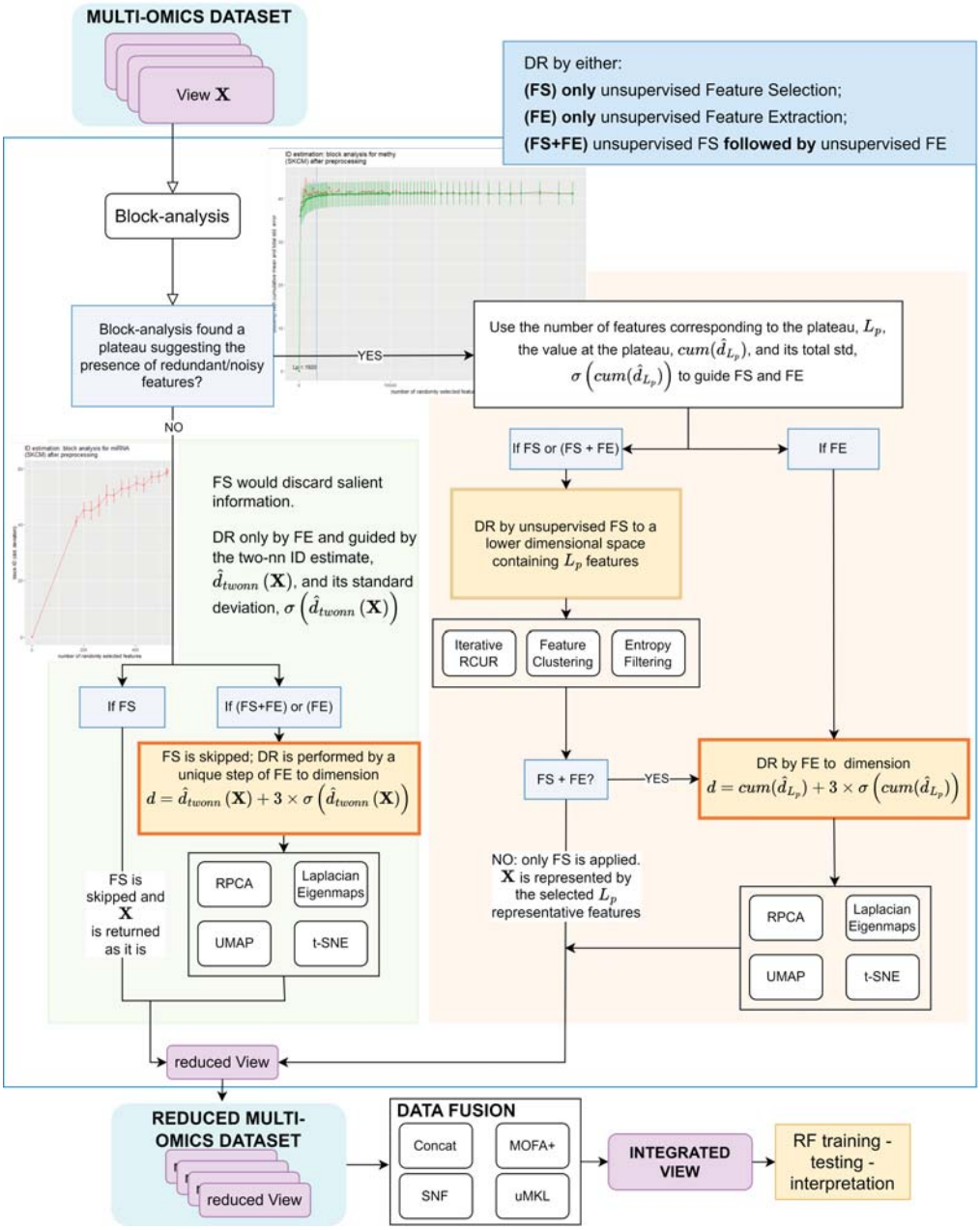
Graphical Abstract

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

Jessica Gliozzo, Valentina Guarino, Arturo Bonometti, Alberto Cabri, Emanuele Cavalleri, Mauricio Soto-Gomez, Justin Reese, Peter N Robinson, Marco Mesiti, Giorgio Valentini, Elena Casiraghi



Highlights

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

Jessica Gliozzo, Valentina Guarino, Arturo Bonometti, Alberto Cabri, Emanuele Cavalleri, Mauricio Soto-Gomez, Justin Reese, Peter N Robinson, Marco Mesiti, Giorgio Valentini, Elena Casiraghi

- We introduce a flexible pipeline to guide in a principled way feature selection and feature extraction methods to reduce the high dimensions and to contrast the curse of dimensionality that affects multi-omics data.
- We harness the power of cutting-edge Intrinsic Dimensionality (`id`) estimation through block-analysis, providing an unbiased estimation of the individual `ids` for each view within a multi-modal dataset.
- We use an exhaustive set of diverse multi-omics cancer datasets from the well-known TCGA dataset to show that the automatic analysis of the distribution of the block-`ids` characterizing each omics-view leverages dimensionality reduction, by (1) evidencing feature noise and redundancy, and (2) providing an unbiased estimate of the `id` for each view, to be used for setting the dimension of the reduced space. This avoids empirical or heuristic choices and allows tailoring the reduction to each data-view.
- The crucial information gained by block-analysis allowed proposing a two-step dimensionality-reduction approach combining feature selection and feature extraction. Our comparative evaluation shows the effectiveness of the proposed technique and its synergy with state-of-the-art data-fusion techniques applied in a multi-omics context.
- We show that the proposed reduction pipeline leverages traditional dimensionality reduction and state-of-the-art data-fusion algorithms. Indeed, it obtains effective performance when predicting overall survival events with simple random forest classifiers, often preferred in the biomedical field due to their robustness, efficiency, and interpretable nature.

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

Jessica Gliozzo[a,b], Valentina Guarino[a], Arturo Bonometti[c,d], Alberto Cabri[a], Emanuele Cavalleri[a], Mauricio Soto-Gomez[a], Justin Reese[e], Peter N Robinson[f], Marco Mesiti[a], 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*

Abstract

The advent of high-throughput sequencing technologies has revolutionized the field of multi-omics patient data analysis. While these techniques offer a wealth of information, they often generate datasets with dimensions far surpassing the number of available cases. This discrepancy in size gives rise to the challenging "small-sample-size" problem, significantly compromising the reliability of any subsequent estimate, whether supervised or unsupervised.

This calls for effective dimensionality reduction techniques to transform high-dimensional datasets into lower-dimensional spaces, making the data manageable and facilitating subsequent analyses. Unfortunately, the definition of a proper dimensionality reduction pipeline is not an easy task; besides the problem of identifying the best dimensionality reduction method, the definition of the dimension of the lower-dimensional space into which each dataset should be transformed is a crucial issue that influences all the subsequent analyses and should therefore be carefully considered.

*corresponding author - email: elena.casiraghi@unimi.it

Further, the availability of multi-modal data calls for proper data-fusion techniques to produce an integrated patient-view into which redundant information is removed while salient and complementary information across views is leveraged to improve the performance and reliability of both unsupervised and supervised learning techniques.

This paper proposes leveraging the intrinsic dimensionality of each view in a multi-modal dataset to define the dimensionality of the lower-dimensional space where the view is transformed by dimensionality reduction algorithms. Further, it presents a thorough experimental study that compares the traditional application of a unique-step of dimensionality reduction with a two-step approach, involving a prior feature selection followed by feature extraction.

Through this comparative evaluation, we scrutinize the performance of widely used dimensionality reduction algorithms. Importantly, we also investigate their impact on unsupervised data-fusion techniques, which are pivotal in biomedical research. Our findings shed light on the most effective strategies for handling high-dimensional multi-omics patient data, offering valuable insights for future studies in this domain.

Keywords: Dimensionality Reduction, Intrinsic Dimensionality, Feature Selection, Feature Extraction, Data Fusion, Multi-omics Datasets

1. Introduction

In the biomedical research field, the emergence of high-throughput technologies has revolutionized the acquisition of 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, 4, 5, 6].

However, the analysis of multi-omics data presents significant challenges due to their high dimensional nature and multi-modality. In particular, the high-dimensional nature of omics data results in high computational costs, data sparsity, and overfitting due to the presence of noisy, uninformative, and redundant features. These problems collectively are referred to as the "curse of dimensionality", and can bias practically all results obtained from these data. This is particularly true in bio-medical datasets, often characterized by high-dimension and small-sample-size, that is by a large number of features relative to the number of samples. Such

datasets may easily reach a level of sparsity that causes samples to appear distributed on the boundaries of the hyperspace, affecting the reliability of subsequent supervised or unsupervised analyses [7, 8, 9].

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 [10]. 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 following data-fusion task, for which several promising algorithms have been already presented in the bio-medical literature [11].

However, while feature selection and feature extraction methods have shown their own advantages and several reviews describe and eventually compare their successful results [12], to the best of our knowledge no paper investigated the following two crucial choices that should be carefully considered when analyzing and reducing high-dimensional datasets, potentially affected by the curse of dimensionality. First, there is no rule of thumb that allows claiming that a dataset is affected by the curse of dimensionality/small-sample-size problem. Second, the choice of the dimension of the reduced space is one of the most crucial choices; too low values would cause the loss of information, while too large values would not consistently reduce the curse of dimensionality. In practice, literature works in the field of bioinformatics either avoid any dimensionality reduction [13] or make some empirical/heuristic decisions [14, 15] not motivated by any theoretical justification. However, the careful design of the DR step affects all the subsequent computations and the reliability of the obtained results [16, 17, 18].

Further, when applying dimensionality reduction in a multi-omics setting, few works consider that different views might carry different amounts of information. Neglecting this fact, most works blindly apply any of the successful data-fusion techniques proposed in literature [19, 11] (supplementary section S. A.3) without any prior view-reduction, or when a reduction is applied, the same (often empirical) dimension is chosen for all the views.

