

## RESEARCH ARTICLE

## CustOmics: A versatile deep-learning based strategy for multi-omics integration

Hakim Benkirane<sup>1,2</sup>, Yoann Pradat<sup>1</sup>, Stefan Michiels<sup>2,3</sup>, Paul-Henry Cournède<sup>1\*</sup>

**1** Université Paris-Saclay, CentraleSupélec, Lab of Mathematics and Informatics (MICS), Gif-sur-Yvette, France, **2** Oncostat U1018, Inserm, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, CESP, Villejuif, France, **3** Bureau de Biostatistique et d'Épidémiologie, Gustave Roussy, Université Paris-Saclay, Villejuif, France

\* [paul-henry.cournede@centralesupelec.fr](mailto:paul-henry.cournede@centralesupelec.fr)

## OPEN ACCESS

**Citation:** Benkirane H, Pradat Y, Michiels S, Cournède P-H (2023) CustOmics: A versatile deep-learning based strategy for multi-omics integration. *PLoS Comput Biol* 19(3): e1010921. <https://doi.org/10.1371/journal.pcbi.1010921>

**Editor:** William Stafford Noble, University of Washington, UNITED STATES

**Received:** September 14, 2022

**Accepted:** February 4, 2023

**Published:** March 6, 2023

**Copyright:** © 2023 Benkirane et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Data used for the study can be downloaded from the NIH Genomic Data Commons Data Portal (<https://portal.gdc.cancer.gov/>) and the cBioPortal (<https://www.cbioportal.org/>). Code for the CustOmics framework is available on GitHub: <https://github.com/HakimBenkirane/CustOmics>.

**Funding:** This work was supported by the Prism project, funded by the Agence Nationale de la Recherche under grant number ANR-18-IBHU-0002 (PHC, SM, YP) and by the Public Health graduate school of Paris-Saclay University (HB).

## Abstract

The availability of patient cohorts with several types of omics data opens new perspectives for exploring the disease's underlying biological processes and developing predictive models. It also comes with new challenges in computational biology in terms of integrating high-dimensional and heterogeneous data in a fashion that captures the interrelationships between multiple genes and their functions. Deep learning methods offer promising perspectives for integrating multi-omics data. In this paper, we review the existing integration strategies based on autoencoders and propose a new customizable one whose principle relies on a two-phase approach. In the first phase, we adapt the training to each data source independently before learning cross-modality interactions in the second phase. By taking into account each source's singularity, we show that this approach succeeds at taking advantage of all the sources more efficiently than other strategies. Moreover, by adapting our architecture to the computation of Shapley additive explanations, our model can provide interpretable results in a multi-source setting. Using multiple omics sources from different TCGA cohorts, we demonstrate the performance of the proposed method for cancer on test cases for several tasks, such as the classification of tumor types and breast cancer subtypes, as well as survival outcome prediction. We show through our experiments the great performances of our architecture on seven different datasets with various sizes and provide some interpretations of the results obtained. Our code is available on (<https://github.com/HakimBenkirane/CustOmics>).

## Author summary

Cancer is a complex disease involving multiple genetic and environmental factors. Those factors affect biological systems on many levels. To better characterize a patient's molecular profile, we need to rely on multiple dimensions simultaneously, for example, genomics, transcriptomics, and epigenomics data. However, those data types are very different, making their integration challenging because of the high heterogeneity between the sources. Moreover, while defining a model architecture that can take any type of input source, we need to tackle the issue of the generalizability of the integration, as different combinations

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

**Competing interests:** The authors have declared that no competing interests exist.

of omic sources can behave differently due to discrepancies in data types and dimensionality. In light of those challenges, we developed a new integration strategy and framework called CustOmics to help scientists integrate multiple omics data. Our results show that this new integration method outperforms the state-of-the-art deep learning methods for multi-omic integration in classification and survival tasks.

This is a *PLOS Computational Biology Methods* paper.

## Introduction

With the advent of high-throughput technologies, multiple omics data are more and more available to characterize the molecular portraits of patients, notably in oncology [1]. Analyzing this mix of data sources and leveraging their information to improve our understanding of the disease's underlying biological phenomena remain challenging.

Several studies and projects have made available cohorts' data characterized by several molecular sources. For example, The Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov/>) has profiled thousands of tumor samples for multiple molecular assays and made available several types of data such as genome sequencing, RNA sequencing, DNA methylation, proteomics, etc.

This high diversity of data comes with two significant challenges. The first one is linked to the high dimensionality of the data. Due to the genetic complexity of the human molecular profile, omics data generally suffers from the 'curse of dimensionality,' which is a phenomenon resulting from the high number of features compared to the smaller number of samples [2]. This high-dimensional space often contains correlated features that result in high redundancy, reducing the prediction performance of algorithms [3]. The second issue lies in data heterogeneity: having its origins from different sources and expressing different phenomena in humans' biological system, omics data is very diverse [4]. For example, transcriptomics and proteomics are normalized differently from other omics data and use different scaling before their analysis, leading to different ranges and data distributions; omics data like metabolomics can also generate sparsity, as some variables can be below the detection limit and hence be assigned null values [5].

To overcome those challenges and alleviate overfitting in prediction tasks, standard methods either manually select a small subset of molecular features based on domain knowledge [6] or use dimensionality reduction techniques before downstream analysis. However, due to the high variability between the sources, those methods can overlook some genome-wide hidden patterns that justify the need for more flexible methods.

In the past few years, multi-omics integration has been a very active research subject in health science and precision medicine [7], giving new insight into biological processes involved with cancer [8, 9]. Multiple statistical learning methods have been proposed from the desire to investigate various complex molecular systems behind cancer. One of the most well-known methods engraved in the statistical field is the Principal Component Analysis, PCA (extensively presented, for example, in Jolliffe [10]). Several variations from this standard model have been explored, including Multiple Factor Analysis [11], Consensus PCA, and multi-block PCA [12]. All those methods try to extend the Principal Component approach to

a multi-source framework by considering each source as a distinct block. Another similar method to the PCA is the Non-negative Matrix Factorization, NMF, first introduced in *Zhang et al.* [13], which uses the same philosophy as PCA but considers a non-negative constraint instead of an orthogonality one. Additionally, we can consider the joint Dimensionality Reduction methods, jDR, as an extension of factorial methods studied in a comprehensive benchmark in *Cantini et al.* [14]. Such methods can be divided into several categories, including Bayesian methods [15] that use assumptions on data distribution and dependencies to build a statistical model and network-based methods that rely on a network representation and can identify modules of the disease-associated mechanisms. Nodes represent the genes, and edges represent the association between those genes [16, 17]. Other more specific methods have been explored, mainly for clustering. For example, *Newman et al.* [18] presented AutoSOME, a framework using the Self-organizing Map approach and couples it with a density equalizer; this equalizer reduces the dimension of the gene expression and preserves its local topology. Another well-known multi-omics method for integrative clustering is iCluster [19], which uses the same idea as NMF but without the non-negative constraint and allows more diversity in data distributions. *Kaspi et al.* explored gene set enrichment analysis for multi-omics data by developing a multi-contrast pathway enrichment for multi-omics and single-cell profiling data, *mitch* [20], for fast and accurate visualization, allowing more straightforward interpretation.

