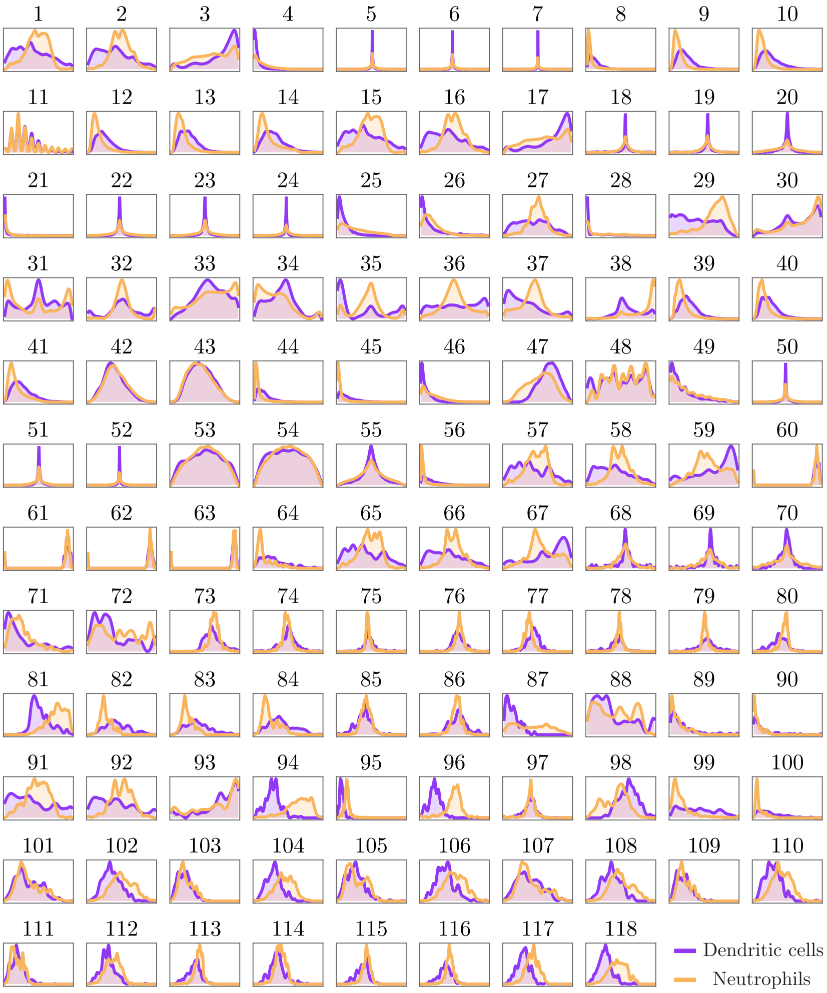

Fig. G.9. Kernel density estimations by variables and groups for the Laser case.

