ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Integrated sources model: A new space-learning model for heterogeneous multi-view data reduction, visualization, and clustering

Paul Fogel[1]*, Christophe Geissler[1], Franck Augé[2], Galina Boldina[3], and George Luta[4]

[1]Data Services, Mazars, Courbevoie, France

[2]Translational Precision Medicine, Sanofi, Vitry-sur-Seine, France

[3]Precision Medicine and Computational Biology, Sanofi, Vitry-sur-Seine, France

[4]Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, D.C., United States of America

Abstract

In machine learning, multi-view data involve multiple distinct sets of attributes ("views") for a common set of observations; when each view has the same attributes considered in different contexts, the data are said to contain multiple views of homogeneous format, which can be conceptualized as a tensor. In this article, we describe a novel approach for integrating multiple views of heterogeneous format into a common latent space using a workflow that involves non-negative matrix and tensor factorization (NMF/NTF). This approach, which we refer to as the integrated sources model (ISM), consists of two main steps: Embedding and analysis. In the embedding step, the views are transformed into matrices with common non-negative components. In the analysis step, the transformed views are combined into a tensor and decomposed using NTF. We also present a variant of ISM; the integrated latent sources model (ILSM), which offers significant advantages over ISM in terms of computational power and in cases where the views are highly unbalanced with regard to the number of attributes per view. Noteworthy, ISM can be extended to process multi-omic and multi-view datasets even in the presence of missing views. We provide a proof-of-concept analysis using five examples, including the UCI Digits (the University of California Irvine Pen-Based Recognition of Handwritten Digits) dataset, a public cell-type gene signatures dataset, and a multi-omic single-cell dataset. These examples demonstrate that, in most cases, multi-view clustering is better achieved with ISM or its variant ILSM than with other latent space approaches. We also show how the non-negativity and sparsity of the ISM model components enable straightforward interpretations, in contrast to other approaches that involve latent factors of mixed signs. Finally, we present potential applications to single-cell multi-omics and spatial mapping, including spatial imaging, spatial transcriptomics, and computational biology, which are currently under evaluation. ISM relies on state-of-the-art algorithms invoked through a simple workflow implemented in Python.

Keywords: Principal component analysis; Non-negative matrix factorization; Non-negative tensor factorization; Multi-view clustering; Canonical correlation analysis; Common principal components; Multidimensional scaling

***Corresponding author:**
Paul Fogel
(paul.fogel@mazars.fr)

Citation: Fogel P, Geissler C, Augé F, Boldina G, Luta G. Integrated sources model: A new space-learning model for heterogeneous multi-view data reduction, visualization, and clustering. *Artif Intell Health*. 2024;1(3):89-113.
doi: 10.36922/aih.3427

Received: April 16, 2024

Accepted: June 5, 2024

Published Online: July 24, 2024

1. Introduction

In machine learning, multi-view data involve multiple distinct sets of attributes ("views") for a common set of observations. In the special case where each view has the same attributes but is considered in different contexts, the data are a multidimensional array of order three that can be conceptualized as a tensor. For example, an RGB image has three color channels: Red, green, and blue, each being a non-negative two-dimensional (2D) matrix in which the intensity of the respective color is stored for each pixel. Non-negative tensor factorization (NTF) is a powerful latent space representation technique designed to analyze non-negative multidimensional arrays of order three or more. In the RGB image example, NTF captures both color and spatial information using non-negative factors, which can be used for various tasks such as image compression, enhancement, segmentation, classification, and fusion.[1]

Unfortunately, NTF cannot be applied to multi-view data when the views have heterogeneous content with distinct sets of attributes. For example, a text document can be mapped to different views, such as bag-of-words, topic modeling, or sentiment analysis, each with a different set of attributes. Another example is the transformed the University of California Irvine Pen-Based Recognition of Handwritten Digits (UCI Digits) dataset analyzed in this article. In this dataset, the original bitmaps of handwritten digits, extracted from a preprinted form, have been subjected to various transformations (e.g., Fourier, profile correlations, Karhunen-Love coefficients, pixel averages of images from 2 × 3 windows, Zernike moments, and morphological features), resulting in views with very different formats unsuitable for the direct application of NTF. Numerous algorithms have been proposed for handling such heterogenous multi-view data, some of which have become popular in the machine learning community. For example, the MVLEARN package uses the scikit-learn API to make it easily accessible to Python users,[2] while the Multi-Omics Factor Analysis (MOFA and MOFA+) Bioconductor packages[3,4] are widely used for the analysis of multi-omics datasets. However, since these algorithms assume a heterogeneous data structure, they do not incorporate NTF's explicit factorization of a three-dimensional (3D) array.

Other methods first convert each view into a similarity matrix between the observations, using techniques such as cosine similarity, Euclidean distance, transition probability, or self-representation learning. Since all views refer to the same observations, the similarity matrices have the same shape regardless of the view they originate from, resulting in a tensor of similarity matrices. Multi-view clustering (MVC) is performed on these similarity matrices, sometimes using tensor-based approaches.[5,6] However, these clustering approaches cannot be applied to other tasks, such as dimensionality reduction. This is because the representations of such similarity matrices do not project the data from multiple views into a common latent space with a small number of common attributes, such as underlying factors or concepts.

Another strategy, which allows the use of tensor decomposition techniques, starts by selecting representative points from the data, known as anchor points. These anchor points act as intermediaries to derive transition probabilities from samples to clusters. Within each view, an anchor graph estimates the probability transition matrix from the observations to the anchor points, typically by imposing a sum-to-one constraint on non-negative similarity indices over all anchor points for each point. Within each view, the probability transition matrices from anchor points to clusters and from observations to clusters need to be estimated, together with the clustering labels of the observations. For this purpose, NTF is applied with an orthogonality constraint on the cluster indicator matrices. A shadow p-norm constraint ensures that the cluster labels are consistent across views.[7-9] This approach is primarily designed for MVC, as it requires a special algorithm to select the anchor points that are best distributed across the clusters. It should be noted that many MVC approaches do not involve tensor decomposition techniques. For example, fuzzy-model-based robust clustering on multivariate *t*-mixture distributions (F-MB-T)[10] uses a *t*-mixture model in the expectation-maximization algorithm, resulting in more robust clustering. Unsupervised multi-view K-means or fuzzy C-means[11,12] consider a K-means-like membership architecture across different views. To eliminate the need for a predefined number of clusters, these methods add penalty terms to construct an unsupervised regularization structure. Starting with each data point forming its own cluster, an agglomerative process allows such approaches to be initialization-free.

This article introduces the integrated sources model (ISM), which allows NTF to analyze non-negative heterogeneous views, albeit indirectly, by means of a preliminary embedding of the data in a latent space common to all views. To this end, each view is subjected to non-negative matrix factorization (NMF), using a simple process that ensures consistency between the NMF components across all views. This consistency ensures that the embedded views share the same (synthetic) attributes, forming a non-negative 3D array that can be analyzed by NTF. Our goal in pursuing this strategy is to directly benefit from the proven performance and convergence properties of the NMF and NTF algorithms,

whose availability in powerful MATLAB, Python, or R packages ensures scalability, as will be shown in the results section (Section 3), and accessibility for the vast majority of the machine learning community. In addition to the NTF components, a view-mapping matrix is estimated to obtain an interpretable link between the dimensions of the latent space and the original attributes from each view. It is worth noting that there are some commonalities between ISM and the anchor-based approaches mentioned above, which are discussed further in the discussion section (Section 4).

The ISM belongs to the class of multi-view latent space representation methods,[13-23] which aim to capture underlying factors or concepts that characterize the data in the latent space while filtering out noise and redundancy. For MVC applications, performing cluster analysis in the latent space generally results in more accurate and consistent cluster partitioning.[24] It is noteworthy that these approaches allow newly collected data (i.e., data that are not part of the data used to train/learn the model) to be embedded in the latent space, thus extending beyond the purpose of MVC. Some of the latent space representation methods generate NMF-based latent factors[21,23] using regularization parameters that ensure sparsity and consistency between model parameters across different views. The originality of ISM lies in its simple workflow involving NMF and NTF steps. As a result, ISM produces latent factors whose interpretation is greatly facilitated by the non-negativity of the attribute loadings that define them, since they cannot cancel each other out. The interpretability of latent factors is of critical importance if they are to be used by an investigator as a follow-up tool, for example, in a clinical trial comprising several surveys with heterogeneous content.

Finally, we show that embedding the views in a 3D array has broader implications in a number of areas, such as parallelization, federated computing, and distributed computing, further illustrating the scalability and versatility of ISM, which extends well beyond the scope of multi-view data analysis.

2. Data and methods

2.1. Data

Five datasets, all with labeled observations, are considered in this article. The labeling will be used for the evaluation of the clustering performance of ISM and other methods. Details of the five datasets are as follows:

(i) UCI Digits dataset: This dataset, available in the UCI machine learning repository[25] (https://archive.ics.uci.edu/dataset/72/multiple+features), contains six heterogeneous views of handwritten digits: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240-pixel averages of the images from 2 × 3 windows, 47 Zernike moments, and six morphological features. Each class of digits (0 – 9) contains 200 labeled examples.

(ii) Signature 915 data: This dataset is available in the GitHub repository (https://github.com/Advestis/adilsm/tree/main/examples/data) in the file "abis_915.csv." It comprises expression data of 915 marker genes in four patients and 16 cell types). There are four views of 915 gene markers (one view per patient) measured across 16 different cell types.[26]

(iii) Reuters dataset: Available in the GitHub repository (https://github.com/mbrbic/Multi-view-LRSSC/tree/master/datasets) in the file "Reuters.mat," Reuters dataset contains features of documents in five different languages over a common set of six categories.[27] All documents are represented in the bag-of-words format. Each of the six classes contains 100 documents, resulting in a dataset of 600 documents. The word counts in each view are 21,526, 24,892, 34,121, 15,487, and 11,539 words, respectively.

(iv) Prokaryotic phyla dataset: Found in the GitHub repository (https://github.com/mbrbic/Multi-view-LRSSC/tree/master/datasets) in the file "prokaryotic.mat," prokaryotic phyla dataset contains 551 prokaryotic species described with heterogeneous multi-view data,[28] including textual data (438 features), proteome composition encoded as relative frequencies of amino acids (three features), and gene repertoire (393 features) encoded as presence/absence indicators of gene families in a genome. Each provided view contains the principal components explaining 90% of the variance.[22] Each species in the dataset is labeled with its phylum, resulting in four unbalanced categories ranging from 35 to 313 species.

(v) TEA-seq multi-omic single-cell dataset: This dataset, available in the figshare repository (https://figshare.com/s/1b13e12f33e83fff7e0e) in the file "tea_preprocessed.h5mu," consists of human peripheral blood mononuclear cells. It includes paired profiling of scRNA-seq (2,500 features), scATAC-seq (15,000 features), and surface proteins (46 features).[29] As the dataset did not come with cell annotations, an annotation was derived from the clustering of cells using MOFA+ with 15 components,[21] resulting in seven major cell types: CD4 effector and memory T cells, B cells, CD4+ naïve T cells, monocytes, CD8+ T cells, Mucosal-associated invariant T (MAIT) cells, and natural killer (NK) cells.

Of note, the UCI Digits and Signature 915 datasets cover both aspects of sparsity (because the Signature 915

dataset contains the expression of marker genes) and redundancy (because the UCI Digits dataset contains redundant information in the nature of the images). For this reason, special emphasis is placed on the analysis of these datasets.

2.2. Methods

2.2.1. Outline of ISM and comparison with other latent space approaches

Before delving into the details of the ISM workflow, we present the main underlying ideas with an illustrative figure (Figure 1A) and compare ISM with other latent space approaches (Figure 1B). The different views are represented by heatmaps on the left side of both panels, with attributes on the vertical axis and observations on the horizontal axis.

(a) ISM

In the central part of Figure 1A, each non-negative view *Xv* is decomposed into the product of two non-negative matrices, H_v and W_v, using NMF. Each W_v matrix corresponds to the transformation of a particular view *v* to a latent space common to all transformed views. ISM ensures that the transformed views, W_v, share the same number and type of latent attributes, as explained in the detailed description. This transforming process, which we call embedding, results in a 3D array, or tensor. The corresponding H_v matrices contain the loadings of the original attributes on each component. We refer to these matrices as the mapping between the original and transformed views.

In the right part of Figure 1A, the 3D array is decomposed into the tensor product of three matrices: W^*, H^*, and Q^* using NTF. W^* contains the meta-scores – the single transformation to the latent space common to all views. H^* and Q^* contain the loadings of the latent attributes and views, respectively, on each NTF component. Each row of Q^* is represented by a diagonal matrix, where the diagonal contains the loadings for a particular view. This allows for each view of the tensor to translate the tensor product into a simple matrix product $W^*(H^*Q_v^*)^T$, as seen in Figure 1A.

(b) Other latent space approaches

In the right part of Figure 1B, each view *v* is decomposed into the product of two matrices, H_v and W, using the latent space method algorithm. As with ISM, W contains the meta-scores – the single transformation in the latent space common to all views.