With the rise of deep learning as the new state-of-the-art approach in many medical applications (like medical imaging and diagnostic [21]), several studies have explored these methods' interest in multi-omics data integration. The primary use of these methods has been with autoencoders [22], notably by *Chaudhary et al.* [23] in the context of survival prediction. The Variational Autoencoder framework for multi-omics integration, OmiVAE [24], introduced new improvements for standard representation learning and OmiEmbed [25] for multitask learning. *Simidjievski et al.* [26] also proposed multiple deep learning architectures based on variational frameworks for multiple tasks. *Huang et al.* [27] developed a neural network architecture to integrate multi-omics data by feeding the eigengene matrix instead of the raw data for survival analysis. Based on these methods, several application studies on specific cancer cases were presented, like *Hira et al.* [28] that used the OmiVAE framework to conduct an in-depth study of the Ovarian Cancer's mechanisms. *Tong et al.* [29] also used standard autoencoders to explore multi-view learning in survival analysis for Breast Cancer. Other methods, more focused on prognosis, rely on prior knowledge to guide their neural networks like *Xie et al.* [30] that associated group lasso regularization with a deep learning framework. To this day, no benchmarking study has explored and compared the different deep learning approaches and strategies for multi-omics data integration in multitask learning.

In this work, we first discuss strategies for integrating high-dimensional multi-source data to learn low-dimensional latent representation from multi-omics datasets. We then introduce a new customizable architecture for multi-omics integration, called CustOmics, that combines the advantages of the different strategies and alleviates some of their limitations. To evaluate the impact of this new method on different test cases on both classification and survival tasks: first, we apply it to a pan-cancer dataset and then study how it handles smaller datasets by assessing it first on a breast cancer dataset for subtype classification and then on five different datasets for survival outcome prediction. We also provide a new package to help bioinformaticians and computational biologists build deep-learning architectures that quickly switch between multiple strategies to adapt to specific use cases.

## Materials and methods

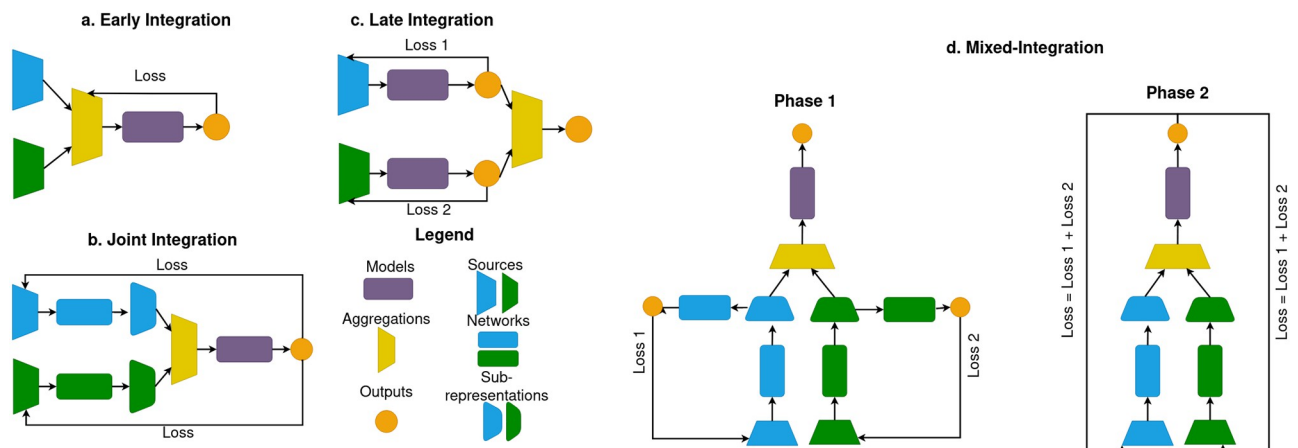
### Representation learning for multi-omics integration

Representation Learning is a field of statistical learning that aims to automatically discover relevant representations of the input data [31]. In the field of multiple omics data integration, it can help synthesize the heterogeneous distributions into a shared space, revealing the interaction between the different sources. From a biological point of view, each omics source represents a specific view of the processes involved. Therefore, the main goal behind integration is to uncover a broader view of the phenomenon of interest and to understand the underlying processes better.

Each patient is characterized by  $K$  omics vectors  $(\mathbf{x}_k)_{1 \leq k \leq K} \in \mathbb{R}^{M_k}$  with  $M_k$  the number of features of the  $k$ -th source. The models presented in this paper aim at mapping the set of omics vectors to a vector  $\mathbf{z} \in \mathbb{R}^m$  with  $m \ll \sum_{k=1}^K M_k$ , the latent representation. The latent features are the components of the vector  $\mathbf{z}$ . We will use factorial methods as a basis of comparison as they have proved to be the reference method for mapping input data into a latent space of lower dimension, notably for multi-omics integration, as illustrated by the recent review of Cantini *et al.* [14].

For factorial analysis, we consider approaches based on the study of principal components using linear operations through Multiple Factor Analysis, MFA. Deep learning methods use autoencoder architectures to build the latent representation by jointly training encoder and decoder functions. The types of architecture can be divided into three main categories described in Fig 1:

- **Early Integration (EI):** Refers to the set of methods that aim at merging the different sources before the dimensionality reduction. Instead of giving as model inputs a set of separate vectors for each source, we first concatenate the vectors of all the sources. It is characterized by its simplicity but finds its limitations when some sources bear more significant signals than others, which makes it difficult for the model to learn interactions between sources.



**Fig 1. a. Early Integration:** Sources are concatenated before being fed to a single model. **b. Joint Integration:** Sub-representations of each source are learned jointly before inputting into the model. **c. Late Integration:** Each source outputs its prediction using its independent model, and the predictions are then aggregated. **d. Mixed Integration:** Representation of the mixed integration approach. In phase 1, a specific model is trained for each source independently and embeds a sub-representation adapted to the source's specificities. In phase 2, the specific models are trained jointly, similarly to the joint integration setup, to create the final output.

<https://doi.org/10.1371/journal.pcbi.1010921.g001>

- **Joint Integration (JI):** In this setting, sub-representations are created inside the same model for each source before learning the output. This is the most promising strategy theoretically and is the most popular in the literature. However, it is challenged by the sources' heterogeneity, as they do not necessarily follow the same learning dynamics and may need different approaches with specific losses.
- **Late Integration (LI):** Consists in learning the output for each source separately, with its own representation model. It can adapt well to the specificities of each source but does not retrieve any cross-modality interaction.

To alleviate the issues brought by those different integration methods, we propose a hybrid strategy, named mixed-integration, that comes halfway between joint and late integration. It consists of two learning phases that switch during a predefined epoch considered a hyperparameter: the first phase independently trains the network of each source with an adapted loss to create sub-representations. Those specific models will then be jointly trained in a second phase to build a global output representation. That way, we can adapt our training to each source with different learning dynamics in the first phase while learning cross-modality interactions in the second phase. This intuition for a two-phase integration has been tackled in [32]. Still, our proposed strategy focuses on optimizing the integration part and thus is novel in the context of variational autoencoders.

## Variational autoencoders

A Variational Autoencoder, VAE, is a deep generative model which can learn meaningful data representations from high-dimensional input data. It is an extension of standard autoencoders in which the encoder encodes the input as a distribution over a latent space instead of a single point.

In this case, the encoding function  $q$  represents a variational distribution (known as encoding distribution)  $q_\phi(\mathbf{z}|\mathbf{x})$  in which  $\phi$  is the parameter to estimate and the decoding function  $p$  represents the posterior  $p_\theta(\mathbf{x}|\mathbf{z})$ .

The particularity of a VAE is its ability to encode a distribution. After the encoding phase, there is a sampling phase in which we sample points from the distribution  $q_\phi(\mathbf{z}|\mathbf{x})$ .

Traditionally, the distributions in the VAE architecture are supposed to be Gaussian: the encoder function will learn the two-parameter vectors  $\mu$  and  $\sigma$  (respectively the means and standard deviations) that are used to generate samples in  $q_\phi(\mathbf{z}|\mathbf{x})$  using the reparametrization trick  $\mathbf{z} = \mu + \sigma \odot \epsilon$  where  $\epsilon \sim \mathcal{N}(0, I)$  and  $\odot$  denotes the element-wise product.