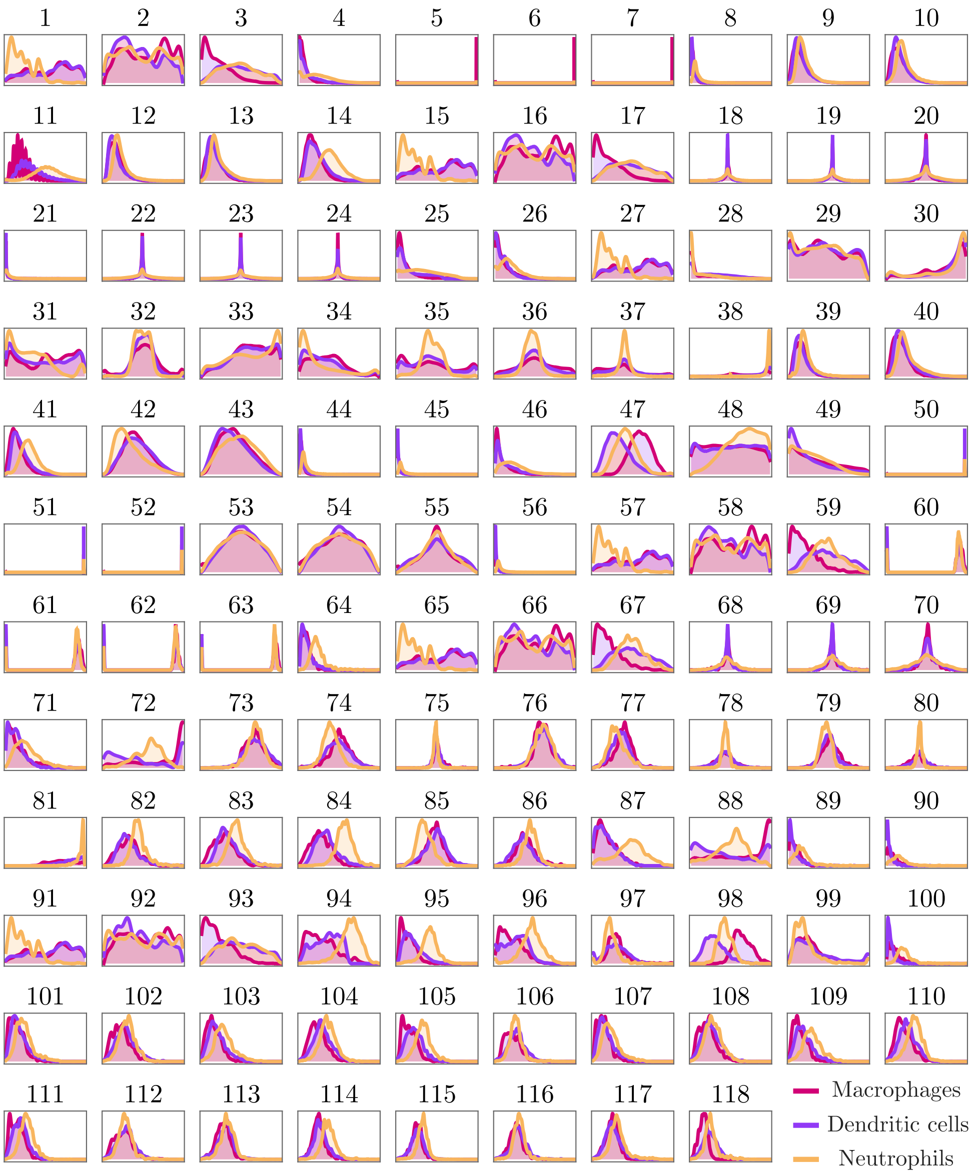

Fig. G.10. Kernel density estimations by variables and groups for the Ischemia case.

Gaussian kernels. The bandwidths were calculated for each case by means of the Silverman method. Results are collected in Figs. G.7, Fig. G.8, Fig. G.9 and Fig. G.10. All variables were rescaled to the range [0,1] for the different analyses performed.

Appendix H. Relevance of each variable

The relevance of variables has been assessed in a supervised way by means of classification models. For binary cases, we have calculated the McFadden likelihood ratio index for each variable (R^2_{McF}), while for ternary ones, we have calculated the Adjusted Rand Index (ARI). Results are shown in Fig. H.11. Notice that, in general, variables measured in Trachea and Cremaster experiments had a higher prediction power than those in Laser and Ischemia datasets. In addition, in Fig. H.12, we show the raw results obtained after bootstrapping of the nonlinear correlation networks obtained for the four datasets analyzed.

Appendix I. Comparison between linear and nonlinear correlation

In this work, we have employed Spearman's nonlinear correlation coefficients instead of Pearson's linear ones. Despite the fact that this increases the computational cost, it shows a better agreement with the relevance of variables in some cases, as measured by R^2_{McF} and ARI (see Fig. I.13). In the figure, each point represents a variable, whose R^2_{McF} or ARI is in the vertical axis. The linear (blue) or nonlinear (red) correlation between the variable and the response is in the horizontal axis. The curves used for the fitting follow the nonlinear equation:

$$y = a\,\rho^{b}\,, \tag{I.1}$$

where y is the R^2_{McF} or the ARI, ρ is the Pearson's or Spearman's correlation coefficient and $\{a, b\}$ are fitting parameters. These parameters have been determined by means of the Levenberg–Marquardt damped least squares method. Observe that the use of nonlinear correlations improves the results over linear correlations in the Trachea

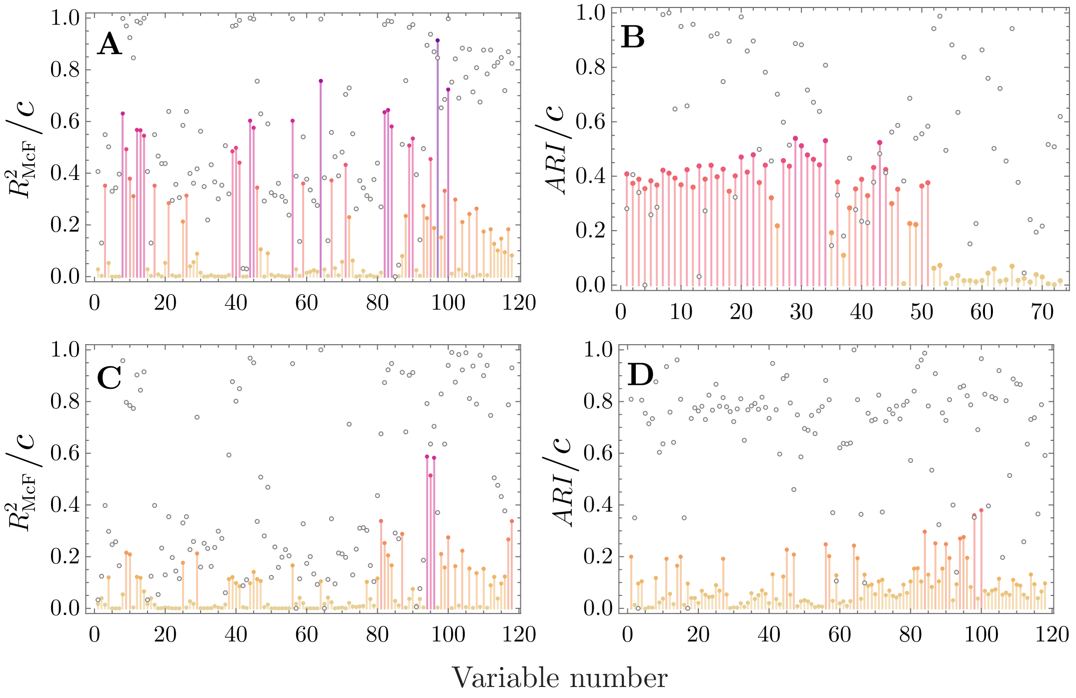

Fig. H.11. Relevance for each variable in the 4 datasets analyzed: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Notice that both R^2_{McF} and ARI are normalized measures in the interval $[0,1]$, with higher values meaning a higher classification capability. Values of the estimated relevance of each variable by unsupervised means (6) are shown in gray open circles for comparison.

