Artificial Intelligence In Medicine 160 (2025) 103049

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Intrinsic-dimension analysis for guiding dimensionality reduction and data fusion in multi-omics data processing

Jessica Gliozzo [a,b,1], Mauricio Soto-Gomez [a,1], Valentina Guarino [a], Arturo Bonometti [c,d], Alberto Cabri [a], Emanuele Cavalleri [a], Justin Reese [e], Peter N. Robinson [f], Marco Mesiti [a,e], Giorgio Valentini [a,g], Elena Casiraghi [a,e,g,h,*]

[a] AnacletoLab, Computer Science Department, Università degli Studi di Milano, Milan, Italy
[b] European Commission, Joint Research Centre (JRC), Ispra, Italy
[c] Department of Biomedical Sciences, Humanitas University, Milan, Italy
[d] Department of Pathology, IRCCS Humanitas Clinical and Research Hospital, Milan, Italy
[e] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[f] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
[g] CINI, Infolife National Laboratory, Roma, Italy
[h] Department of Computer Science, Aalto University, Espoo, Finland

ARTICLE INFO

Keywords:
Dimensionality reduction
Intrinsic dimensionality
Feature selection
Feature extraction
Data fusion
Multi-omics datasets

ABSTRACT

Multi-omics data have revolutionized biomedical research by providing a comprehensive understanding of biological systems and the molecular mechanisms of disease development. However, analyzing multi-omics data is challenging due to high dimensionality and limited sample sizes, necessitating proper data-reduction pipelines to ensure reliable analyses. Additionally, its multimodal nature requires effective data-integration pipelines.

While several dimensionality reduction and data fusion algorithms have been proposed, crucial aspects are often overlooked. Specifically, the choice of projection space dimension is typically heuristic and uniformly applied across all omics, neglecting the unique high dimension small sample size challenges faced by individual omics.

This paper introduces a novel multi-modal dimensionality reduction pipeline tailored to individual views. By leveraging intrinsic dimensionality estimators, we assess the curse-of-dimensionality impact on each view and propose a two-step reduction strategy for significantly affected views, combining feature selection with feature extraction.

Compared to traditional uniform reduction pipelines in a crucial and supervised multi-omics analysis setting, our approach shows significant improvement. Additionally, we explore three effective unsupervised multi-omics data fusion methods rooted in the main data fusion strategies to gain insights into their performance under crucial, yet overlooked, settings.

1. Introduction

In the biomedical research field, high-throughput technologies allow acquiring vast and diverse omics data types such as genomic, transcriptomic, proteomic, and methylomic data [1,2]. These distinct modalities (views) provide valuable insights into the intricate molecular landscape governing biological processes and diseases; if appropriately processed and integrated they can uncover crucial disease triggers and enhance our understanding of various health conditions [3–6].

However, the analysis of multi-omics data presents significant challenges due to their High-Dimensional Low Sample Size (HDLSS) characteristic and multi-modality. In particular, the HDLSS nature of omics data results in high computational costs, data sparsity, and overfitting due to the presence of noisy, uninformative, and redundant features [7]. These problems, collectively referred to as the "curse of dimensionality", can bias practically all results obtained from these data [8–10].

* Correspondence to: AnacletoLab, Department of Computer Science "Giovanni degli Antoni", Via Celoria 18, 20133 Milan, Italy.
E-mail address: elena.casiraghi@unimi.it (E. Casiraghi).
[1] JG and MSG equally contributed to the work

https://doi.org/10.1016/j.artmed.2024.103049
Received 24 October 2023; Received in revised form 3 December 2024; Accepted 4 December 2024
Available online 11 December 2024

To address these issues, unsupervised Dimensionality Reduction (DR) gained a lot of interest over the past decade and is now recognized as being a crucial preliminary phase in various fields [11]. DR techniques, including feature selection and feature extraction methods, mitigate the curse of dimensionality by reducing the dimension of the input dataset so that it concisely conveys similar information. In case of bio-medical multi-modal datasets, DR may be individually applied to reduce each input modality (view) and better expose its characterizing informative content. This would aid the subsequent data fusion task, for which several promising algorithms have been presented [12].

While feature selection and feature extraction methods have shown their own advantages and several reviews describe and eventually compare their successful results [13], to the best of our knowledge no paper investigated the following two crucial choices when treating multimodal HDLSS data, where each view is individually impacted by the curse of dimensionality. First, there is no rule of thumb to decide whether and how much the available HDLSS data is impacted by the curse of dimensionality. Second, the choice of the dimension for reduced space is challenging; too low a dimension causes information loss, while too high a dimension may retain noise and redundancies, failing to solve the problem. In practice, literature works in the field of bioinformatics either avoid any dimensionality reduction [14] or make some empirical/heuristic decisions [15,16] not motivated by any theoretical justification. Other interesting works specifically designed for HDLSS data reduction, evaluate different values [7]. However, the careful design of the DR step affects all the subsequent computations and the reliability of the obtained results [17–19].

Additionally, few studies in the multi-omics context account for the fact that different views may carry varying amounts of information and should therefore be treated differently to better highlight and expose the information within each view, thereby facilitating and potentially improving the effectiveness of the subsequent data fusion task. Instead, they either neglect view reduction prior to data fusion or uniformly reduce each view using often unmotivated heuristic decisions.

The main novelty of this work is a block-analysis technique that leverages any of the most promising and recent Intrinsic Dimensionality (`id`) estimators (supplementary section S.A.1) and exploits it to tailor reduction of each view in a multi-modal datasets, by determining when and how a DR could improve representation of individual views. If curse of dimensionality is detected, block-analysis allows defining a robust `id` estimate to be used as the dimension of the lower-dimensional space where the view should be transformed using any of the promising feature selection or feature extraction approaches proposed in literature (supplementary section S.A.2). By exploiting the information provided by block-analysis, we further propose reducing views more impacted by HDLSS by a novel two-step dimensionality reduction process that combines the advantages of initial feature selection followed by feature extraction.

Considering the paramount importance of accurate prognosis for cancer patients [12,20–23], we assessed the proposed DR technique in the context of survival prediction on multi-omics data analysis and showcased experiments where three representative and effective multi-omics data integration algorithms are assessed: MOFA+ [24], an input data fusion technique rooted in the stochastic variational inference field; uMKL [25], a kernel-fusion approach employing unsupervised multiple kernel learning; and SNF [14], a kernel-fusion approach exploiting a diffusion process to fuse similarity graphs (see supplementary section S.A.3). By using them we investigated the effect of the following multi-omics data fusion settings that are often overlooked: (1) using and integrating subsets of the available multi-omics views; (2) adding not-omics patients' views, e.g. patients' (demographics) views carrying completely different semantics.

Results on nine varied multi-omics datasets from the well-known TCGA repository (Section 2) show that, in the considered predictive task, the proposed DR approach improves prediction performance; the robustness of prediction is further improved when all the available omics and non-omics views are considered.

The key contributions and novelties of our DR framework are:

1. *Unbiased `id` Estimation.* We leverage block-analysis to utilize state-of-the-art `id` estimators, ensuring unbiased `id` estimates for real-world datasets. Unlike existing techniques, which are often biased by small-sample-size issues, noise, outliers, or boundary points, our approach directly addresses these challenges to provide reliable `id` estimation.
2. *Principled dimensionality reduction.* In the literature, DR is often:
   - *Neglected*, leading to biases caused by small-sample-size problems.
   - *Performed heuristically*, relying on unmotivated or empirical decisions often omitted in publications.
   - *Applied uniformly across all views in a multi-modal dataset*, disregarding the unique characteristics of each view, such as its dimensionality, complexity, or noise.

   In contrast, our methodology provides a principled and reproducible approach to determining the reduced space dimensionality. By tailoring the reduction to better characterize each view, it ensures robust and reliable results.
3. *View-specific and principled DR.* As outlined in point 2, the proposed block-analysis technique evaluates each view individually to determine its susceptibility to the curse of dimensionality. Unlike existing works, which either apply or omit DR based on unmotivated choices, our framework provides a principled approach to decide whether DR should be applied and, if so, to what extent. For views significantly impacted by the curse of dimensionality, we introduce a novel two-step DR process that combines feature selection and feature extraction. This approach is uniquely guided by insights derived from block-analysis, ensuring an effective and tailored reduction strategy.
4. *Robust and exhaustive evaluation with insights into multi-omics data integration.* We rigorously evaluated the proposed DR pipeline using nine diverse multi-omics cancer datasets from the publicly available TCGA repository. Multi-omics approaches have gained significant attention in recent years, particularly in medicine, due to their potential to uncover complex biological mechanisms and improve disease understanding. In addition to assessing the effectiveness of our DR approach and comparing it to state-of-the-art methods that determine dimensionality based on heuristics or empirical choices, our experiments provided an opportunity to explore critical and often overlooked challenges in multi-omics data integration. These challenges include integrating subsets of omics data and incorporating demographic variables to enhance analysis.
5. *Enhancing data fusion and predictive performance:* We showed that the proposed DR pipeline significantly improves the predictive performance of state-of-the-art data fusion methods. Applied to the critical binary task of predicting overall survival in cancer patients, our approach leverages random forest classifiers – a widely trusted and interpretable model in biomedical research.

2. TCGA datasets

Cancer remains the leading cause of death worldwide, which is the rationale behind our choice of assessing our proposal on cancer data. To obtain reliable results, we mined the following nine multi-omics datasets from the TCGA cancer repository[2] (see Tables 1 and 2): the BLadder urothelial Carcinoma dataset (**BLCA**); the BReast infiltrating ductal CArcinoma (**BRCA1**) and the BReast infiltrating lobular CArcinoma (**BRCA2**) datasets, composed by splitting all the samples in the BReast nvasive CArcinoma dataset (BRCA); the KIdney Renal Clear cell

[2] The R package ''curatedTCGAData'' [26] was used to download the tumor datasets from the TCGA repository (dataset version 2.0.1).