(c) Comparison between ISM and other latent space approaches

If we multiply each mapping matrix H_v by H^*Q^* in Figure 1A, we obtain a representation similar to that in Figure 1B. This shows that ISM belongs to the family of latent space decomposition methods. However, view loadings are a constitutive part of ISM, whereas in other models, they are derived separately. For example, the MOFA+ method uses variance decomposition by factor.[3]

(d) Important implications of ISM's preliminary embedding

As will be detailed in the workflow description, ISM begins by applying NMF to the concatenated views. Importantly, NMF can be applied to each view X_v separately, leading to view-specific decompositions $X_v = W_v^{nmf} H_v^{nmf\,T}$ before ISM itself is applied to the *m* NMF-transformed views W_v^{nmf}. In this case, the view mapping returned by ISM, H_v^{ism}, refers to the NMF components of each W_v^{nmf}. However, by embedding the W_v^{nmf} in a 3D array, ISM allows H_v^{ism} to be mapped back to the original views through simple chained matrix multiplication such that: $X_v = W^* H_v^T$ with $H_v = H_v^{nmf} H_v^{ism} H^* Q_v^*$. We refer to this alternative approach as integrated latent source model (ILSM). As shown in the results (Section 3) and discussion (Section 4) sections, ILSM offers important advantages in several respects.

2.2.2. Compared methods

In this article, we compare ISM and ILSM with multi-view multidimensional scaling (MVMDS),[2,14] MOFA+,[3,4] group factor analysis (GFA),[18] and Multi-Omics Wasserstein inteGrative anaLysIs (MOWGLI).[21] Below is a brief description of each of these methods.

(a) MVMDS: After computing and double-centering the Euclidean distance matrices for each of the views, MVMDS estimates the common principal components of the matrices in a manner similar to the generalization of principal component analysis (PCA) for multiple covariance matrices

(b) MOFA+ and GFA: Both models are formulated in a probabilistic Bayesian framework, where prior distributions are placed on all unobserved variables of the model, using a standard normal prior for the factors W and sparsity priors for the mapping matrices H_v

(c) MOWGLI: This model is a multi-view generalization of NMF, using optimal transport instead of the Frobenius cost function and regularization parameters that ensure sparsity and consistency between model parameters across different views. A sum-to-one constraint is applied to the common factors W to give them a probabilistic interpretation.

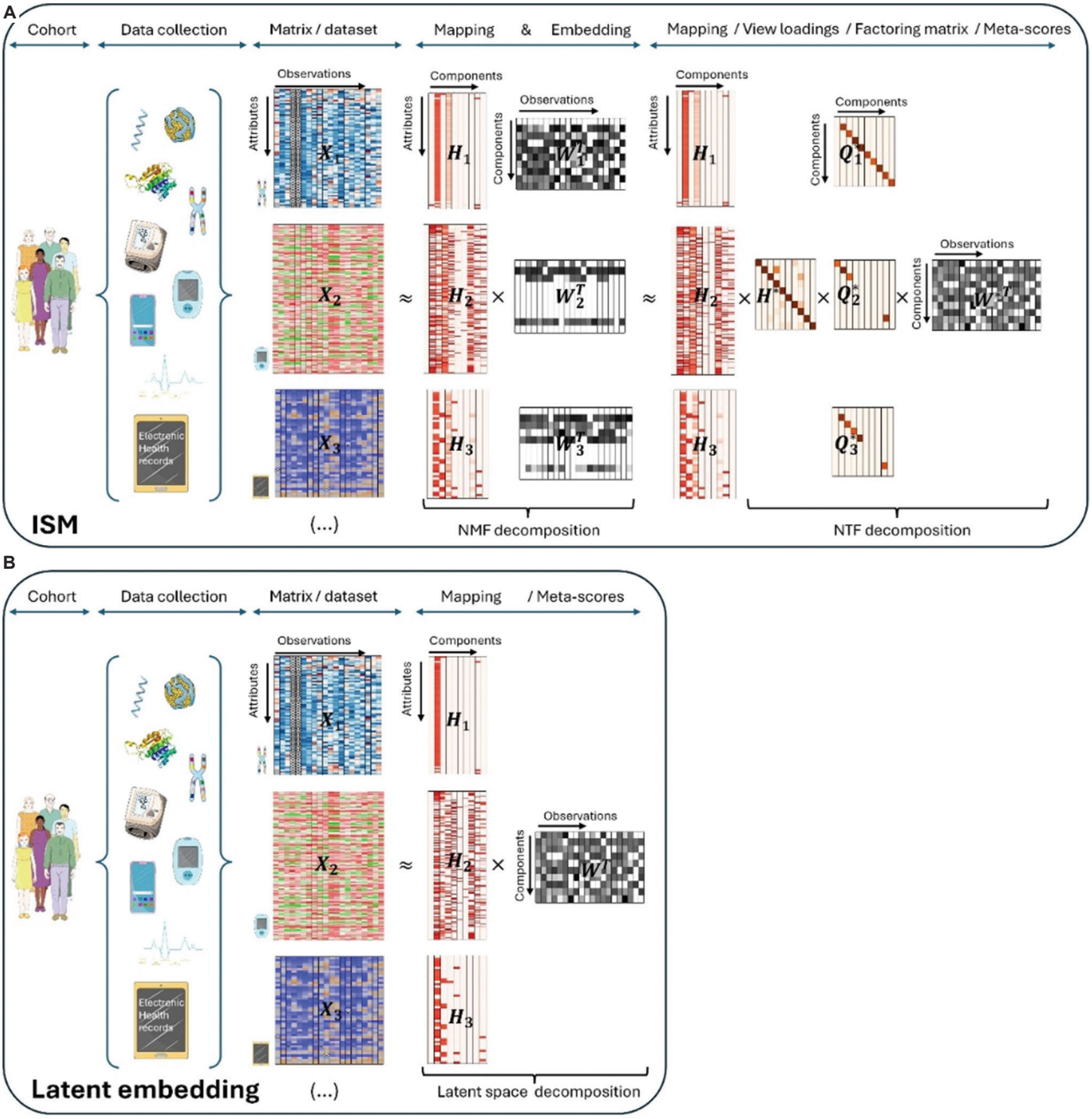

Figure 1. Comparison between integrated sources model (A) and other latent space approaches (B). Note: The image was created using elements provided by Servier Medical Art (https://smart.servier.com/citation-sharing/).
Abbreviations: NMF: Non-negative matrix factorization; NTF: Non-negative tensor factorization.

2.2.3. Detailed workflows

In this section, we present three workflows. The first workflow consists of training the ISM model to generate a latent space representation and view-mapping. The second workflow enables the projection of new observations obtained in multiple views into the latent space. The third workflow contains the detailed analysis steps for each example.

(a) Workflow 1: Latent space representation and view-mapping

The training of the ISM model can be divided into five units, as described in Figure 2. The first four process

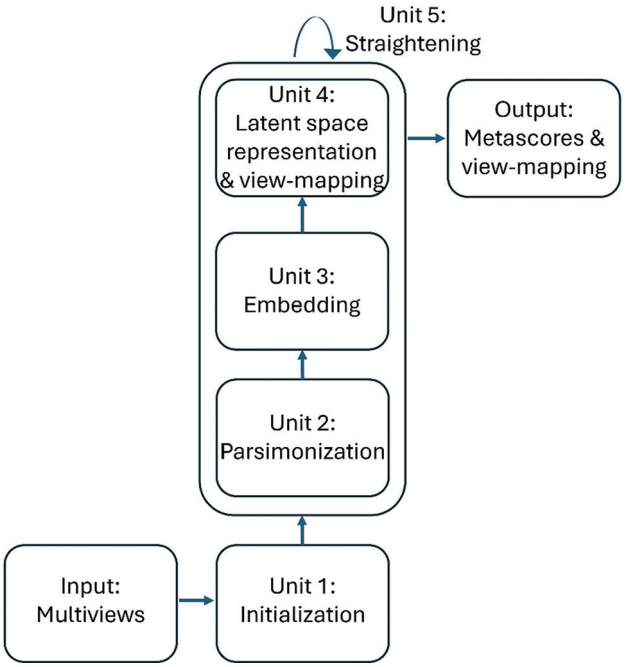

Figure 2. Training of the integrated sources model

units enable the discovery of the latent space within an "embedding" space. Once the latent space is identified, it is assimilated with the embedding space. During the fifth "straightening" unit, the latent space remains fixed, while the sequence of Units 3, 4, and 2 is repeated to further parsimonize the view-mapping until the degree of sparsity remains unchanged. The theoretical foundations of combining NMF and NTF during the embedding and latent space discovery steps are discussed at the end of this section. The sizes of the embedding space and the latent space are discussed in the section describing the third workflow.

(i) Unit 1: Initialization

An NMF is first performed on the matrix X of the m concatenated views X_v, $1 \le v \le m$, resulting in the decomposition: $X = WH^T + E, W \in \mathbb{R}_+^{n \times d_e}, H \in \mathbb{R}_+^{d \times d_e}$, $E \in \mathbb{R}^{n \times d}$ where W represents the transformed data, the columns of H contain the loadings of the $d = \sum_{v \le m} d_v$ attributes across all views on each component, d_e is the embedding size, and n is the total number of observations.

Unit 1. Initialization

Input: m views $\{X_1, , X_m\}$, $X_v \in \mathbb{R}_+^{n \times d_v}$ where n is the number of rows common to all views and d_v is the number of columns in the v^{th} view (it is assumed for each column that its values lie between 0 and 1 after normalization by the maximum row value).

Output: Factoring matrices $W \in \mathbb{R}_+^{n \times d_e}$, $H \in \mathbb{R}_+^{d \times d_e}$ where d_e is the embedding dimension, and is the sum of the number of columns in all views and $d = \sum_{v \le m} d_v$ the matrix of concatenated views X.

1: Concatenate the m views: $X = \left[X_1, \cdots, X_m\right], X \in \mathbb{R}_+^{n \times d}$;

2: Factorize X using NMF with d_e components:

$X = WH^T + E, W \in \mathbb{R}_+^{n \times d_e}, H \in \mathbb{R}_+^{d \times d_e}, E \in \mathbb{R}^{n \times d}$;

(ii) Unit 2: Parsimonization

The initial degree of sparsity of H is crucial to prevent the embedding dimensions from being overly distorted between the different views during the embedding process, as will be seen in the next section. This is achieved by applying a hard threshold to each column of the H matrix. The threshold is based on the reciprocal of the Herfindahl-Hirschman index (HHI),[30] which provides an estimate of the number of non-negligible values in a non-negative vector.

For columns with strongly positively skewed values, the use of the L2 norm for the estimate's denominator can lead to excessively sparse factors, which in turn can lead to an overly large approximation error during embedding. Therefore, the estimate is multiplied by a coefficient whose default value was set at 0.8, after testing with simulated datasets using the simulation framework described in Fogel *et al.*[31]

Unit 2. Parsimonization

Input: Factoring matrix

Output: Parsimonized factoring matrix $H \in \mathbb{R}_+^{d \times d_e}$ (since the initial H is not used outside parsimonization, we use the same symbol for the sake of simplicity).

1: for each component h_k of H do

2: Calculate the reciprocal of the Herfindahl-Hirschman Index to estimate the number of non-negligible entries in h_k:

$$\tau_k = \frac{\left(\sum_{i \le d} h\left[i,k\right]\right)^2}{\sum_{i \le d} h\left[i,k\right]^2} = \frac{h_{k1}^2}{h_{k2}^2};$$

3: Enforce sparsity on hk using hard thresholding:
If rank $(h\,[i,k]) < \tau_k \times \lambda$ then set $h\,[i,k] = 0$ where λ is a sparsity parameter ($0 < \lambda < 1$, the default value $\lambda = 0.8$ was chosen as it led in many trials to better results than the original index τ_k, which may be a too strict filter);

4: end for

(iii) Unit 3: Embedding

The matrices W and H are further updated along each view, yielding matrices $W_v \in \mathbb{R}_+^{n \times d_e}$ of common shape (number of observations n × factorization rank d_e) corresponding to the transformed views.

NMF multiplicative updates are used during view matching to leave the zeros in the primary H matrix unchanged. Further optimizations of the simplicial cones $H_v \in \mathbb{R}_+^{d_v \times d_e}$ for each view v are therefore limited to the non-zero loadings so that they remain tightly connected. This ensures that the transformed views W_v, $v \leq m$, form a tensor. Multiplicative updates usually start with a linear rate of convergence, which becomes sublinear after a few hundred iterations.[32] By default, the number of iterations is set to 200 to ensure a reasonable approximation to each view, as required for the latent space representation described in the next section.

Unit 3. Embedding

Input: m views $\{X_1, \ldots, X_m\}$ and factoring matrices $W \in \mathbb{R}_+^{n \times d_e}$, $H \in \mathbb{R}_+^{d \times d_e}$.

Output: view-specific factoring matrices $W_v \in \mathbb{R}_+^{n \times d_e}$, $H_v \in \mathbb{R}_+^{d_v \times d_e}$ and tensor $\boldsymbol{\mathcal{A}}$.

1: for each view v do

2: Define $H_v \in \mathbb{R}_+^{d_v \times d_e}$ as the part of H corresponding to view v;

3: Factorize X_v into view-specific $W_v \in \mathbb{R}_+^{n \times d_e}$ and using $H_v \in \mathbb{R}_+^{d_v \times d_e}$ NMF multiplicative updating rules and initialization matrices $W \in \mathbb{R}_+^{n \times d_e}, H_v \in \mathbb{R}_+^{d_v \times d_e}$:

$X_v = W_v H_v^T + E_v, W_v \in \mathbb{R}_+^{n \times d_e}, H_v \in \mathbb{R}_+^{d_v \times d_e}, E_v \in \mathbb{R}^{n \times d_v}$;

4: Normalize each component of W_v by its maximum value and update H_v accordingly;

5: Define tensor slice: $\boldsymbol{\mathcal{A}}(:,:,v) = W_v$;

6: end for

(iv) Unit 4: Latent space representation and view-mapping

The resulting tensor $\boldsymbol{\mathcal{A}}$ is analyzed using NTF, which leads to the decomposition: $\boldsymbol{\mathcal{A}} = W^* \otimes H^* \otimes Q^* + \varepsilon$ where $W^* \in \mathbb{R}_+^{n \times d_l}$, $H^* \in \mathbb{R}_+^{d_e \times d_l}$, $Q^* \in \mathbb{R}_+^{m \times d_l}$, $\varepsilon \in \mathbb{R}_+^{n \times d_e \times m}$, and d_l is the dimension of the latent space. The components W^*, H^*, and Q^* enable the reconstruction of the horizontal, lateral, and frontal slices of the embedding tensor. The loadings of the views on each component are contained in the matrix Q^*. The integrated multiple views, or meta-scores, are contained in the matrix W^*. The matrix H^* represents the latent space in the form of a simplicial cone contained in the embedding space. Finally, the view-mapping matrix H is updated by applying steps 3 – 8 of Unit 4. Its sparsity is ensured by further applying Unit 2 (parsimonization).