Instead, a prior view-specific reduction would better emphasize and expose the information within each view, therefore improving the effectiveness of the following data-fusion task, whose aim is to uncover the salient information across views, while removing the (between-)view redundancy [20, 21].

The aim of this work is to propose a novel block-analysis technique leveraging one of the most promising and recent Intrinsic Dimensionality (`id`) estimators

([22], supplementary section S. A.2), namely the *two-nn* estimator (supplementary section S. A.2.1) to understand when and how a feature selection or feature extraction method could improve the data representation. If curse of dimensionality is detected, the block-analysis allows defining the dimension of the lower-dimensional space where each view should be transformed by any of the promising feature selection or feature extraction approaches proposed at the state of the art (supplementary section S. A.1). Further, by exploiting the information provided by block-analysis, we propose and experiment with a novel two-step DR process that improves results by combining the advantages of the first application of feature selection followed by feature extraction.

The proposed DR technique is applied in the context of multi-omics data analysis, where effective multi-omics data-fusion algorithms have been recently developed. In particular, we compared some of the most promising and effective unsupervised multi-omics data-fusion techniques (i.e. MOFA+ [21], uMKL [23], and SNF [13], all summarized in supplementary section S. A.3) to assess their strengths and compare their robustness across different settings. Indeed, while the effectiveness of these data integration approaches is undoubted, we wanted to investigate (1) the effect of using subsets of the input multi-omics views, to understand whether a subset of the input views could suffice to provide effective results, or (2) whether the integration of not-omics patients' views, e.g. patients' demographics views carrying a completely different semantic, would enhance the salient and discriminative information, therefore facilitating the following supervised/unsupervised analysis.

To perform our comparative evaluation we selected nine (high-dimensional) multi-omics datasets from the well-known TCGA repository (more details in section 2) and designed a supervised machine learning pipeline that reduces all the omics views in the input dataset, fuses them (by eventually integrating also the demographic view or concatenating it to the integrated view), and finally uses a random forest classifier for analyzing the fused information to predict patients' survival. By using a supervised classification task the obtained performance can be compared using well-established performance measures, such as the area under the ROC curve (AUROC) and the area under the Precision-Recall curve (AUCPR).

Results show that, in our classification problem, DR guided by block-analysis outperforms traditional DR approaches that use heuristics to set the dimensionality of the reduced space. Further, the robustness of DR improves when a two-step DR approach is applied, or when non-omics patients' descriptors are also integrated into the analysis. On the other hand, when a prior (and properly designed) step of DR is applied to effectively remove intra-view redundancy and noise, there

is no need to spare computational time for testing the usage of subsets of the input views, because the data-fusion algorithms can produce effective integrated representations that achieve robust classification results.

2. TCGA datasets

To obtain reliable results, we mined the following nine multi-omics datasets from the TCGA cancer repository[1] (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 carcinoma dataset (**KIRC**); the LUng ADenocarcinoma dataset (**LUAD**); the LUng Squamous Cell carcinoma datset (**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. B.1-S. B.3 for further details). Patients in the TCGA dataset may be classified based on their Overall Survival (OS) event, which is available from the TCGA-CDR [25] dataset. We used the overall survival label to perform a supervised classification task.

Note that some literature studies using TCGA datasets for testing classification models [26, 27, 28, 29, 30] already exist. However, these studies typically restrict their analysis to a maximum of four TCGA datasets, without providing clear justification for their choices. In contrast, our approach involved the selection of nine diverse datasets, that were chosen to encompass a wide range of heterogeneity, not only in terms of different tumor types being investigated, but also in the ratio between the number of cases and variables within each dataset, and the balance between positive (patients with OS event equal to 1) and negative

[1]The R package "curatedTCGAData" [24] was used to download the tumor datasets from the TCGA repository (dataset version 2.0.1).

patients (OS = 0). By adopting this comprehensive approach, we aimed to capture a more nuanced and representative perspective in our analysis.

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		12276	12276	0.027			
	proteins		183	183	1.831			
	methy		315551	30000	0.012			
BRCA1	miRNA	317	496	496	0.6391	42	275	0.13
	mRNA		12242	12242	0.026			
	proteins		202	202	1.569			
	methy		289962	30000	0.011			
BRCA2	miRNA	128	502	502	0.255	14	114	0.11
	mRNA		8128	8128	0.016			
	proteins		192	192	0.667			
	methy		278099	30000	0.004			
KIRC	miRNA	169	364	364	0.464	48	121	0.28
	mRNA		7942	7942	0.021			
	proteins		186	186	0.909			
	methy		319740	30000	0.006			
LUAD	miRNA	300	465	465	0.645	120	180	0.40
	mRNA		11131	11131	0.027			
	proteins		179	179	1.676			
	methy		331828	30000	0.01			

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 file 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}	$\frac{N_{pos}}{N}$
LUSC	miRNA	228	491	491	0.464	93	135	0.41
	mRNA		11473	11473	0.02			
	proteins		178	178	1.281			
	methy		273884	30000	0.008			
OV	miRNA	226	308	308	0.734	143	83	0.63
	mRNA		11731	11731	0.019			
	proteins		186	186	1.215			
	methy		13296	13296	0.017			
PRAD	miRNA	337	457	457	0.737	6	331	0.02
	mRNA		8887	8887	0.038			
	proteins		169	169	1.994			
	methy		301920	30000	0.011			
SKCM	miRNA	334	523	523	0.639	150	184	0.45
	mRNA		13050	13050	0.026			
	proteins		186	186	1.796			
	methy		311405	30000	0.011			

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

3. Dimensionality reduction approach

In this section, we describe the block-analysis we propose to provide unbiased estimates of the `id` of a data-view (subsection 3.1).

The automated analysis of the block-`id` distribution provides a quantitative information about the amount of feature noise and redundancy affecting the view (subsection 3.2) and, based on that, it allows tailoring the dimensionality reduction of the analyzed view.

To guide the reader, Figure 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.

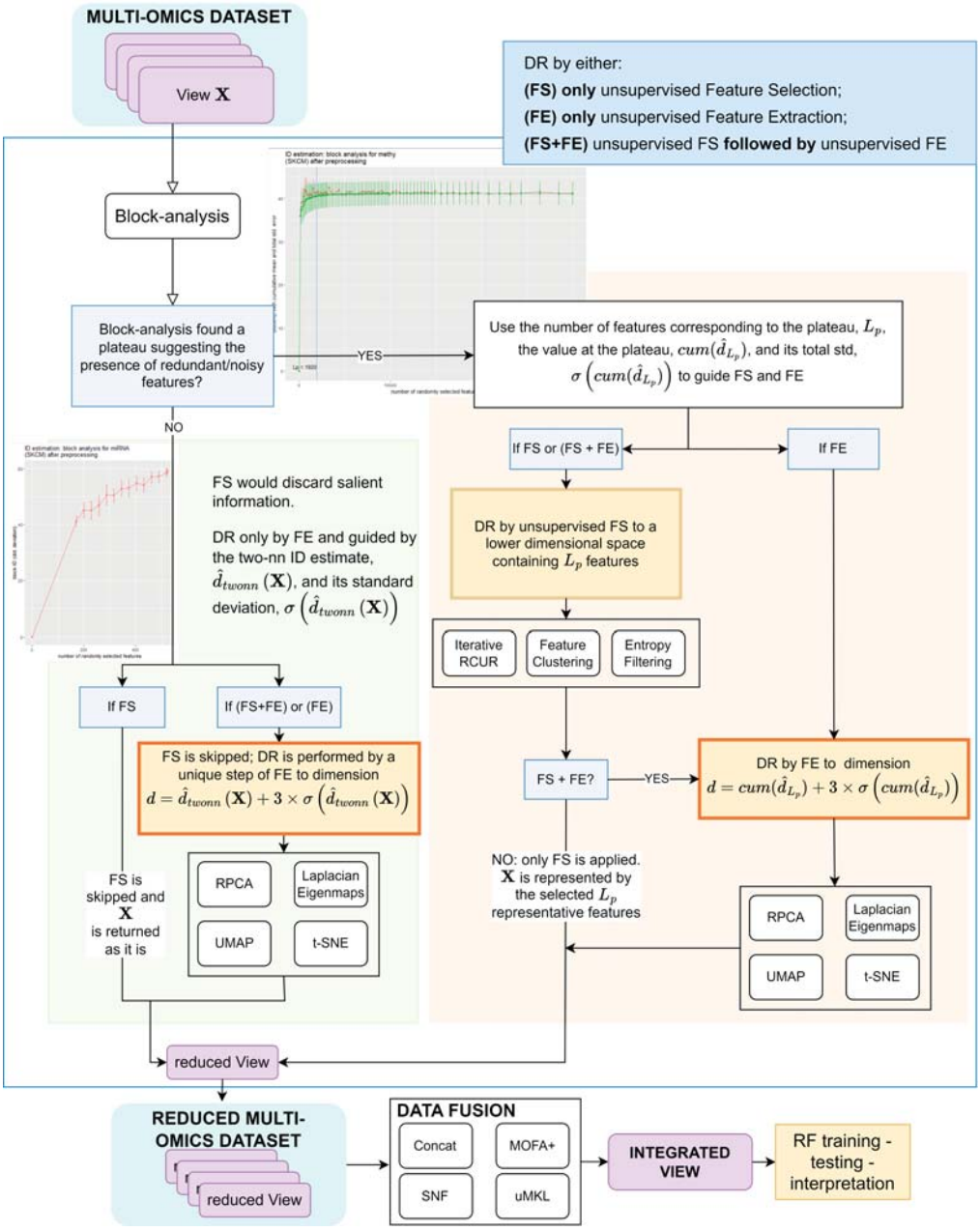


Figure 1: Experimented DR and data-integration pipelines.