Table 1
Descriptive statistics for BLCA, BRCA1, BRCA2, KIRC, and LUAD datasets. Column N reports the number of cases; column D (raw) reports the original dimension of each view; column D reports the dimension of each view after data pre-filtering to remove noise and high pairwise-redundancy (see supplementary section S.B); column $\frac{N}{D}$ reports the ratio between the number of cases and the dimension of each view; columns N_{neg}, N_{pos}, and $\frac{N_{pos}}{N}$ report, respectively, the number of negative (OS = 0) and positive (OS = 1) patients, and the balance ratio, measured as the ratio between the number of positive cases and all the cases in the dataset. "methy" stands for DNA methylation data.

Dataset	View	N	D (raw)	D	$\frac{N}{D}$	N_{pos}	N_{neg}	$\frac{N_{pos}}{N}$
BLCA	miRNA	335	469	469	0.7143	151	184	0.45
	mRNA		12 276	12 276	0.027			
	Proteins		183	183	1.831			
	methy		315 551	30 000	0.012			
BRCA1	miRNA	317	496	496	0.6391	42	275	0.13
	mRNA		12 242	12 242	0.026			
	Proteins		202	202	1.569			
	methy		289 962	30 000	0.011			
BRCA2	miRNA	128	502	502	0.255	14	114	0.11
	mRNA		8128	8128	0.016			
	Proteins		192	192	0.667			
	methy		278 099	30 000	0.004			
KIRC	miRNA	169	364	364	0.464	48	121	0.28
	mRNA		7942	7942	0.021			
	Proteins		186	186	0.909			
	methy		319 740	30 000	0.006			
LUAD	miRNA	300	465	465	0.645	120	180	0.40
	mRNA		11 131	11 131	0.027			
	Proteins		179	179	1.676			
	methy		331 828	30 000	0.01			

carcinoma dataset (**KIRC**); the LUng ADenocarcinoma dataset (**LUAD**); the LUng Squamous Cell carcinoma dataset (**LUSC**); the PRostate ADenocarcinoma dataset (**PRAD**); the OVarian serous cystadenocarcinoma dataset (**OV**); the SKin Cutaneous Melanoma dataset (**SKCM**).

For each dataset, we considered miRNA and mRNA (RNA-Sequencing expression values), protein expression (Reverse Phase Protein Arrays), and DNA methylation (Methylation Array) views, which were pre-processed to filter variables mainly carrying noise or highly redundant information (see supplementary section S.B for further details).

We also complemented the omics information with demographic patient data (age at first pathological diagnosis, gender, race, ethnicity, see supplementary tables S.1-S.3 for further details). Patients in the TCGA dataset may be classified based on their Overall Survival (OS) event, a binary label available from the TCGA-CDR [27] dataset.

Accurate prediction of overall survival is a crucial and challenging task in oncological research, as extensively documented in the literature [21–23]. We therefore chose to assess our proposal by performing a supervised classification task.

Table 2
Descriptive statistics for LUSC, OV, PRAD, and SKCM datasets.

Dataset	View	N	D (raw)	D	$\frac{N}{D}$	N_{pos}	N_{neg}	$\frac{N_{pos}}{N}$
LUSC	miRNA	228	491	491	0.464	93	135	0.41
	mRNA		11 473	11 473	0.02			
	Proteins		178	178	1.281			
	methy		273 884	30 000	0.008			
OV	miRNA	226	308	308	0.734	143	83	0.63
	mRNA		11 731	11 731	0.019			
	Proteins		186	186	1.215			
	methy		13 296	13 296	0.017			
PRAD	miRNA	337	457	457	0.737	6	331	0.02
	mRNA		8887	8887	0.038			
	Proteins		169	169	1.994			
	methy		301 920	30 000	0.011			
SKCM	miRNA	334	523	523	0.639	150	184	0.45
	mRNA		13 050	13 050	0.026			
	Proteins		186	186	1.796			
	methy		311 405	30 000	0.011			

3. Dimensionality reduction approach

In this section, we describe the block-analysis we propose to, first, provide unbiased estimates of the `id` of a data-view (Section 3.1) and, second, automatically assess the amount of feature noise and redundancy affecting the view (Section 3.2).

The collected information allows building a DR tailored to the HDLSS characteristics specific to each view. To guide the reader, Fig. 1 sketches the DR pipeline guided by block-analysis. In the whole section, we consider an input view (dataset), $\mathbf{X} \in \Re^{N \times D}$, with N being the number of cases, and D the number of features (dimension) of the view.

3.1. Block analysis and block-ID estimate

Several of the most promising `id` estimators, which use a Nearest-Neighbor estimation approach, produce unstable global estimates on real HDLSS data due to sample-sparsity, outlier points, and noise [28–30] (see supplementary section S.A.1 for further details).

3.1.1. An undersampling strategy to reduce the variance and bias due to noisy and outlier points in real datasets

Nearest-neighbor `id` estimators base their estimation on the analysis of the distribution of points withing small data-neighborhoods. Due to the unreliability of pairwise-distances in datasets characterized by the small-sample-size, these estimators often suffer from high variance or overestimation when, e.g., the considered point-neighborhood size increases. Furthermore, since all the `id` estimators contain some randomness, most of them suffer from an added factor of variance, particularly evident when working in high dimensions.

To account for such variance as well as the presence of outlier and boundary points that could bias the estimates, authors of *two-nn* [30] proposed experiments on simulated datasets (with a large number of samples, i.e. not affected by the small-sample-size) where they apply a classic block-analysis [31]. In particular, they compute (sub-optimal) `id` estimates (and their standard deviation) by averaging the estimates obtained on under-sampled, non-intersecting datasets composed of a number $n < N$ of samples. By plotting the distribution of the obtained estimates for increasing values of n, a plateau is found, corresponding

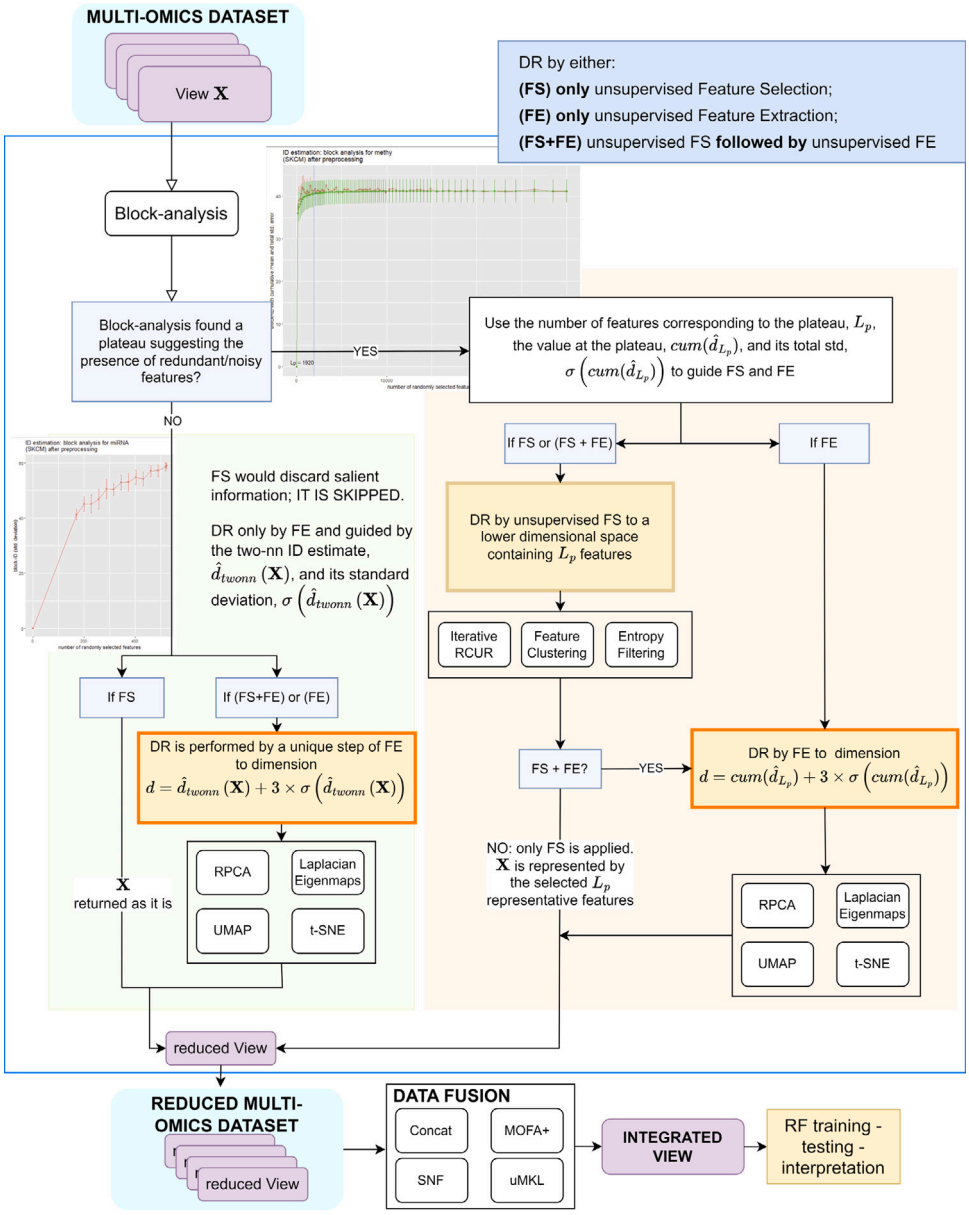

Fig. 1. Experimented DR and data-integration pipelines.

to an unbiased (optimal) estimate of the `id` characterizing the dataset's informative content.