Unit 4. Latent space representation and view-mapping

Input: view-specific factoring matrices $H_v \in \mathbb{R}_+^{d_v \times d_e}$ and tensor $\boldsymbol{\mathcal{A}}$.

Output: NTF factors $W^* \in \mathbb{R}_+^{n \times d_l}, H^* \in \mathbb{R}_+^{d_e \times d_l}, Q^* \in \mathbb{R}_+^{m \times d_l}$ and view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$.

1: Define view-mapping matrix $H \in \mathbb{R}_+^{d \times d_e}$ as the concatenation of $H_v \in \mathbb{R}_+^{d_v \times d_e}$;

2: Factorize $\boldsymbol{\mathcal{A}}$ using NTF with d_l components: $\boldsymbol{\mathcal{A}} = W^* \otimes H^* \otimes Q^* + \varepsilon$ where $W^* \in \mathbb{R}_+^{n \times d_l}, H^* \in \mathbb{R}_+^{d_e \times d_l}, Q^* \in \mathbb{R}_+^{m \times d_l}, \mathrm{E} \in \mathbb{R}_+^{n \times d_e \times m}$;

3: Update view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$: $H \leftarrow HH^*$;

4: for each view v do

5: Update H_v: $\mathrm{H}_v \leftarrow \mathrm{H}_v \circ \mathrm{Q}^*$ [v, :];

6 end for

7: Update view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$ as the concatenation of updated H_v;

8: Parsimonize view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$ by applying Unit 2;

(v) Unit 5: Straightening

The sparsity of the view-mapping matrix H can be further optimized together with the meta-scores W^* and the view-loadings Q^* by repeating Units 3, 4, and 2 until the number of zero entries in H remains unchanged. To achieve this, the embedding is restricted to the latent space defined by the simplicial cone formed by H^*. In this simplified embedding space, H^* becomes the identity matrix I_{d_l} when the updating process of W^*, H^*, and Q^* starts. In other words, the embedding and latent spaces are assimilated during the straightening process. Optionally, for faster convergence, H^* can be fixed to I_{d_l}, at the cost of a slightly higher approximation error, as observed in simulated experiments, due to only small deviations from I_{d_l}.

Unit 5. Straightening

Input: X, $\boldsymbol{\mathcal{A}}$, H, W^*, H^*, Q^*.

Output: NTF factors W^*, H^*, Q^* and updated view-mapping matrix H.

1: $H^* = I_{d_l}$ Set where d_l is the size of the latent space;

2: do until the number of 0-entries in H remains unchanged

3: Apply Unit 3 to embed X using the embedding size $d_e = d_l$, initialization matrices $W^* \in \mathbb{R}_+^{n \times d_l}$ and view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$ found in the previous iteration;

4: Apply Unit 4 to factorize $\boldsymbol{\mathcal{A}}$ and update the view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$, using embedding size $d_e = d_l$, initialization matrices W^*, H^*, Q^* obtained in the previous iteration and fixed $H^* = I_{d_l}$;

5: end for

(vi) Theoretical foundations of combining non-negative matrix and tensor factorization

From a more theoretical perspective, NMF estimates, for each view, the transformed data in the form of a matrix W_v and a view-mapping matrix H_v, which allows the reconstruction of the original view. Following a geometrical interpretation from Donoho and Stodden,[33] we consider

the simplicial cone Γ_{W_v} contained in the positive orthant of R^n and generated by the columns of W_v. In this simplicial cone, each view-attribute corresponds to a point, with coordinates found in the corresponding row of H_v. To identify a consensus simplicial cone between the Γ_{W_v}, NTF decomposes the tensor formed by the W_v into a sum of rank-1 tensors (Unit 4). However, for such a decomposition to be meaningful, the dimensions defined by the columns of the W_v must be consistent from one view to another. This implies a strong overlap between the simplicial cones Γ_{W_v}. Such consistency is achieved by the multiple zeros found across the columns of H_v when starting the embedding process (Unit 3). These are "inactive" attributes, as their zero status cannot be changed by multiplicative update rules. They can be interpreted as anchors ensuring that the W_v do not deviate significantly from their common ancestor W, estimated in the preliminary NMF over concatenated views (Unit 1). The parsimonization process (Unit 2) is designed to ensure that there will be a sufficient number of anchor attributes to rein in the multiplicative updates.

(b) Workflow 2: Projection of new observations

For new observations Y comprising k views, $k \leq m$, ISM parameters H^*, Q^*, and the view-mapping matrix H can be used to project Y onto the latent ISM components, as described in Workflow 2.

Workflow 2. Projection of new observations

Input: New observations Y (k views, $k \leq m$),

NTF factors H^*, Q^* and mapping matrix H.

Output: Estimation of Y^*.

1: Disregard any views in Q^*, H that are absent in Y;

2: Apply Unit 3 of Workflow 1 to embed Y with W initialized with ones and with fixed mapping matrix H;

3: Apply step 2 of Unit 4 of Workflow 1 to calculate W^* with fixed NTF factors H^*, Q^* and define the projection of Y on the latent space as $Y^* = W^*$;

Abbreviation: NTF: Non-negative tensor factorization.

(c) Workflow 3: Proof of concept analysis

Each dataset is analyzed using ISM, ILSM, NMF, MVMDS, GFA, MOFA+, and MOWGLI. PCA is also applied to the concatenated views of the UCI Digits and Signature 915 datasets, mainly to show the added value of alternative approaches over this widely used method.

To facilitate interpretation, the transformed data are projected onto a 2D map before being subjected to K-means clustering, where k is the known number of classes (K-means clustering was chosen for its versatility and simplicity, as it only requires the number of clusters to be found, and this number is known for our example datasets). Within each cluster, the class that contains the majority of the points, that is, the main class is identified. If two clusters share the same main class, they are merged unless they are not contiguous (the ratio of the distance between the centroids to the intra-cluster distance between points >1). In this case, the non-contiguous clusters are excluded because they are assigned to the same class, which should appear homogeneous in the representation. Similarly, any cluster that does not contain an absolute majority is not considered clearly representative of the class to which it is assigned and is excluded from the study. A global purity index is then calculated for the remaining clusters using Workflow 3. To enhance clarity, the clusters are visualized using 95% confidence ellipses, while the classes are represented using distinct colors. In addition to the proportion of classes retrieved and the global purity index, the adjusted rand index (ARI),[34] normalized mutual information (NMI) index,[35] and Fowlkes-Mallows score (FMS)[36] are also included, along with the factor specificity index (FSI) and view-mapping sparsity (VSI) defined as follows:

The FSI reflects the level of factor specificity with respect to a given class: A value close to 1 means that only one factor contributes significantly to the explanation of the class; while a value close to 0 means that the class is explained by a large number of factors. This index was proposed in Huizing *et al.*,[21] but in its original definition, it measures the level of specificity of each factor relative to the class. The FSI is defined as the ratio of the maximum specificity observed across all factors over the number of significant factors. To estimate the number of significant factors, we use the inverse HHI of all factor indices.

The VSI reflects the level of the sparsity of the mapping matrix H. To obtain the VSI: (i) Estimate, for each view and each ISM component, the number of significant loadings, using the inverse HHI; (ii) for each view, define the view-sparsity as the average sparsity over all ISM components; and (iii) define VSI as the average view-sparsity over all views.

Multidimensional scaling is applied to achieve the 2D map projection. MDS uses a simple metric objective to find a low-dimensional embedding that accurately represents the distances between points in the latent space.[37] MDS is, therefore, agnostic to the intrinsic clustering performances of the methods that we want to evaluate. Effective embedding methods, for example, uniform manifold approximation and projection (UMAP) or *t*-distributed stochastic neighbor embedding, are not as optimal for preserving the global geometric structure in the latent space.[38] For example, a resolution parameter needs to

be defined for the UMAP embedding of single-cell data, whereby a higher resolution leads to a higher number of clusters. In addition, the subtle differences between some cell types from one family can be smoothed out if the dataset contains transcriptionally distinct cell types from multiple families, as is the case with immune cells for the Signature 915 dataset.

Latent space methods require that the rank of the factorization is determined in advance. ISM benefits from the advantages of the NMF and NTF workflow components, that is, the choice of the correct rank is less critical than with other methods (we will come back to this point in the results [Section 3] and discussion [Section 4] sections). This allows, even if we expect some redundancy in the latent factors – for instance, due to the proximity of certain digits in the first dataset – to set the rank to the number of known classes.

The dimension of the ISM embedding space must also be determined during the discovery step. A natural choice is the dimension of the latent space since both spaces are merged at the end of the ISM workflow. Nevertheless, by examining the approximation error for an embedding dimension in the neighborhood of the chosen rank, it is possible to further optimize the ISM representation.

The rank for PCA, MVMDS, GFA, and MOFA+ is set by inspecting the scree plot of the variance ratio.

The analysis of the Signature 915 dataset also examines the biological relevance of the distance between clusters in each latent multi-view space. Of the five datasets analyzed in this article, only the Signature 915 dataset is a 3D array; therefore, NTF is also directly applied to this particular dataset.

Detailed analysis steps are provided in Workflow 3.

Workflow 3. Analysis steps

Input: 2D map projection of the data transformation in the latent space.

Output: Cluster purity index.

1: Perform K-means with k equal to the number of known classes;

2: For each cluster, identify the main class related to the cluster, that is, the class corresponding to the majority of observations in the cluster;

3: Merge contiguous clusters that refer to the same class or ignore them if not contiguous;

4: for each cluster do

5: p_1=proportion of the main class in relation to all elements in the cluster;
p_2=proportion of the main class in cluster c in relation to all elements of the same class;

6: If $p_2 < 0.5$ then

7: Disregard cluster as the main class does not constitute an absolute majority in relation to all elements of the same class;

8: Else

9: $p=p_1 \times p_2$=purity corrected for cluster representativity for the main class;

10: end for

11: Calculate the global purity=sum of corrected purities over all retained clusters, divided by the number of known classes;

2.3. Implementation

Scikit-learn[39] was used for K-means, ARI, NMI, MDS, and PCA. The mvlearn (https://pypi.org/project/mvlearn/) package was used for MVMDS. NMF and NTF were performed with the package adnmtf (https://pypi.org/project/adnmtf/). ISM was implemented in Python and was invoked from a Jupyter Python notebook available on the Advestis GitHub (https://github.com/Advestis). GFA was performed with the Python package gfa-python (https://github.com/mladv15/gfa-python). MOFA+ was performed with the Python package mofapy2 (https://github.com/bioFAM/mofapy2). Matplotlib (https://matplotlib.org/stable/tutorials/pyplot.html) was used to create the clustering figures. Treemaps were obtained with the Graph Builder platform from JMP® (Version 17.2.0. SAS Institute Inc., USA). The distinctipy package (https://pypi.org/project/distinctipy/) was used to generate colors that are visually distinct from one another.

3. Results

We first present a synthesis of the calculated metrics across all datasets (Table 1) and provide some general observations. We then present more detailed results for each dataset.