3.1. Block analysis and block-ID estimate

Several of the most promising `id` estimators produce unstable global estimates on real datasets, often affected by the small-sample-size causing sample-sparsity, outliers, and noise [31, 32, 22] (see supplementary section S. A.2 for further details). This is particularly true for nearest-neighbor `id` estimators, which 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* [22] 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 [33]. In particular, authors 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. Plotting the distribution of the obtained estimates for increasing values of n, a plateau is found, corresponding to an unbiased (optimal) estimate of the `id` characterizing the informative content of the dataset.

The above-mentioned approach is effective on simulated experiments, where enough samples can be generated to avoid the curse of dimensionality. On the other hand, when dealing with real bio-medical datasets, often limited in sample-size and potentially affected by the curse of dimensionality, we propose to reduce the bias due to the presence of noisy and outlier points by 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[2]. Choosing a proper value for the percentage t allows to have enough samples in each sub-dataset M, so that the average (and the standard deviation) of all the M `id`-estimates may be a first, more robust, *two-nn* `id`-estimate (and standard deviation of the estimate) of the input dataset.

[2]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.

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.

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.

In particular, we start by using the *two-nn* `id`-estimate for the input view, $\hat{d}_{twonn}(\mathbf{X})$, to set the dimension L_0 of the smaller block as $L_0 = 3 \times \hat{d}_{twonn}(\mathbf{X})$. Though we are aware that this `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-`id`s.

More precisely, at the j^{th} iteration (j-th block $\mathbf{B}_j$), when the block size is $L_j = L_0 + j \times L_0$, $L_j \leq D$, 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[3].

3.2. Automatic analysis of Block-`id`s allows tailoring dimensionality reduction

The red dotted lines in figure 2 (and supplementary figures S. C.1-S. C.4) plot the block-`id`s, $\hat{d}_{L_j}$ for increasing block dimensions in the miRNA and pro-

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

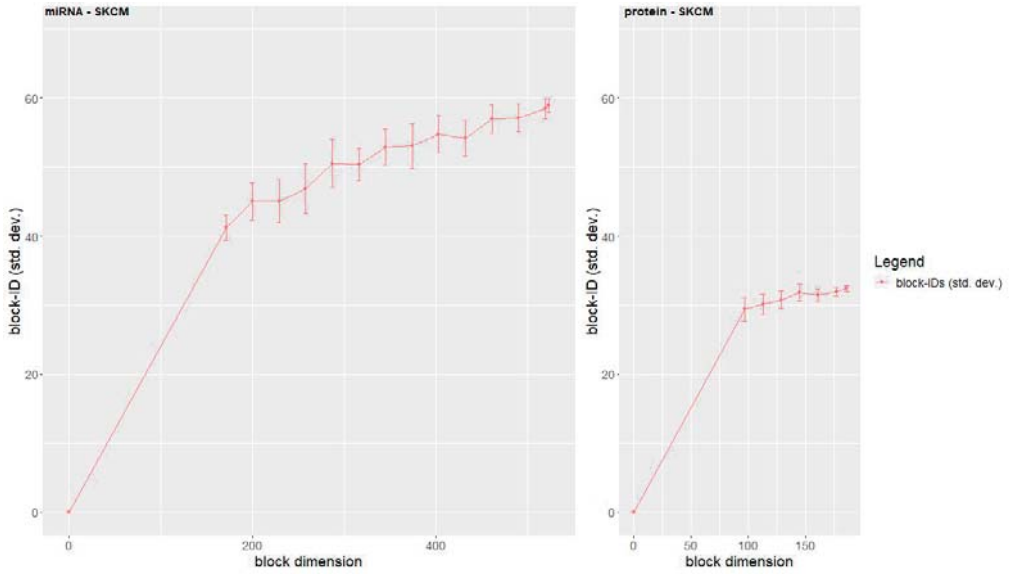


Figure 2: **Block-analysis performed by using the *two-nn* estimator on the SKCM dataset**. Left: miRNA view (SKCM dataset). Right: protein view (SKCM dataset). Point L_j of the red-dotted line (and the vertical bars) represents $\hat{d}_{L_j}$ (and its standard deviation $\sigma\left(\hat{d}_{L_j}\right)$), that is the estimated block-`id` for block $\mathbf{B}_j$, computed as the mean (and standard deviation) of the `id`s estimated on n_{try} blocks with dimension L_j (and its standard deviation). The block-`id` increases as the block dimension increases, suggesting that each of the added features increases the amount of information. 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 (figure 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 (figure 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.

tein views (SKCM dataset); red bars in the figure represent standard deviations, computed as the square root of the variance $\sigma\left(\hat{d}_{L_j}\right) = \sqrt{var\left(\hat{d}_{L_j}\right)}$.

When observing the miRNA and protein views, which are characterized by higher ratios $\frac{N}{D}$ when compared to the mRNA and methylation data-views, we note that the block-`id` keeps increasing until the block size includes all the features in the view, $L_j = D$, that is, the block-`id` equals the *two-nn* `id`-estimate of the whole view: $\hat{d}_{L_j=D} = \hat{d}_{twonn}(\mathbf{X})$ and $\sigma\left(\hat{d}_{L_j}\right) = \sigma\left(\hat{d}_{twonn}\left(\mathbf{X}\right)\right)$.

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).

In practice, *when no plateau is automatically detected by block-analysis* (see figure 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, which considers (and combines) all the features in the dataset (all the original information in the dataset) while computing the reduced view. In this case, the dimension of the reduced space, d where the view is transformed is computed by using the `id` estimate of the whole view: $d = \hat{d}_{twonn}(\mathbf{X}) + 3\sigma\left(\hat{d}_{twonn}(\mathbf{X})\right)$.

On the other hand, for the mRNA and the methylation view (figure 3) the distribution of the block-`ids` (red-dotted line) is more noisy, and increases until it reaches a (noisy) plateau. The dimension L_p of the block where the plateau starts (horizontal axis in figure 3, automatically detected as described in supplementary section S. C) can be regarded as an estimate of the minimum number of (salient) 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.

In practice, L_p is the number of the original features to be selected by an unsupervised feature selection algorithm to reduce noise and redundancy. While this step reduces the curse of dimensionality effects, the value of L_p is often high. To reduce the computational costs of the following algorithms and compute a data-representation concisely conveying similar information, we therefore propose applying a **two-step DR** approach where the reduced L_p-dimensional view is input to a feature extraction algorithm[4].

The feature extraction transforms the data into a space whose dimension, d, is computed based on an unbiased estimate of the view `id`, computed by considering that the block-`id` distribution is very noisy for views affected by the curse of dimensionality. To reduce noise effects by averaging, we compute the cumulative mean of the block-`ids` (green-dotted line in figure 3).

More precisely, the cumulative mean for block $\mathbf{B}_j$, $cum(\hat{d}_{L_j})$, is computed

[4]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.

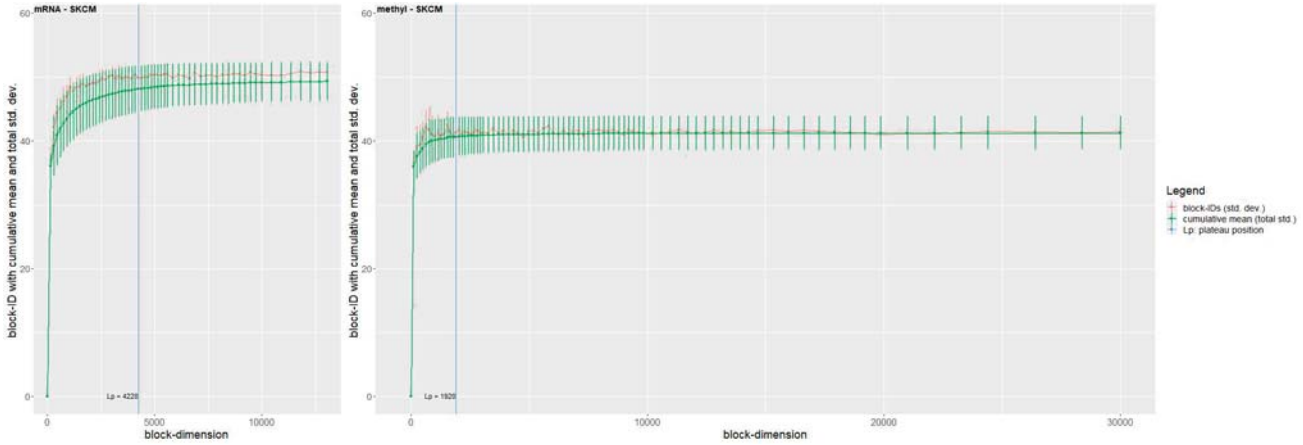


Figure 3: **Block-analysis performed by using the *two-nn* estimator on the SKCM dataset**. Left: mRNA; right: methylation data. The block-ids (red-dotted line) are more noisy than those in the miRNA and protein views. The effect of noise is reduced by the computation of the cumulative mean (green-dotted 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.

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 [34] allows computing the total variance of $cum(\hat{d}_{L_j})$, $var\left(cum(\hat{d}_{L_j})\right)$, as the 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[5].

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 $\mathbf{B}_j$.

Further, assuming some features are mostly carrying noise and/or redundant

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

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.

Summarizing, *when a plateau is automatically detected in position L_p of the block-`id` distribution* (see figure 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

In this section we first summarize the DR+data-fusion pipelines we devised and experimented (subsection 4.1); next, we detail the experimental settings of the supervised classification task we exploited to objectively compare the different pipelines (subsection 4.2); finally we report and discuss the results obtained by our comparative evaluation (subsection 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, figure 1 sketches all of them.