This approach is effective on simulated experiments, where enough samples can be generated to avoid the curse of dimensionality. However, when dealing with real bio-medical datasets, often limited in sample-size and potentially affected by the curse of dimensionality, such a strategy cannot be applied. To reduce the bias due to the presence of noisy and outlier points we propose averaging all the *two-nn* `id` estimates computed on M under-sampled versions of the dataset, where the under-sampling randomly selects (with repetition) a fixed percentage, t, of the dataset points.[3] Choosing a proper value for the percentage t ensures enough samples in each of the M sub-dataset, so that the average (and the standard deviation) of all the M `id`-estimates provides a first, more robust, *two-nn* `id`-estimate (and standard deviation of the estimate) of the input dataset. In the following, any reference to the *two-nn* `id`-estimate of a dataset $\mathbf{X}$, $\hat{d}_{twonn}(\mathbf{X})$ (and its standard deviation $\sigma\left(\hat{d}_{twonn}(\mathbf{X})\right)$) refers to this unbiased estimate.

Note that the undersampling strategy we propose is general enough to be applied for leveraging the estimates produced on HDLSS data by any state-of-the-art `id`-estimator.

3.1.2. Block-ID estimate

While the aforementioned procedure mitigates the problems affecting real, noisy datasets, it still cannot cope with the possible curse of dimensionality, which practically shows up with a large number of features being noisy or redundant. Unfortunately, given an input view $\mathbf{X} \in \Re^{N \times D}$ there is no rule of thumb for deciding when a dataset characterized by low values of the ratio $\frac{N}{D}$ is affected by the curse of dimensionality. To provide such understanding and to obtain an unbiased `id`-estimate of the view even in the presence of noisy and redundant features we propose applying the block-analysis feature-wise, as detailed in this section.

[3] In all our experiments we set $M = 11$ and $t = 90\%$. The low value of M limits the computational-time costs of the algorithm; however, the higher this value, the lower the variability of the estimate and the higher the precision of the estimate. The value of $t = 90\%$ is chosen to obtain under-sampled datasets with enough samples.

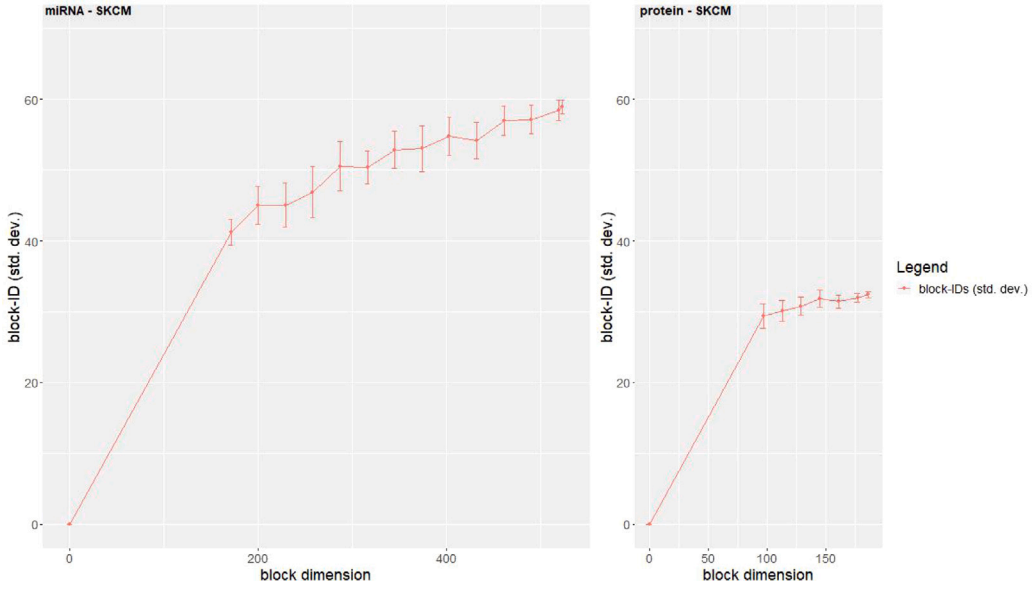

Fig. 2. Block-analysis performed by using the *two-nn* estimator on the SKCM dataset. Left: miRNA view (SKCM dataset). Right: protein view (SKCM dataset). The X axis reports the block dimension L_j; the Y axis the estimated `id` averaged across the n_{try} blocks $\mathbf{B}_j$ with dimension L_j. Bars represent the estimated standard deviation, $\sigma\left(\hat{d}_{L_j}\right)$, for the n_{try} blocks $\mathbf{B}_j$. The block-`id` increases as the block dimension increases, suggesting that each added features increases the information content. Therefore, the `id` of the whole view, i.e. the `id` of the block covering all the features, is a reliable estimate of the dimensionality of the space where the data should be transformed by a feature extraction algorithm (Fig. 1 - light green box - FE option). On the other hand, considering that each feature adds novel information, if feature selection is the chosen dimensionality reduction approach (Fig. 1 - light green box - FS option), the view is not reduced and it is returned as it is; in other words, feature selection is avoided because it would necessarily spare information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In particular, we start by using the *two-nn* `id`-estimate (Section 3.1.1) for the input view, $\hat{d}_{twonn}(\mathbf{X})$, to set the dimension L_0 of the smallest block as $L_0 = c \times \hat{d}_{twonn}(\mathbf{X})$, being c a user-set constant we set at $c = 3$. Though we are aware that the `id` estimate might still be biased by redundant and noisy features, if any, it can be a valid aid to guarantee that even smaller blocks can contain enough information to produce reliable estimates.

Once L_0 is set, we perform the block-analysis by iterating over blocks with increasing dimensions, estimating the *two-nn* `id` of each block, and then analyzing the distribution of all the block-`ids`.

More precisely, at the jth iteration (jth block $\mathbf{B}_j$), when the block size is $L_j = c_0 \times L_0 + j \times L_0$, $L_j \leq D$, being c_0 a user-set constant we set to $c_0 = 3$, we estimate the `id` (and its fluctuations) for $\mathbf{B}_j$ by:

(I) creating n_{try} blocks, $\mathbf{B}_j(i) \in \Re^{N \times L_j}$, $i \in [1, \ldots, n_{try}]$, each representing all the samples in the input view with L_j randomly sampled features;
(II) estimating the *two-nn* `id` of each $\mathbf{B}_j(i)$, $\hat{d}_{twonn}(\mathbf{B}_j(i))$ and then computing the mean (and variance) of all the computed estimates to obtain the block-`id` estimate, $\hat{d}_{L_j}$ (and its variance, $var\left(\hat{d}_{L_j}\right)$) for $\mathbf{B}_j$, being var the variance operator.

This step essentially provides an estimate of the `id` (and its variance) that would be obtained if the data was represented by L_j randomly selected features.[4]

3.2. Automatic analysis of block-ids allows tailoring dimensionality reduction

The red dotted lines in Fig. 2 (and supplementary figures S.1-S.4) plot the block-`ids`, $\hat{d}_{L_j}$ for increasing block dimensions in the miRNA and protein views (SKCM dataset); red bars in the figure represent standard deviations of the block-`ids` $\sigma\left(\hat{d}_{L_j}\right)$.[5]

The miRNA and protein views are characterized by higher ratios $\frac{N}{D}$ when compared to the mRNA and methylation data-views; in this case, we note that the block-`id` keeps increasing until the block size includes all the features in the view.

This suggests that each new feature adds novel information. In other words, the view contains a limited amount of noise and redundancy, supposedly due to the data-view belonging to a real dataset. In this case, the *two-nn*-`id` of the whole view, $\hat{d}_{twonn}(\mathbf{X})$ (and its standard deviation, $\sigma\left(\hat{d}_{twonn}(\mathbf{X})\right)$) is an unbiased estimate of the `id` of the whole view (and its fluctuations).

This bring to the following decision in the tailored reduction of each view: *when no plateau is automatically detected by block-analysis* (Fig. 1 - light green box) no DR via feature selection is applied because the selection of a subset of features would surely cause loss of information. Instead, we allow performing DR via feature extraction, to combine the information carried by all the features and obtain a compressed but still informative representation to a space with dimension $d = \hat{d}_{twonn}(\mathbf{X}) + 3\sigma\left(\hat{d}_{twonn}(\mathbf{X})\right)$.

On the other hand, for the mRNA and the methylation view (Fig. 3) the distribution of the block-`ids` (red-dotted line) is more noisy, and increases until it reaches a plateau. To reduce noise effects by averaging, we compute the cumulative mean of the block-`ids` (green-dotted line in Fig. 3).

More precisely, the cumulative mean for block $\mathbf{B}_j$, $cum(\hat{d}_{L_j})$, is computed as the average of all the block-`ids` computed for blocks $\mathbf{B}_0, \ldots, \mathbf{B}_j$: $cum(\hat{d}_{L_j}) = mean\left(\hat{d}_{L_t}\right)$, $t = [1, \ldots, j]$ Eve's law of total variance [32] allows computing the total variance of $cum(\hat{d}_{L_j})$, $var\left(cum(\hat{d}_{L_j})\right)$, as the

[4] We set $n_{try} = 31$ to reduce time costs of the algorithm; however, the higher this value, the higher the precision of the estimate.

[5] The standard deviations of each block-`id` is computed as the square root of the variance $\sigma\left(\hat{d}_{L_j}\right) = \sqrt{var\left(\hat{d}_{L_j}\right)}$.