3.1. Synthesis of calculated metrics over all datasets

Based on the average index across all seven indices, ISM ranks first in the UCI Digits, Signature 915, and Reuters datasets, while ILSM ranks first in the prokaryotic dataset and the TEA-seq multi-omic single-cell dataset (although very close to ISM for the latter dataset, 0.80 vs. 0.79, respectively). It is easy to explain why ILSM performed much better than ISM on the prokaryotic dataset (0.52 vs. 0.37, respectively): Since ISM first performs a global factorization over concatenated views (Unit 1 of Workflow 1), it tends to ignore the smallest views when they are extremely unbalanced, as is the case in the prokaryotic dataset. However, when using ILSM, separate factorizations are applied to each view, and ISM itself is applied to transformed views of equal size. As a result, the original views with the smallest size are given equal weight. Among the criteria used, the proportion of classes retrieved, purity, and sparsity indices are the most discriminative. It is noteworthy that NMF performs as

Table 1. Metrics comparing latent-space methods on five datasets

Dataset	Method	Nr classes	Embedding (ISM) rank	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
UCI DIGITS	MVMDS	10	10	0.70	0.41	0.49	0.61	0.54	0.62	0.21	0.51
	ISM	10	(9,10)	1.00	0.58	0.57	0.67	0.62	0.87	0.43	0.68
	ILSM	10	(10,10)	0.80	0.41	0.45	0.58	0.51	0.50	0.48	0.53
	GFA	10	10	0.90	0.45	0.48	0.61	0.54	0.32	0.15	0.49
	MOFA+	10	10	0.70	0.29	0.36	0.46	0.44	0.34	0.13	0.39
	MOWGLI	10	10	0.80	0.46	0.51	0.65	0.57	0.60	0.58	0.60
	PCA	10	10	0.40	0.19	0.44	0.57	0.51	0.73	0.38	0.46
	NMF	10	10	0.90	0.58	0.59	0.68	0.63	0.46	0.34	0.60
Signature 915	MVMDS	16	10	0.75	0.70	0.97	0.95	0.97	0.56	0.21	0.73
	ISM	16	(16,16)	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90
	ILSM	16	(16,16)	0.75	0.62	0.93	0.91	0.94	0.93	0.74	0.83
	GFA	16	12	0.81	0.73	0.98	0.96	0.98	0.30	0.08	0.69
	MOFA+	16	13	0.81	0.76	0.94	0.93	0.95	0.56	0.19	0.73
	MOWGLI	16	16	0.63	0.44	0.87	0.89	0.89	0.89	0.82	0.77
	PCA	16	10	0.56	0.40	0.94	0.89	0.95	0.57	0.23	0.65
	NMF	16	16	0.81	0.55	0.94	0.89	0.95	0.91	0.88	0.85
	NTF	16	16	0.69	0.52	0.94	0.89	0.95	0.98	0.75	0.82
Reuters	MVMDS	6	4	0.50	0.19	0.19	0.30	0.37	0.93	0.28	0.39
	ISM	6	(6,6)	0.50	0.23	0.25	0.34	0.41	0.98	0.37	0.44
	ILSM	6	(6,6)	0.33	0.16	0.21	0.31	0.39	0.97	0.30	0.38
	GFA	6	3	0.17	0.03	0.01	0.09	0.39	0.94	0.21	0.26
	MOFA+	6	-	-	-	-	-	-	-	-	-
	MOWGLI	6	6	0.08	0.04	0.01	0.20	0.28	0.10	0.86	0.22
	NMF	6	6	0.33	0.14	0.21	0.32	0.41	0.96	0.36	0.39
Prokaryotic	MVMDS	4	4	0.50	0.22	0.18	0.23	0.50	0.68	0.67	0.43
	ISM	4	(4,4)	0.25	0.14	0.00	0.00	0.63	0.66	0.88	0.37
	ILSM	4	(4,4)	0.75	0.36	0.28	0.31	0.54	0.55	0.88	0.52
	GFA	4	6	-	-	-	-	-	-	-	-
	MOFA+	4	4	0.75	0.36	0.29	0.32	0.55	0.53	0.42	0.46
	MOWGLI	4	4	0.25	0.14	0.10	0.10	0.60	0.39	0.63	0.32
	NMF	4	4	0.25	0.14	0.00	0.00	0.63	0.47	0.88	0.34
TEA-seq	MVMDS	7	7	0.71	0.60	0.89	0.86	0.92	0.67	0.48	0.73
	ISM	7	(7,7)	0.71	0.57	0.87	0.84	0.90	0.76	0.88	0.79
	ILSM	7	(7,7)	0.86	0.72	0.88	0.85	0.91	0.75	0.67	0.80
	GFA	7	15	0.71	0.61	0.91	0.89	0.93	0.45	0.25	0.68
	MOWGLI	7	7	0.43	0.23	0.52	0.60	0.64	0.39	0.62	0.49
	NMF	7	7	0.71	0.61	0.88	0.86	0.91	0.70	0.90	0.80

Abbreviations: ARI: Adjusted rand index; GFA: Group factor analysis; FMS: Fowlkes-Mallows score; ILSM: Integrated latent sources model; ISM: Integrated sources model; MOWGLI: Multi-Omics Wasserstein inteGrative anaLysIs; MVMDS: Multi-view multidimensional scaling; NMF: Non-negative matrix factorization; NMI: Normalized mutual information index.

well as ILSM in the TEA-seq multi-omic single-cell data in terms of average performance (0.80). However, we will show in the detailed analysis of this dataset that ISM finds a superior representation in terms of biology. In addition, NMF retrieves only one class in the prokaryotic data due to the extreme imbalance in the number of features per view

and the fact that NMF runs on the concatenated views, thus tending to ignore the smallest ones. Although GFA and MOFA+ are closely related, MOFA+ fails to recover common factors in the Reuters dataset, while GFA fails in the prokaryotic dataset. MVMDS performs relatively well on all datasets, in most cases with lower factor sparsity and specificity than ISM or ILSM. MOWGLI could only be run on a fraction of the data for the Reuters and TEA-seq multi-omic single-cell data due to its extremely high computational time. The poor performance observed can, therefore, be attributed to the sampling itself.

3.2. Detailed results

3.2.1. UCI digits dataset

PCA, MVMDS, MOFA+, and MOWGLI use a 10-factorization rank, while GFA uses a 9-factorization rank. ISM uses a primary embedding of dimension 9 and a 10-factorization rank. The Karhunen-Love coefficients contain data with mixed signs, so the corresponding view is split into its positive part and the absolute value of its negative part when applying the non-negative approaches ISM, NMF, and MOWGLI. The clusterings of the digits are shown in Figure 3. ISM outperforms the other methods with 10-digit-specific clusters. It should be noted that NMF performs slightly better than ISM in terms of purity index, ARI, NMI, and FMS. However, digits 5 and 3 are mixed together, resulting in one less digit being recognized. PCA is far behind all other approaches, recognizing only four-digit classes.

Figure 4 shows how the views affect the individual ISM components using a treemap chart. For each component, each view corresponds to a rectangle within a rectangular display, where the size of the rectangle represents the loading of the view. It is noteworthy that some components are supported by only a few views, for example, component 1 (2 views) and component 8 (3 views), while others involve most views, for example, component 5 (6 views). As each component is associated with a digit, this emphasizes the specifics and complementarity of the image representations that are dependent on the respective digit. It is also interesting to note that for some components, the loadings of the views are diametrically opposed to the respective number of attributes. For example, for component 8, the view of 240-pixel averages has the lowest loading, while the view of six morphological features has the highest loading. This clearly shows that the views are evenly balanced regardless of their respective number of attributes when using ISM.

3.2.2. Signature 915 data

Before the analysis, each marker gene was normalized using the mean of the four highest expression values. PCA and MVMDS use a 10-factorization rank, GFA uses a 12-factorization rank, and MOFA+ uses a 13-factorization rank. ISM uses a primary embedding of dimension 16 and a 16-factorization rank. The clusterings of the marker genes are shown in Figure 5. ISM outperforms the other methods with 14 cell type-specific clusters and higher metrics.

Regarding the positioning of the clusters on the 2D map, MVMDS places classical monocyte (monocyte C) and non-classical and intermediate monocytes (monocyte NC+I) opposite of each other, contrary to all other approaches and, more importantly, against biological intuition. ISM and GFA methods outperform other methods on this dataset as they reveal close proximity between transcriptionally and functionally similar cell types of the major immune cell families. Indeed, three cell types from the myeloid lineage, including monocytes C, monocytes NC+I, and myeloid dendritic cells (mDC), are grouped together. A similar trend is observed for three cell types from the B cell family, where only ISM and GFA revealed close proximity of naïve B cells, memory B cells, and plasmablasts, out of the eight methods considered. The most challenging cell types were in the T cell family, where only ISM was able to identify clusters for three cell types (CD4+ effectors, naïve T cells, and Vδ2+ T cells [VD+] gamma delta non-conventional T cells) and place them in close proximity. VD+ gamma delta non-conventional T cells share some similarities with NK cells in terms of the expression of certain receptors, and only the ISM method was able to recognize both cell types and place them in close proximity, highlighting their similarity. The ISM method also captured subtle similarities between two types of dendritic cells, mDC, and plasmacytoid dendritic cells (pDC), which correspond to antigen-presenting cells.

Figure 6 shows the impact of the four patients on the individual ISM components using a treemap chart. In this chart, each patient corresponds to a rectangle within a rectangular display, where the size of the rectangle represents the loading of the patient. In contrast to the UCI Digits data, most components are supported by three patients (three components) or four patients (11 components). Two components involve only two patients.

The loadings of the view-mapping matrix are shown in Figure 7 using a treemap chart. Recall that each attribute of this dataset is a combination of a patient and a cell type, in which the expressions of 915 marker genes were measured. For each component, such a combination corresponds to a rectangle within a rectangular display, where the size of the rectangle represents the loading of the combination. ISM components 1 and 2 are both associated with the same cell type, pDC, while component 15 is simultaneously associated with CD8-activated, VD2-, and VD2+ cells. In the final clustering, the cluster comprising these three

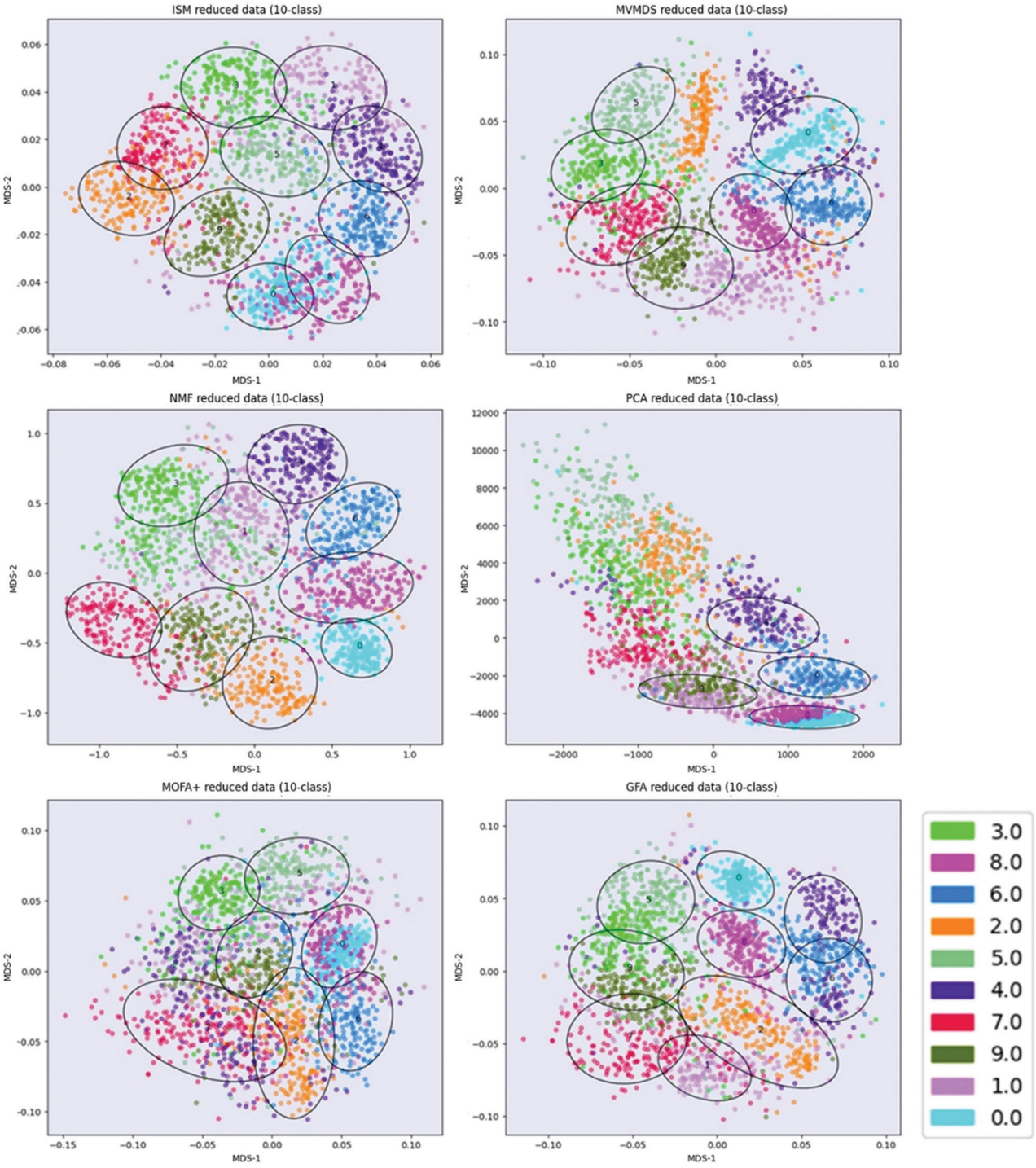

Figure 3. UCI Digits data: Clustering of digit images along ISM, MVMDS, NMF, PCA, MOFA+, and GFA components in 2D scatterplots of the MDS projection of transformed data.
Abbreviations: ISM: Integrated sources model; MDS: Multidimensional scaling; MOFA+: Multi-Omics factor analysis; MVMDS: Multi-view multidimensional scaling; NMF: Non-negative matrix factorization; NTF: Non-negative tensor factorization; PCA: Principal component analysis.

cell types has no main type and is therefore discarded, resulting in 14 identified cell types. All other components are associated with only one cell type, illustrating the sparsity and interpretability of ISM components. Notably,