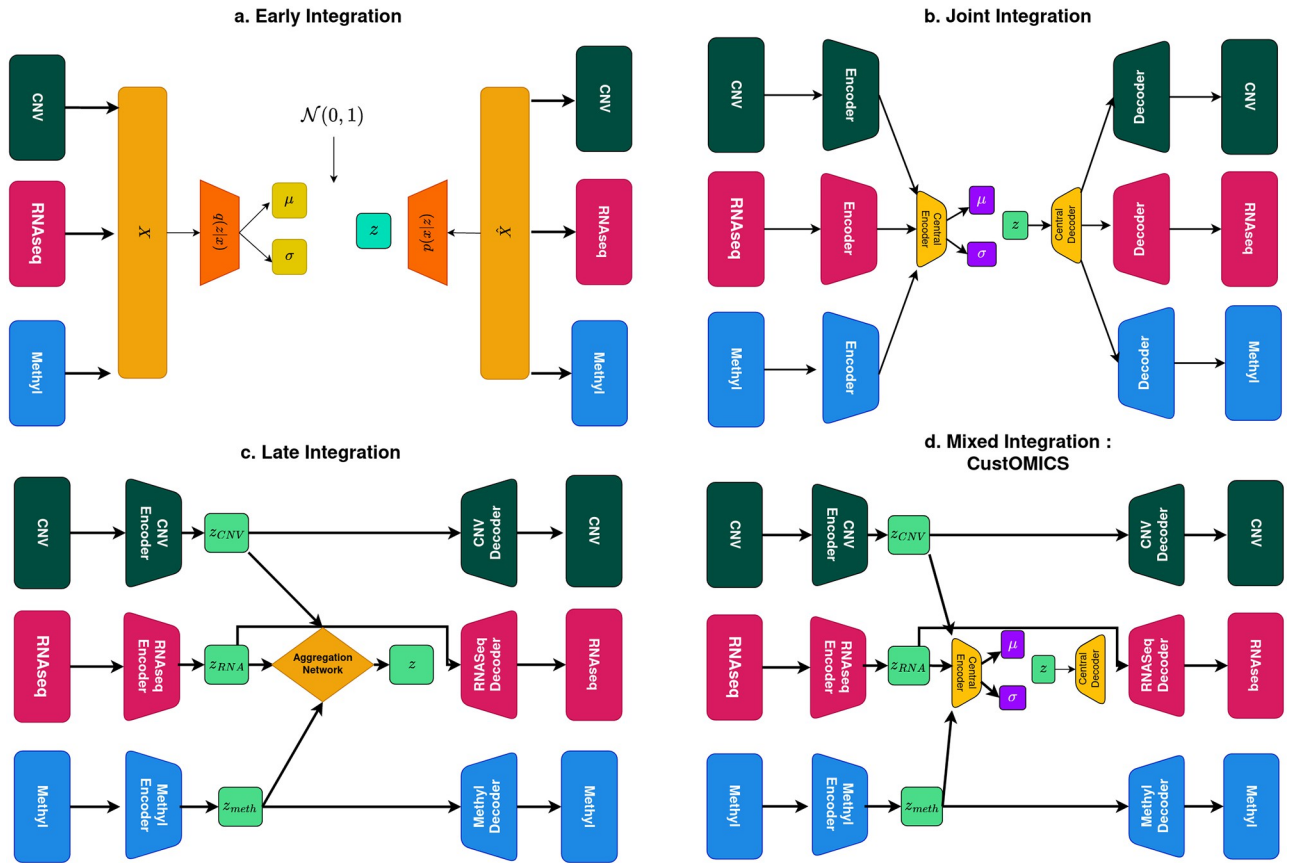
The loss function for this architecture can be written as the sum of two distinct losses. First, a reconstruction loss that focuses on the autoencoder's ability to reconstruct the data:

$\mathcal{L}_{recon} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [p_\theta(\mathbf{x}|\mathbf{z})]$ . It can be interpreted as the conditional entropy of  $x$  over  $z$ , which quantifies the uncertainty one has over the joint distribution  $(x, z)$ , knowing  $z$ . More practically, it is related to the quantity of information of  $x$  that is retained by  $z$  (thus qualifying a reconstruction potential).

Second, a regularization loss aims at getting the encoding distribution as close as possible to the theoretical distribution of the latent vector. Traditionally, a Kullback-Leibler divergence,

KL, is used:  $\mathcal{L}_{reg} = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) = \mathbb{E}_{q_\phi} \left[ -\log \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right]$ .

The total loss will benefit from the framework introduced in [33],  $\mathcal{L} = \mathcal{L}_{recon} + \beta \mathcal{L}_{reg}$ , adding a weight on the regularization term to balance between the two parts of the loss function, with a hyperparameter  $\beta$  to optimize.



**Fig 2. a. Early Integration VAE:** Variational Autoencoder architecture with early integration strategy. **b. Joint Integration VAE:** Variational Autoencoder architecture with joint integration strategy. **c. Late Integration VAE:** Variational Autoencoder architecture with late integration strategy. **d. Mixed-Integration/CustOmics:** This is a hierarchical architecture composed of specific per-source autoencoders that converges into a central variational autoencoder.

<https://doi.org/10.1371/journal.pcbi.1010921.g002>

### Early integration

An early integration autoencoder is an autoencoder applied to the concatenated data from various sources. *Gautier et al.* [34] reported this type of architecture as the SDAE (Standard Deep Autoencoder). The drawback of such an integration is that the resulting concatenation of every dataset results in a noisier and more complex high-dimensional matrix that makes the training more difficult [35]. Moreover, this type of integration does not consider the specific data distribution of each source separately and thus leads the model to learn irrelevant patterns that can, in the worst case, mask the essential signals. For these reasons, even though this strategy has been introduced in the benchmark, it is not recommended in a real setting for complex multi-source datasets such as omics data. For comparison purposes, we adapt the SDAE architecture to a Variational Autoencoder instead of a standard autoencoder as described in Fig 2.

### Joint integration

This is a standard way to approach multi-source integration using Variational Autoencoders. This model has been used in multiple designs in literature ([24–26]) and is probably the most used as it is relatively straightforward. The idea is to reduce through a series of the hidden layer the dimension of each source separately, then concatenate them in a central layer before

sampling. In this architecture, we would learn the mean and standard deviation of the concatenated data such that:

- $\mathbf{z} = \mathcal{N}(\mu, \sigma^2 I)$
- $\mu = q_\mu(T(q^{(1)}(\mathbf{x}^{(1)}; \theta_{e,1}), \dots, q^{(K)}(\mathbf{x}^{(K)}; \theta_{e,K})))$
- $\sigma = q_\sigma(T(q^{(1)}(\mathbf{x}^{(1)}; \theta_{e,1}), \dots, q^{(K)}(\mathbf{x}^{(K)}; \theta_{e,K})))$

where  $\mathbf{x}^{(k)}$  is the vector for the  $k^{th}$  source and  $q^{(i)}$  its corresponding encoder,  $T$  is a concatenation operation and  $\theta_{e,k}$  are the trainable parameters of the encoder  $k$ .

### Late integration

A late integration approach for variational autoencoders is a variation of the architecture proposed in [26] with the Hierarchical VAE architecture. The idea is to use variational inference on each layer separately and input the resulting representations in an aggregation network learned independently of the other VAEs, which is the main difference from the joint integration approach. This architecture is described in Fig 2.