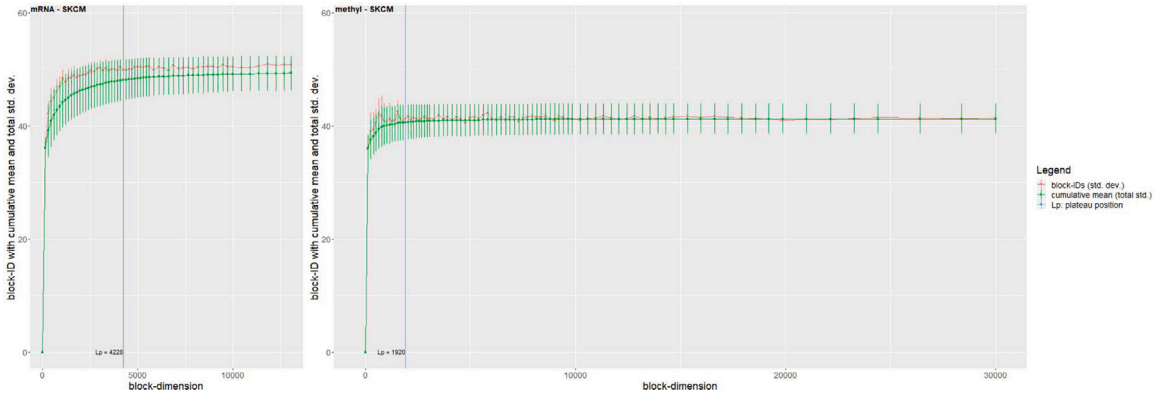

Fig. 3. Block-analysis performed by using the *two-nn* estimator on the SKCM dataset. Left: mRNA; right: methylation data. The block-ids (red-dotted line) are more noisy than those in the miRNA and protein views. The effect of noise is reduced by the computation of the cumulative mean (green-dotted line), soon reaching stability (plateau), providing a reliable estimate of the view id. The analysis of the cumulative mean allows the automatic detection of the position of the plateau ($L_p = 4228$ and $L_p = 1920$ for, respectively, the mRNA and the methylation view), corresponding to the number of features that may be selected from the dataset to reduce the information loss. In other words, the block-analysis of the mRNA and methylation views allows to detect signs of the curse of dimensionality in terms of the presence of feature redundancy. Moreover, it provides an unbiased estimate of the id characterizing the information content of view, and an estimate of the number of features that could be retained by any unsupervised feature selection algorithm to avoid information loss. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sum of the (unexplained) variance UV_{L_j} due to the id-estimator, and the (explained) variance EV_{L_j} due to the sampling process.[6]

In practice, each point of the cumulative mean represents the average block-id that would be obtained on a view composed by randomly sampling a number of features that is equal to (or lower than) the dimension of block $\mathbf{B}_j$. Assuming some features are mostly carrying noise and/or redundant information, the random under-sampling of features that is performed to compose blocks with varied and increasing dimensions, as well as the evaluation of the id for increasing block dimensions, is able to reduce (by averaging) biasing effects due to noise and redundancy. This is also visible in the plot of the cumulative mean, which approaches the block-id (red-dotted line) plot and is less noisy. This suggests that the (cumulative) value corresponding to the plateau of the block-id, $cum(\hat{d}_{L_p})$, and its total standard deviation, $\sigma\left(cum(\hat{d}_{L_p})\right) = \sqrt{UV_{L_p} + EV_{L_p}}$, can be considered as an unbiased estimate of the id (and its fluctuations) of the whole view.

Note that the dimension L_p of the block where the plateau starts (horizontal-axis in Fig. 3, automatically detected as described in supplementary section S.C) can be regarded as an estimate of the minimum number of features that can be used to represent the salient information in the data-view, and after which the addition of extra features mainly adds redundancy and/or noise.

This allows devising a **two-step** DR approach for views impacted by the curse of dimensionality. It first applies an unsupervised feature selection approach to select the most representative L_p features in the dataset, therefore removing noise and redundancy. Since L_p is often a large value, to further reduce the computational costs and compute a data representation concisely conveying similar information, the reduced L_p-dimensional view is next input to an unsupervised feature extraction algorithm that transforms it into a space with dimension $d = cum(\hat{d}_{L_p}) + 3\sigma\left(cum(\hat{d}_{L_p})\right)$.

Note that most feature extraction techniques are based on the computation of pairwise sample distances, which are biased under the curse of dimensionality due to the high level of sample-sparsity. Besides the reduction of computational costs, the prior application of a feature selection algorithm reducing the amount of redundancy and noise facilitates the task of the following feature extraction algorithm by allowing the computation of more reliable pairwise sample-distances.

This led to the following decision in the tailored reduction of each view: *when a plateau is automatically detected in position L_p of the block-id distribution* (see Fig. 1 - light orange box) we reduce the curse of dimensionality by applying any of the following three DR options: (1) if feature selection is the preferred approach, we select L_p salient features; (2) if feature extraction is the preferred DR approach, the data-view is transformed to a lower dimensional space with dimension $d = cum(\hat{d}_{L_p}) + 3\sigma\left(cum(\hat{d}_{L_p})\right)$; (3) if a two phase DR is chosen, feature selection is applied to select L_p features and the reduced view is input to a feature extraction algorithm that transforms the dataset into a space with dimension $d = cum(\hat{d}_{L_p}) + 3\sigma\left(cum(\hat{d}_{L_p})\right)$.

Tables 3 and 4 report, for each dataset and view used in our experiments (Section 2), the number of features L_p corresponding to the plateau, if any is found, the id estimate $\hat{d}$, which equals either $\hat{d}_{twonn}(\mathbf{X})$ - when no plateau is found - or $cum(\hat{d}_{L_p})$ - when a plateau is found, and its total standard deviation, $\sigma\left(\hat{d}\right)$, computed as the square root of the total variance, $var\left(cum(\hat{d}_{L_p})\right)$.

4. Results

To compare our DR approach with methods commonly used in the literature (such as uniformly applying the same DR algorithm across all views of a multi-modal dataset and defining the latent space dimension based on heuristics), we applied it to nine clinically relevant binary prediction tasks (prediction of overall survival using TCGA datasets). Specifically, we applied either our DR approach or common heuristic methods from the literature, obtained reduced views, which were then integrated by four different data-fusion algorithms, and we trained random forest classifiers for the chosen binary prediction task. We measured the difference in prediction performance between our DR approach and the methods commonly used in the literature.

In this section we first summarize the DR+data fusion pipelines we devised and experimented (Section 4.1); next, we detail the experimental settings of the supervised classification task we exploited to objectively compare the different pipelines (Section 4.2); finally we report and discuss the results obtained by our comparative evaluation (Section 4.3).

4.1. Dimensionality reduction guided by block-analysis and multi-omics data fusion

In this section, we detail the (one-step or two-step) DR+data fusion pipelines we designed and compared by their application for supervised prediction. To help readers' comprehension, Fig. 1 sketches all of them.

[6] UV_{L_j} is computed as the mean of the variances of block-ids for blocks $\mathbf{B}_0, \ldots, \mathbf{B}_j$: $UV_{L_j} = mean\left(var\left(\hat{d}_{L_t}\right)\right)$, $t = 1, \ldots, j$. EV_{L_j} is computed as the variance of block-ids for blocks $\mathbf{B}_0, \ldots, \mathbf{B}_j$: $EV_{L_j} = var\left(\hat{d}_{L_t}\right)$, $t = 1, \ldots, j$.

Table 3
Block-`id` estimates for BLCA, BRCA1, BRCA2, KIRC and LUAD datasets. For each dataset-view the table reports the number of features L_p corresponding to the plateau (if any is found) of the blocking plot, and the `id` estimate $\hat{d}$ with its corresponding standard deviation $\sigma(\hat{d})$. Column *time* reports the empirical computational time (min s) required to perform the block analysis on a standard laptop with Intel I9 processor (11th Generation), 2.50 GHz, using 3 Core(s) for parallel computation.

Dataset	View	L_p	$\hat{d}$	$\sigma(\hat{d})$	Time
BLCA	miRNA		53.72	1.81	03.31
	mRNA	2835	42.75	2.85	12.38
	Proteins		28.83	1.19	02.30
	methy	4238	52	3.48	21.01
BRCA1	miRNA		51.11	1.68	03.23
	mRNA	4640	46.61	3.18	18.15
	Proteins		35.04	1.26	01.35
	methy	5738	48.57	3.55	27.41
BRCA2	miRNA		35.21	3.69	05.03
	mRNA	4216	42.26	3.57	10.46
	Proteins		27.75	1.23	02.27
	methy	3000	39.87	3.04	08.09
KIRC	miRNA		53.26	5.55	02.00
	mRNA	4340	39.28	3.35	12.41
	Proteins		31.1	1.6	01.37
	methy	2208	47.35	4.74	06.06
LUAD	miRNA		44.15	1.12	03.48
	mRNA	3168	43.32	2.72	10.31
	Proteins		33.26	1.44	01.20
	methy	1677	43.27	3.36	07.33

Table 4
Block-`id` estimates for LUSC, OV, PRAD and SKCM datasets. For each dataset-view the table reports the number of features L_p corresponding to the plateau (if any is found), the `id` estimate $\hat{d}$ with its corresponding standard deviation $\sigma(\hat{d})$, and the empirical time complexity (min s) of the block analysis.

Dataset	View	L_p	$\hat{d}$	$\sigma(\hat{d})$	Time
LUSC	miRNA		50.77	1.41	03.31
	mRNA	4350	48.38	3.55	17.48
	Proteins		35.79	0.84	01.12
	methy	3400	43.19	3.26	08.43
OV	miRNA		33.97	2.22	02.55
	mRNA	3173	55.19	4.78	08.47
	Proteins		38.07	2.08	01.17
	methy	3614	44.14	2.93	11.21
PRAD	miRNA		54.04	1.95	03.11
	mRNA	3645	43.57	2.66	15.41
	Proteins		27.43	0.97	01.47
	methy	7760	61.92	4.23	27.29
SKCM	miRNA		57.11	3.14	03.00
	mRNA	4228	48.6	3.39	17.05
	RPPA		32.45	1.07	01.45
	methy	1920	40.86	2.88	08.11

DR is performed by either a unique step of unsupervised feature selection (FS in Fig. 1), a unique step of unsupervised feature extraction (FE in Fig. 1), or by a two-step DR process where the output of feature selection is input to feature extraction (FS+FE in Fig. 1).