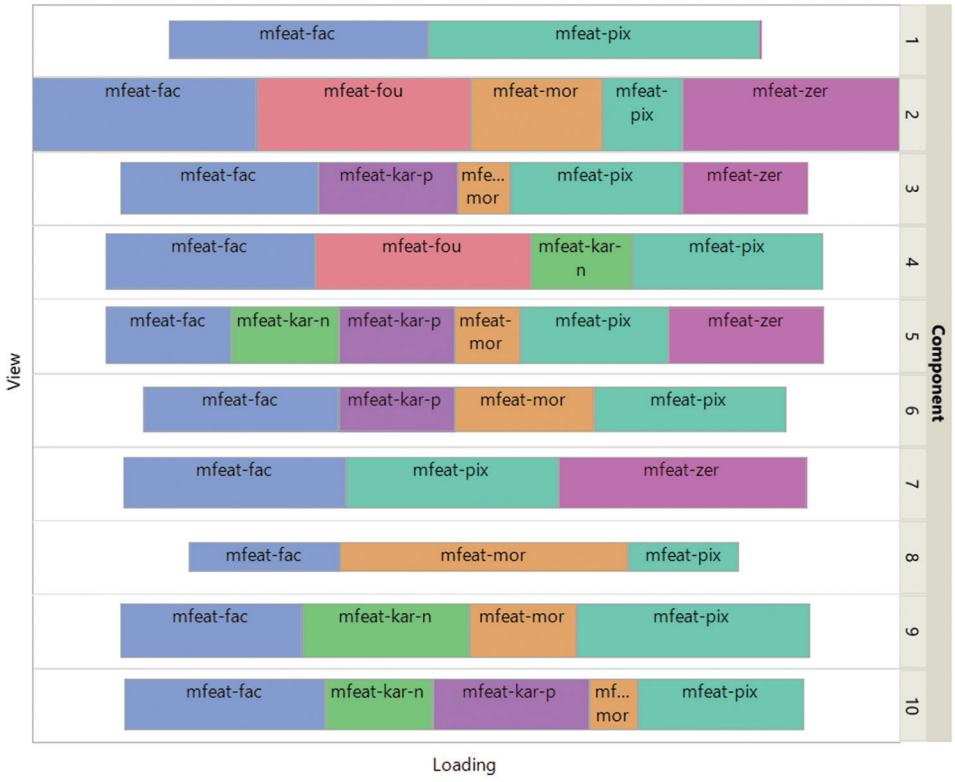

Figure 4. UCI Digits data: Treemap of integrated sources model view weights

all non-negative approaches yield a high view-mapping sparsity index (i.e., 0.93, 0.89, 0.91, and 0.98 for ISM, MOWGLI, NMF, and NTF, respectively), as opposed to mixed-sign approaches (i.e., 0.56, 0.30, 0.56, and 0.57 for MVMDS, GFA, MOFA+, and PCA, respectively).

3.2.3. Reuters data

MVMDS and MOWGLI use a 6-factorization rank, while MOFA+ and GFA use a 10-factorization rank. ISM uses a primary embedding of dimension 6 and a 6-factorization rank (equal to the number of known categories). Overall, ISM outperforms the other methods, identifying three out of six categories and achieving higher metrics, followed by MVMDS. However, all performance indices are relatively low, as previously observed in Brbic and Kopriva.[22] MOFA+ fails to identify a common structure between the different views. It should be noted that MOWGLI was performed on only 20% of the samples due to its extremely high computational time, despite using an activated graphics processing unit (GPU). The poor performance observed can, therefore, be attributed to the sampling itself.

3.2.4. Prokaryotic data

MVMDS, MOFA+, and MOWGLI use a 4-factorization rank, while GFA uses a 6-factorization rank. ILSM uses a primary embedding of dimension 4 and a 4-factorization rank (equal to the number of known categories). Since the provided views contain the principal components explaining 90% of the variance, they need to be split into their positive part and the absolute value of their negative part when applying the non-negative approaches ISM, NMF, and MOWGLI. Overall, ILSM outperforms the other methods, identifying three out of four categories (missing the category which the smallest size) and achieving higher metrics.

3.2.5. TEA-seq multi-omic single-cell data

MVMDS and MOWGLI use a 7-factorization rank, while GFA uses a 15-factorization rank. ISM uses a primary embedding of dimension 7 and a 7-factorization rank (equal to the number of categories). MOFA+ metrics are not presented for this particular dataset since the corresponding clustering was used to annotate the cells.[21] We used a UMAP projection because the size of the dataset makes MDS impractical (Figure 8).

The ILSM outperforms the other methods, identifying six out of seven cell types, missing only MAIT T cells, which are too close to CD4 effector and memory T cells to be identified as a separate cluster.

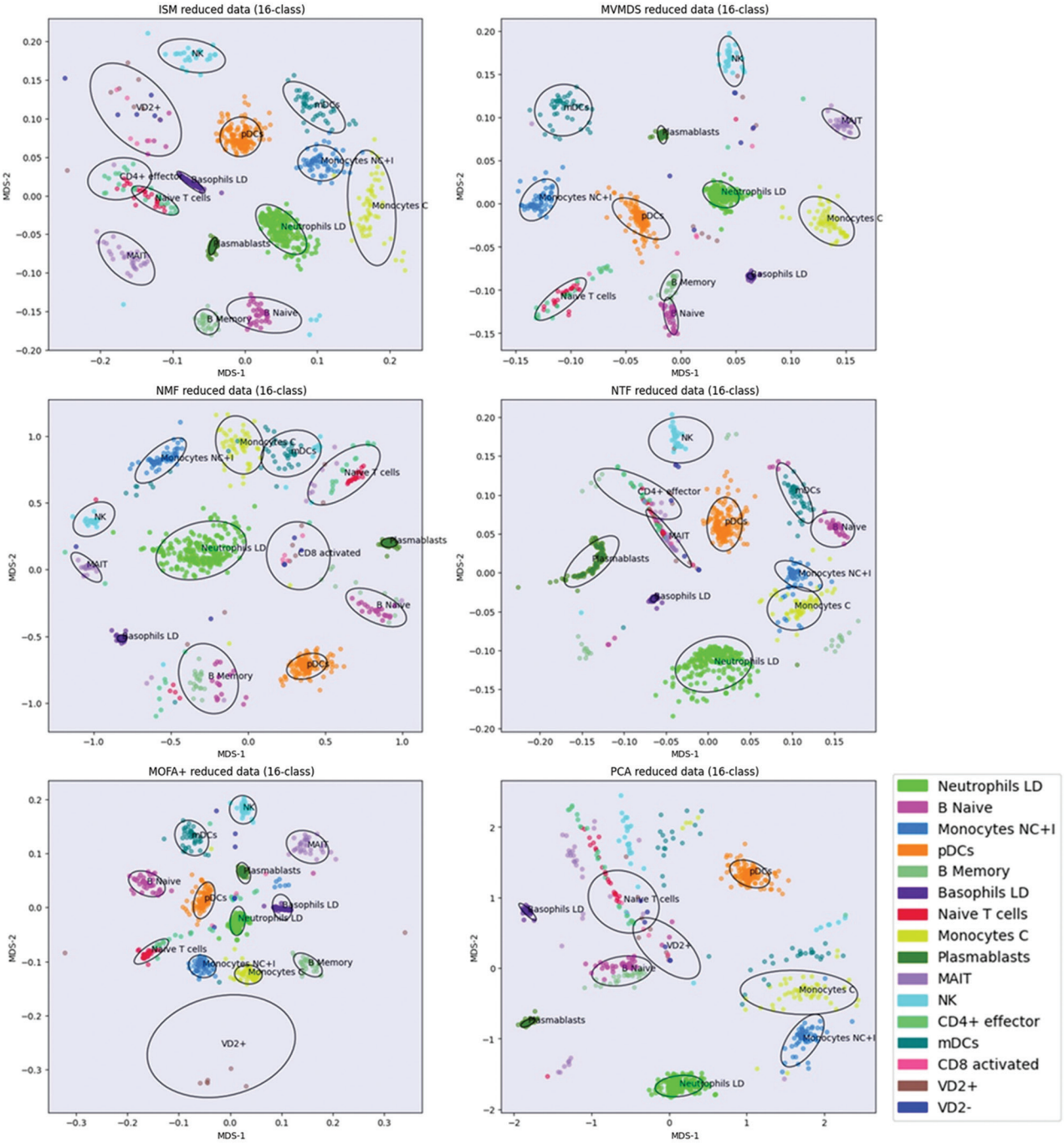

Figure 5. Signature 915 Data: Clustering of cell type marker genes along ISM, MVMDS, NMF, NTF, MOFA+, and PCA components in the 2D scatterplots of the MDS projection of the transformed data.
Abbreviations: Basophil LD: Low-density basophil; ISM: Integrated sources model; MAIT: Mucosal-associated invariant T cell; mDCs: Myeloid dendritic cells; MOFA+: Multi-Omics Factor Analysis; Monocyte C: Classical monocyte; Monocyte NC+I: Non-classical + intermediate monocytes; MVMDS: Multi-view multidimensional scaling; Neutrophil LD: Low-density neutrophil; NK: Natural killer cell; NMF: Non-negative matrix factorization; NTF: Non-negative tensor factorization; PCA: Principal component analysis; pDCs: Plasmacytoid dendritic cells; VD2: Vδ2+ T cells.

Other methods, including ISM, revealed heterogeneity in the CD4+ naive T cell population, which appeared to be split in the corresponding UMAP projections. In the ISM UMAP projection, we annotated the smaller split near

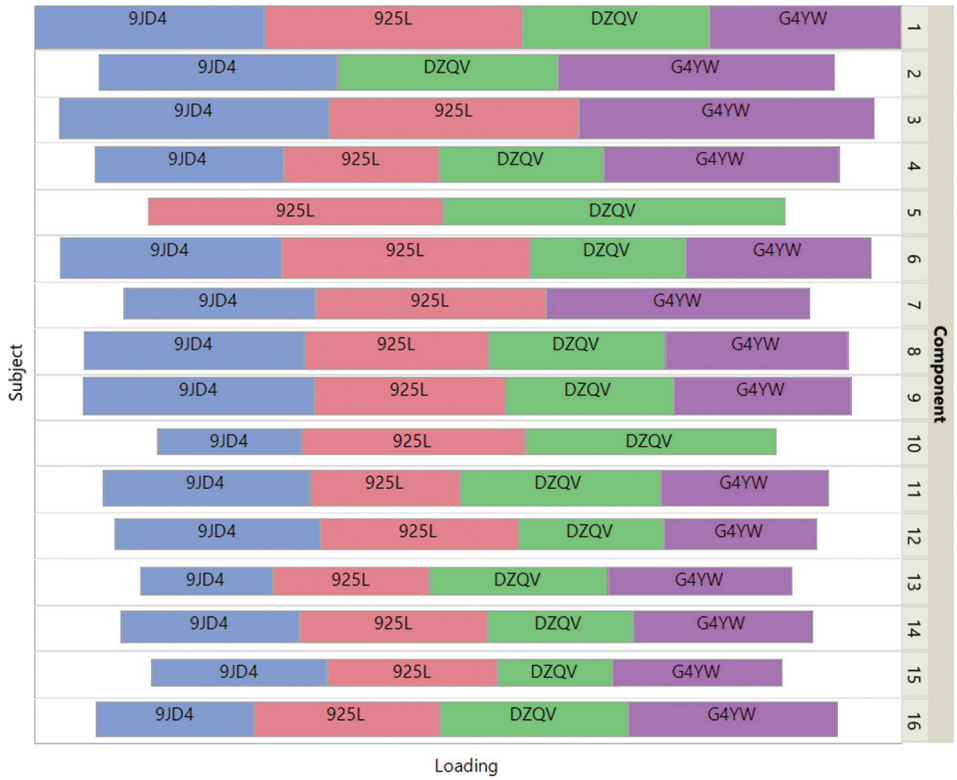

Figure 6. Signature 915 data: Treemap of integrated sources model view weights

CD8+ T cells and found that it actually represents CD8+ naive T cells. We then verified that the split observed in the other UMAP projections was also consistent with this cell type. In particular, the distance between CD8+ naive T cells and CD8+ T cells is minimal with ISM, consistent with biology. In contrast, CD8+ naive T cells are closer to CD4+ naive T cells and not to CD8+ T cells in the NMF UMAP projection, contrary to biological intuition. One possible reason is that NMF does not take advantage of the complementarity of different views by indiscriminately concatenating them.

Interestingly, and in contrast to other multi-view approaches, ISM allows the direct identification of factors and views that are discriminative with respect to a particular cell type (e.g., CD8+ naive T cells). From the factor specificities, we found two specific ISM latent factors with positive factor specificities with respect to CD8+ naive T cells (0.50 and 55, respectively). In all three multi-omic modalities, these factors have close loadings in the view-weights matrix Q^* (13.09/9.00 in the RNA-seq view, 10.82/8.04 in the ATAC-seq view, and 11.45/8.94 in the ADT view, respectively), highlighting the contribution of the three modalities to specifically distinguish CD8+ and CD4+ cell subpopulations among naive T cells.

It should be noted that MOWGLI was performed on only 20% of the samples and 20% of the scATAC-seq features due to its extremely high computational time, despite using an activated GPU. The poor performance observed can, therefore, be attributed to the sampling itself.

3.3. Further insights regarding the model

3.3.1. Model's potential dependency on embedding dimension and rank

In this section, we evaluate how ISM performance might be affected by changing the embedding dimension and the rank in the neighborhood of the chosen values. First, we examine the relative approximation error for an embedding dimension in the neighborhood of the chosen rank to select an optimal value, as described in the analysis workflow. Second, we examine the relative error, number of found classes, and purity for a rank in the neighborhood of the chosen embedding dimension.