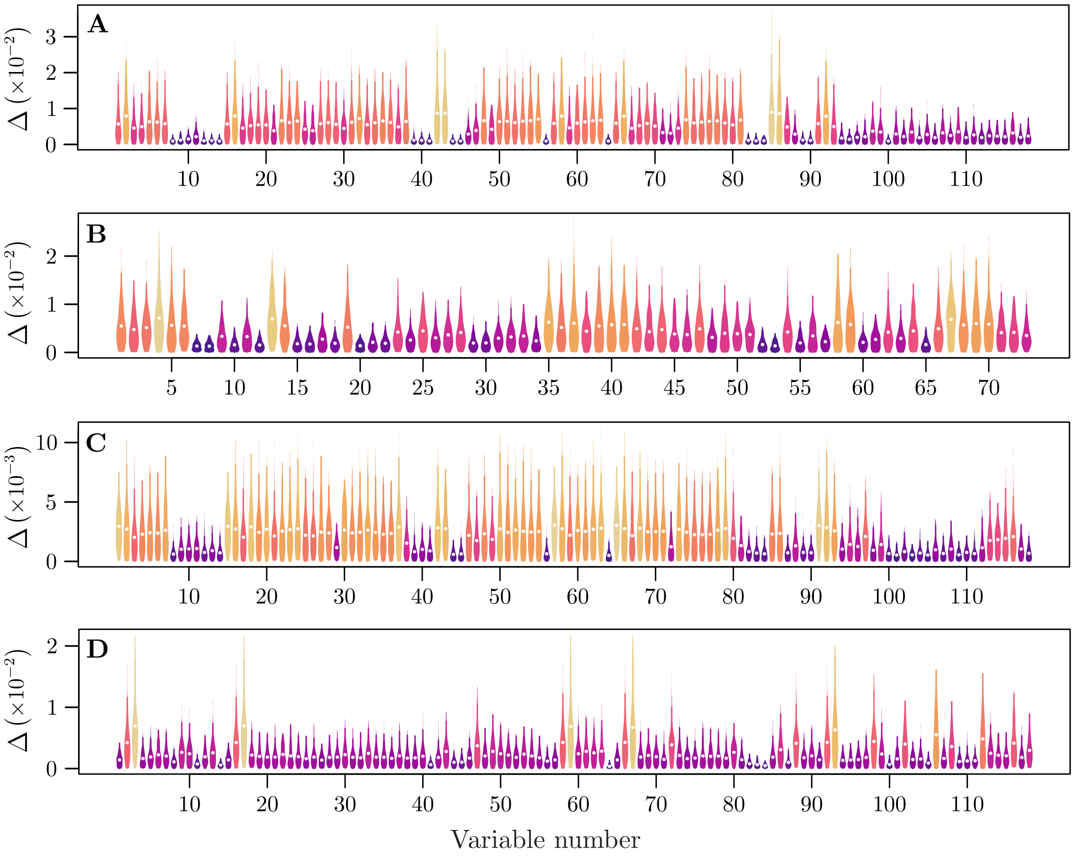

Fig. H.12. Displacements obtained from bootstrapping of nonlinear correlation networks, for which we employed 100 iterations, in the 4 datasets analyzed: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Kernel density estimations have been performed using standard Gaussian kernels. The color scale highlights the mean value (from dark purple for the lower displacements to light yellow for the higher), which is shown with white circles.