The unsupervised feature selection algorithms we adopt were chosen based on their documented promising results, their limited computational costs, and considering preliminary experiments we ran, which showed their robustness with respect to datasets characterized by a limited cardinality. For interested readers, a brief literature background about unsupervised feature selection is reported in supplementary section S.A.2.1. In particular, the following algorithms were selected, and eventually optimized to reduce their computational costs:

- A parallel feature clustering algorithm returning the features that are the centroids of the identified clusters. Given the high computational costs of feature clustering methods at the state-of-the-art, the algorithm we implemented splits the input view into non-intersecting feature subsets that are distributed on multiple cores. Each core applies the Genie agglomerative clustering algorithm [33] to cluster the input feature subset, and then returns the features that are centroids of each cluster (feature medoids). The main algorithm recollects and concatenates all the feature medoids and iterates the algorithm on the concatenated feature medoids to perform a further selection until a number L_p of feature medoids is reached. More details are reported in supplementary section S.D.1.
- An iterative version of the RCUR [34] algorithm, whose parallel schema is similar to the one applied for feature clustering (more details are reported in supplementary section S.E); it allows selecting an L_p-dimensional subset of the original features, based on their potential to represent the information in the input view.
- A simple entropy filtering algorithm that selects the L_p features with the highest entropy.

When a unique step of unsupervised feature selection is applied to reduce all the multi-omics views in the input dataset, only views for which the block-analysis identified a plateau (that is, views affected by feature redundancy - light red box in Fig. 1) are reduced by selecting a number of features corresponding to the position L_p of the plateau of the block-`id`. The other views are kept as they are to avoid loss of information (light-green box in Fig. 1).

Feature extraction algorithms were similarly chosen based on their promising and successful results (supplementary section S.A.2.2). In detail, we compared Randomized PCA (RPCA, alias RSVD), Laplacian eigenmaps, UMAP, and t-SNE.

When a unique step of unsupervised feature extraction is applied to reduce all the multi-omics views in the input dataset, all views are transformed to a space whose dimension is $d = \hat{d} + 3 \times \sigma(\hat{d})$, where $\hat{d}$ (and $\sigma(\hat{d})$) is the estimate of the view-`id` (and its standard deviation) computed by block-analysis,.[7]

When we apply the two-step DR pipelines we simply perform a preliminary unsupervised feature selection method among those listed above followed by any of the unsupervised feature extraction methods listed above. This practically means that only views where a plateau is found (light-red box in Fig. 1) undergo feature selection and then feature extraction; the other views undergo only feature extraction.

Once all the views in the input dataset have been individually reduced, they are input to a data fusion algorithm to leverage the related and complementary information across views and produce an integrated view that may be input to any further analysis. Besides the basic concatenation of the reduced views, which can be considered as a simple benchmark for comparison, we exploited and compared data fusion approaches that showed their promise in several multi-omics data analysis tasks and are applied in different stages of the data analysis [12,35]. In particular, we experimented with: (1) an input-data fusion technique, namely MOFA+ [24], which applies a Bayesian approach to derive a set of latent factors capturing and representing the information content of the input multi-modal representation; two Patient Similarity Network (PSN) fusion techniques, that are (2a) the widely used Similarity Network Fusion algorithm (SNF, [14]), which applies a smart diffusion process that merges the similarities between pairs of samples that have "shared" neighbors across views, and (2b) an unsupervised Multiple Kernel Learning technique (uMKL, [25]), which outputs the (integrated) kernel that best aligns with all the unimodal kernels (i.e. the Gram matrices) representing the topological structure of each input view. More details about the three algorithms are reported in supplementary section S.A.3.

Overall, we experimented with nineteen DR methods; seven of them (one-step DR approach) applied either one of the three unsupervised feature selection methods (feature clustering, iterative RCUR - *par_rcur*

[7] As detailed in Section 3.1 the `id`-estimate computed for the view by block-analysis depends on the impact of HDLSS.

in the following, entropy filtering) or four unsupervised feature extraction methods (RPCA, Laplacian Eigenmaps, t-SNE, and UMAP); twelve were two-step DR pipelines obtained by all the combinations of the three feature selection algorithms and the four feature extraction algorithms. Considering that the reduced data is input to any of the four data integration methods we experimented (SNF, uMKL, MOFA+, and the simple concatenation), for each multi-omics dataset we run about eighty different DR+data fusion pipelines (experiments).

4.2. Experimental settings

More precisely, each DR+data fusion pipeline was tested on a binary classification task across all the nine multi-omics cancer datasets (Section 2). To this aim, a random forest classifier (RF) [36] was trained and tested to predict the overall survival event of patients.

Besides their interpretable nature [37], their often superior effectiveness with respect to even the (less efficient) deep neural network models [38], and their capability of handling a set of heterogeneous variables [36], we chose RF classifiers due to their robustness to the input feature set and the choice of hyper-parameter values. This makes it easier to apply them consistently across different datasets, DR, and data fusion approaches, allowing an objective assessment of the informativeness of the (reduced) input-data representation and the effectiveness of the data fusion algorithms, without the confounding effects due to the prior application of supervised feature selection or hyper-parameter tuning steps.[8]

To obtain an unbiased evaluation, the RF training and testing phase was repeated across fifteen stratified holdouts (80:20 train:test ratio) that obviously differed for each dataset but were kept fixed across all the experiments performed on the same dataset. To avoid confounding effects that could hamper an objective comparison, we avoided the application of any supervised feature selection algorithm and we set all the RF parameters to their default values.

4.2.1. Statistical analysis

Paired samples Wilcoxon test, alias Wilcoxon signed-rank test, at the 95% of confidence (i.e. $\alpha = 0.05$) was used for comparison. If not specified, the test was performed by pooling the results obtained on all the nine TCGA datasets; for each comparison, we considered AUCPR and AUC for hypothesis testing and exploited win-tie-loss tables to summarize the statistical comparison between each method against all the others.[9] In particular, when two specific DR+data fusion experiments were compared, we paired the results obtained on each of the nine TCGA datasets and the fifteen stratified holdouts. When, instead, we performed more generic comparisons to assess each DR approach (or each data fusion method), we paired the results obtained across the nine different datasets, the fifteen holdouts, and the four data fusion methods (or nineteen DR pipelines).

Wilcoxon signed-rank tests summarize and compare the performance of different pipelines across multiple settings. Therefore, pipelines that achieve the highest/lowest number of wins/losses can be regarded as being, on the average of all the experimented settings, the top-performing and most robust.

However, under specific settings, some other pipelines may achieve promising results; to provide a more exhaustive and detailed description of the obtained results, for each of the considered comparative evaluations we collected and analyzed the list of the top-performing experiments; in simpler words, for each of the nine TCGA datasets, we collected the three (DR+data fusion) experiments that obtained the highest AUC or AUCPR values. This allowed counting the frequency of the DR and data fusion pipelines occurring among the top-performers.

4.3. Comparative evaluation results

Our first experiment regarded the straightforward application of data fusion algorithms on the nine TCGA datasets (Section 2). When we avoided hyper-parameter tuning and supervised feature selection prior to RF training, results were far from being satisfactory (supplementary figure S.5), with AUCs often lower than 0.6 or even 0.5; when supervised feature selection and hyper-parameter tuning were applied prior to RF training (as described in supplementary section S.F.4), the performance improvement was almost unnoticeable (supplementary figure S.6), probably due to the high number of variables relative to the limited cardinality of the samples in the internal training holdouts. We obtained similar discouraging results even when concatenating patients' demographic descriptors (supplementary figure S.7).

This cleared that the high-data sparsity, affecting especially the DNA methylation and mRNA views, was a source of large bias misleading even supervised tuning steps; an effective DR approach was therefore necessary.

Then, we conducted tests to compare the tailored DR-approach proposed in this work to methods that exploit heuristics or empirical measures to uniformly set the dimension of the lower dimensional space (Section 4.3.1). Next, we compared the results obtained by the block-analysis guided DR+data fusion pipelines to gain insights about different data fusion settings, ranging from the traditional multi-omics fusion setting (Section 4.3.2), to those settings where we fused subsets of omics (Section 4.3.3), and omics plus non-omics views (Section 4.3.4).

4.3.1. When compared to heuristics, the usage of block-analysis and the id estimate obtain better results

The first aim of our comparative evaluation is to assess the effectiveness of using the `id` estimate to set the dimension of the lower dimensional space.

To this aim, we initially experimented with DR pipelines exploiting a unique step of either feature selection or feature extraction (1-step DR, Section 4.1) to choose the better performing among two heuristically set dimensions, $\bar{d}_{HD1}$ and $\bar{d}_{HD2}$. In particular, the heuristic dimensions we chose to compare are based on the rationale that most of the feature extraction algorithms allow to compute a reduced space whose number of dimensions is lower or equal than $min(N, D) - 1$ [34,39,40]. Based on this consideration, we run all the one-step DR+data fusion pipelines by using two heuristics, HD_1 and HD_2. In particular, HD_1 sets the dimension of the reduced space to $\bar{d}_{HD1} = min(N, D) - 1$; HD_2 halves $\bar{d}_{HD1}$, i.e. for HD_2 we used $\bar{d}_{HD2} = \frac{min(N,D)}{2}$.

In supplementary section S.F.2 we report details about the experiments we performed to compare the results obtained by using HD_1, HD_2, and our block-analysis to guide the reduction. First, the comparison between HD_1 and HD_2 showed the superiority of the second heuristic. Next, exhaustive comparison between HD_2 and reduction guided by block-analysis evidenced the superiority of our proposal. Further, the obtained results hint that the most robust and effective results are obtained by a two-step DR pipeline combining the iterative version of RCUR we implemented with RPCA. On the other hand, when comparing the performance and robustness of the data fusion algorithms, SNF is undoubtedly among the most promising techniques in all the comparative evaluations; however, also uMKL and MOFA+ show their promise.

[8] If the input data contains discriminative information and a limited amount of redundancy, default RF parameters can achieve decent results. Moreover, while it is undoubted that supervised feature selection and hyper-parameter tuning increase RF performance, it is also true that the increase also depends on some randomness; by avoiding preliminary feature selection and hyper-parameter tuning, we could more directly focus on comparing the performance of the DR+data fusion pipelines, which are not confounded by the effects of feature selection and hyper-parameter tuning choices.

[9] When using sided-hypothesis tests to compare the performance of two methods A and B, a win/loss (or tie) is assigned if the sided-test is below (above) the α-value. When assessing multiple methods, all pairwise comparisons are performed, and a three-column table is computed that lists, for each method, the number of wins, ties, and losses.