### Mixed-integration / custOmics

This integration strategy will be the foundation for building a customizable architecture for multi-omics integration, CustOmics. The proposed method is a hierarchical mixed-integration that consists of an autoencoder for each source that creates a sub-representation that will then be fed to a central variational autoencoder. This new integration strategy benefits from two training phases. The first phase will act as a normalization process: each source will train separately to learn a more compact representation that synthesizes its information with less noise. This will help the integration as we will lose all imbalance issues between the sources and avoid losing focus when a source has an inferior dimensionality or weaker signal than the others. The second phase will constitute a simple joint integration between the learned sub-representations, while still training all the encoders to fine-tune those representations as some signals are enhanced in the presence of other sources.

Regarding the regularization loss for the central layer, the KL divergence can be an obstacle to generalization. As stated in [36], the KL divergence suffers from various problems. The first is that the model can fail to learn a meaningful representation of the input. Indeed, the KL divergence can sometimes be too restrictive and naturally tends to make the latent code a random sample from  $p_\theta(\mathbf{z})$ . The second is that the KL divergence can make the model overfit and learn a latent code with variance tending to infinity.

So instead, we will use the Maximum Mean Discrepancy (MMD) to assess the distance between the distributions. This distance stands on the foundation that two distributions are identical if and only if their moments are the same. Let  $p, q$  be two distributions, the MMD distance is given as follows:

$$MMD(p(x)||q(x)) = \mathbb{E}_{p(x),p(x')}(\kappa(x, x')) + \mathbb{E}_{q(x),q(x')}(\kappa(x, x')) - 2\mathbb{E}_{p(x),q(x')}(\kappa(x, x'))$$

where  $\kappa$  is a kernel function and where  $x$  and  $x'$  are two sample points. We will choose a Gaussian kernel  $\kappa(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$ .

Whereas all the models mentioned previously build the latent representation in an unsupervised way, we also create latent features adapted to specific tasks like classification or survival. This idea has been used multiple times in the literature, for example, in [25]. The solution

relies on adding a task-related loss to the autoencoder objective function. Therefore, we denote by  $\mathcal{L}_{task}$  the loss such that the total loss function would be expressed as follows:  $\mathcal{L}_{tot} = \mathcal{L}_{AE} + \alpha\mathcal{L}_{task}$ , where  $\mathcal{L}_{AE}$  is the autoencoder loss.

For the classification task, we use a categorical cross-entropy loss defined by  $\mathcal{L}_{class} = \sum_i y_i \log(\hat{y}_i)$ , where  $y_i$  is the ground truth for the  $i^{th}$  sample, and  $\hat{y}_i$  its estimation with the downstream model.

We use the deep learning survival framework, DeepSurv, loss function for the survival task. This nonlinear proportional hazard model has been introduced in [37]. The model is built by using the negative partial log-likelihood formula that translates, in our case, into:

$$L(\theta) = - \sum_{i:E_i=1} (\hat{\mu}(x_i; \theta) - \log \sum_{j \in \mathcal{R}(t)} e^{\hat{\mu}(x_j; \theta)}) \quad (1)$$

where  $E_i$  is the event for patient  $i$ ,  $\hat{\mu}(x; \theta)$  is the risk function associated with the risk score estimated by the output layer of the network,  $\mathcal{R}(t)$  is the risk set, that is the set of patients still at risk of failure after time  $t$ .

Additionally, this loss will be assigned to each omic-specific network in the first training phase to create adequate sub-representations before the joint integration phase.

## Interpretability

To build a more interpretable architecture, we adapt the method introduced in [38] to compute SHapley Additive exPlanations values, SHAP values [39], for deep variational autoencoders. Whereas this method has only been conceived for single source inputs, we expand it to the multi-source setting of CustOmics by adapting it to any deep autoencoder and applying it to each source autoencoder thanks to the dual-phase approach characterizing the mixed-integration strategy. After training the CustOmics network, we pass each intermediate autoencoder to a SHAP Explainer for deep learning architectures, DeepSHAP explainer that was modified to make sure that it could take either each component of the latent vector or the prediction output for the contribution analysis. For each prediction, SHAP values corresponding to each gene or latent dimension are computed to estimate the overall contribution by averaging them over a group of samples with similar features. When combined with the mixed-integration approach through CustOmics, we can compute for each source two types of feature importance. During the first training phase, we compute the feature importance of each source separately to assess the importance of genes when only considering this source. During the second training phase, the features' importance can change as it will then consider other sources, showing how some genes can have increasing or decreasing importance when introducing complementary views of the biological system.

## Test cases and datasets

In this study, we use datasets extracted from the Genomic Data Commons (GDC) pan-cancer multi-omics study [40]. It is one of the most comprehensive datasets for multi-omics analysis, with high-dimensional omics data and corresponding phenotype data from The Cancer Genome Atlas (TCGA). Our experiments use three types of omics data: Copy Number Variations (CNV), RNA-Seq gene expressions, and DNA methylation. The CNV dataset comprises Gistic2 measurements on a total of 19,729 genes. The RNA-Seq expression dataset profile comprises around 60,484 identifiers referring to corresponding exons and measuring log2 transformed Fragments Per Kilobase of transcript per Million mapped reads, FPKM. Finally, the DNA methylation dataset was produced using the Infinium HumanMethylation450 BeadChip

(450K) arrays with 485,578 probes in which beta values of probes indicate the methylation ratio of corresponding CpG sites. Moreover, we also evaluate our method on five smaller cohorts from TCGA: Bladder Urothelial Carcinoma (BLCA,  $n = 437$ ), Breast Invasive Carcinoma (BRCA,  $n = 1022$ ), Lung Adenocarcinoma (LUAD,  $n = 498$ ), Glioblastoma & Lower Grade Glioma (GBMLGG,  $n = 515$ ) and Uterine Corpus Endometrial Carcinoma (UCEC,  $n = 538$ ). More details on the different datasets used are available in [S1 Table](#). We will perform in this study 4 different evaluations on 4 test cases for classification and survival. The first task in our set of experiments is classifying the different tumor types in the pan-cancer study. The classification performance was measured using five metrics: Accuracy, Macro-averaged F1 score, precision, recall, and ROC-AUC [41]. We also perform a second classification task to validate our findings on a smaller dataset and test the robustness of our method. This task aims at performing a tumor subtype classification based on the PAM50 classification (LuminalA, LuminalB, Basal, and HER2). We use the same setup as the pan-cancer case. The third test case will be a survival study of the Pancancer dataset. Finally, the fourth test case will evaluate the survival performances on the five datasets presented earlier. For all those test cases, We compare the CustOmics model to several reference methods for multi-omics integration: first with a combination of dimensionality reduction methods, Multiple Factor Analysis, MFA [11], Uniform Manifold Approximation and Projection, UMAP [42], and non-negative matrix factorization, NMF [13], and also with different deep learning methods corresponding to the various strategies described in the Materials and Methods section.

## Data preprocessing

The data required some preprocessing before analysis.

- For RNA gene expression profiles, 594 exons located on the Y chromosome were removed, along with 1,904 ones with zero expression and 248 with missing values.
- For DNA methylation data, the same strategy as with gene expression profiles was used, in addition to removing probes that cannot be mapped to the human reference genome. It leaves us with 438,831 CpG sites.

Afterward, we intersected each combination of omics data in order to retrieve the maximum number of samples for each test case. We then identified and removed features with missing/consistently zero/NA values for other omic files. Finally, for non-normalized datasets such as CNVs and RNA-Seq data, we applied a min-max normalization to ensure that each omic source was scaled identically and thus would have the same importance during integration.

## Implementation details

The CustOmics framework is based on the Pytorch deep-learning library [43]. It can be applied to any combination of high-dimensional datasets and incorporates different integration strategies depending on the type of data and task to perform. As done in *Zhang et al.* [25], DNA methylation data can be divided into 23 separate blocks, each feeding a hidden layer corresponding to a chromosome to avoid overfitting and save GPU memory.