Dataset	view	L_p	$\hat{d}$	$\sigma(\hat{d})$
BLCA	miRNA		53.72	1.81
	mRNA	2835	42.75	2.85
	proteins		28.83	1.19
	methy	4238	52	3.48
BRCA1	miRNA		51.11	1.68
	mRNA	4640	46.61	3.18
	proteins		35.04	1.26
	methy	5738	48.57	3.55
BRCA2	miRNA		35.21	3.69
	mRNA	4216	42.26	3.57
	proteins		27.75	1.23
	methy	3000	39.87	3.04
KIRC	miRNA		53.26	5.55
	mRNA	4340	39.28	3.35
	proteins		31.1	1.6
	methy	2208	47.35	4.74
LUAD	miRNA		44.15	1.12
	mRNA	3168	43.32	2.72
	proteins		33.26	1.44
	methy	1677	43.27	3.36

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), and the `id` estimate $\hat{d}$ with its corresponding standard deviation $\sigma(\hat{d})$.

DR is performed by either a unique step of unsupervised feature selection (FS in figure 1), a unique step of unsupervised feature extraction (FE in figure 1), or by a 2-step DR process where the ouput of feature selection is input to feature extraction (FS+FE in figure 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

Dataset	**view**	L_p	$\hat{d}$	
LUSC	miRNA		50.77	1.41
	mRNA	4350	48.38	3.55
	proteins		35.79	0.84
	methy	3400	43.19	3.26
OV	miRNA		33.97	2.22
	mRNA	3173	55.19	4.78
	proteins		38.07	2.08
	methy	3614	44.14	2.93
PRAD	miRNA		54.04	1.95
	mRNA	3645	43.57	2.66
	proteins		27.43	0.97
	methy	7760	61.92	4.23
SKCM	miRNA		57.11	3.14
	mRNA	4228	48.6	3.39
	RPPA		32.45	1.07
	methy	1920	40.86	2.88

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), and the `id` estimate $\hat{d}$ with its corresponding standard deviation $\sigma(\hat{d})$.

literature background about unsupervised feature selection is reported in supplementary section S. A.1.1. 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 [35] 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 [36] 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 figure 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 figure 1).

Feature-extraction algorithms were similarly chosen based on their promising and successful results (supplementary section S. A.1.2). In detail, we compared Randomized PCA (RPCA, alias RSVD), laplacian eigenmaps, UMAP, and t-SNE and defined the dimension of the lower-dimensional space as $d = \hat{d} + 3 \times \sigma\left(\hat{d}\right)$, where $\hat{d}$ (and $\sigma\left(\hat{d}\right)$) is the estimated `id` (and its standard deviation). We recall that (section 3.1) a reliable `id` estimate (and standard deviation) of views not affected by feature redundancy is obtained as the mean (and standard deviation) of the *two-nn* estimates computed on M undersampled sets (light-green box in figure 1). When instead the block-analysis detects feature redundancy (light-red box in figure 1), the cumulative value at the plateau of the block-`id` distribution (and its total standard deviation) provides a reliable `id` estimate.

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 figure 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 [37, 11] (details in supplementary section S. A.3). In particular, we experimented with: (1) an input-data fusion technique, namely MOFA+ [21], 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, [13]), 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, [23]), 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.

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* 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 four feature-selection algorithms and the three 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

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, [38]) was trained and tested to predict the overall survival event of patients.

Besides their interpretable nature [39], their often superior effectiveness with respect to even the (less efficient) deep neural network models [40], and their capability of handling a set of heterogeneous variables [38], 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, and allows 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[6].

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 run 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 instead set all the RF parameters to their default values.

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 (supplementary files report also the details per dataset); 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[7]. 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 asses 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 test 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 promis-

[6]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.

[7]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

ing 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

After performing data-fusion tests with no prior DR that evidenced the need of a properly designed DR approach (supplementary section S. F.1), we conducted tests to compare the proposed DR approach (guided by block-analysis) to methods that exploit heuristics or empirical measures to 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 (subsection 4.3.2), to those settings where we fused subsets of omics (subsection 4.3.3), and omics plus non-omics views (subsection 4.3.4).

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

Besides comparing the described DR+data-fusion pipelines, in our experiments we also aimed 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 $min(N,D)-1$ [36, 41, 42]. 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. Besides avoiding empirical or heuristic choices, the assessment showed the promise 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; howver, also uMKL and MOFA+ show their promise.

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

To assess and compare 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, hold-outs, and data-fusion algorithms; (2) each data-fusion algorithm against each other, by pairing all results across datasets, holdouts, and DR pipelines. Figures 4 and figure 5 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 [18]. 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. F.13 and S. F.14 and supplementary tables S7 and S8).

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. F.14), 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). Figure 6 shows that the majority of top-performers use a two-step reduction schema, including RPCA and fusing data by means of SNF.

DR pipeline	data_integration	wins	ties	losses	auc	std(auc)
par_rcur		14	5	0	0.574	0.136
par_rcur+rpca		14	5	0	0.56	0.172
rpca		14	5	0	0.559	0.163
feature_clustering+rpca		13	6	0	0.564	0.157
entropy+rpca		13	6	0	0.559	0.152
feature_clustering		13	3	3	0.558	0.155
feature_clustering+umap		9	4	6	0.551	0.15
par_rcur+umap		8	5	6	0.552	0.147
entropy		8	5	6	0.542	0.153
umap		8	5	6	0.533	0.163
entropy+umap		8	4	7	0.544	0.128
par_rcur+laplacianEigenmaps		1	7	11	0.534	0.126
entropy+laplacianEigenmaps		0	8	11	0.534	0.13
laplacianEigenmaps		0	8	11	0.533	0.127
feature_clustering+laplacianEigenmaps		0	8	11	0.529	0.115
tsne		0	8	11	0.523	0.158
entropy+tsne		0	8	11	0.521	0.144
feature_clustering+tsne		0	8	11	0.513	0.158
par_rcur+tsne		0	7	12	0.51	0.16
	SNF	3	1	0	0.566	0.141
	uMKL	1	2	1	0.536	0.159
	concatenation	1	2	1	0.536	0.157
	MOFA+	0	1	3	0.529	0.133

Figure 4: Win-tie-loss tables obtained by Wilcoxon signed-rank test when using the AUC measure to compare (top table) all the DR pipelines exploiting the information provided by block-analysis to guide the reduction of each of the four omics views, and (bottom table) the data-fusion algorithms that fuse the reduced views (bottom table).

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. Win-tie-loss tables comparing results obtained when using multi-omics combinations, DR pipelines, and data-fusion algorithms are shown in figures 7 (when the AUC is used for evaluation) and figure 8 (when the AUCPR is used for evaluation). Extracts of win-tie-loss tables comparing the whole DR+data-fusion pipelines (also specified by the input multi-omics combination) are shown in sup-

DR pipeline	data_integration	wins	ties	losses	aucpr	std(aucpr)
par_rcur+rpca		15	4	0	0.512	0.142
rpca		15	4	0	0.503	0.16
feature_clustering+rpca		14	5	0	0.509	0.143
entropy+rpca		13	6	0	0.502	0.154
par_rcur		13	4	2	0.508	0.138
feature_clustering		13	3	3	0.494	0.137
entropy		11	2	6	0.495	0.134
feature_clustering+umap		10	3	6	0.483	0.152
umap		10	2	7	0.484	0.157
tsne		4	6	9	0.478	0.151
entropy+tsne		4	6	9	0.475	0.149
feature_clustering+tsne		4	6	9	0.474	0.149
par_rcur+tsne		4	6	9	0.474	0.149
par_rcur+umap		4	6	9	0.467	0.161
entropy+umap		4	6	9	0.455	0.173
laplacianEigenmaps		0	4	15	0.436	0.169
par_rcur+laplacianEigenmaps		0	4	15	0.434	0.166
entropy+laplacianEigenmaps		0	4	15	0.426	0.164
feature_clustering+laplacianEigenmaps		0	4	15	0.42	0.178
	SNF	3	1	0	0.502	0.15
	concatenation	1	2	1	0.489	0.134
	uMKL	1	2	1	0.487	0.154
	MOFA+	0	1	3	0.422	0.175

Figure 5: Win-tie-loss tables computed by Wilcoxon signed-rank test when using the AUCPR measure to compare all the DR pipelines guided by block-analysis (top table) and the data-fusion algorithms (bottom table). All the four omics are reduced and then integrated.

plementary figures S. F.15 and S. F.16, while supplementary files S10 and S11 report the complete win-tie-loss tables.

Regarding the comparison of the input-views (top tables in figures 7 and 8) win-tie-loss tables obtained with AUC seem suggesting that specific combinations of (mainly) three omics achieve results that are comparable and slightly better than those obtained when using four omics. However, when observing the win-tie-loss table obtained by using the AUCPR we note that the usage of four omics is comparable to only one combination of three omics (which, again, scores slightly better) and one combination of two omics. Considering that most of the datasets used for our tests are unbalanced, i.e. the AUCPR measure is less biased and more informative while the AUC might be over-optimistic [43], we can infer that

Figure 6: 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.

the usage of a superior number of omics, i.e. a superior number of features potentially adding more informative content but also some noise and redundancy, does not affect performance but instead achieves results comparable to specific combinations of omics and guarantees robustness with respect to data unbalance. 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 (supplementary section S. F.1). Therefore, provided that a proper DR is applied, the usage of all the available omics allows to obtain robust results and to avoid costly experiments to choose the most suitable combination of omics given the problem at hand.

Among the DR pipelines, we again note that 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. On the other hand, 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.