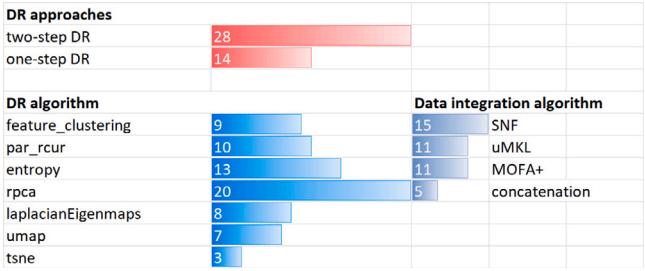

Fig. 4. Frequencies of DR algorithms, DR approaches (one-step versus two-step), and data integration methods that appear among the best models when block-analysis guides the DR of the four omics composing the multimodal dataset.

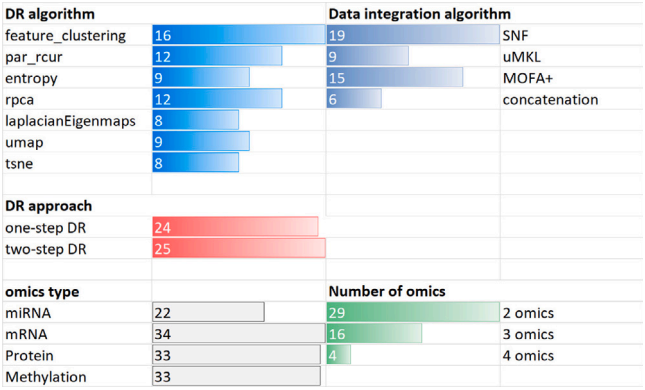

Fig. 5. Frequencies of multi-omics combinations, DR pipelines, and data fusion methods appearing among the best models when at least two omics are fused and the block-analysis guides the DR.

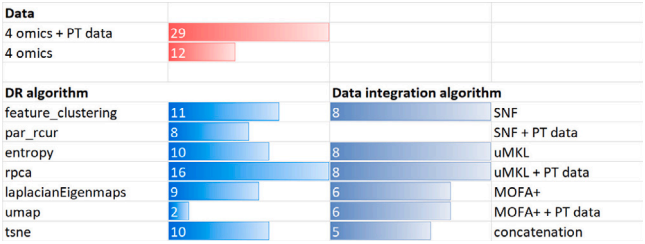

Fig. 6. Frequencies of DR methods, and data integration algorithms appearing among the top-performing models when all the views (omics and non-omics) are analyzed.

4.3.2. Comparison of DR and data integration pipelines guided by the block-analysis

To assess the robustness of the DR+data fusion pipelines guided by block analysis we applied the paired samples Wilcoxon test to compare: (1) each DR pipeline against each other, by pairing all results across datasets, holdouts, and data fusion algorithms; (2) each data fusion algorithm against each other, by pairing all results across datasets, holdouts, and DR pipelines. Supplementary figures S.16 and S.17 show the win-tie-loss tables obtained when the AUC or the AUCPR measures are used for comparison.

For what regards DR approaches, the two-step DR approaches using RPCA are in the list of top-winning pipelines that have zero losses (three up to five approaches when the AUC is used, and, most importantly, three up to four approaches when the AUCPR measure is used); generally speaking, all DR methods exploiting RPCA are the top-winners, confirming the experiments reported in [19]. The superiority of two-step DR approaches using RPCA was also confirmed by the Wilcoxon signed-rank tests we ran to compare each DR+data fusion pipeline against each other (supplementary figures S.18 and S.19 and supplementary files S7.xlsx and S8.xlsx).

Among the data integration methods, SNF seems the most robust algorithm with respect to different settings. However, when observing the pairwise DR+data fusion comparisons where the AUCPR is used as the evaluation measure (supplementary figure S.19), uMKL also shows its promise.

Further, for each dataset, we collected the top-performing pipelines, that is the list of DR+data fusion pipelines that obtain the three highest AUCs or AUCPRs. Next, we counted the frequency of occurrence of each DR and data fusion algorithm in the top-performing list (the detailed list of top-performers is reported in supplementary file S9). Fig. 4 shows that the majority of top-performers use a two-step reduction schema, including RPCA and fusing data by means of SNF.

4.3.3. The usage of all the available omics improves robustness with respect to noise and data unbalance

While performing the experiments we reasoned that the usage of all four omics might provide redundant and/or misleading information for the problem at hand. Moreover, some data fusion algorithms might profit when fewer views are integrated. Therefore, we ran experiments to compare the usage of all the available (four) omics to the usage of multi-omics datasets containing at least two omics.

While AUC results are higher when only three omics are integrated, AUCPR results evidence that using all four omics guarantees robustness with respect to class imbalance, as it is the case for most of the used benchmark datasets (Supplementary figures S.20-S.23 and files S10.xlsx and S11.xlsx).

Note that, this performance is indirectly related to the effectiveness of the prior DR step. Indeed, we recall that, when no DR is applied at all (supplementary section S.F.1), the algorithms that fuse all the omics and then apply supervised feature selection and hyper-parameter tuning achieve poor results. Therefore, provided that a proper DR is applied, the usage of all the available omics allows obtaining robust results, avoiding costly experiments to choose the most suitable combination of omics given the problem at hand.

Among the DR pipelines, feature clustering followed by RPCA, the iterative RCUR (either alone or in combination with RPCA), or RPCA alone were still the most robust DR approaches. When instead the paired samples Wilcoxon test was used to compare the four data fusion algorithms, SNF confirmed its superiority for both AUC and AUCPR measures. Note that, the simple integration via concatenation seemed to benefit from the reduction of omics, which is probably due to the lower number of features when less than four omics are considered.

When observing the frequency of multi-omics combination, DR pipeline, and data fusion algorithm used in experiments obtaining the highest AUC or AUCPR (Fig. 5 - supplementary file S12.xlsx) we note that all views but the miRNA-view equally contributed in obtaining a good performance. Moreover, combinations of two omics could appear among the top-performers; this suggests that, when having enough samples and computing power, the combination of input views could be regarded as a further hyper-parameter to be tuned to maximize performance.

4.3.4. Integration of demographic patient data may further improve results

The available TCGA datasets contain demographic descriptions (gender, age at diagnosis, ethnicity, race) that might provide further useful information to improve the performance of the supervised classification task.

Since several bio-medical studies provide both omics and non-omics patients' descriptors, and considering the documented literature interest about challenges regarding the integration of omics and non-omics views [41] (supplementary section S.F.3.1), it was interesting to understand not only if patients' descriptors other than multi-omics may improve results of our supervised analysis, but also if a simple approach

that concatenates patients' descriptors to the fused multi-omics view could be more effective than using the patients' descriptors as a further view to be integrated.

In our classification pipeline, once the multi-omics views are integrated, concatenation of demographic descriptors is possible because RF can process heterogeneous data. On the other hand, to use SNF, uMKL for integrating multi-omics and demographic views we used the Gower similarity[10] to compute pairwise similarities, and then used them as the fifth kernel to be integrated by SNF and uMKL. When using MOFA+ we simply provided the demographic view as a further view to guide the discovery of the latent components.

Based on the results from the previous experiment (Section 4.3.3), in this comparative evaluation we limited the number of experiments to those that used all the available omics.

Fig. 6 plots the frequency of each data view, DR algorithm, and data fusion method appearing in the top-performing experiments (listed in supplementary file S15.xlsx). In the figures, "4 omics+ pt" refers to the usage of omics and non-omics variables; "MOFA+ + PT data", "SNF + PT data", and "uMKL + PT data" refer to the data fusion algorithms also integrating the demographic view; "MOFA+", "SNF", "uMKL", and "concatenation" refer to the traditional application of the data fusion algorithms for integrating multi-omics, followed by concatenation with the demographics views.

Observing the results (Supplementary figures S.24-S.27) we understand that surely the inclusion of demographic data improves performance; indeed, for both AUC and AUCPR, combinations of views including patient data always win with respect to combinations neglecting demographic predictors. The comparative performance of DR pipelines remains unaltered with respect to previous experiments; indeed, iterative RCUR (par_rcur) and feature clustering followed by RPCA, or the usage of a unique step of par_rcur or RPCA are the DR methods achieving the most robust results for both AUC and AUCPR.

Regarding the data integration models, SNF is still among the top-performing models. When comparing the two (multi-omics plus demographic) integration approaches, we note that both "SNF + PT data" and "MOFA + PT data" (using the demographic descriptors as a further view to be integrated) score better than their counterparts ("SNF" and "MOFA+") that simply concatenate the demographic variables to the integrated multi-omics views. This is not true for uMKL, which, according to the win-tie-loss tables, obtained more robust results when we first integrated omics descriptors and then concatenated demographic variables to the fused kernel representation. This might be due to the fact that the Gower similarity measure is not a proper kernel similarity, as required by uMKL. Despite this fact, we note that "uMKL" and "uMKL + PT data" appear with the highest frequency in the list of top-performing models (supplementary file S15.xlsx and supplementary Fig. 6), which evidence the potentials of the uMKL data fusion strategy and suggests that the transformation of Gower similarity into a kernel matrix might further improve results.

Overall, these results suggest that integration of omics and non-omics variables, when opportunely designed, might be a promising way. Indeed, when considering the more detailed comparison between DR+data fusion pipelines (extracts of the top twenty-five winners in supplementary figures S.26 and S.27) we note that the best fusion algorithm is MOFA+ integrating demographic descriptors. Considering that one of the advantages of MOFA+ relies on its ability to integrate heterogeneous data type, its superiority is not a surprise and further supports our belief that the integration of omic and non-omics data is a promising way that needs a careful design.

[10] Gower distance/similarity is a measure of dissimilarity or similarity between two individuals (or data points) described by a set of heterogeneous variables, including categorical, binary, ordinal, and numerical variables. The Gower distance/similarity is computed as the average of all the distances/similarities measured on each variable, taking into account their different data types.