For the UCI Digits data, where the chosen ISM rank is 10, an embedding dimension of 9 clearly minimizes the relative error—0.52 versus 0.72 or higher for other dimensions. The number of classes found and the purity index are also significantly higher (Table 2, upper part). The relative

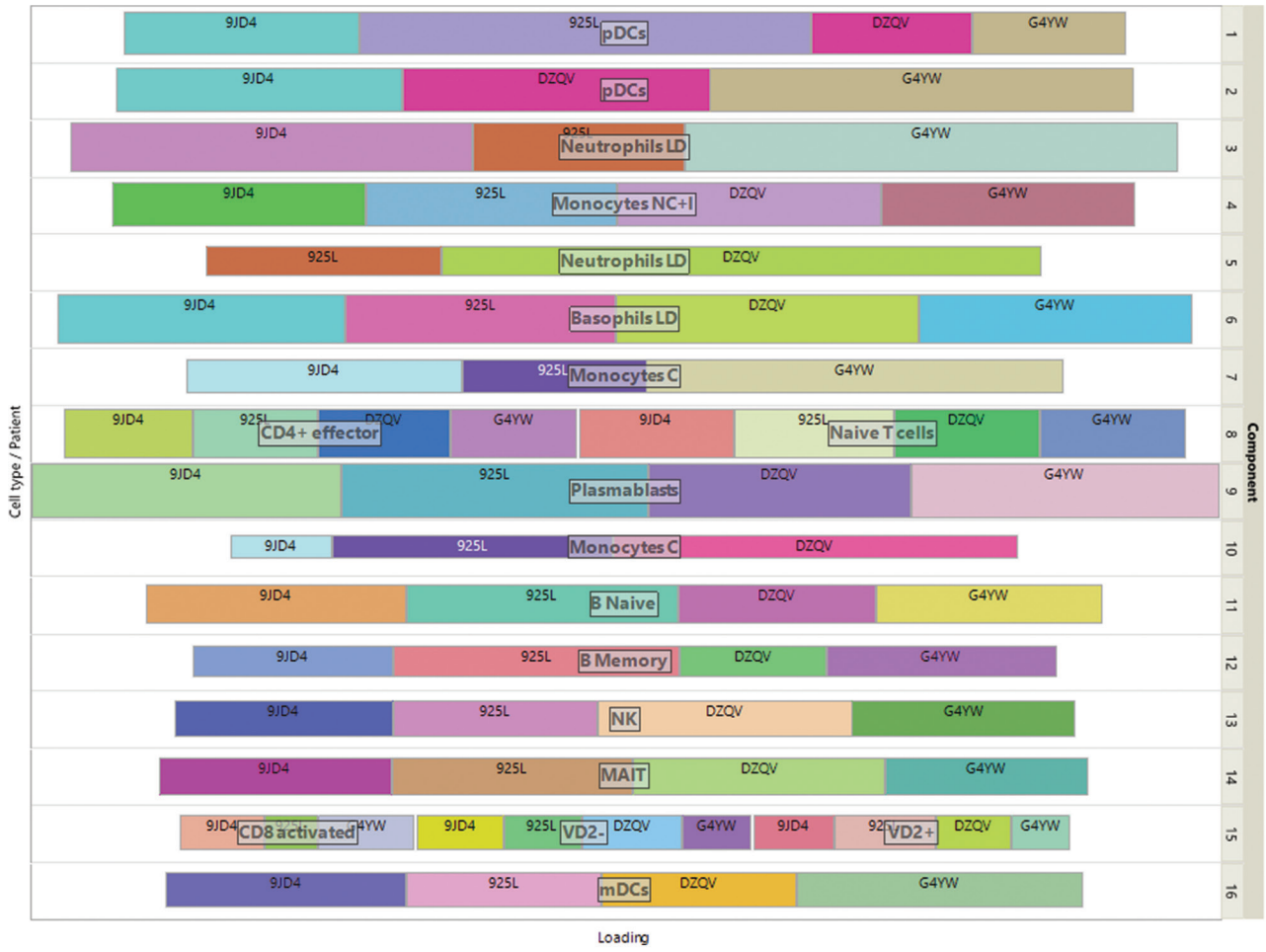

Figure 7. Signature 915 data: treemap of integrated sources model loadings of the view-mapping matrix

error associated with a rank is not as critical if it exceeds the number of known classes. Compared to a 10-rank ISM model, a 12-rank model also finds 10 classes and gives a slightly higher purity index (6.24 vs. 5.81), despite a larger relative error (0.60 vs. 0.52) (Table 2, bottom part). The final part of this section discusses this point further.

For the Signature 915 dataset, where the chosen ISM rank is 16, the relative error does not change significantly for neighboring embedding dimensions: 0.33 for a 15-embedding and 0.34 for a 17-embedding (Table 3, upper part). Choosing an embedding dimension equal to the rank is more consistent with the ISM workflow, where the embedding and latent spaces are united during the straightening process. Therefore, we chose an embedding dimension of 16. In terms of purity, a 17-rank ISM model gives results that are slightly superior to the 16-rank ISM model (Table 3, bottom part).

Overall, these results confirm that ISM provides relatively stable estimates in the neighborhood of the chosen rank, in line with its parent methods, NMF and NTF.

3.3.2. About changing the sparsity coefficient

We have already mentioned that the initial degree of sparsity of H returned by NMF is a critical part of ISM, as zero-loading attributes are anchors that maintain consistency between view components during the embedding process. However, it is extremely difficult to predict how sparse an NMF representation will be, as this depends on the dataset under analysis.[33] To ensure that a sufficient number of anchors will guide the embedding, only significant loadings are retained, while other loadings are set to 0. ISM uses the inverse of the HHI to identify significant loadings, but an additional sparsity parameter is provided to allow this index to be relaxed. This parameter is set to 0.8 by default. In this section, we examine the effect of changing this parameter in the UCI Digits and Signature 915 experiments (Tables 4 and 5, respectively).

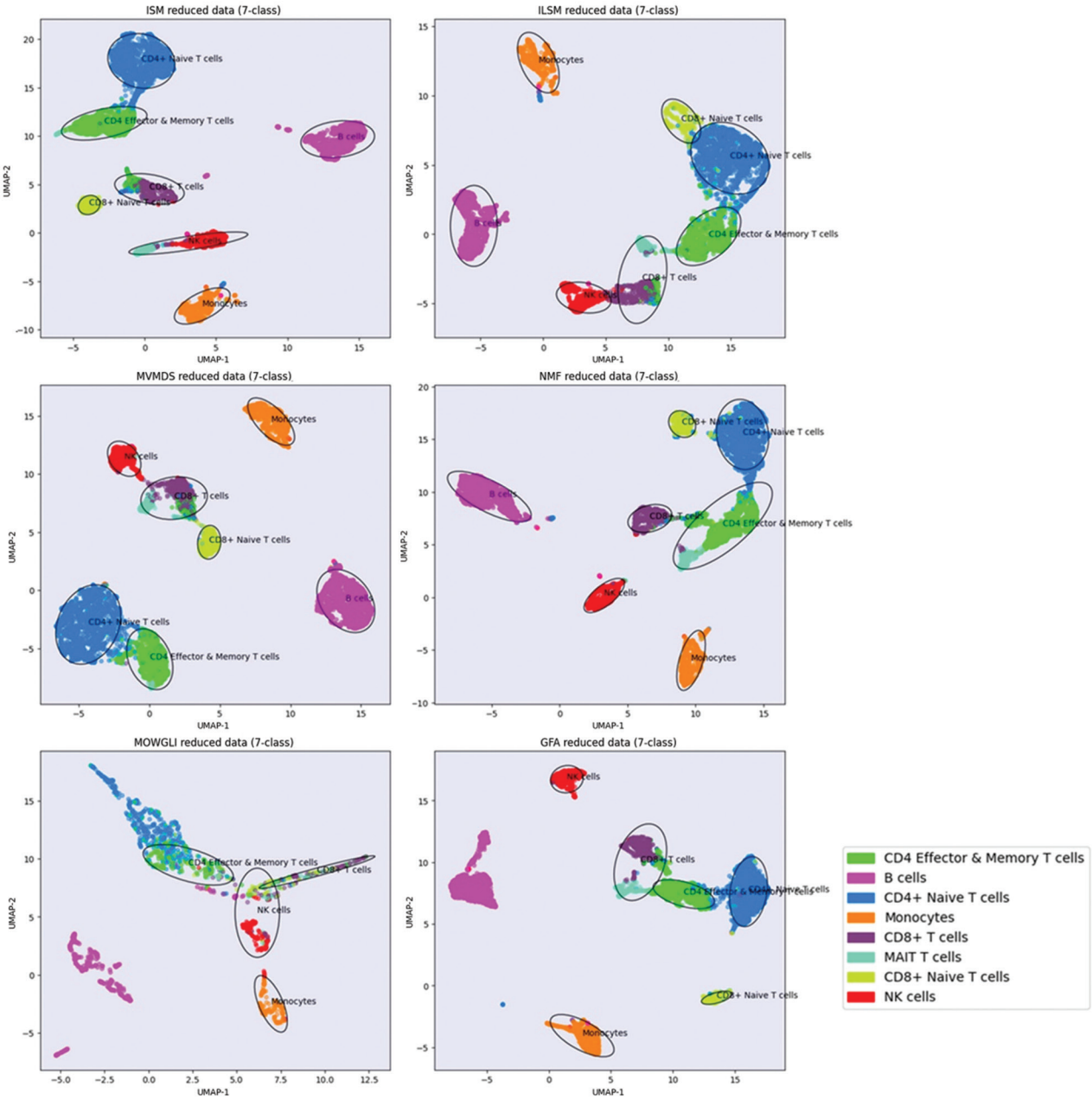

Figure 8. TEA-seq multi-omic single-cell data: Clustering of cells along ISM, ILSM, MVMDS, NMF, MOWGLI, and GFA components in the 2D scatterplots of the UMAP projection of the transformed data, with an additional cell type CD8+ naïve T cells as identified by ISM.
Abbreviations: GFA: Group factor analysis; ILSM: Integrated latent source model; ISM: Integrated sources model; MAIT: Mucosal-associated invariant T cell; MOWGLI: Multi-Omics Wasserstein inteGrative anaLysIs; MVMDS: Multi-view multidimensional scaling; NK: Natural killer cell; NMF: Non-negative matrix factorization; UMAP: Uniform manifold approximation and projection.

Several key observations are summarized as follows:

(i) The use of a sparsity parameter slightly >1 (e.g., 1.1) severely degrades performance across all metrics due to increased relative error and can even lead to computational errors, as observed in the UCI Digits dataset. Therefore, using a sparsity parameter too close to 1 poses a significant risk. To err on the conservative side, we choose the default value of 0.8.

(ii) The proportion of classes retrieved is not significantly affected by a low sparsity parameter. For example, eight out of 10-digit classes and 11 out of 16 cell types are still recognized with a sparsity parameter of 0.

Table 2. Relative error and other metrics as functions of the embedding dimension and rank (UCI Digits dataset, 10 classes)

Embedding (ISM only) rank used	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
(8,10)	0.84	0.60	0.27	0.38	0.50	0.44	0.86	0.24	0.47
(9,10)	0.52	1.00	0.58	0.57	0.67	0.61	0.87	0.20	0.64[a]
(10,10)	0.72	0.30	0.14	0.32	0.45	0.40	0.84	0.20	0.38
(11,10)	1.20	0.80	0.47	0.51	0.62	0.56	0.91	0.32	0.60
(12,10)	0.91	0.70	0.32	0.37	0.53	0.44	0.90	0.21	0.50
(9,8)	0.7	0.70	0.32	0.40	0.55	0.47	0.86	0.22	0.50
(9,9)	0.52	0.80	0.36	0.41	0.55	0.47	0.86	0.21	0.52
(9,10)	0.52	1.00	0.58	0.57	0.67	0.61	0.87	0.20	0.64[a]
(9,11)	0.62	0.90	0.42	0.44	0.58	0.50	0.87	0.19	0.56
(9,12)	0.61	1.00	0.62	0.60	0.67	0.64	0.88	0.21	0.66

Note: [a]The most performant combinations.
Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; ISM: Integrated sources model; NMI: Normalized mutual information index.

Table 3. Relative error and other metrics as functions of the embedding dimension and rank (Signature 915 dataset, 16 classes)

Embedding (ISM only) rank used	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
(14,16)	0.36	0.63	0.55	0.92	0.91	0.93	0.94	0.70	0.80
(15,16)	0.33	0.75	0.69	0.95	0.94	0.96	0.94	0.74	0.85
(16,16)	0.34	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90[a]
(17,16)	0.34	0.81	0.67	0.92	0.91	0.93	0.94	0.74	0.85
(18,16)	0.38	0.81	0.75	0.95	0.94	0.96	0.94	0.74	0.87
(16,14)	0.39	0.69	0.60	0.96	0.94	0.97	0.92	0.76	0.83
(16,15)	0.34	0.88	0.73	0.98	0.96	0.98	0.93	0.84	0.90
(16,16)	0.34	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90[a]
(16,17)	0.31	0.88	0.76	0.96	0.94	0.96	0.93	0.79	0.89
(16,18)	0.31	0.63	0.55	0.95	0.92	0.95	0.93	0.76	0.81

Note: [a]The most performant combinations.
Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; ISM: Integrated sources model; NMI: Normalized mutual information index.

Table 4. Effect of changing the sparsity parameter in the UCI Digits experiment (10 classes), with embedding dimension=9 and rank=10

Sparsity parameter	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
1.1	0.98	0.10	0.00	0.07	0.13	0.22	1.00	0.17	0.24
1	0.70	0.70	0.27	0.35	0.52	0.42	0.94	0.16	0.48
0.9	0.55	0.90	0.37	0.39	0.54	0.46	0.90	0.17	0.53
0.8	0.52	1.00	0.58	0.57	0.67	0.61	0.87	0.20	0.64
0.7	0.62	1.00	0.59	0.59	0.68	0.63	0.84	0.24	0.65
0.6	0.69	1.00	0.55	0.54	0.65	0.59	0.80	0.20	0.62
0.5	0.71	0.80	0.53	0.56	0.67	0.61	0.73	0.24	0.59
0.4	0.71	0.90	0.51	0.54	0.65	0.58	0.39	0.23	0.54
0.3	0.48	0.9	0.53	0.56	0.67	0.61	0.62	0.27	0.59
0.2	0.53	0.9	0.60	0.60	0.71	0.64	0.59	0.27	0.62
0.1	0.56	0.8	0.41	0.47	0.60	0.54	0.56	0.28	0.52
0	0.76	0.8	0.50	0.54	0.64	0.59	0.51	0.31	0.55

Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; NMI: Normalized mutual information index.