The whole architecture is built using fully-connected blocks with weights initialized following a uniform distribution  $\mathcal{U}(-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}})$  where  $k$  is the number of weight parameters. We use a batch normalization technique in each layer composing the neural network to address the internal covariate shift problem [44]. Also, to avoid overfitting problems, we use dropout [45]; its rate is considered a hyperparameter.

The input dataset was randomly split into training, validation, and testing set (60–20–20%) by using stratified 5-fold cross-validation so that the proportion of samples in each tumor type between the different sets is preserved in all the folds. We perform Bayesian optimization [46] using the validation set to find our model's best possible combination of hyperparameters.

All models were trained using an Nvidia Tesla V100S with 32GB memory.

## Results

### Classification results

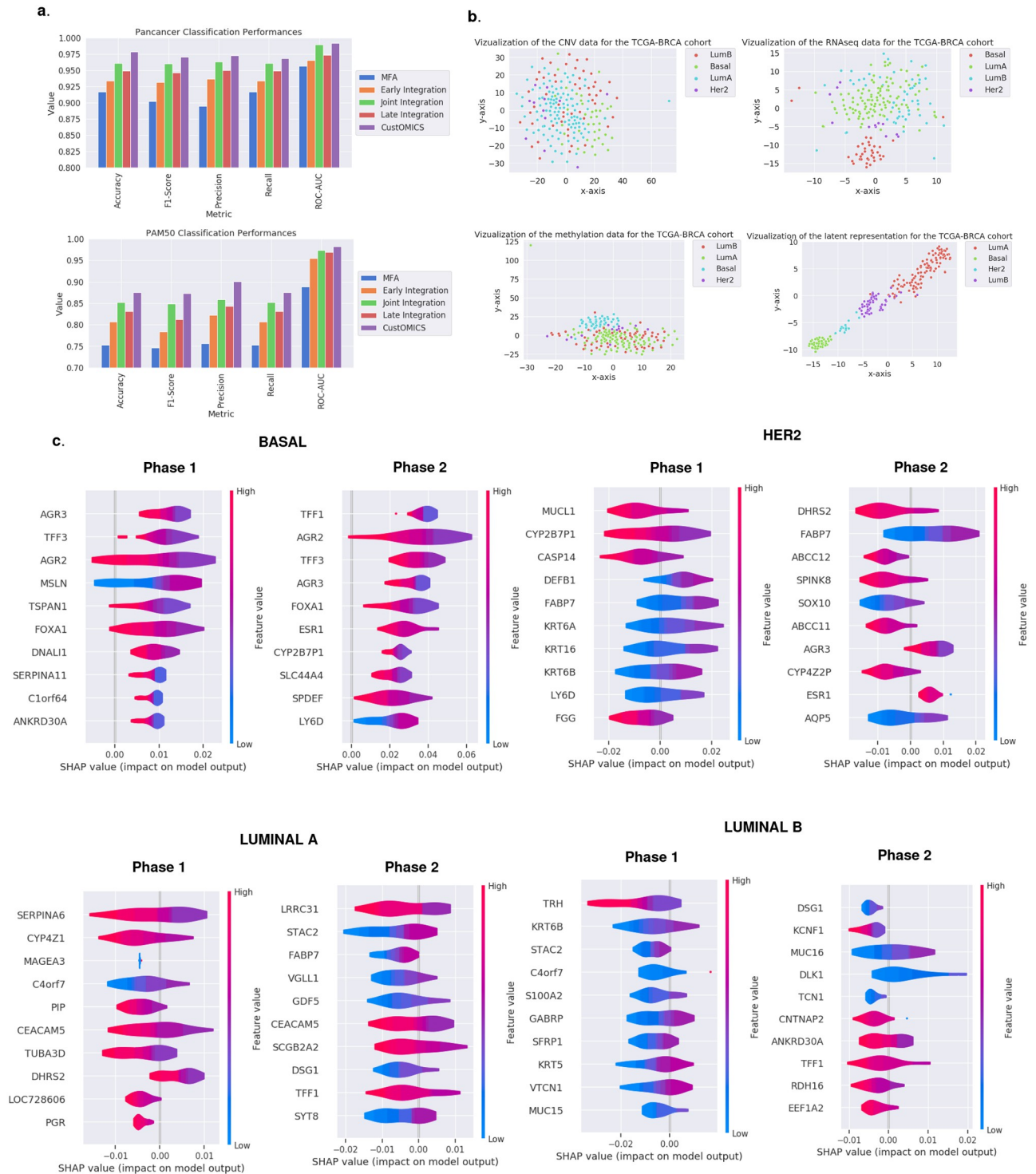
We first perform the classification task on the pan-cancer dataset. Each architecture is coupled with an artificial neural network classifier composed of two hidden layers with 256 and 128 neurons. This network is trained using a categorical cross-entropy loss with ReLU activation function on the hidden layer and a Softmax activation function on the output layer.

Fig 3 and Tables 1 and 2 show the overall classification results (More details can be found in S3 Table). Among the factorial methods, the MFA achieved the best results, so we coupled this method with the same ANN classifier used with the deep-learning representation methods as a basis for comparison. However, it does not perform as well as most deep-learning methods. It is because MFA cannot uncover nonlinear relationships between different sources, unlike deep-learning architectures. Moreover, as the MFA is an early integration, it suffers from problems of signal imbalance related to early integration.

We also assessed the performance of an unsupervised representation given by our CustOmics network by plugging it into the same ANN classifier as used for the factorial methods. This unsupervised setting performs quite well compared to other similar unsupervised methods like the factorial ones showing the robustness of the representation learned by CustOmics, even without adding the task loss. Moreover, we can see in Table 1 that, in general, Variational Autoencoders perform better than standard autoencoders, which comforts us in choosing a variational setup for CustOmics.

As hinted earlier, we can see that for the deep learning strategies, early integration is behind the others in terms of performance. It can be explained by the fact that RNA-Seq data hold more signals when determining tumor types or subtypes. Thus, concatenating the sources before feeding them to the VAE overshadows the other sources, and the learned representation depends mostly on RNA-Seq data without leveraging the other modalities. It is illustrated in S3 Table, showing that the classification results using RNA-Seq data only are very close to those obtained with early integration, indicating that the model may overlook the interactions between sources. Late integration is not optimal either since interactions between sources are not properly learned as the predictors are trained separately. Joint integration performs well in most cases, but we see that the best results displayed in S4 Table are achieved by the combination of only two sources, RNA-Seq and Methylation data, as it seems that CNV data only adds noise to the latent representation, meaning that its information is not handled well with this strategy.

These results confirm the interest of the CustOmics architecture, as it gives the best performances for all the test cases, while being able to converge without apparent overfitting as suggested by S2 Fig. Moreover, it also takes advantage of the complementarity and interactions between sources: As shown in S3 Table, all sources bring additional information. We can witness that even though transcriptomics data gives the best performances, other omics sources succeed at bringing additional information. The high performance of transcriptomics data is predictable as most information about tumor types and molecular subtypes is expressed in RNA data. It is also interesting to see that in S1 Fig, transcriptomics data don't need much layers to converge to their best results whereas it is not the case for other types of data. Regarding



**Fig 3. a. pan-cancer and PAM50 classification results:** Overall classification results for the pan-cancer tumor classification test case and the PAM50 subtype classification for breast cancer. **b. T-SNE visualization** for each omic source separately, along with the latent representation constructed by CustOmics. We see that the constructed layer representation succeeds at separating the data into four distinct clusters that we couldn't distinguish with each omic source alone. **c. PAM50 gene importance:** Computed SHAP values on the RNA-Seq data of the most relevant genes responsible for discriminating between subtypes against the others using CustOmics for both integration phases.