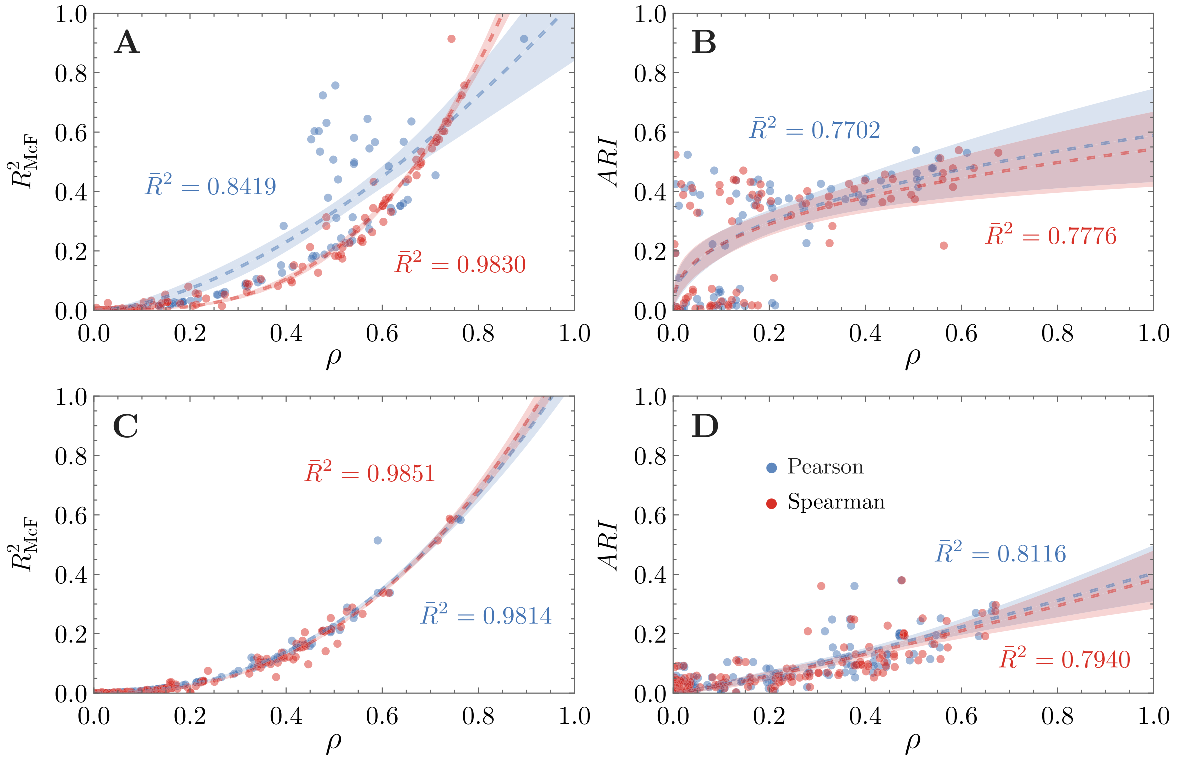

Fig. I.13. Comparison between Pearson's linear correlation (blue) and Spearman's nonlinear correlation (red) as a proxy for relevance of variables in: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Dashed lines represent the nonlinear model fitted to the data, while mean confidence bands have been highlighted in shades at 95% confidence level. $\bar{R}^2$ denotes the adjusted coefficient of determination.

case (Fig. I.13.A). For the rest of the datasets, there were no significant differences between both. Still, we preferred nonlinear correlations as it provided improvements over linear ones in the Trachea experiment, with a similar performance on the rest.

Appendix J. Nonlinear correlation matrices

In this appendix, we represent the Spearman's nonlinear correlation matrices for the four datasets analyzed (see Fig. J.14). Dependencies between groups of variables have been highlighted by ordering them following the first principal component of the dataset.

Appendix K. Elastic Net results

In this appendix, we show additional results from EN obtained for variable selection (see Fig. K.15). First, we set the value of λ_1, which is the parameter that weights the regularized model between Lasso ($\lambda_1 = 1$) and Ridge ($\lambda_1 = 0$). This was done employing 10-fold cross-validation. The best performance was found for Lasso regularization. Still, we set $\lambda_1 = 0.99$ with the aim of improving numerical stability, as the performance is very similar to Lasso, but removes any degeneracies caused by very high correlations. Then, the value of λ_2 was set by calculating the full regularization path at a grid of values on the logarithmic scale, evaluating their performance by means of their Cohen likelihood ratio (R^2_L):

$$R^2_L = 1 - \frac{\ln\left(\mathcal{L}_s/\mathcal{L}_m\right)}{\ln\left(\mathcal{L}_s/\mathcal{L}_0\right)},$$

where $\mathcal{L}_m$ represents the likelihood of the model, $\mathcal{L}_0$ represents the likelihood of the null model, and $\mathcal{L}_s$ is the likelihood of the saturated model. This is a measure of the improvement in model fit when adding a predictor variable. We stopped either when $R^2_L > 0.999$ or when the relative change was below 10^{-5}. Notice that, in all cases, we obtained values of $R^2_L > 0.93$, that is, we obtained a high degree of classification power in the subsets selected.

Appendix L. Sparsity evaluation

To complement the results presented in the main text, we present a comprehensive assessment of the sparsity achieved by our proposed method compared to PCA and EN across the four datasets analyzed in this study. Fig. L.16 displays a heatmap illustrating the variables selected by each method, along with their corresponding R^2_{McF} or ARI values, indicating their relevance to the analysis. This heatmap provides a visual representation of the effectiveness of our method in reducing the number of variables, particularly in comparison to PCA, which retains all original variables, and EN. A detailed examination reveals that our approach significantly bolsters sparsity by selectively retaining only the most pertinent variables for analysis, thereby achieving a more substantial reduction than both PCA and EN.

Appendix M. Stability evaluation

In this appendix, we present an assessment of the consistency in the results obtained from our analysis of the four distinct datasets. Each dataset was divided into ten separate folds, using a process akin to 10-fold cross-validation. For each fold, while maintaining the same set of variables identified in the main body of the manuscript, we computed the silhouette score (c_s). This evaluation aimed to gauge the stability and consistency of our outcomes across different data segments. The findings, as detailed in Table M.4, show a low level of variance across the various methods used in our study. Moreover, these results closely align with those in the main text, where the complete datasets comprising all observations were analyzed.