4.3.5. Feature selection and hyper-parameter tuning further improves results

To avoid confounding effects that would bias the comparative assessment, all the results reported in our experiments (besides being averaged across all the nine TCGA datasets) were obtained with RF classifiers without applying internal supervised feature selection or hyper-parameter tuning. Consequently, we conducted a final experiment where we used all four omics as input to data fusion, concatenated the fused representation with demographic descriptors, and then optimized the RF performance by supervised feature selection through RF importance [37] and hyper-parameter tuning through internal stratified holdout validation (more details are reported in supplementary section S.F.4).

Under this setting, we compared the results obtained when avoiding any DR step prior to data fusion (supplementary figures S.6 and S.7) to those obtained when using the proposed DR pipeline. While results obtained when avoiding DR evidence the effect of the curse of dimensionality, the application of the proposed approach obtained more than satisfactory results (supplementary figures S.8 and S.9), with AUCs often greater than 0.70/0.75 and large AUCPRs even for datasets characterized by the largest class unbalance (positive:negative ratio = 0.25).

5. Discussion and conclusions

In this paper, we described a novel application of block-analysis to leverage the most promising `id` estimators for computing an unbiased and more robust `id` estimate of the views in a multi-modal dataset, especially those impacted by the HDLSS issue. We also proposed an automated analysis of the block-`id` distribution for each view, enabling the detection of views highly impacted by HDLSS. This led to the development of a tailored DR pipeline guided by block-`id` analysis. Instead of uniformly applying a single DR algorithm to all views, our approach automatically defines a specific latent space dimension for each view and selectively applies either a single step of feature selection or feature extraction, or a two-phase DR process where feature selection is followed by feature extraction.

The two-phase DR process was devised to handle views where the impact of HDLSS is higher. These views require a more complex and careful reduction; a single step of feature selection is often not sufficient to reduce the curse of dimensionality, while a single step of feature extraction suffers from the fact that most feature extraction techniques rely on the computation of pairwise sample distances, which are biased under the curse of dimensionality due to high sample sparsity. The prior application of a feature selection algorithm reduces the curse of dimensionality and allows for more reliable distance computations, thereby facilitating the task of the subsequent feature extraction algorithm.

Cancer remains the leading cause of death worldwide, and its heterogeneicity represents a great challenge in the understanding of tumor drivers, cancer prognosis, and treatment design. This is one of the reasons why we assessed our proposal and compared it to classic DR pipelines that use heuristics to uniformly define the latent space for reduction by a single selection or extraction strategy, by utilizing nine tumor datasets from the TCGA repository. While some literature studies have used TCGA datasets for testing classification models [42–46], they typically limit their analysis to a maximum of four TCGA datasets without providing clear justification for their choices. In contrast, our approach involved selecting nine diverse datasets, chosen to encompass a wide range of heterogeneity. This heterogeneity includes different tumor types, varying HDLSS ratios, and differing imbalance ratios characterizing the supervised prediction tasks. By adopting this comprehensive approach, we aimed to capture a more nuanced and representative perspective in our analysis.

After demonstrating the superiority of our proposal using such diverse benchmark datasets, we aimed to further advance knowledge in the field of multi-omics data fusion and analysis by evaluating the

effects of the proposed DR pipeline on subsequent data fusion techniques that have shown promise in this context. Our particular focus was on the critical task of prognosis prediction (overall survival events), which is gaining increasing interest [21,22,47]. For this purpose, we employed RF classifiers, often preferred in the biomedical field due to their interpretability, relative robustness to hyper-parameter settings, superior efficiency, and effectiveness in many tasks [38]. This analysis allowed us to investigate the impact of crucial yet often overlooked settings in multi-omics data integration.

Specifically, our results first demonstrated that a properly designed DR step is crucial and should never be neglected when processing complex multi-omics data. Secondly, we showed that DR approaches guided by block-analysis were superior to traditional DR approaches that set the dimension of the reduced space using heuristics or preliminary empirical experiments. This was expected, as our strategy tailors DR – and therefore the latent reduction space – to ensure the individual informative content of views is maintained as much as possible after reduction. A uniform reduction process, on the other hand, might cause information loss in some views while failing to address the curse of dimensionality in others. This is particularly true in multi-omics datasets, where the impact of HDLSS is highly evident in some views (e.g., methylation or mRNA with up to 450,000 features) and less so in others (e.g., miRNA and proteins with hundreds of features).

When we compared the results achieved by all the tailored DR pipelines we experimented with, we noticed that our two-step reduction strategy is particularly effective when all four omics are used. This is likely due to the inclusion of views highly impacted by HDLSS. Conversely, when only two views are used, tailored DR approaches employing a single step of feature selection or extraction can also achieve promising results.

When comparing the performance of the individual DR algorithms within the tailored (one-step or two-step) DR pipelines, we found that algorithms using feature clustering, RCUR, or RPCA (alias RSVD) consistently achieved better performance across all our experiments. The superior performance of the feature clustering method can be explained by its ability to select features that are representative of feature clusters, thus covering all the main information in the data. The good performance of RPCA and RCUR aligns with the findings of other relevant works on multi-omics data [19] and is supported by the robust theoretical foundations of both RCUR and RPCA. On the other hand, the lower performance of models such as t-SNE and UMAP is likely due to their reliance on optimization processes, which are not robust with a limited sample size. This issue was practically evident when we observed the higher variability of results obtained by DR pipelines using UMAP and t-SNE.

When observing the robustness and performance of the experimented data fusion algorithms, we confirmed the efficiency and effectiveness of SNF, which demonstrated robustness across different settings. MOFA+ also showed promise, although its robustness and efficiency were often lower than those of SNF. MOFA+ is based on a stochastic variational inference framework optimized by gradient descent, which is effective when the number of samples is much larger than the number of features [24]. While MOFA+ was given reduced data, the limited number of cases might still impact its robustness. Nonetheless, MOFA+ has the advantage of handling heterogeneous data types, effectively integrating omics and non-omics views, making it a promising approach.

On the other hand, the lower performance of uMKL might be due to the computation of an integrated kernel that is equally "distant" from all the input kernels. Although multi-omics data are somewhat related, their different semantics can result in kernels with different topological characteristics. In such cases, the integrated kernel might fail to accurately represent the individual information carried by each view.

We further noted that comparable results can be obtained when using all four omics or specific subsets of at least two omics. This suggests that in our experiments, the DR and data fusion steps can handle the presence of potentially redundant information within and between the four omics views, allowing all available omics types to be used without the need to test all possible combinations. In other words, a well-designed DR can facilitate the subsequent data fusion task by effectively removing redundancies within individual data types while better exposing their informative content. This simplifies the data fusion algorithm task, which then only needs to address redundancies across views while uncovering the shared and unique informative content of the multiple omics. Consequently, this approach eliminates the need for costly empirical experiments to select the optimal subset of omics for integration.

Additionally, we assessed the impact of including patients' demographic descriptors in the analysis and demonstrated that this addition effectively increases classification performance. We however warn that, when non-omics data is integrated to omics views, the interpretation and explanation of the obtained predictions to, e.g., uncover triggers of mortality risks, should be carefully performed due to ascertain bias [3], data quality issues [48], and dataset/model fairness [49] (supplementary section S.F.3.1).

To conclude, the key findings of our work are:

- **Importance of a properly designed DR approach:** Our results demonstrate that neglecting a DR step leads to suboptimal outcomes, particularly in complex multi-omics data. A well-designed DR approach, guided by block-analysis-based `id` estimation, ensures the preservation of the individual informative content of views, effectively addressing the curse of dimensionality and minimizing information loss in multi-modal datasets.
- **Superiority of block-analysis-guided DR:** DR approaches guided by block-analysis consistently outperformed traditional heuristic-based methods. This tailored strategy preserves the individual content of each view while effectively mitigating HDLSS effects, especially in views with extreme feature counts, such as methylation or mRNA.
- **Effectiveness of the two-step reduction strategy:** The proposed two-step DR strategy demonstrated particular effectiveness when all four omics were used, consistently outperforming single-step DR pipelines. However, tailored single-step DR approaches also yielded promising results when applied to subsets of two views.
- **Performance of DR algorithms:** Among the DR algorithms tested, feature clustering, RCUR, and RPCA (RSVD) consistently outperformed other methods, demonstrating theoretical robustness and an ability to retain critical information, making them highly suitable for multi-omics data fusion tasks. In contrast, optimization-reliant methods such as t-SNE and UMAP showed higher variability and lower performance, likely due to their sensitivity to limited sample sizes.
- **Performance of data-fusion algorithms:** SNF demonstrated robust performance across varying settings, while MOFA+ showed potential but was less robust due to its reliance on stochastic variational inference, which can be sensitive to small sample sizes. uMKL exhibited lower performance, likely due to challenges in kernel integration for multi-modal data.
- **Insights on crucial data-fusion settings:**

  (a) Comparable results were observed when using all four omics or subsets of at least two omics. This highlights the ability of a well-designed DR pipeline to handle within-view redundancies, simplifying data fusion algorithms, which can then focus on cross-view redundancies and shared informative content.

  (b) **Impact of demographic data integration:** Including demographic descriptors improved classification performance. However, caution is needed when integrating non-omics data due to potential biases, data quality issues, and fairness considerations.

These findings highlight the utility of our tailored DR pipeline for multi-omics data analysis and its capacity to enhance downstream tasks such as data fusion and predictive modeling.

While the experiments exhaustively show the effectiveness of the proposed approach, we evidence some limitations that we are investigating as future work.