<https://doi.org/10.1371/journal.pcbi.1010921.g003>

**Table 1. The classification performance for the pan-cancer dataset is evaluated with 5 standard metrics for UMAP, NMF, MFA, Unsupervised Customomics with SVM, and supervised deep-learning methods.** We evaluate the performances on the final predicted output of the downstream classifier.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC
UMAP	0.7598 ± 0.0036	0.7149 ± 0.0029	0.7200 ± 0.0031	0.7598 ± 0.0032	0.8740 ± 0.0012
NMF	0.8599 ± 0.0017	0.8406 ± 0.0013	0.8460 ± 0.0018	0.8599 ± 0.0021	0.9266 ± 0.0019
MFA	0.9167 ± 0.0012	0.9025 ± 0.0014	0.8945 ± 0.0008	0.9167 ± 0.0013	0.9565 ± 0.0003
Unsup. Cust.	0.9335 ± 0.0038	0.9323 ± 0.0039	0.9342 ± 0.0043	0.9335 ± 0.0038	0.9689 ± 0.0019
Early Int. VAE	0.9337 ± 0.0079	0.9314 ± 0.0086	0.9367 ± 0.0067	0.9337 ± 0.0079	0.9655 ± 0.0041
Joint Int. VAE	0.9610 ± 0.0032	0.9600 ± 0.0032	0.9631 ± 0.0043	0.9610 ± 0.0032	0.9898 ± 0.0005
Late Int. VAE	0.9492 ± 0.0115	0.9464 ± 0.0111	0.9498 ± 0.0079	0.9492 ± 0.0115	0.9737 ± 0.0060
Mix Int. AE	0.9453 ± 0.0056	0.9423 ± 0.0063	0.9452 ± 0.0050	0.9453 ± 0.0056	0.9717 ± 0.0029
<b>CustOmics</b>	<b>0.9788 ± 0.0025</b>	<b>0.9705 ± 0.0033</b>	<b>0.9728 ± 0.0041</b>	<b>0.9685 ± 0.0034</b>	<b>0.9918 ± 0.0001</b>

<https://doi.org/10.1371/journal.pcbi.1010921.t001>

the computational cost of all those methods, we see in [S2 Table](#) that they all have around the same number of trainable parameters. The slight increase in the number of parameters in CustOmics is to the intermediate networks necessary for phase 1, similar to the late integration set-up. [Fig 3](#) gives a visualization of the different sources: Even though the initial sources are quite entangled, the CustOmics latent representation succeeds in separating the clusters using the mutual information between modalities. We also use the interpretability property of CustOmics introduced in the Methods section to highlight the most relevant features for the discrimination between PAM50 subtypes by computing their respective SHAP values for each source. We do it for both phases: in phase 1, we retrieve the relevant genes considered when using a single omic source, whereas, in phase 2, we investigate how the addition of other sources' signals changes the genes' importance. [Fig 3](#) references the results of such explanations on RNA-Seq data, [S3](#) and [S4](#) Figs show the results for CNV and methylation data. We witness some well-referenced biomarkers for breast cancer like TFF1 [47], suggesting that our method can retrieve relevant biological information.

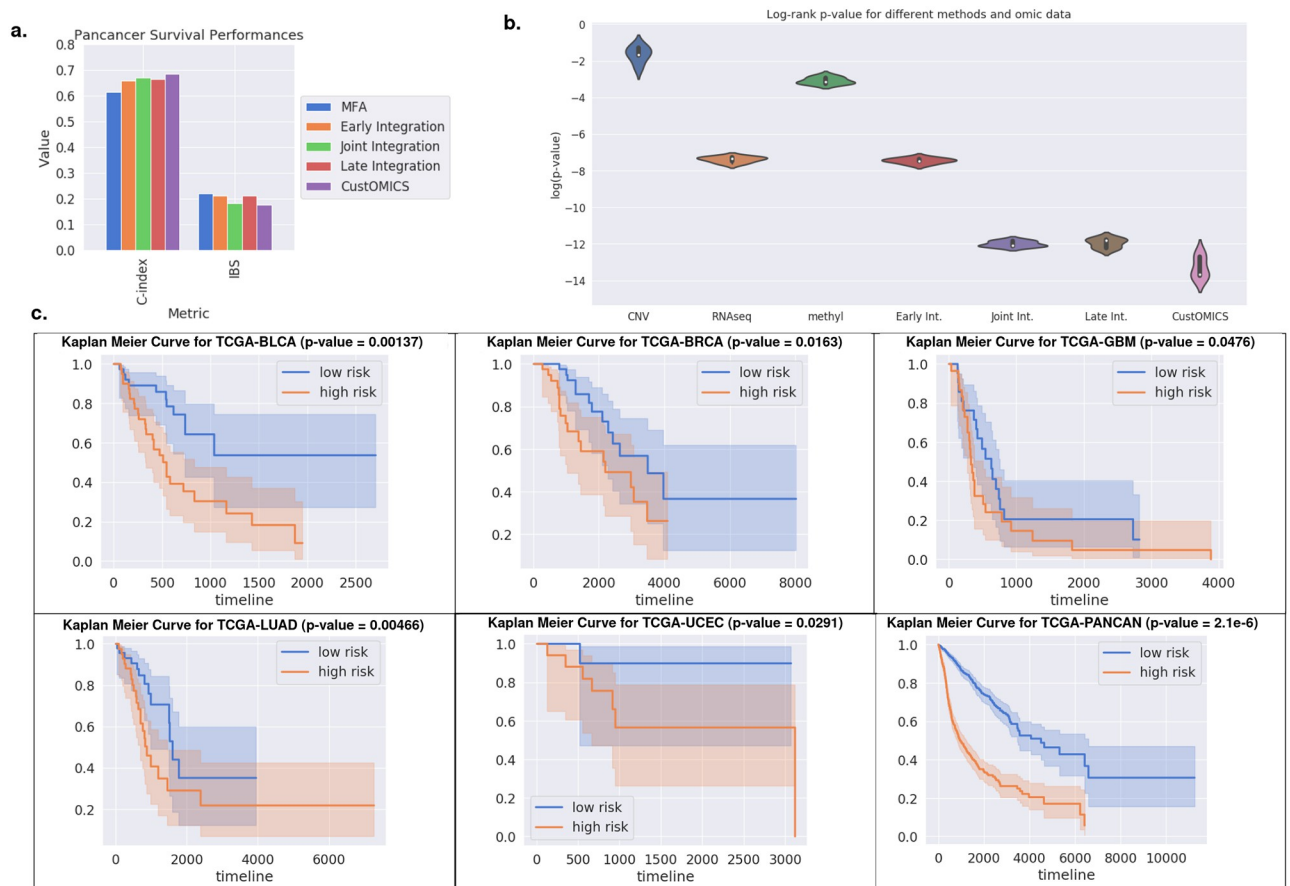
## Survival analysis

The second task in this study is survival analysis. The goal is to predict the risk score associated with each patient from the corresponding high-dimensional omics data. Two standard metrics evaluate the performance of this downstream task: The C-index, which is a generalization of the AUC metric for censored data [48], and the Integrated Brier Score, IBS [49].