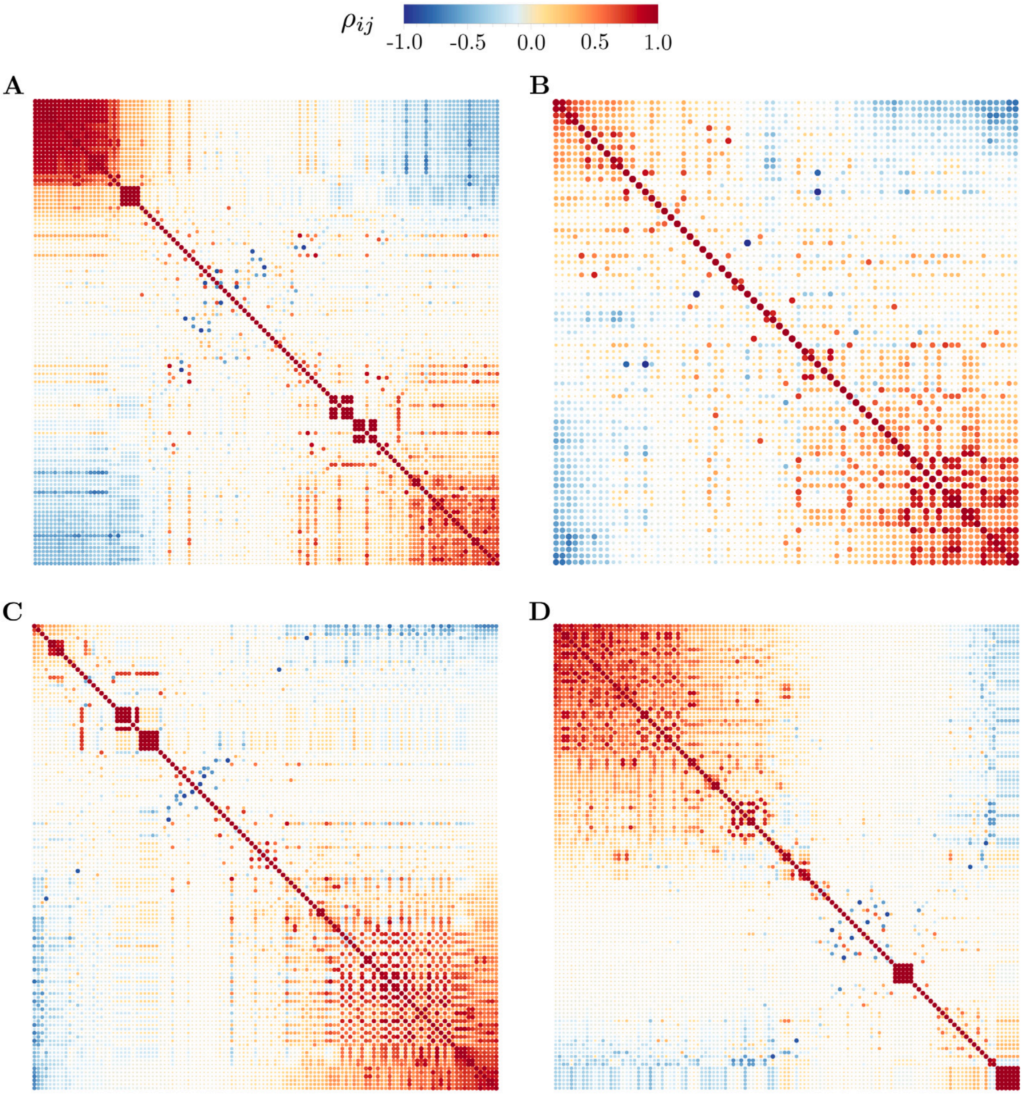

Fig. J.14. Nonlinear correlation matrices for the four datasets analyzed: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Large clusters of highly correlated variables are present, particularly in the Laser and Ischemia datasets.

Appendix N. Efficiency evaluation

We also investigated the computational efficiency of the employed methods, focusing on runtime and memory usage for each approach. Our analysis, detailed in Table N.5, reveals several key insights. Notably, the Laser and Ischemia datasets, which contain a larger number of observations, exhibited higher runtimes and increased memory usage (see Appendix A). Additionally, applying EN to the Cremaster and Ischemia datasets was more resource-intensive, primarily due to the response variable's multinomial distribution, requiring a more complex multiclass classification model. Furthermore, a third trend was observed: our proposed method, in both its unsupervised and supervised formats, demonstrated significantly higher computational costs

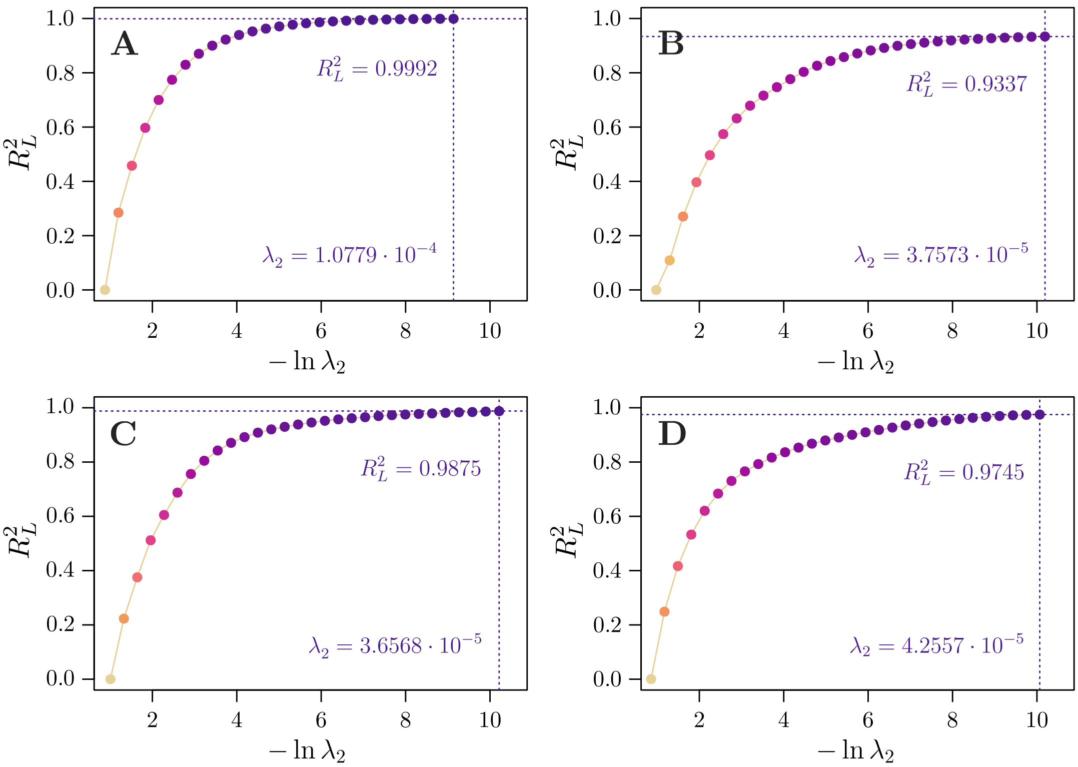

Fig. K.15. EN results in the four datasets analyzed: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. Dotted lines correspond to the λ_2 parameter and the R_L^2 of the subset selected.