At first, we focus our investigation on a specific biomedical research field, i.e., multi-omics data analysis, and we limited our analysis to a binary prediction task because the limited cardinality of our datasets hampered multi-class classifications, as the cardinality of each class would often be too scarce to allow model generalization. Considering the generality of the proposed DR pipeline, future works are planned to test our reduction strategy on multimodal data acquired in different and varied settings and having enough samples to assess performance on multi-class classification problems. Secondly, we tested the proposed pipeline by applying classic unsupervised dimensionality-reduction algorithms, due to their widespread usage in the bioinformatics field. The relative simplicity of such methods allowed a more reliable assessment of the proposed DR pipeline avoiding confounding effects that would bias the comparison. However, variational autoencoders [7,50] are now showing their potential in the HDLSS field, with interesting applications in the multi-omics field [51]. As shown by the experiments reported in [7], the choice of the latent space dimension is crucial for this powerful learning machines; however, automated hyperparameter optimization is hampered by the limited sample size, so this hyperparameter is often empirically chosen (see experiments reported by [7]). The technique we propose is therefore useful to engineer variational autoencoders and construct, eventually multimodal, variational autoencoder models where the size of the bottleneck layer is set by the `id`-estimate computed via block-analysis. The insights gained by block-analysis could be further used to design more complex architectures for views more impacted by HDLSS effects.

We finally note that, even with parallel computation, the proposed DR pipeline may encounter scalability issues. Considering recent advancements in effective unsupervised feature selection algorithms designed to scale on massive datasets [52,53], our future work will focus on integrating these algorithms within our DR pipeline. Notably, compared to feature extraction algorithms that combine input features through often non-linear operations, an effective feature selection offers the advantage of maintaining interpretability in the reduced set.

CRediT authorship contribution statement

Jessica Gliozzo: Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mauricio Soto-Gomez:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Valentina Guarino:** Software, Methodology, Conceptualization. **Arturo Bonometti:** Data curation. **Alberto Cabri:** Writing – review & editing. **Emanuele Cavalleri:** Writing – review & editing. **Justin Reese:** Writing – review & editing, Conceptualization. **Peter N. Robinson:** Writing – review & editing, Formal analysis, Conceptualization. **Marco Mesiti:** Writing – review & editing. **Giorgio Valentini:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Elena Casiraghi:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Funding

This paper is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

Elena Casiraghi acknowledges travel support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951847.

We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work was realized with the collaboration of the European Commission Joint Research Centre under the Collaborative Doctoral Partnership Agreement $N°35454$.

The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of code

All the code for the blocking analysis and for running the experiments is available at https://github.com/AnacletoLAB/DR_omics_integration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jessica Gliozzo reports financial support was provided by European Commission Joint Research Centre Ispra. Elena Casiraghi reports travel support was provided by Horizon Europe - ELISE Project. Elena Casiraghi reports computing power was provided by Interuniversity Consortium Cineca. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Authors would like to thank Prof. Juho Rousu (lead of KEPACO Lab - Department of Computer Science, Aalto University, Espoo, Finland) for his invaluable support, and PhD Riikka Huusari (Department of Computer Science, Aalto University, Espoo, Finland) for her precious comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.103049.

References

[1] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18(1):1–15.
[2] Dai X, Shen L. Advances and trends in omics technology development. Front Med 2022;9:911861.
[3] Athieniti E, Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. Comput Struct Biotechnol J 2023;21.
[4] Conesa A, Beck S. Making multi-omics data accessible to researchers. Sci Data 2019;6(1):251.
[5] Babu M, Snyder M. Multi-omics profiling for health. Mol Cell Proteomics 2023;22(6).
[6] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14:1177932219899051.
[7] Mahmud MS, Huang JZ, Fu X, Ruby R, Wu K. Unsupervised adaptation for high-dimensional with limited-sample data classification using variational autoencoder. Comput Inform 2021;40(1):1–28.
[8] Trunk GV. A problem of dimensionality: A simple example. IEEE Trans Pattern Anal Mach Intell 1979;PAMI-1(3):306–7.
[9] Lv J. Impacts of high dimensionality in finite samples. Ann Statist 2013;41(4):2236–62.
[10] Hughes G. On the mean accuracy of statistical pattern recognizers. IEEE Trans Inf Theory 1968;14(1):55–63.
[11] Nanga S, Bawah AT, Acquaye BA, Billa M-I, Baeta FD, Odai NA, Obeng SK, Nsiah AD. Review of dimension reduction methods. J Data Anal Inf Process 2021;9(3):189–231.

[12] Gliozzo J, Mesiti M, Notaro M, Petrini A, Patak A, Puertas-Gallardo A, Paccanaro A, Valentini G, Casiraghi E. Heterogeneous data integration methods for patient similarity networks. Brief Bioinform 2022.
[13] Xiang R, Wang W, Yang L, Wang S, Xu C, Chen X. A comparison for dimensionality reduction methods of single-cell rna-seq data. Front Genet 2021;12:646936.
[14] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333–7.
[15] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci 2013;110(11):4245–50.
[16] Rappoport N, Shamir R. Nemo: cancer subtyping by integration of partial multi-omic data. Bioinformatics 2019;35(18):3348–56.
[17] Nguyen H, Shrestha S, Draghici S, Nguyen T. Pinsplus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics 2019;35(16):2843–6.
[18] Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLoS Comput Biol 2019;15(6):e1006907.
[19] Nguyen H, Tran D, Tran B, Roy M, Cassell A, Dascalu S, Draghici S, Nguyen T. Smrt: Randomized data transformation for cancer subtyping and big data analysis. Front Oncol 2021;11.
[20] Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. Front Oncol 2020;10:1030.
[21] Ramirez R, Chiu Y-C, Zhang S, Ramirez J, Chen Y, Huang Y, Jin Y-F. Prediction and interpretation of cancer survival using graph convolution neural networks. Methods 2021;192:120–30.
[22] Sun B, Chen L. Interpretable deep learning for improving cancer patient survival based on personal transcriptomes. Sci Rep 2023;13(1):11344.
[23] Jiang L, Xu C, Bai Y, Liu A, Gong Y, Wang Y-P, Deng H-W. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. NPJ Precis Oncol 2024;8(1):4.
[24] Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol 2020;21(1):1–17.
[25] Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics 2018;34(6):1009–15.
[26] Ramos M, Geistlinger L, Oh S, Schiffer L, Azhar R, Kodali H, de Bruijn I, Gao J, Carey VJ, Morgan M, et al. Multiomic integration of public oncology databases in bioconductor. JCO Clin Cancer Inform 2020;1:958–71.
[27] Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018;173(2):400–16.
[28] Ceruti C, Bassis S, Rozza A, Lombardi G, Casiraghi E, Campadelli P. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. Pattern Recognit 2014;47(8):2569–81.
[29] Campadelli P, Casiraghi E, Ceruti C, Rozza A. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. Math Probl Eng 2015;2015.
[30] Facco E, d'Errico M, Rodriguez A, Laio A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Sci Rep 2017;7(1):1–8.
[31] Badii R, Politi A. Hausdorff dimension and uniformity factor of strange attractors. Phys Rev Lett 1984;52(19):1661.
[32] Blitzstein JK, Hwang J. Introduction to probability. Crc Press; 2019.
[33] Gagolewski M. Genieclust: Fast and robust hierarchical clustering. SoftwareX 2021;15:100722. http://dx.doi.org/10.1016/j.softx.2021.100722, URL https://www.sciencedirect.com/science/article/pii/S2352711021000649.
[34] Mahoney MW, Drineas P. Cur matrix decompositions for improved data analysis. Proc Natl Acad Sci 2009;106(3):697–702.
[35] Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nature Rev Genet 2015;16(2):85–97.
[36] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[37] Casiraghi E, Malchiodi D, Trucco G, Frasca M, Cappelletti L, Fontana T, Esposito AA, Avola E, Jachetti A, Reese J, et al. Explainable machine learning for early assessment of covid-19 risk prediction in emergency departments. Ieee Access 2020;8:196299–325.
[38] Zhou Z-H, Feng J. Deep forest. Natl Sci Rev 2019;6(1):74–86.
[39] Rokhlin V, Szlam A, Tygert M. A randomized algorithm for principal component analysis. SIAM J Matrix Anal Appl 2010;31(3):1100–24.
[40] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 2003;15(6):1373–96.
[41] López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, Calle ML, Malats N. Challenges in the integration of omics and non-omics data. Genes 2019;10(3):238.
[42] Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. Netdx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol 2019;15(3):e8497.
[43] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, Huang K. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Commun 2021;12(1):3445.
[44] Moon S, Lee H. Moma: a multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics 2022;38(8):2287–96.
[45] Zhong Y, Peng Y, Lin Y, Chen D, Zhang H, Zheng W, Chen Y, Wu C. Modilm: towards better complex diseases classification using a novel multi-omics data integration learning model. BMC Med Inform Decis Mak 2023;23(1):1–18.
[46] Ouyang D, Liang Y, Li L, Ai N, Lu S, Yu M, Liu X, Xie S. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Comput Biol Med 2023;164:107303.
[47] Jiang L, Xiao Y, Ding Y, Tang J, Guo F. Discovering cancer subtypes via an accurate fusion strategy on multiple profile data. Front Genet 2019;10:20.
[48] Callen J, Georgiou A, Li J, Westbrook JI. The impact for patient outcomes of failure to follow up on test results, how can we do better? EJIFCC 2015;26(1):38.
[49] Casiraghi E, Wong R, Hall M, Coleman B, Notaro M, Evans MD, Tronieri JS, Blau H, Laraway B, Callahan TJ, et al. A method for comparing multiple imputation techniques: A case study on the us national covid cohort collaborative. J Biomed Inform 2023;139:104295.
[50] Mahmud MS, Huang JZ, Fu X. Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. Int J Comput Intell Appl 2020;19(01):2050002.
[51] Doncevic D, Herrmann C. Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations. Bioinformatics 2023;39(6):btad387.
[52] Luo C, Wang S, Li T, Chen H, Lv J, Yi Z. Large-scale meta-heuristic feature selection based on bpso assisted rough hypercuboid approach. IEEE Trans Neural Netw Learn Syst 2023;34(12):10889–903. http://dx.doi.org/10.1109/TNNLS.2022.3171614.
[53] Luo C, Wang S, Li T, Chen H, Lv J, Yi Z. Rhdofs: A distributed online algorithm towards scalable streaming feature selection. IEEE Trans Parallel Distrib Syst 2023;34(6):1830–47. http://dx.doi.org/10.1109/TPDS.2023.3265974.