Table 5. Effect of changing the sparsity parameter in the Signature 915 experiment (16 classes) with embedding dimension=16 and rank=16

Sparsity parameter	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
1.1	-	-	-	-	-	-	-	-	-
1	0.36	0.81	0.70	0.98	0.95	0.98	0.93	0.80	0.88
0.9	0.34	0.88	0.71	0.98	0.95	0.98	0.93	0.83	0.89
0.8	0.34	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90
0.7	0.34	0.63	0.50	0.96	0.92	0.96	0.93	0.83	0.82
0.6	0.33	0.81	0.67	0.97	0.93	0.97	0.93	0.79	0.87
0.5	0.34	0.63	0.54	0.96	0.92	0.96	0.92	0.83	0.82
0.4	0.34	0.69	0.61	0.96	0.94	0.97	0.93	0.72	0.83
0.3	0.33	0.75	0.65	0.97	0.94	0.97	0.93	0.72	0.85
0.2	0.34	0.75	0.67	0.95	0.94	0.96	0.93	0.72	0.85
0.1	0.33	0.81	0.70	0.94	0.93	0.95	0.92	0.72	0.85
0	0.31	0.69	0.61	0.91	0.90	0.92	0.92	0.74	0.81

Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; NMI: Normalized mutual information index.

This is due to the inherently high percentage of zero loadings in H when running standard NMF (on an average of 51% for the UCI Digits dataset and 92% for the Signature 915 dataset).

(iii) No metric shows an advantage of running ISM with a low sparsity parameter. For example, with the default sparsity parameter, higher sparsity and higher factor specificity are observed in the UCI Digits and Signature 915 experiments, respectively. To ensure a sufficient percentage of zero loadings regardless of the dataset, we recommend maintaining the default value of 0.8.

3.3.3. Evolution of the relative error over the course of model training

In this section, we evaluate how each main factorization step performed in the ISM workflow contributes to the final approximation error. Specifically, we examine the relative error obtained after (i) the preliminary NMF, (ii) the first call to NTF before the straightening process, and (iii) the last iteration of NTF in the straightening process.

While the increase in relative error is very small for the Signature 915 dataset (0.35 vs. 0.30), we observe a large increase for the UCI Digits dataset (0.53 vs. 0.36). This increase is mainly due to the straightening process (0.53 vs. 0.39 before). Recall that this process iteratively parsimonizes the view-mapping matrix H. The highly sparse nature of the Signature 915 dataset explains the difference in behavior between the two datasets: for the denser UCI Digits dataset, the increased sparsity of the view-mapping matrix induced by the straightening process significantly inflates the relative error, as more of the smaller values in the original views are filtered out. Unless the zero attribute loadings in some of the ISM components are relevant to digit class identification, this is not an issue. In fact, if we bypass the straightening process to achieve a smaller relative error, the performance of ISM is reduced; only nine-digit classes are found instead of 10, and the purity is 0.18 instead of 0.17, indicating that the model becomes overfit. This illustrates how ISM manages to filter out the specific part of the signal that is irrelevant to the main mechanisms in the data and hinders their recovery.

3.3.4. Computational time

In this section, we discuss the computation time required to analyze the TEA-seq dataset, which is a very large dataset (Table 6). The processing times for NMF, ISM, ILSM, and MVMDS are relatively short (0.55, 1.17, 1.31, and 5.31 min, respectively, on a computer equipped with an 11th Gen Intel® Core™ i7 processor and 16 GB of RAM, without the GPU activation). In contrast, GFA and MOFA+ require about 20 min with the GPU activated (7.9 GB available). MOWGLI is extremely slow, even with the GPU activated. For this reason, we had to consider only a 20% random sample of the Reuters and TEA-seq multi-omic single-cell datasets.

4. Discussion

The performance metrics used for our proof-of-concept analysis demonstrate that ISM performs as well as or better than other methods. The ISM workflow uses algorithms with proven performance and convergence properties, such as NMF and NTF, which is consistent with the good performance of ISM observed in our examples. In addition, the low computational time for large datasets indicates that this approach is highly scalable.

Table 6. Computational time observed in the TEA-seq multi-omic single-cell data

Method	Time (min)
MVMDS	5.31
ISM	1.17
ILSM	1.31
GFA	23.14
MOFA+	19.38
MOWGLI (20%)	82.31
NMF	0.55

Note: The parallelization of separate factorizations was not activated for ILSM, hence the slightly higher computational time compared to ISM. Abbreviations: GFA: Group factor analysis; ILSM: Integrated latent sources model; ISM: Integrated sources model; MOFA+: Multi-Omics Factor Analysis+; MOWGLI: Multi-Omics Wasserstein inteGrative anaLysIs; MVMDS: Multi-view multidimensional scaling; NMF: Non-negative matrix factorization.

However, the proportion of known categories retrieved and other metrics depend on the data being analyzed. For example, for the Reuters data, only three out of six categories are recognized at best using ISM or MVMDS, suggesting that latent-space-based methods may not be the most effective approaches with bag-of-words data.

In contrast to the other approaches studied, MVMDS and ISM are the only approaches that perform relatively well on all the datasets analyzed, demonstrating their versatility. The main advantages of ISM over MVMDS are its speed and increased sparsity in the latent-space representation. Regarding missing data, the ISM implementation uses an NTF package that can handle missing data, unlike MVMDS.

To the best of our knowledge, ISM is the first approach that uses NMF to transform heterogeneous views into a 3D array and then uses NTF to extract consistent information from the transformed views. However, apparent commonalities with anchor-based MVC methods (A-MVC) are worth mentioning to further illustrate the originality of ISM:

(i) In the first step, ISM relies on anchors, akin to A-MVC. ISM anchors correspond to zero-loading attributes in the latent spaces defined by the H_v, whereas A-MVC anchors are observations well distributed over existing clusters. Both act as intermediaries to derive either a latent space or cluster labels shared by all views.

(ii) In the second step, ISM applies NTF on the embedded views. A-MVC applies NTF on a tensor of anchor graphs, albeit with added constraints that ensure orthogonality and consistency in the cluster labels across all views.

A-MVC requires a specialized algorithm to select the anchor points that are best distributed across clusters. Since clusters must be sufficiently populated with A-MVC anchors for the method to work, the number of anchors must be set higher than the number of clusters. In contrast, ISM attribute anchors are found automatically through the process of parsimonization. This process requires the setting of a sparsity parameter to relax the reciprocal of the HHI, which may otherwise lead to excessive sparsity. In the examples considered in this article, this value is experiment independent and is set to 0.8. Further reducing the sparsity parameter risks a lack of overlap between the simplicial cones, potentially rendering the tensor decomposition ineffective. Therefore, until more experience is gained with ISM, we do not recommend changing this parameter.

Just as NMF and NTF factors are more interpretable and meaningful due to the non-negativity of their loadings, ISM produces latent factors whose interpretation is greatly facilitated by the non-negativity and sparsity of the attribute loadings. This is illustrated by the example of the Signature 915 dataset. It is noteworthy that all non-negative approaches result in a high sparsity index of the view-mapping, in contrast to the mixed-sign approaches.

The ISM has only three hyperparameters, which are very few compared to alternative methods: The sparsity coefficient, the embedding dimension, and the rank dimension. As mentioned, the sparsity coefficient should be kept at its default value of 0.8. Regarding the rank and embedding dimensions chosen for the ISM model, an objective and natural choice was the known number of classes for our examples, as we expect each factor to be distinctly assigned to a particular class. The only exception was the UCI digit dataset, where reducing the embedding dimension by one unit significantly reduced the error rate. However, this is only possible in a supervised setting where classes are known. More generally, as with all factorization methods, the factorization rank must be determined in advance.

This raises the issue of the subjectivity of the choice made, especially in an unsupervised setting where cross-validation cannot be used. For PCA, MVMDS, MOFA+, and GFA, setting the rank by inspecting the scree plot of the variance ratio is indeed a subjective choice due to the variety of possible criteria that can be used to identify an "elbow" in the scree plot. We tried a range of values around the "observed" elbow. The observed changes in the close neighborhood metric had no impact on the conclusions about the performance of ISM relative to other approaches (Tables S1 and S2). Since GFA and MOFA+ include automatic rank detection (ARD), increasing the rank should not adversely affect performance, as

it can be automatically reduced if the ARD criteria are met. Notably, for both experiments, increasing the chosen rank decreased performance in terms of cluster association with known classes. This again illustrates the difficulty of choosing the "right" rank. However, non-negative factorization-based methods, including ISM, are not subject to orthogonality constraints and can, therefore, create a new dimension by, for example, splitting a given component into two parts to disentangle close mechanisms that are otherwise intertwined in that component.[40] For this reason, the rank could be set to the number of known classes in a more logical and objective way. Finding the correct rank is, therefore, less critical than with mixed signed factorization approaches such as singular value decomposition (SVD), where low-variance components tend to represent the noisy part of the data. However, multiple solutions have been proposed, among which the cophenetic correlation coefficient is widely used to estimate a rank that provides the most stable clustering derived from the NMF components.[41] A similar criterion, named concordance, has been proposed,[31] where extensive simulations showed that NMF finds the most stable solutions around the correct rank, even if the latent factors are strongly correlated. While such an approach could be used with ISM to determine the best combination for the preliminary embedding and latent space dimensions, it would become too computationally intensive. However, in line with the fact that embedding and latent spaces are later merged in the ISM workflow, it can still be applied in the case where the model imposes the same dimensions for both parameters. As demonstrated in the proof-of-concept analysis of our examples, the embedding dimension can be further optimized by examining the approximation error in the neighborhood of the chosen rank.

Redundancy in the latent factors is a known issue for NMF-based techniques, as identified and illustrated early on with Donoho's swimmer dataset, where a ghost torso appeared in all basis vectors representing body parts in different orientations.[32] L1 regularization techniques, such as using Hoyer's sparsity index[42,43] or appropriate initialization like non-negative SVD (NNSVD),[44] can help mitigate these problems. Notably, in our ISM workflow implementation, the HHI used in the embedding step is mathematically equivalent to Hoyer's sparsity index, and NNSVD is used for NMF and NTF initialization.

ISM's intrinsic view loadings also enable the automatic weighting of views within each latent factor. This allows the simultaneous analysis of views of very different sizes without the need for prior normalization to give each view the same importance, as is necessary with methods like consensus PCA. However, this property reaches its limits when view sizes are extremely unbalanced, as seen in the prokaryotic dataset. In such cases, it is recommended to use ILSM, as ISM is applied to transformed views of equal size, giving equal weight to the original views with the smallest size, whereas global factorization tends to ignore them at initialization. In addition, ILSM requires significantly less computational time due to parallelizable view factorizations.

Recently, graph transformers and deep learning approaches have been proposed for the inference of biological single-cell networks.[45] The preliminary NMF in Unit 1 of Workflow 1, which combines the data before the application of NTF, is somewhat reminiscent of the "attention" mechanism used in transformers before the application of a lightweight neural network.[46] This could explain why ISM can outperform NTF when applied to a multidimensional array, even if the data structure is suitable for the direct application of NTF, as shown by the clustering of marker genes achieved in the Signature 915 dataset example. This also explains why, in the first two examples, although NMF is close to ISM in terms of purity index and other metrics, ISM outperforms NMF in terms of the number of classes detected and, in the second example, by generating a better positioning of the detected cell types on the 2D map projection. Likewise, in the multi-omic single-cell TEA-seq dataset, only ISM identifies and places a naïve cell subtype next to the most biologically relevant one.

Like other latent space methods, ISM is not limited to the purpose of MVC. The ISM components and the view-mapping matrix can be used for data reduction on newly collected data (i.e., data that is not part of the data used to train/learn the model) by fixing these components in the ISM model. Data reduction for newly collected data remains feasible even if some of the views contained in the training data are missing, as the ISM parameters are compartmentalized by views.

The ISM is not limited to views with non-negative data. Each mixed-signed view can be split into its positive part and the absolute value of its negative part, resulting in two different non-negative views, as illustrated in the UCI Digits and prokaryotic data examples.

An important limitation of ISM and other multi-view latent space approaches is the requirement for the availability of multi-view data for all observations in the training set. For financial or logistical reasons, a particular view may be missing in a subset of the observations, and this subset may vary depending on the view under consideration. We are currently developing a variant of ISM that can process multi-view data with missing views.

In this approach, sets of views with enough common observations are integrated with ISM separately. Using the model parameters, the transformation into the latent ISM space can be expanded to all views over all observations in the set, resulting in much larger transformed views than the original intersection would allow. This expansion process enables the integration of the ISM-transformed data from the different view sets, again using the ISM. Interestingly, a similar integrated latent space approach has already been proposed to study the influence of social networks on human behavior.[47] After masking a large number of views, the dataset of UCI Digits dataset was analyzed using this approach. A more detailed description of the expansion process (Workflow S1, Figure S1) and preliminary results (Figure S2) can be found in the Supplementary Materials.

Important issues such as the handling of highly dynamic or rapidly updating datasets have not yet been investigated. This will be addressed in a future article.