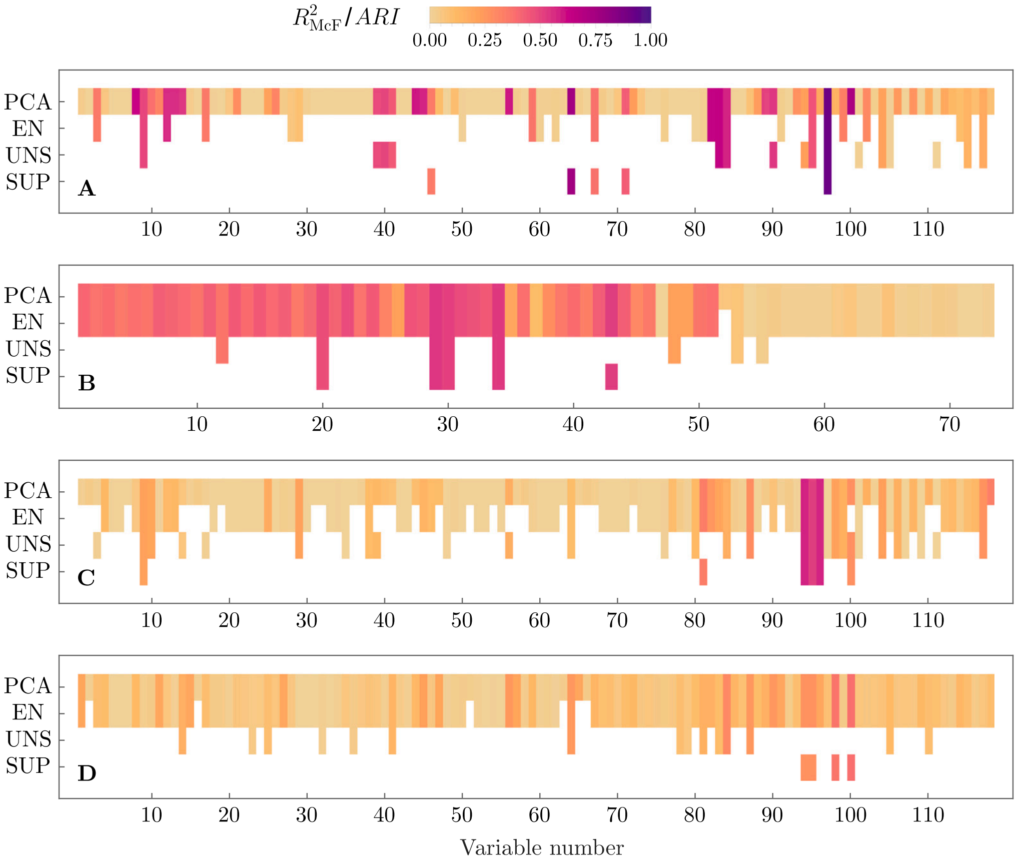

Fig. L.16. Heatmap showcasing the variables selected with the four methods utilized in the four distinct datasets analyzed: **(A)** Trachea; **(B)** Cremaster; **(C)** Laser; and **(D)** Ischemia. The color gradient signifies the R^2_{McF} or ARI values of the selected variables, with unselected variables omitted from the display.

Table M.4
Silhouette scores obtained from 10-fold cross-validation. We show the results obtained using the full datasets together with the mean and the standard deviations derived from cross-validation.

		(A) Trachea	(B) Cremaster	(C) Laser	(D) Ischemia
PCA	Full	0.3005	0.2067	0.2779	0.2157
	Mean	0.3000	0.2050	0.2805	0.2147
	SD	0.0220	0.0048	0.0071	0.0043
EN	Full	0.2537	0.1366	0.3736	0.2225
	Mean	0.2533	0.1354	0.3730	0.2225
	SD	0.0188	0.0120	0.0084	0.0052
UNS	Full	0.5201	0.3120	0.5007	0.4262
	Mean	0.5204	0.3113	0.5031	0.4221
	SD	0.0104	0.0143	0.0080	0.0072
SUP	Full	0.6521	0.2966	0.7039	0.4297
	Mean	0.6521	0.2961	0.7034	0.4267
	SD	0.0107	0.0169	0.0097	0.0069

Table N.5
Runtimes and total memory usage for each of the methods and the datasets analyzed in this work. All calculations were done using a 28×2.5 GHz CPU.

		(A) Trachea	(B) Cremaster	(C) Laser	(D) Ischemia
PCA	Runtime	8.4 s	6.4 s	5.2 min	8.1 min
	Memory	4.1 GB	2.2 GB	11.8 GB	19.2 GB
EN	Runtime	14.7 s	13.9 min	2.5 min	6.5 h
	Memory	1.1 GB	2.4 GB	7.5 GB	23.6 GB
UNS	Runtime	56.8 min	49.1 min	42.0 h	56.3 h
	Memory	124.9 GB	151.4 GB	708.6 GB	1152.9 GB
SUP	Runtime	49.1 min	41.3 min	35.3 h	47.3 h
	Memory	105.1 GB	126.2 GB	590.0 GB	960.3 GB

compared to PCA and EN. This was particularly evident in scenarios where supervised information was not utilized.