To provide an exhaustive description of the obtained results, for each dataset we collected the list of experiments that had the highest AUC or AUCPR. Figure 9 plots the frequency of appearance of each multi-omics combination, DR pipeline, and data-fusion algorithm (supplementary table S12 lists all the best models and their performance). For what regards the composition of the input multi-omics combinations, all views but the miRNA-view equally contributed in obtaining

data	DR pipeline	data_integration	wins	ties	losses	auc	std(auc)
Proteins_methy			5	6	0	0.543	0.143
miRNA_Proteins_methy			5	6	0	0.543	0.144
miRNA_mRNA_Proteins			5	6	0	0.542	0.148
miRNA_mRNA_Proteins_methy			4	7	0	0.542	0.149
mRNA_Proteins_methy			4	7	0	0.542	0.149
mRNA_Proteins			4	7	0	0.541	0.152
miRNA_Proteins			4	4	3	0.542	0.143
miRNA_methy			2	2	7	0.538	0.141
miRNA_mRNA_methy			2	2	7	0.537	0.144
mRNA_methy			0	2	9	0.536	0.142
miRNA_mRNA			0	2	9	0.535	0.148
	par_rcur		18	1	0	0.571	0.147
	feature_clustering		16	2	1	0.566	0.152
	feature_clustering+rpca		16	2	1	0.565	0.156
	par_rcur+rpca		15	1	3	0.559	0.157
	entropy+rpca		13	2	4	0.558	0.149
	rpca		13	2	4	0.557	0.158
	entropy		12	1	6	0.555	0.153
	feature_clustering+umap		11	1	7	0.544	0.143
	par_rcur+umap		10	1	8	0.545	0.141
	umap		9	1	9	0.537	0.149
	entropy+umap		8	1	10	0.536	0.134
	laplacianEigenmaps		7	1	11	0.534	0.124
	par_rcur+laplacianEigenmaps		6	1	12	0.531	0.128
	feature_clustering+laplacianEigenmaps		5	1	13	0.529	0.118
	entropy+laplacianEigenmaps		4	1	14	0.529	0.124
	tsne		3	1	15	0.517	0.152
	entropy+tsne		2	1	16	0.513	0.146
	feature_clustering+tsne		0	2	17	0.507	0.152
	par_rcur+tsne		0	2	17	0.507	0.151
		SNF	3	1	0	0.559	0.147
		concatenation	2	1	1	0.537	0.157
		uMKL	1	1	2	0.535	0.152
		MOFA+	0	1	3	0.529	0.122

Figure 7: Win-tie-loss tables computed by Wilcoxon signed-rank test when using the AUC measure to compare models guided by block-analysis, integrating at least two omics, and neglecting demographic data.

a good performance. Moreover, while less robust when compared to combinations of three/four omics by Wilcoxon signed-rank tests, also 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 optimize performance.

data	DR pipeline	data_integration	wins	ties	losses	aucpr	std(aucpr)
mRNA_Proteins_methy			10	1	0	0.477	0.158
miRNA_mRNA_Proteins_methy			8	2	1	0.475	0.157
Proteins_methy			8	2	1	0.475	0.163
miRNA_Proteins			5	3	3	0.475	0.157
miRNA_mRNA_Proteins			5	3	3	0.472	0.16
miRNA_Proteins_methy			5	3	3	0.472	0.161
mRNA_Proteins			4	1	6	0.47	0.162
miRNA_mRNA_methy			3	1	7	0.467	0.154
miRNA_methy			1	2	8	0.463	0.156
mRNA_methy			1	2	8	0.463	0.156
miRNA_mRNA			0	1	10	0.461	0.155
	par_rcur		18	1	0	0.508	0.14
	feature_clustering+rpca		17	1	1	0.504	0.149
	entropy		14	3	2	0.503	0.14
	feature_clustering		13	4	2	0.503	0.139
	par_rcur+rpca		13	4	2	0.501	0.151
	rpca		12	4	3	0.499	0.154
	entropy+rpca		12	2	5	0.499	0.148
	umap		11	1	7	0.47	0.164
	feature_clustering+umap		10	1	8	0.469	0.155
	tsne		8	2	9	0.464	0.148
	par_rcur+umap		8	2	9	0.461	0.162
	entropy+umap		7	1	11	0.457	0.167
	entropy+tsne		5	2	12	0.46	0.15
	feature_clustering+tsne		5	2	12	0.459	0.15
	par_rcur+tsne		4	1	14	0.457	0.152
	laplacianEigenmaps		3	1	15	0.435	0.172
	par_rcur+laplacianEigenmaps		2	1	16	0.432	0.167
	entropy+laplacianEigenmaps		0	2	17	0.425	0.166
	feature_clustering+laplacianEigenmaps		0	2	17	0.424	0.176
		SNF	3	1	0	0.5	0.147
		concatenation	2	1	1	0.489	0.136
		uMKL	1	1	2	0.481	0.154
		MOFA+	0	1	3	0.41	0.177

Figure 8: Win-tie-loss tables computed by Wilcoxon signed-rank test when using the AUCPR measure to compare models guided by block-analysis, integrating at least two omics, and neglecting demographic data.

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'

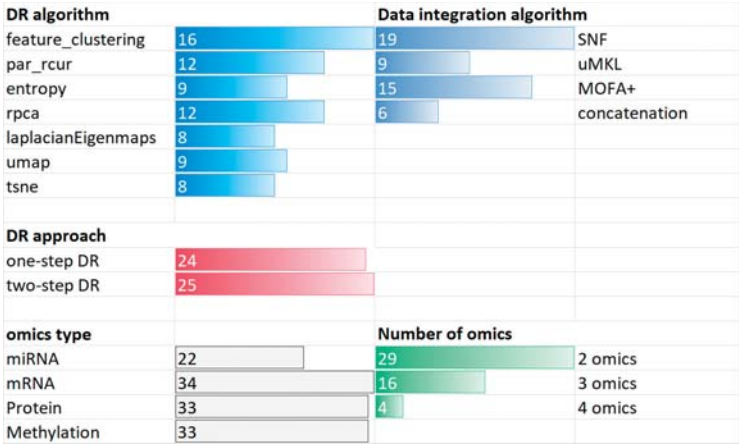


Figure 9: Frequencies of multi-omics combinations, DR pipelines, and data-fusion methods appearing among the best models when at least two omics are fused and the block-analysis guides the DR.

descriptors, and considering the documented literature interest about challenges regarding the integration of omics and non-omics views [44] (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[8] 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 com-

[8]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 the data types of those variables.

data	DR pipeline	data_integration	wins	ties	losses	auc	std(auc)
4 omics + pt			1	1	0	0.552	0.147
4 omics			0	1	1	0.542	0.149
	par_rcur		16	3	0	0.575	0.141
	feature_clustering+rpca		15	4	0	0.57	0.151
	rpca		15	4	0	0.568	0.157
	par_rcur+rpca		15	3	1	0.565	0.166
	entropy+rpca		14	1	4	0.565	0.148
	feature_clustering		13	1	5	0.559	0.156
	feature_clustering+umap		12	1	6	0.553	0.146
	par_rcur+umap		10	2	7	0.552	0.139
	entropy		9	3	7	0.546	0.155
	entropy+umap		9	2	8	0.547	0.143
	umap		8	1	10	0.536	0.161
	par_rcur+laplacianEigenmaps		5	3	11	0.543	0.127
	entropy+tsne		5	3	11	0.537	0.144
	tsne		5	3	11	0.536	0.158
	entropy+laplacianEigenmaps		0	5	14	0.539	0.128
	laplacianEigenmaps		0	5	14	0.538	0.127
	feature_clustering+laplacianEigenmaps		0	5	14	0.535	0.114
	feature_clustering+tsne		0	5	14	0.526	0.159
	par_rcur+tsne		0	5	14	0.525	0.16
		SNF + PT data	6	1	0	0.582	0.146
		SNF	5	1	1	0.568	0.141
		uMKL	4	1	2	0.541	0.159
		MOFA+ + PT data	2	2	3	0.547	0.14
		concatenation	2	2	3	0.538	0.157
		MOFA+	1	1	5	0.535	0.138
		uMKL + PT data	0	1	6	0.532	0.135

Figure 10: Top table: Win-tie-loss tables obtained when assessing the integration of omics and non-omics descriptors ("4 omics+pt") by comparing it to the usage of only the omics descriptors ("4 omics"). Center table: DR pipelines are compared when 4 omics and demographic descriptors are integrated. Bottom table: data-fusion algorithms are compared when 4 omics and demographic descriptors are integrated either by concatenating demographic descriptors to the integrated representation ("SNF", "uMKL", "MOFA+", and "concatenation"), or by using the non-omics descriptors as a further view to be integrated ("SNF + PT data", "uMKL + PT data", "MOFA+ + PT data"). AUC is used when computing the Wilcoxon signed-rank test.

parative evaluation we limited the number of experiments to those that used all the available omics.

Figures 10 and 11 report the win-tie-loss tables computed by Wilcoxon signed-rank test when comparing the integration of omics and non-omics descriptors to the usage of only omics-views. Supplementary figures S. F.17 and S. F.18 show the win-tie-loss tables obtained when comparing full pipelines character-

data	DR pipeline	data_integration	wins	ties	losses	aucpr	std(aucpr)
4 omics + pt			1	1	0	0.475	0.161
4 omics			0	1	1	0.475	0.157
	par_rcur+rpca		16	3	0	0.511	0.147
	feature_clustering+rpca		15	4	0	0.508	0.148
	rpca		14	5	0	0.507	0.157
	par_rcur		14	4	1	0.507	0.145
	entropy+rpca		14	3	2	0.504	0.149
	feature_clustering		12	2	5	0.494	0.141
	entropy		12	2	5	0.493	0.138
	umap		10	2	7	0.482	0.147
	feature_clustering+umap		10	2	7	0.481	0.151
	tsne		6	4	9	0.478	0.162
	entropy+tsne		6	4	9	0.476	0.158
	par_rcur+tsne		6	4	9	0.475	0.159
	feature_clustering+tsne		5	5	9	0.475	0.158
	entropy+umap		5	2	12	0.468	0.157
	par_rcur+umap		4	1	14	0.463	0.159
	laplacianEigenmaps		2	2	15	0.435	0.175
	par_rcur+laplacianEigenmaps		2	2	15	0.434	0.172
	entropy+laplacianEigenmaps		1	1	17	0.423	0.167
	feature_clustering+laplacianEigenmaps		0	1	18	0.415	0.181
		SNF + PT data	6	1	0	0.51	0.141
		SNF	5	1	1	0.504	0.149
		uMKL	4	1	2	0.491	0.152
		concatenation	3	1	3	0.492	0.135
		MOFA+ + PT data	1	2	4	0.441	0.178
		MOFA+	1	2	4	0.433	0.172
		uMKL + PT data	0	1	6	0.435	0.175

Figure 11: Win-tie-loss tables obtained when assessing the integration of omics and non-omics descriptors ("4 omics+pt") by comparing it to the usage of only the omics descriptors ("4 omics"). AUCPR is used when computing the Wilcoxon signed-rank test.

ized by a specific set of views being integrated, and a specific DR+data-fusion pipeline (complete tables can be found in supplementary tables S13 and S14). 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.