It is worth noting that by replacing NMF with NTF in the initialization unit of the ISM workflow, ISM can be easily extended to multi-view data where the views are themselves tensors of order three or higher, provided that all dimensions except the attribute dimension are shared between the views. Interesting applications include the analysis of longitudinal multi-view data or the integration of multiple X-ray views. These topics will be the subject of dedicated articles.

Finally, the extension of ISM to the ILSM approach, as described in the methods section (Section 2), is achieved by a simple chained matrix multiplication – an example of ISM inheriting the simplicity and compactness of the NTF model, made possible by embedding views in a 3D array. This has important advantages:

i. Performance
   - Independent view factorizations can be achieved using parallel computing.
   - The number of attributes in each transformed view is reduced to its factorization rank, allowing ISM to be performed on a much smaller dataset.

ii. Versatility
   - ILSM can be applied to compute NMF on big data in a federated or distributed way. To this end, smaller slices are constructed at random, with each slice considered a particular view that is submitted to ISM. Preliminary results indicate significant performance improvements (Workflow S2 and example in the Supplementary Materials).

While ILSM does not claim to outperform all alternative approaches in every context, this illustrates the scalability and versatility of ILSM, extending far beyond the scope of multi-view data analysis.

5. Conclusion

The proof-of-concept analysis results provide strong preliminary support for the proposed new method. As a next step, we will perform a comprehensive comparison of ISM with state-of-the-art alternative methods, including those considered in this article, and report the findings in a follow-up article.

To further illustrate ISM's key benefits and broad applicability, we will conclude by presenting some potential applications currently under evaluation, with results to be published in future articles.

In longitudinal clinical studies, where participants are followed up later, the ISM model can be trained at baseline and applied to subsequent data to calculate meta-scores. The interpretability of the associated components makes ISM meta-scores more appealing to clinicians compared to the mixed-sign latent factors from other factorization methods.

Consider complex multidimensional multi-omics data from one and the same set of cells (single-cell technology). There is a growing amount of single-cell data corresponding to different molecular layers of the same cell. Data integration is a challenge as each modality can provide a different clustering stemming from a specific biological signal. Therefore, data integration and its projection into a space must: (i) preserve the consensus between two clusterings and (ii) highlight the differences each modality may bring. ISM view loadings can address these two key requirements: components with similar contributions from each molecular layer highlight a consensus that can be inferred from clustering based on the ISM meta-scores of such components. In contrast, components with differing contributions from each molecular layer highlight each modality's specificities, which can be inferred from clustering based on the ISM meta-scores of such components.

The area of spatial mapping, including spatial imaging and spatial transcriptomics, is expanding at an unprecedented pace. An effective method for integrating different levels of information, such as gene or protein expression and spatial organization of cell phenotypes, is an unmet methodological need. We believe that ISM can integrate these different levels of information, as shown in the analysis of the UCI Digits data, to capture the constituents that allow spatial patterns to be distinguished across all levels.

The identification of new chemotypes with biological activity similar to that of a known active

molecule is an important challenge in drug discovery, known as "scaffold hopping."[48] In this context, we are currently analyzing the fingerprints of the docking of 10 of 1000 of molecules to dozens of proteins, with protein-associated fingerprints forming the different views of each molecule. The goal is to use the ISM-transformed fingerprints to predict scaffold-hopping chemotypes. Given the enormous size of the dataset – each fingerprint contains more than 100 binary digits – the ILSM strategy is being evaluated as a possible way to reduce computational problems, as smaller sets of views can be analyzed on smaller subsets of observations before integrating them in their entirety.

Acknowledgments

Our sincere thanks to Prasad Chaskar, Translational Medicine Senior Expert Data Science Lead at Galderma, for stimulating discussions, especially on potential limitations arising from missing views when training latent models with multiple views. We also thank Philippe Pinel from the Center for Computation Biology, Mines Paris/PSL, and Iktos SAS, Paris France, for discussions on addressing ISM calculation challenges in Computational Biology.

Funding

None.

Conflict of interest

Franck Augé and Galina Boldina are employees of Sanofi and may hold shares and/or stock options in the company. All other authors declare no conflicts of interest.

Author contributions

Conceptualization: Paul Fogel, George Luta
Investigation: Franck Augé, Galina Boldina
Writing-original draft: Paul Fogel, Christophe Geissler, Galina Boldina
Writing-review & editing: George Luta, Christophe Geissler, Franck Augé

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data used in this article and the ISM Jupyter Python notebook can be downloaded from the Advestis part of Mazars GitHub repository (https://github.com/Advestis/adilsm).

Further disclosure

The paper has been uploaded to or deposited in a preprint server (https://www.preprints.org/manuscript/202402.1001/v3).

References

1. Cichocki A, Zdunek R, Phan AH, Amari S. *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons; 2009.

   doi: 10.1002/9780470747278

2. Perry R, Mischler G, Guo R, *et al*. mvlearn: Multiview Machine Learning in Python. *arXiv*. Preprint posted online 2020.

   doi: 10.48550/arXiv.2005.11890

3. Argelaguet R, Velten B, Arnol D, *et al*. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14(6):e8124.

   doi: 10.15252/msb.20178124

4. Argelaguet R, Arnol D, Bredikhin D, *et al*. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 2020;21(1):111.

   doi: 10.1186/s13059-020-02015-1

5. Wu J, Lin Z, Zha H. Essential tensor learning for multi-view spectral clustering. *IEEE Trans Image Process*. 2019;28(12):5910-5922.

   doi: 10.1109/tip.2019.2916740

6. Guo W, Che H, Leung MF. Tensor-based adaptive consensus graph learning for multi-view clustering. *IEEE Trans Consumer Electron*. 2024;70(2):4767-4784.

   doi: 10.1109/tce.2024.3376397

7. Li J, Gao Q, Wang Q, Xia W, Gao X. Multi-View Clustering via Semi-non-negative Tensor Factorization. *arXiv*. Preprint posted online 2023.

   doi: 10.48550/arXiv.2303.16748

8. Wang S, Cao J, Lei F, Jiang J, Dai Q, Ling BW. Multiple kernel-based anchor graph coupled low-rank tensor learning for incomplete multi-view clustering. *Appl Intell*. 2022;53(4):3687-3712.

   doi: 10.1007/s10489-022-03735-6

9. Zhao W, Gao Q, Li G, Deng C, Yang M. One-Step Multi-View Clustering Based on Transition Probability. *arXiv*. Preprint posted online 2024.

   doi: 10.48550/arXiv.2403.01460

10. Ali W, Yang M, Ali M, Ud-Din S. Fuzzy model-based sparse clustering with multivariate t-mixtures. *Appl Artif Intell*. 2023;37(1):2169299.

    doi: 10.1080/08839514.2023.2169299

11. Yang M, Hussain I. Unsupervised multi-view k-means

clustering algorithm. *IEEE Access*. 2023;11:13574-13593.

doi: 10.1109/access.2023.3243133

12. Hussain I, Sinaga KP, Yang M. Unsupervised multiview fuzzy C-means clustering algorithm. *Electronics*. 2023;12(21):4467-4467.

doi: 10.3390/electronics12214467

13. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J Chemometr*. 2003;17(6):323-337.

doi: 10.1002/cem.811

14. Trendafilov NT. Stepwise estimation of common principal components. Comput Stat Data Anal. 2010;54(12):3446-3457.

doi: 10.1016/j.csda.2010.03.010

15. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. Eur J Oper Res. 2014;238(2):391-403.

doi: 10.1016/j.ejor.2014.01.008

16. Zhang C, Hu Q, Fu H, Zhu PF, Cao X. Latent Multi-View Subspace Clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*); 2017:4333-4341.

doi: 10.1109/cvpr.2017.461

17. Chen M, Huang L, Wang C, Huang D. Multi-view clustering in latent embedding space. *Proc AAAI Conf Artif Intell*. 2020;34(4):3513-3520.

doi: 10.1609/aaai.v34i04.5756

18. Leppäaho E, Ammad-ud-din M, Kaski S. GFA: Exploratory analysis of multiple data sources with group factor analysis. *J Mach Learn Res.* 2017;18(39):1-5.

19. Zhao S, Gao C, Mukherjee S, Engelhardt BE. Bayesian group factor analysis with structured sparsity. *J Mach Learn Res.* 2016;17(196):1−47.

20. Zhang X, Zhao L, Zong L, Liu X, Yu H. Multi-view Clustering via Multi-Manifold Regularized Nonnegative Matrix Factorization. In: *IEEE International Conference on Data Mining*; 2014:1103-1108.

doi: 10.1109/icdm.2014.19

21. Huizing G, Deutschmann IM, Peyré G, Cantini L. Paired single-cell multi-omics data integration with Mowgli. *Nat Commun*. 2023;14(1):7711.

doi: 10.1038/s41467-023-43019-2

22. Brbic M, Kopriva I. Multi-view low-rank sparse subspace clustering. *Pattern Recognit*. 2018;73:247-258.

doi: 10.1016/j.patcog.2017.08.024

23. Dong Y, Che H, Leung MF, Liu C, Yan Z. Centric graph regularized log-norm sparse non-negative matrix factorization for multi-view clustering. *Signal Process.* 2024;217:109341.

doi: 10.1016/j.sigpro.2023.109341

24. Fu L, Lin P, Vasilakos AV, Wang S. An overview of recent multi-view clustering. *Neurocomputing*. 2020;402:148-161.

doi: 10.1016/j.neucom.2020.02.104

25. Duin R. Multiple Features. UC Irvine Machine Learning Repository; 1998.

doi: 10.24432/C5HC70

26. Boldina G, Fogel P, Rocher C, Bettembourg C, Luta G, Augé F. A2Sign: Agnostic algorithms for signatures-a universal method for identifying molecular signatures from transcriptomic datasets prior to cell-type deconvolution. *Bioinformatics*. 2021;38(4):1015-1021.

doi: 10.1093/bioinformatics/btab773

27. Lewis DD, Yang Y, Rose TG, Li F. RCV1: A new benchmark collection for text categorization research. *J Mach Learn Res.* 2004;5:361-397.

28. Brbic M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Res*. 2016;44:10074-10090.

doi: 10.1093/nar/gkw964

29. Swanson E, Lord C, Reading J, *et al*. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife*. 2021;10:e63632.

doi: 10.7554/eLife.63632

30. Hirschman AO. The paternity of an index. *Am Econ Rev.* 1964;54(5):761-762.

31. Fogel P, Geissler C, Morizet N, Luta G. On rank selection in non-negative matrix factorization using concordance. *Mathematics*. 2023;11(22):4611.

doi: 10.3390/math11224611

32. Badeau R, Bertin N, Vincent E. Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization. *IEEE Trans Neural Netw*. 2010;21(12):1869-1881.

doi: 10.1109/tnn.2010.2076831

33. Donoho DL, Stodden VC. *When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?* Columbia University. 2004.

doi: 10.7916/D88D05N7

34. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2(1):193-218.

doi: 10.1007/BF01908075

35. Strehl A, Ghosh J. Cluster ensembles-A knowledge reuse framework for combining multiple partitions. *J Mach Learn Res.* 2002;3:583-617.

doi: 10.1162/153244303321897735

36. Fowlkes EB, Mallows CL. A method for comparing two

hierarchical clusterings: Rejoinder. *J Am Stat Assoc*. 1983;78(383):584.

doi: 10.2307/2288123

37. Demaine E, Hesterberg A, Koehler F, Lynch J, Urschel J. Multidimensional Scaling: Approximation and Complexity. *arXiv*. Preprint posted online 2021.

doi: 10.48550/arXiv.2109.11505

38. Zhai Z, Lei YL, Wang R, Xie Y. Supervised capacity preserving mapping: A clustering guided visualization method for scRNA-seq Data. *Bioinformatics*. 2022;38(9):2496-2503.

doi: 10.1093/bioinformatics/btac131

39. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine Learning in Python. *arXiv*. Preprint posted online 2012.

doi: 10.48550/arXiv.1201.0490

40. Fogel P, Hawkins DM, Beecher C, Luta G, Young SS. A tale of two matrix factorizations. *Am Stat*. 2013;67(4):207-218.

doi: 10.1080/00031305.2013.845607

41. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164-4169.

doi: 10.1073/pnas.0308531101

42. Hoyer PO. Non-negative matrix factorization with sparseness constraints. *arXiv*. Preprint posted online 2004.

doi: 10.48550/arXiv.CS/0408058

43. Potluru VK, Plis S, Le Roux J, Pearlmutter BA, Calhoun VD, Hayes TP. Block Coordinate Descent for Sparse NMF. International Conference on Learning Representations (ICLR); 2013.

44. Boutsidis C, Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit*. 2008;41(4):1350-1362.

doi: 10.1016/j.patcog.2007.09.010

45. Ma A, Wang X, Li J, *et al*. Single-cell biological network inference using a heterogeneous graph transformer. *Nat Commun*. 2023;14(1):964.

doi: 10.1038/s41467-023-36559-0

46. Vaswani A, Shazeer NM, Parmar N, *et al*. Attention is all you need. In: NIPS'17: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017:6000-6010.

47. Park J, Jin IH, Jeon M. How social networks influence human behavior: An integrated latent space approach for differential social influence. *Psychometrika*. 2023;88:1529-1555.

doi: 10.1007/s11336-023-09934-5

48. Pinel P, Guichaoua G, Najm M, *et al*. Exploring isofunctional molecules: Design of a benchmark and evaluation of prediction performance. *Mol Inform*. 2023;42(4):e2200216.

doi: 10.1002/minf.202200216