References

[1] D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A.J. Carr, C. Burdziak, K.R. Moon, C.L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, Recovering gene interactions from single-cell data using data diffusion, Cell 174 (2018) 716–729.

[2] V.Y. Kiselev, T.S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data, Nature Rev. Genet. 20 (2019) 273–282.

[3] J. Zhou, O.G. Troyanskaya, An analytical framework for interpretable and generalizable single-cell data analysis, Nature Methods 18 (2021) 1317–1321.

[4] X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, D. Huang, Y. Xu, W. Huang, M. Jiang, X. Jiang, J. Mao, Y. Chen, C. Lu, J. Xie, Q. Fang, Y. Wang, R. Yue, T. Li, H. Huang, S.H. Orkin, G.-C. Yuan, M. Chen, G. Guo, Mapping the mouse cell atlas by microwell-seq, Cell 172 (2018) 1091–1107.

[5] K. Davie, J. Janssens, D. Koldere, M. De Waegeneer, U. Pech, L. Kreft, S. Aibar, S. Makhzami, V. Christiaens, C. Bravo Gonzalez-Blas, S. Poovathingal, G. Hulselmans, K.I. Spanier, T. Moerman, B. Vanspauwen, S. Geurs, T. Voet, J. Lammertyn, B. Thienpont, S. Liu, N. Konstantinides, M. Fiers, P. Verstreken, S. Aerts, A single-cell transcriptome atlas of the aging drosophila brain, Cell 174 (2018) 982–998.

[6] D.A. Cusanovich, J.P. Reddington, D.A. Garfield, R.M. Daza, D. Aghamirzaie, R. Marco-Ferreres, H.A. Pliner, L. Christiansen, X. Qiu, F.J. Steemers, C. Trapnell, J. Shendure, E.E.M. Furlong, The cis-regulatory dynamics of embryonic development at single-cell resolution, Nature 555 (2018) 538–542.

[7] J. Cao, D.R. O'Day, H.A. Pliner, P.D. Kingsley, M. Deng, R.M. Daza, M.A. Zager, K.A. Aldinger, R. Blecher-Gonen, F. Zhang, M. Spielmann, J. Palis, D. Doherty, F.J. Steemers, I.A. Glass, C. Trapnell, J. Shendure, A human cell atlas of fetal gene expression, Science 370 (2020) eaba7721.

[8] M. Greenacre, P.J. Groenen, T. Hastie, A.I. D'Enza, A. Markos, E. Tuzhilina, Principal component analysis, Nat. Rev. Methods Primers 2 (2022) 100.

[9] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nature Biotechnol. 37 (2019) 38–44.

[10] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.

[11] L. van der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245.

[12] L. McInnes, J. Healy, N. Saul, L. Grosberger, UMAP: Uniform manifold approximation and projection, J. Open Source Software 3 (2018) 861.

[13] A.E. Teschendorff, Avoiding common pitfalls in machine learning omic data science, Nature Mater. 18 (2019) 422–427.

[14] E.H. Houssein, D. Oliva, E. Çelik, M.M. Emam, R.M. Ghoniem, Boosted sooty tern optimization algorithm for global optimization and feature selection, Expert Syst. Appl. 213 (2023) 119015.

[15] M. Zanin, D. Papo, P. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, S. Boccaletti, Combining complex networks and data mining: Why and how, Phys. Rep. 635 (2016) 1–44.

[16] D.M. Camacho, K.M. Collins, R.K. Powers, J.C. Costello, J.J. Collins, Next-Generation machine learning for biological networks, Cell 173 (2018) 1581–1592.

[17] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375.

[18] F. Karimi, M.B. Dowlatshahi, A. Hashemi, SemiACO: A semi-supervised feature selection based on ant colony optimization, Expert Syst. Appl. 214 (2023) 119130.

[19] S. Solorio-Fernandez, J.A. Carrasco-Ochoa, J.F. Martinez-Trinidad, A review of unsupervised feature selection methods, Artif. Intell. Rev. 53 (2020) 907–948.

[20] G. Bidkhori, R. Benfeitas, M. Klevstig, C. Zhang, J. Nielsen, M. Uhlen, J. Boren, A. Mardinoglu, Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes, Proc. Natl. Acad. Sci. USA 115 (2018) E11874–E11883.

[21] M. Zanin, J.M. Tunas, E. Menasalvas, Understanding diseases as increased heterogeneity: A complex network computational framework, J. R. Soc. Interface 15 (2018) 20180405.

[22] C. Liu, Y. Ma, J. Zhao, R. Nussinov, Y.-C. Zhang, F. Cheng, Z.-K. Zhang, Computational network biology: Data, models, and applications, Phys. Rep. 846 (2020) 1–66.

[23] J. Ding, A. Condon, S.P. Shah, Interpretable dimensionality reduction of single cell transcriptome data with deep generative models, Nature Commun. 9 (2018) 2002.

[24] J.M. Zhang, J. Fan, H.C. Fan, D. Rosenfeld, D.N. Tse, An interpretable framework for clustering single-cell RNA-Seq datasets, BMC Bioinformatics 19 (2018) 93.