Figure 12 plots the frequency of each data view, DR algorithm, and data-fusion method appearing in the top-performing experiments (listed in supplementary ta-

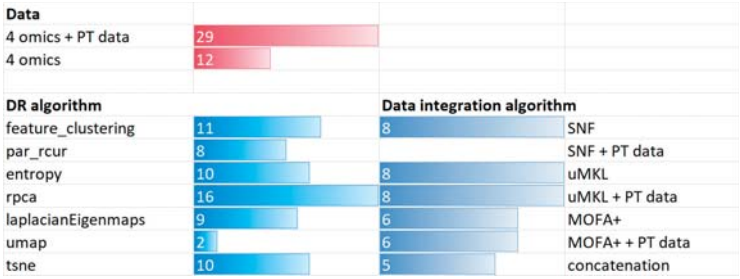


Figure 12: Frequencies of DR methods, and data integration algorithms appearing among the top-performing models when all the views (omics and non-omics) are analyzed.

ble S15).

Observing the results we understand that surely the inclusion of patient 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 table S15 and figure 12), 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 figures S. F.17 and S. F.18) 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.

5. Discussion and Conclusions

In this paper, we have described a novel application of block-analysis to leverage any of the most promising `id` estimators and obtain an unbiased `id`-estimate of the views in a multi-modal dataset. We also proposed an automatic analysis of the block-`id` distribution computed by the block-analysis to detect feature noise and redundancy contributing to the curse of dimensionality and therefore evidence the need to apply a view-specific dimensionality reduction phase (guided by the `id` estimate) prior to any subsequent analysis to reduce curse-of-dimensionality effects.

Using the proposed `id` analysis we can therefore automatically take view-specific decisions so that views containing higher levels of noise and redundancy can undergo a two-step dimensionality reduction approach combining the advantages of feature-selection and feature-extraction; on the other hand, views less affected by the curse of dimensionality may be reduced by traditional feature-extraction approaches.

Besides assessing our proposal by using nine heterogeneous multi-omics cancer datasets from the TCGA repository, we analyzed and compared the DR effects on the subsequent application of data-fusion techniques that have shown their promise in the field of multi-omics. To this aim, we used the fused view to predict overall survival events by means of RF classifiers, often preferred in the biomedical field due to their interpretable nature, relative robustness to hyper-parameter settings, superior efficiency, and effectiveness in many competitions [40].

The results we obtained first evidenced that a properly designed DR step is crucial and should never be neglected when complex multi-omics data is processed. Secondly, we showed that DR approaches guided by block-analysis were superior to traditional DR approaches setting the dimension of the reduced space by some heuristics or by preliminary empirical experiments.

When analyzing the impact of the proposed DR approach on different multi-omics fusion settings we first noticed that the two-step DR approach can be an effective solution. Particularly, in our classification task the most robust and effective results were achieved when combining the iterative version of RCUR we implemented (or feature clustering) with RPCA, whose formulation is simpler and more explainable than, e.g., UMAP and t-SNE.

When observing the robustness and performance of the experimented data-fusion algorithm we confirmed the efficiency and effectiveness of SNF, which showed its robustness with respect to different settings. MOFA+ also showed its promise, though its robustness and efficiency were lower than that of SNF. However, the advantage of MOFA+ relies on its ability to deal with heterogeneous data types, so that it can effectively integrate omics and non-omics views.

Regarding two different and crucial data-fusion settings we were interested in investigating, we 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 are able to cope with the presence of potentially redundant information within and between the four omics views, so that all the available omics types can be used without the need to try all their different combinations. In other words, the design of a proper DR can facilitate the following data-fusion task by effectively removing redundancies within individual data types, while better exposing their characterizing informative content. This facilitates the task of the following data-fusion algorithms, which must only deal with redundancies across views while uncovering the shared and individual informative content of the multiple omics. This allows to avoid costly empirical experiments to choose the subset of omics to be integrated.

We further assessed the addition of patients' demographic descriptors in the analysis and showed that it effectively increases the classification performance.

Dataset	N	d	unsup feature selection	unsup feature extraction	data_integration	aucpr	std(aucpr)	auc	std(auc)
BLCA	335	21	par_rcur		MOFA+	0.720	0.055	0.720	0.064
		21		rpca	MOFA+	0.705	0.062	0.673	0.064
		341	feature_clustering	rpca	uMKL	0.704	0.102	0.719	0.095
		21	feature_clustering		MOFA+	0.687	0.047	0.652	0.043
		21	feature_clustering	rpca	MOFA+	0.671	0.071	0.659	0.063
		21	par_rcur		MOFA+	0.715	0.062	0.723	0.060
		21		rpca	MOFA+	0.698	0.074	0.677	0.060
		21	feature_clustering		MOFA+	0.682	0.056	0.662	0.034
BRCA1	317	322	par_rcur		SNF	0.533	0.127	0.749	0.095
		322	par_rcur		uMKL	0.523	0.123	0.763	0.142
		6861	par_rcur		concatenat	0.491	0.142	0.694	0.117
		322		rpca	SNF	0.489	0.172	0.739	0.134
		20		rpca	MOFA+	0.433	0.150	0.698	0.098
BRCA2	128	20	par_rcur		MOFA+	0.917	0.144	0.961	0.072
		14	feature_clustering	rpca	MOFA+	0.867	0.207	0.886	0.177
		133	par_rcur		uMKL	0.814	0.184	0.870	0.165
		20	feature_clustering		MOFA+	0.775	0.166	0.864	0.157
		133	feature_clustering		uMKL	0.720	0.166	0.788	0.157
KIRC	169	7110	feature_clustering		concatenat	0.751	0.086	0.809	0.083
		175	feature_clustering	rpca	uMKL	0.703	0.155	0.748	0.127
		4848	par_rcur		concatenat	0.699	0.098	0.777	0.110
		20	par_rcur	rpca	MOFA+	0.682	0.169	0.758	0.132
		175	feature_clustering		SNF	0.674	0.153	0.764	0.111
		7110	feature_clustering		concatenat	0.728	0.118	0.823	0.067
		175	feature_clustering	rpca	uMKL	0.700	0.160	0.753	0.121
		4848	par_rcur		concatenat	0.675	0.139	0.786	0.096
LUAD	300	197	feature_clustering	rpca	concatenat	0.678	0.066	0.660	0.053
		305	feature_clustering	rpca	uMKL	0.638	0.077	0.643	0.059
		305	par_rcur	rpca	uMKL	0.636	0.036	0.608	0.028
		20	par_rcur	rpca	MOFA+	0.631	0.061	0.659	0.055
		20	par_rcur		MOFA+	0.619	0.088	0.672	0.070
		20	par_rcur	rpca	MOFA+	0.624	0.075	0.660	0.051
LUSC	228	20	feature_clustering	rpca	MOFA+	0.670	0.105	0.651	0.085
		233	feature_clustering		uMKL	0.652	0.106	0.685	0.086
		215		rpca	concatenat	0.634	0.083	0.626	0.038
		233	par_rcur		uMKL	0.632	0.057	0.692	0.051
		233		rpca	SNF	0.631	0.111	0.642	0.080

Figure 13: Classification performance obtained one the BLCA, BRCA1, BRCA2, KIRC, LUAD, and LUSC datasets when using the block-analysis to guide the DR of all four omics. After reduced data-fusion the demographics views are concatenated and used as input for supervised feature selection, RF tuning, RF training, and sample classification.

OV	226	229		rpca	uMKL	0.861	0.046	0.732	0.088
		211	par_rcur	rpca	concatenat	0.859	0.042	0.756	0.051
		229	feature_clustering	rpca	uMKL	0.854	0.046	0.727	0.069
		18	par_rcur		MOFA+	0.835	0.070	0.718	0.097
		229	par_rcur	rpca	uMKL	0.829	0.045	0.665	0.077
		229		rpca	uMKL	0.859	0.049	0.735	0.085
		229	feature_clustering	rpca	uMKL	0.852	0.048	0.729	0.065
		18	par_rcur		MOFA+	0.835	0.070	0.721	0.093
		229	par_rcur	rpca	uMKL	0.824	0.052	0.669	0.071
PRAD	337	220		rpca	concatenat	0.652	0.233	0.777	0.212
		18	par_rcur	rpca	MOFA+	0.617	0.097	0.860	0.191
		18	par_rcur		MOFA+	0.600	0.084	0.834	0.161
		340	feature_clustering	rpca	uMKL	0.577	0.078	0.833	0.129
		220	feature_clustering	rpca	concatenat	0.560	0.051	0.825	0.100
SKCM	334	218		rpca	concatenat	0.748	0.066	0.756	0.059
		340	par_rcur	rpca	SNF	0.735	0.084	0.715	0.083
		340	par_rcur		SNF	0.732	0.100	0.724	0.088
		340	feature_clustering		SNF	0.724	0.094	0.714	0.090
		340		rpca	SNF	0.722	0.085	0.706	0.083
		340	par_rcur		SNF	0.729	0.104	0.726	0.083
		340	par_rcur	rpca	SNF	0.726	0.096	0.718	0.077
		340	feature_clustering		SNF	0.722	0.097	0.715	0.087
		218	par_rcur	rpca	concatenat	0.714	0.106	0.725	0.079

Figure 14: Classification performance obtained on the OV, PRAD, and SKCM datasets when using the block-analysis to guide the DR of all four omics. After reduced data-fusion the demographics views are concatenated and used as input for supervised feature selection, RF tuning, RF training, and sample classification.