**Table 2. Classification performance for PAM50 classification on the TCGA-BRCA dataset, with 5 standard metrics.** We compare machine-learning methods like UMAP, NMF, and MFA with deep-learning methods. We evaluate the performances on the final predicted output of the downstream classifier.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC
UMAP	0.6815 ± 0.0152	0.6612 ± 0.0140	0.6637 ± 0.0157	0.6815 ± 0.0151	0.8482 ± 0.0055
NMF	0.7025 ± 0.0132	0.6955 ± 0.0135	0.7032 ± 0.0110	0.7025 ± 0.0131	0.8576 ± 0.0119
MFA	0.7532 ± 0.0164	0.7460 ± 0.0160	0.7562 ± 0.0162	0.7531 ± 0.0165	0.8884 ± 0.0037
Unsup. Cust.	0.8133 ± 0.0172	0.8044 ± 0.0168	0.8322 ± 0.0144	0.8112 ± 0.0150	0.9674 ± 0.0031
Early Int. VAE	0.8063 ± 0.0152	0.7840 ± 0.0167	0.8228 ± 0.0167	0.8063 ± 0.0150	0.9552 ± 0.0077
Joint Int. VAE	0.8518 ± 0.0184	0.8488 ± 0.0189	0.8589 ± 0.0161	0.8518 ± 0.0151	0.9734 ± 0.0035
Late Int. VAE	0.8312 ± 0.0174	0.8124 ± 0.0201	0.8429 ± 0.0215	0.8312 ± 0.0176	0.9689 ± 0.0066
Mix Int. AE	0.8452 ± 0.0168	0.8322 ± 0.0214	0.8477 ± 0.0234	0.8417 ± 0.0181	0.9709 ± 0.0051
<b>CustOmics</b>	<b>0.8758 ± 0.0162</b>	<b>0.8728 ± 0.0141</b>	<b>0.9012 ± 0.0137</b>	<b>0.8758 ± 0.0130</b>	<b>0.9828 ± 0.0022</b>

<https://doi.org/10.1371/journal.pcbi.1010921.t002>



**Fig 4. a. Survival Analysis Performances:** We evaluate the performances of the survival model for the pan-cancer dataset using both the C-index and the Integrated Brier Score (IBS). Here again, our model outperforms the other integration strategies on both metrics. **b. Log-rank test:** We compute the p-value associated with the log-rank test between high and low-risk groups for every integration strategy on a validation set for the pan-cancer survival test case and compare it to mono-omic survival predictions. **c. Kaplan Meier Curves:** We draw the Kaplan Meier curves and display the p-value associated with the log-rank test as computed previously for each dataset using the predicted hazard from the CustOmics model and stratify the population into high and low risk on the test set for the predicted hazard ratio. This figure shows that our method successfully stratifies the patients into risk subgroups.

<https://doi.org/10.1371/journal.pcbi.1010921.g004>

Fig 4 and Tables 3 and 4 show the results for the different methods (more details can be found in S5 Table) for the survival task. The same observations as for the classification task can be made regarding the differences between integration strategies. Here again, we also evaluated the performance of the CustOmics method for each combination of omics sources (S5 Table).

The last task consists in evaluating the model performances for survival analysis for several specific cancer types of the TCGA datasets described in the dataset section. The objective is to evaluate the robustness of the models when dealing with smaller datasets.

Finally, we perform a more thorough analysis on the survival results that we display in Fig 4. We leave out 20% of our datasets for validation purposes, and we perform 5-fold cross-validation on the remaining 80% to compute the p-values associated with the log-rank test for different combinations of the Pan-cancer test case. We show the ability of CustOmics to stratify the patients into distinct risk groups using the predicted hazard ratio. This stratification ability was also measured quantitatively using the p-value associated with the log-rank test between the different categories. Even though the comparison between joint, late, and mixed integration is not significant in this case, it is interesting to note that the addition of multiple omics

**Table 3. The survival analysis performance for the pan-cancer dataset is evaluated with two standard metrics, C-index and IBS.** We compare classical methods like UMAP, NMF, and MFA with deep-learning methods and evaluate the performances on the final predicted output of the downstream survival network.

Model	C-index	IBS
UMAP	0.5948 ± 0.0231	0.2486 ± 0.0327
NMF	0.6012 ± 0.0204	0.2207 ± 0.0264
MFA	0.6127 ± 0.0164	0.2192 ± 0.0203
Unsup. Cust.	0.6329 ± 0.0144	0.2087 ± 0.0207
Early Int. VAE	0.6578 ± 0.0103	0.2106 ± 0.0117
Joint Int. VAE	0.6709 ± 0.0041	0.1802 ± 0.0072
Late Int. VAE	0.6629 ± 0.0086	0.2112 ± 0.0088
Mix Int. AE	0.6618 ± 0.0051	0.1815 ± 0.0074
<b>CustOmics</b>	<b>0.6841 ± 0.0033</b>	<b>0.1745 ± 0.0052</b>

<https://doi.org/10.1371/journal.pcbi.1010921.t003>

**Table 4. Survival performances of state-of-the-art integration methods for survival analysis, using concordance index on 5 TCGA cohorts: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma & Lower Grade Glioma (GBMLGG), Lung Adenocarcinoma (LUAD) and Uterine Corpus Endometrial Carcinoma (UCEC).**

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
UMAP	0.527 ± 0.048	0.524 ± 0.039	0.557 ± 0.028	0.530 ± 0.017	0.543 ± 0.025	0.536
NMF	0.553 ± 0.060	0.560 ± 0.036	0.584 ± 0.067	0.589 ± 0.052	0.570 ± 0.068	0.571
MFA	0.591 ± 0.052	0.599 ± 0.043	0.605 ± 0.036	0.597 ± 0.058	0.586 ± 0.020	0.596
Unsup. Cust.	0.599 ± 0.062	0.624 ± 0.048	0.633 ± 0.032	0.606 ± 0.049	0.622 ± 0.023	0.617
Early Int. VAE	0.603 ± 0.054	0.618 ± 0.039	0.628 ± 0.056	0.612 ± 0.041	0.609 ± 0.032	0.614
Joint Int. VAE	0.616 ± 0.072	0.627 ± 0.030	0.635 ± 0.020	0.608 ± 0.038	0.630 ± 0.021	0.624
Late Int. VAE	0.610 ± 0.055	0.620 ± 0.057	0.621 ± 0.067	0.595 ± 0.012	0.627 ± 0.022	0.615
Mix Int. AE	0.611 ± 0.049	0.625 ± 0.061	0.609 ± 0.047	0.594 ± 0.031	0.615 ± 0.020	0.611
<b>CustOmics</b>	<b>0.637 ± 0.050</b>	<b>0.633 ± 0.018</b>	<b>0.642 ± 0.028</b>	<b>0.625 ± 0.037</b>	<b>0.667 ± 0.022</b>	<b>0.640</b>

<https://doi.org/10.1371/journal.pcbi.1010921.t004>

sources has dramatically affected the p-value as it was nearly multiplied by a factor  $10^{-5}$  for CNV data and  $10^{-2}$  for RNA-Seq and methylation data. Those results also show that early integration is the strategy with minimal enrichment from other sources. This corroborates our previous intuition and is coherent with the results found in the different experiments.

## Discussion

In this work, we presented a range of integration strategies for multi-source data that can handle both the high dimensionality and the heterogeneity of the data. To take the best of all those strategies, we presented the mixed-integration and the CustOmics framework to alleviate the limitations of the existing methods. This new framework can achieve better latent representations and lead to a more robust and generalizable architecture, as shown by the systematic better results than alternative strategies. Importantly, our method can adapt to each omic source by handling the training independently in the first phase, which solves the issue of unbalanced signals between the sources by standardizing the representations before learning cross-modality interactions. Our fusion model demonstrated better performances for both classification and survival outcome prediction across all test cases. It not only achieved great results on pan-cancer data, but it also did on smaller datasets for particular cohorts, showing the robustness of our method in situations with fewer samples. Interestingly, the performances of CustOmics in comparison to other integration strategies for different combinations of omic sources show that the mixed-integration handles better the fusion between CNV, RNA-Seq, and methylation

data. This aspect could probably be further enhanced by adding prior knowledge of those sources in the intermediate autoencoders, for example, introducing a negative binomial prior in the RNA-Seq autoencoder. Furthermore, by adapting the SHAP method to our architecture, we were able to highlight genes of importance for specific tasks. However, there is room for improvement from a computational point of view since the interpretability method is very expensive.