[25] T. Tian, J. Wan, Q. Song, Z. Wei, Clustering single-cell RNA-seq data with a model-based deep learning approach, Nat. Mach. Intell. 1 (2019) 191–198.

[26] R. Qi, A. Ma, Q. Ma, Q. Zou, Clustering and classification methods for single-cell RNA-sequencing data, Brief. Bioinform. 21 (2020) 1196–1208.

[27] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. Mauck, S. Zheng, A. Butler, M.J. Lee, A.J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E.P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L.M. Fleming, B. Yeung, A.J. Rogers, J.M. McElrath, C.A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data, Cell 184 (2021) 3573–3587.

[28] F.C. Koch, G.J. Sutton, I. Voineagu, F. Vafaee, Supervised application of internal validation measures to benchmark dimensionality reduction methods in scRNA-seq data, Brief. Bioinform. 22 (2021) bbab304.

[29] J.M. Perkel, Single-cell analysis enters the multiomics age, Nature 595 (2021) 614–616.

[30] R. Argelaguet, A.S.E. Cuomo, O. Stegle, J.C. Marioni, Computational principles and challenges in single-cell data integration, Nature Biotechnol. 39 (2021) 1202–1215.

[31] W. Kopp, A. Akalin, U. Ohler, Simultaneous dimensionality reduction and integration for single-cell ATAC-seq data using deep learning, Nat. Mach. Intell. 4 (2022) 162–168.

[32] G. Crainiciuc, M. Palomino-Segura, M. Molina-Moreno, J. Sicilia, D.G. Aragones, J.L.Y. Li, R. Madurga, J.M. Adrover, A. Aroca-Crevillen, S. Martin-Salamanca, A.S. del Valle, S.D. Castillo, H.C.E. Welch, O. Soehnlein, M. Graupera, F. Sanchez-Cabo, A. Zarbock, T.E. Smithgall, M. Di Pilato, T.R. Mempel, P.-L. Tharaux, S.F. Gonzalez, A. Ayuso-Sacido, L.G. Ng, G.F. Calvo, I. Gonzalez-Diaz, F. Diaz-de Maria, A. Hidalgo, Behavioural immune landscapes of inflammation, Nature 601 (2022) 415–421.

[33] M. Molina-Moreno, I. Gonzalez-Diaz, J. Sicilia, G. Crainiciuc, M. Palomino-Segura, A. Hidalgo, F. Diaz-de Maria, ACME: Automatic feature extraction for cell migration examination through intravital microscopy imaging, Med. Image Anal. 77 (2022) 102358.

[34] M. Palomino-Segura, Automated Cell Migration Examination (ACME) - v1.0, 2021, http://dx.doi.org/10.5281/zenodo.5638537.

[35] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, S. Boyd, OSQP: An operator splitting solver for quadratic programs, Math. Program. Comput. 12 (2020) 637–672.

[36] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M.G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, C. Yau, Bayesian statistics and modelling, Nat. Rev. Methods Primers 1 (2021) 1.

[37] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103 (2006) 8577–8582.

[38] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.

[39] V.A. Traag, L. Waltman, N.J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities, Sci. Rep. 9 (2019) 5233.

[40] R. Satija, J.A. Farrell, D. Gennert, A.F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nature Biotechnol. 33 (2015) 495–502.

[41] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nature Biotechnol. 36 (2018) 411–420.

[42] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data, Cell 177 (2019) 1888–1902.e21.

[43] J.H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010) 1–22.

[44] Y. Liu, V. Ročková, Variable selection via Thompson sampling, J. Amer. Statist. Assoc. 118 (541) (2023) 287–304.

[45] Z. Chen, L. Xinxian, R. Guo, L. Zhang, S. Dhahbi, S. Bourouis, L. Liu, X. Wang, Dispersed differential hunger games search for high dimensional gene data feature selection, Comput. Biol. Med. 163 (2023) 107197.

[46] X. Guo, J. Hu, H. Yu, M. Wang, B. Yang, A new population initialization of metaheuristic algorithms based on hybrid fuzzy rough set for high-dimensional gene data feature selection, Comput. Biol. Med. 166 (2023) 107538.

[47] C. Zhong, G. Li, Z. Meng, H. Li, W. He, A self-adaptive quantum equilibrium optimizer with artificial bee colony for feature selection, Comput. Biol. Med. 153 (2023) 106520.

[48] A. Moslemi, M. Bidar, A. Ahmadian, Subspace learning using structure learning and non-convex regularization: Hybrid technique with mushroom reproduction optimization in gene selection, Comput. Biol. Med. 164 (2023) 107309.

[49] X. Tang, Z. Mo, C. Chang, X. Qian, Group-shrinkage feature selection with a spatial network for mining DNA methylation data, Comput. Biol. Med. 154 (2023) 106573.

[50] X. Nie, D. Qin, X. Zhou, H. Duo, Y. Hao, B. Li, G. Liang, Clustering ensemble in scRNA-sq data analysis: Methods, applications and challenges, Comput. Biol. Med. 159 (2023) 106939.

[51] J. Chen, Y.K. Ng, L. Lin, X. Zhang, S. Li, On triangle inequalities of correlation-based distances for gene expression profiles, BMC Bioinformatics 24 (2023) 40.

[52] Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik, Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization, J. Mach. Learn. Res. 22 (2021) 1–73.

[53] B. MacDonald, P. Ranjan, H. Chipman, GPfit: An r package for fitting a Gaussian process model to deterministic simulator outputs, J. Stat. Softw. 64 (2015) 1–23.