Concluding, all the experiments led us to the definition of a DR+data-fusion pipeline that obtains promising results without the need for empirical and heuristically based choices. In particular, considering that all the results reported in our experiments (besides being averaged across all the nine TCGA datasets) were obtained when applying neither supervised feature selection nor hyper-parameter tuning to avoid confounding effects that would bias the comparative assessment, we ran the last experiment where we used all the four omics as input to data fusion, we concatenated the fused representation with demographic descriptors and then optimized the RF performance by supervised feature selection through RF importance [39] and hyper-parameter tuning through internal stratified holdout validation (more details are reported in supplementary section S. F.4).

This procedure allowed obtaining more than satisfactory results (see figures 13 and 14), with AUCs often greater than $0.70/0.75$ and large AUCPRs even for datasets characterized by a low (≤ 0.2) ratio between the number of positive

Dataset	N	d	data fusion	aucpr	std(aucpr)	auc	std(auc)
BLCA	335	341	SNF	0.58	0.05	0.62	0.04
		21	MOFA+	0.54	0.05	0.59	0.05
		341	uMKL	0.55	0.05	0.62	0.05
BRCA1	317	322	SNF	0.21	0.03	0.50	0.05
		20	MOFA+	0.23	0.04	0.53	0.07
		322	uMKL	0.29	0.08	0.57	0.09
BRCA2	128	133	SNF	0.35	0.05	0.46	0.16
		20	MOFA+	0.43	0.18	0.53	0.25
		133	uMKL	0.46	0.21	0.48	0.26
KIRC	169	175	SNF	0.45	0.10	0.64	0.09
		21	MOFA+	0.44	0.07	0.65	0.07
		175	uMKL	0.46	0.08	0.62	0.09
LUAD	300	305	SNF	0.49	0.06	0.57	0.06
		20	MOFA+	0.54	0.05	0.60	0.06
		305	uMKL	0.46	0.05	0.54	0.05
LUSC	228	233	SNF	0.48	0.07	0.57	0.07
		20	MOFA+	0.42	0.05	0.45	0.06
		233	uMKL	0.50	0.07	0.57	0.08
OV	226	229	SNF	0.69	0.04	0.52	0.07
		18	MOFA+	0.65	0.06	0.47	0.07
		229	uMKL	0.69	0.06	0.54	0.09
PRAD	337	340	SNF	0.52	0.00	0.53	0.08
		18	MOFA+	0.51	0.01	0.26	0.28
		340	uMKL	0.51	0.00	0.25	0.09
SKCM	334	340	SNF	0.65	0.05	0.67	0.04
		21	MOFA+	0.54	0.06	0.61	0.06
		340	uMKL	0.65	0.06	0.65	0.04

Figure 15: Classification performance obtained when integrating all four omics and then concatenating the demographics views. No prior DR is performed but supervised feature selection and hyper-parameter tuning are applied prior to RF training.

cases and the total number of cases. Of note, these results greatly outperform the results we obtained when we avoided any DR step prior to data-fusion, eventual concatenation of demographic descriptors, and RF optimization via supervised feature selection and hyper-parameter tuning (supplementary figure S. F.6 and figure 15).

Acknowledgements

Authors would like to thank Prof. Juho Rouso (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.

Funding

This research was supported by the National Center for Gene Therapy and Drugs based on RNA Technology, PNRR-NextGenerationEU program (G43C22001320007). It 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.

6. Conflict of interest statement

The authors declare no conflict of Interest

References

[1] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome biology 18 (1) (2017) 1–15.

[2] X. Dai, L. Shen, Advances and trends in omics technology development, Frontiers in Medicine 9 (2022) 911861.

[3] E. Athieniti, G. M. Spyrou, A guide to multi-omics data collection and integration for translational medicine, Computational and Structural Biotechnology Journal 21 (2023).

[4] A. Conesa, S. Beck, Making multi-omics data accessible to researchers, Scientific data 6 (1) (2019) 251.

[5] M. Babu, M. Snyder, Multi-omics profiling for health, Molecular & Cellular Proteomics 22 (6) (2023).

[6] I. Subramanian, S. Verma, S. Kumar, A. Jere, K. Anamika, Multi-omics data integration, interpretation, and its application, Bioinformatics and biology insights 14 (2020) 1177932219899051.

[7] G. V. Trunk, A problem of dimensionality: A simple example, IEEE Transactions on pattern analysis and machine intelligence PAMI-1 (3) (1979) 306–307.

[8] J. Lv, Impacts of high dimensionality in finite samples, The Annals of Statistics 41 (4) (2013) 2236–2262.

[9] G. Hughes, On the mean accuracy of statistical pattern recognizers, IEEE transactions on information theory 14 (1) (1968) 55–63.

[10] S. Nanga, A. T. Bawah, B. A. Acquaye, M.-I. Billa, F. D. Baeta, N. A. Odai, S. K. Obeng, A. D. Nsiah, Review of dimension reduction methods, Journal of Data Analysis and Information Processing 9 (3) (2021) 189–231.

[11] J. Gliozzo, M. Mesiti, M. Notaro, A. Petrini, A. Patak, A. Puertas-Gallardo, A. Paccanaro, G. Valentini, E. Casiraghi, Heterogeneous data integration methods for patient similarity networks, Briefings in Bioinformatics (2022).

[12] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, X. Chen, A comparison for dimensionality reduction methods of single-cell rna-seq data, Frontiers in genetics 12 (2021) 646936.

[13] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature methods 11 (3) (2014) 333–337.

[14] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, R. Shen, Pattern discovery and cancer gene identification in integrated cancer genomic data, Proceedings of the National Academy of Sciences 110 (11) (2013) 4245–4250.

[15] N. Rappoport, R. Shamir, Nemo: cancer subtyping by integration of partial multi-omic data, Bioinformatics 35 (18) (2019) 3348–3356.

[16] H. Nguyen, S. Shrestha, S. Draghici, T. Nguyen, Pinsplus: a tool for tumor subtype discovery in integrated genomic data, Bioinformatics 35 (16) (2019) 2843–2846.

[17] L. H. Nguyen, S. Holmes, Ten quick tips for effective dimensionality reduction, PLoS computational biology 15 (6) (2019) e1006907.

[18] H. Nguyen, D. Tran, B. Tran, M. Roy, A. Cassell, S. Dascalu, S. Draghici, T. Nguyen, Smrt: Randomized data transformation for cancer subtyping and big data analysis, Frontiers in oncology 11 (2021).

[19] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, R. Bellazzi, Integrated multi-omics analyses in oncology: a review of machine learning methods and tools, Frontiers in oncology 10 (2020) 1030.

[20] E. F. Lock, K. A. Hoadley, J. S. Marron, A. B. Nobel, Joint and individual variation explained (jive) for integrated analysis of multiple data types, The annals of applied statistics 7 (1) (2013) 523.

[21] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, O. Stegle, Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data, Genome biology 21 (1) (2020) 1–17.

[22] E. Facco, M. d'Errico, A. Rodriguez, A. Laio, Estimating the intrinsic dimension of datasets by a minimal neighborhood information, Scientific reports 7 (1) (2017) 1–8.

[23] J. Mariette, N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, Bioinformatics 34 (6) (2018) 1009–1015.

[24] M. Ramos, L. Geistlinger, S. Oh, L. Schiffer, R. Azhar, H. Kodali, I. de Bruijn, J. Gao, V. J. Carey, M. Morgan, et al., Multiomic integration of public oncology databases in bioconductor, JCO Clinical Cancer Informatics 1 (2020) 958–971.

[25] J. Liu, T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee, et al., An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics, Cell 173 (2) (2018) 400–416.

[26] S. Pai, S. Hui, R. Isserlin, M. A. Shah, H. Kaka, G. D. Bader, netdx: interpretable patient classification using integrated patient similarity networks, Molecular systems biology 15 (3) (2019) e8497.

[27] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, K. Huang, Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, Nature Communications 12 (1) (2021) 3445.

[28] S. Moon, H. Lee, Moma: a multi-task attention learning algorithm for multi-omics data interpretation and classification, Bioinformatics 38 (8) (2022) 2287–2296.

[29] Y. Zhong, Y. Peng, Y. Lin, D. Chen, H. Zhang, W. Zheng, Y. Chen, C. Wu, Modilm: towards better complex diseases classification using a novel multi-omics data integration learning model, BMC Medical Informatics and Decision Making 23 (1) (2023) 1–18.

[30] D. Ouyang, Y. Liang, L. Li, N. Ai, S. Lu, M. Yu, X. Liu, S. Xie, Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification, Computers in Biology and Medicine 164 (2023) 107303.

[31] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, P. Campadelli, Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration, Pattern recognition 47 (8) (2014) 2569–2581.

[32] P. Campadelli, E. Casiraghi, C. Ceruti, A. Rozza, Intrinsic dimension estimation: Relevant techniques and a benchmark framework, Mathematical Problems in Engineering 2015 (2015).

[33] R. Badii, A. Politi, Hausdorff dimension and uniformity factor of strange attractors, Physical review letters 52 (19) (1984) 1661.

[34] J. K. Blitzstein, J. Hwang, Introduction to probability, Crc Press, 2019.

[35] M. Gagolewski, genieclust: Fast and robust hierarchical clustering, SoftwareX 15 (2021) 100722. doi:https://doi.org/10.1016/j.softx.2021.100722.
URL https://www.sciencedirect.com/science/article/pii/S2352711021000649

[36] M. W. Mahoney, P. Drineas, Cur matrix decompositions for improved data analysis, Proceedings of the National Academy of Sciences 106 (3) (2009) 697–702.

[37] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype–phenotype interactions, Nature Reviews Genetics 16 (2) (2015) 85–97.

[38] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[39] E. Casiraghi, D. Malchiodi, G. Trucco, M. Frasca, L. Cappelletti, T. Fontana, A. A. Esposito, E. Avola, A. Jachetti, J. Reese, et al., Explainable machine learning for early assessment of covid-19 risk prediction in emergency departments, Ieee Access 8 (2020) 196299–196325.

[40] Z.-H. Zhou, J. Feng, Deep forest, National science review 6 (1) (2019) 74–86.

[41] V. Rokhlin, A. Szlam, M. Tygert, A randomized algorithm for principal component analysis, SIAM Journal on Matrix Analysis and Applications 31 (3) (2010) 1100–1124.

[42] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural computation 15 (6) (2003) 1373–1396.

[43] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 233–240.

[44] E. López de Maturana, L. Alonso, P. Alarcón, I. A. Martín-Antoniano, S. Pineda, L. Piorno, M. L. Calle, N. Malats, Challenges in the integration of omics and non-omics data, Genes 10 (3) (2019) 238.

[45] J. Xie, M. Wang, S. Xu, Z. Huang, P. W. Grant, The unsupervised feature selection algorithms based on standard deviation and cosine similarity for genomic data analysis, Frontiers in Genetics 12 (2021) 684100.

[46] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC bioinformatics 18 (1) (2017) 1–14.

[47] B. F. Darst, K. C. Malecki, C. D. Engelman, Using recursive feature elimination in random forest to account for correlated variables in high dimensional data, BMC genetics 19 (1) (2018) 1–6.

[48] M. B. Kursa, A. Jankowski, W. R. Rudnicki, Boruta–a system for feature selection, Fundamenta Informaticae 101 (4) (2010) 271–285.

[49] E. Hancer, B. Xue, M. Zhang, A survey on feature selection approaches for clustering, Artificial Intelligence Review 53 (2020) 4519–4545.

[50] S. Solorio-Fernández, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, A review of unsupervised feature selection methods, Artificial Intelligence Review 53 (2) (2020) 907–948.

[51] N. Halko, P.-G. Martinsson, J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM review 53 (2) (2011) 217–288.

[52] S. Voronin, P.-G. Martinsson, Efficient algorithms for cur and interpolative matrix decompositions, Advances in Computational Mathematics 43 (3) (2017) 495–516.

[53] S. Vaithyanathan, B. Dom, Generalized model selection for unsupervised learning in high dimensions, Advances in neural information processing systems 12 (1999).

[54] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 1151–1157.

[55] N. B. Erichson, S. Voronin, S. L. Brunton, J. N. Kutz, Randomized matrix decompositions using r, arXiv preprint arXiv:1608.02148 (2016).

[56] P. Drineas, M. W. Mahoney, Randnla: randomized numerical linear algebra, Communications of the ACM 59 (6) (2016) 80–90.

[57] L. Jiang, Y. Xiao, Y. Ding, J. Tang, F. Guo, Discovering cancer subtypes via an accurate fusion strategy on multiple profile data, Frontiers in genetics 10 (2019) 20.

[58] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11) (2008).

[59] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[60] Q. Hu, C. S. Greene, Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell rna transcriptomics, in: BIOCOMPUTING 2019: Proceedings of the Pacific Symposium, World Scientific, 2018, pp. 362–373.

[61] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning, Nature methods 14 (4) (2017) 414–416.

[62] K. Johnsson, Manifold Dimension Estimation for Omics Data Analysis: Current Methods and a Novel Approach, Lund University, 2011.

[63] R. N. Shepard, A. K. Romney, S. B. Nerlove, Multidimensional scaling: Theory and applications in the behavioral sciences: II. Theory., Seminar press, 1972.

[64] R. N. Shepard, A. K. Romney, S. B. Nerlove, Multidimensional scaling: Theory and applications in the behavioral sciences: I. Theory., Seminar press, 1972.

[65] I. T. Jolliffe, Principal component analysis for special types of data, Springer, 2002.

[66] P. Mordohai, G. Medioni, Tensor voting: a perceptual organization approach to computer vision and machine learning, Morgan & Claypool Publishers, 2006.

[67] C.-G. Li, J. Guo, B. Xiao, Intrinsic dimensionality estimation within neighborhood convex hull, International Journal of Pattern Recognition and Artificial Intelligence 23 (01) (2009) 31–44.

[68] I. M. James, History of topology, Elsevier, 1999.

[69] K. Falconer, Fractal Geometry-Mathematical Foundations and Applications 2e: Mathematical Foundations and Applications, Wiley, 2003.

[70] N. Tatti, T. Mielikainen, A. Gionis, H. Mannila, What is the dimension of your binary data?, in: Sixth International Conference on Data Mining (ICDM'06), IEEE, 2006, pp. 603–612.

[71] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D: nonlinear phenomena 9 (1-2) (1983) 189–208.

[72] B. Kégl, Intrinsic dimension estimation using packing numbers, Advances in neural information processing systems 15 (2002).

[73] V. Guarino, J. Gliozzo, F. Clarelli, B. Pignolet, K. Misra, E. Mascia, G. Antonino, S. Santoro, L. Ferré, M. Cannizzaro, et al., Intrinsic-dimension analysis for guiding dimensionality reduction in multi-omics data, in: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies. 3: Bioinformatics, Scitepress, 2023, pp. 243–251.

[74] M. Gönen, E. Alpaydın, Multiple kernel learning algorithms, The Journal of Machine Learning Research 12 (2011) 2211–2268.

[75] Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, Y.-C. F. Wang, A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection, IEEE Transactions on multimedia 14 (3) (2012) 563–574.

[76] M. Žitnik, B. Zupan, Data fusion by matrix factorization, IEEE transactions on pattern analysis and machine intelligence 37 (1) (2014) 41–53.

[77] P. Chalise, B. L. Fridley, Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm, PloS one 12 (5) (2017) e0176278.

[78] F. Vitali, S. Marini, D. Pala, A. Demartini, S. Montoli, A. Zambelli, R. Bellazzi, Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia, JAMIA open 1 (1) (2018) 75–86.

[79] K. Tomczak, P. Czerwińska, M. Wiznerowicz, Review the cancer genome atlas (tcga): an immeasurable source of knowledge, Contemporary Oncology/Współczesna Onkologia 2015 (1) (2015) 68–77.

[80] J. Hedegaard, K. Thorsen, M. K. Lund, A.-M. K. Hein, S. J. Hamilton-Dutoit, S. Vang, I. Nordentoft, K. Birkenkamp-Demtröder, M. Kruhøffer, H. Hager, et al., Next-generation sequencing of rna and dna isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue, PloS one 9 (5) (2014) e98187.

[81] R. Cai, Z. Zhang, A. K. Tung, C. Dai, Z. Hao, A general framework of hierarchical clustering and its applications, Information Sciences 272 (2014) 29–48. doi:https://doi.org/10.1016/j.ins.2014.02.062.
URL https://www.sciencedirect.com/science/article/pii/S0020025514001686

[82] R. J. G. B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu (Eds.), Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 160–172.

[83] F. Nielsen, Introduction to HPC with MPI for Data Science, Undergraduate Topics in Computer Science, Springer, 2016. doi:10.1007/978-3-319-21903-5.
URL https://doi.org/10.1007/978-3-319-21903-5

[84] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLOS ONE 10 (3) (2015) 1–21. doi:10.1371/journal.pone.0118432.
URL https://doi.org/10.1371/journal.pone.0118432

[85] S. Spiegl-Kreinecker, D. Lötsch, B. Ghanim, C. Pirker, T. Mohr, M. Laaber, S. Weis, A. Olschowski, G. Webersinke, J. Pichler, et al., Prognostic quality of activating tert promoter mutations in glioblastoma: interaction with the rs2853669 polymorphism and patient age at diagnosis, Neuro-oncology 17 (9) (2015) 1231–1240.

[86] J. Callen, A. Georgiou, J. Li, J. I. Westbrook, The impact for patient outcomes of failure to follow up on test results. how can we do better?, EJIFCC 26 (1) (2015) 38.

[87] S. B. Naeem, R. Bhatti, A. Khan, An exploration of how fake news is taking over social media and putting public health at risk, Health Information & Libraries Journal 38 (2) (2021) 143–149.

[88] E. Casiraghi, R. Wong, M. Hall, B. Coleman, M. Notaro, M. D. Evans, J. S. Tronieri, H. Blau, B. Laraway, T. J. Callahan, et al., A method for comparing multiple imputation techniques: A case study on the us national covid cohort collaborative, Journal of Biomedical Informatics 139 (2023) 104295.

[89] C. K. Ormiston, J. Chiangong, F. Williams, The covid-19 pandemic and hispanic/latina/o immigrant mental health: Why more needs to be done (2023).

[90] F. Zhao, B. Copley, Q. Niu, F. Liu, J. A. Johnson, T. Sutton, G. Khramtsova, E. Sveen, T. F. Yoshimatsu, Y. Zheng, et al., Racial disparities in survival outcomes among breast cancer patients by molecular subtypes, Breast cancer research and treatment 185 (2021) 841–849.

[91] W. Fang, Z.-Y. Yang, T.-Y. Chen, X.-F. Shen, C. Zhang, Ethnicity and survival in bladder cancer: a population-based study based on the seer database, Journal of Translational Medicine 18 (2020) 1–11.

[92] L. A. G. Ries, Ovarian cancer: survival and treatment differences by age, Cancer 71 (S2) (1993) 524–529.

[93] S. Chen, C. Gao, Q. Du, L. Tang, H. You, Y. Dong, A prognostic model for elderly patients with squamous non-small cell lung cancer: a population-based study, Journal of Translational Medicine 18 (1) (2020) 1–11.

[94] O. D'Ecclesiis, S. Caini, C. Martinoli, S. Raimondi, C. Gaiaschi, G. Tosti, P. Queirolo, C. Veneri, C. Saieva, S. Gandini, et al., Gender-dependent specificities in cutaneous melanoma predisposition, risk factors, somatic muta-

tions, prognostic and predictive factors: a systematic review, International Journal of Environmental Research and Public Health 18 (15) (2021) 7945.