As clinical applications often suffer from a lack of samples to build efficient deep learning models and as all the modalities are not always extracted for all patients, it would be beneficial to study the benefits of per-source transfer learning during phase 1 to pick up weaker signals and eliminate the noise. In fact, one of the advantages of phase 1 learning the sub-representations independently is that it is not necessary for all patients to have all omic sources available at all times, which helps us take the best advantage of the data at hand.

In conclusion, our generic and interpretable multi-source deep learning framework improves on state-of-the-art integration strategies by proposing a hybrid approach that fits well with multi-omics data. The framework is available on Github: <https://github.com/HakimBenkirane/CustOmics>. Possible steps to refine CustOmics are a more thorough study of the interpretability part of the framework and adding prior knowledge to individual autoencoders to enhance the sub-representations of omic sources during phase 1 of mixed-integration.

## Supporting information

**S1 Table. Data description.** Description of the datasets used in this study. We specify the number of samples overall and per data type. The censorship rate for survival is also highlighted for all the datasets.  
(XLSX)

**S2 Table. Computational cost.** Number of trainable parameters for each model used for the downstream tasks.  
(XLSX)

**S3 Table. Classification performances.** for multiple combinations of omics data using CustOmics on the pan-cancer dataset. We can see that Transcriptomics data bring the best performances, but adding other omics data increases the performances, showing that the integration is relevant.  
(XLSX)

**S4 Table. Classification performances.** for multiple combinations of omics data using joint integration on the pan-cancer dataset.  
(XLSX)

**S5 Table. Survival performances.** for multiple combinations of omics data using CustOMICS on the pan-cancer dataset. We can see that Transcriptomics data bring the best performances, but adding other omics data increases the performances, showing that the integration is relevant.  
(XLSX)

**S1 Fig. Evolution of the AUC score with the network's depth.** We first assess the evolution of the performance on the tumor classification task for each source using the intermediate autoencoders, then we evaluate the effect of the depth on the central encoder using the best results for the intermediate autoencoders for each source. We see that RNAseq data does not

need as many layers as CNV and methylation data, suggesting that its convergence may be simpler as it holds most of the signal for tumor-type prediction.

(PNG)

**S2 Fig. Evolution of the loss function.** We display the evolution of both training and validation loss before and after the phase switch for the tumor classification task.

(PNG)

**S3 Fig. PAM50 gene importance for CNV.** Computed SHAP values on the CNV data of the most relevant genes responsible for the discrimination between each subtype against the others using CustOmics for both integration phases.

(PNG)

**S4 Fig. PAM50 gene importance for methylation.** Computed SHAP values on the methylation data of the most relevant genes responsible for the discrimination between each subtype against the others using CustOmics for both integration phases.

(PNG)

**S1 Text. SHAP values.** Explanations for SHAP Values and the computations behind.

(PDF)

**S2 Text. Survival analysis.** Definition of survival analysis and the notations used in the study.

(PDF)

## Author Contributions

**Conceptualization:** Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède.

**Data curation:** Hakim Benkirane.

**Formal analysis:** Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède.

**Investigation:** Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède.

**Methodology:** Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède.

**Project administration:** Paul-Henry Cournède.

**Resources:** Stefan Michiels, Paul-Henry Cournède.

**Software:** Hakim Benkirane.

**Supervision:** Stefan Michiels, Paul-Henry Cournède.

**Validation:** Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède.

**Visualization:** Hakim Benkirane.

**Writing – original draft:** Hakim Benkirane, Stefan Michiels, Paul-Henry Cournède.

**Writing – review & editing:** Hakim Benkirane, Yoann Pradat, Stefan Michiels, Paul-Henry Cournède.

## References

1. Dai X, Shen L. Advances and Trends in Omics Technology Development. *Frontiers in Medicine*. 2022; 9:911861. <https://doi.org/10.3389/fmed.2022.911861> PMID: 35860739
2. Misra B, Langefeld C, Olivier M, Cox L. Integrated Omics: Tools, Advances, and Future Approaches. *Journal of Molecular Endocrinology*. 2018; 61(1):R21–45. <https://doi.org/10.1530/JME-18-0055> PMID: 30006342

3. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Springer; 2013.
4. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*. 2016; 17(2):S15. <https://doi.org/10.1186/s12859-015-0857-9> PMID: 26821531
5. van Wyk M, Westerhuis J, Reinecke C, Venter J. Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp. *BMC Bioinformatics*. 2017; 18:83. <https://doi.org/10.1186/s12859-017-1480-8>
6. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In: *Biocomputing 2009*. World Scientific; 2009.
7. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*. 2019; 50:71–91. <https://doi.org/10.1016/j.inffus.2018.09.012> PMID: 30467459
8. Ho W, Erbe R, Danilova L, Phyo Z, Bigelow E, Stein-O'Brien G, et al. Multi-omic profiling of lung and liver tumor microenvironments of metastatic pancreatic cancer reveals site-specific immune regulatory pathways. *Genome Biology*. 2021; 22. <https://doi.org/10.1186/s13059-021-02363-6> PMID: 33985562
9. Fernandez-Rozadilla C, Timofeeva M, Chen Z, Law P, Thomas M, Schmit S, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nature Genetics*. 2022; p. 1–11.
10. Jolliffe IT. *Principal Component Analysis*. Springer; 2002.
11. Sánchez A, Fernández-Real J, Vegas E, Carmona F, Amar J, Burcelin R, et al. Multivariate methods for the integration and visualization of omics data. In: *Spanish Symposium on Bioinformatics*. Springer; p. 382.
12. Hassani S, Hanafi M, Qannari EM, Kohler A. Deflation strategies for multi-block principal component analysis revisited. *Chemometrics and Intelligent Laboratory Systems*. 2013; 120:154–68. <https://doi.org/10.1016/j.chemolab.2012.08.011>
13. Zhang S, Li Q, Liu J, Zhou X. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*. 2011; 27(13): i401–09. <https://doi.org/10.1093/bioinformatics/btr206> PMID: 21685098
14. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*. 2021; 12(1):124. <https://doi.org/10.1038/s41467-020-20430-7> PMID: 33402734
15. Lock E, Dunson D. Bayesian Consensus Clustering. *Bioinformatics*. 2013; 29(20):2610–16. <https://doi.org/10.1093/bioinformatics/btt425> PMID: 23990412
16. Vaske C, Benz S, Sanborn J, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010; 26(12): i237–45. <https://doi.org/10.1093/bioinformatics/btq182> PMID: 20529912
17. Bonnet E, Calzone L, Michoel T. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLoS computational biology*. 2014; 11(2):e1003983. <https://doi.org/10.1371/journal.pcbi.1003983>
18. Newman A, Cooper J. AutoSOME: A clustering method for identifying gene expression modules without prior knowledge of cluster number. *BMC bioinformatics*. 2010; 11(1):117. <https://doi.org/10.1186/1471-2105-11-117> PMID: 20202218
19. Shen R, Mo Q, Schultz N, Seshan V, Olshen A, Huse J, et al. Integrative Subtype Discovery in Glioblastoma Using iCluster. *PloS one*. 2012; 7(4):e35236. <https://doi.org/10.1371/journal.pone.0035236> PMID: 22539962
20. Kaspi A, Ziemann M. Mitch: Multi-contrast pathway enrichment for multi-omics and single-cell profiling data. *BMC Genomics*. 2020; 21. <https://doi.org/10.1186/s12864-020-06856-9> PMID: 32600408
21. Yang S, Zhu F, Ling X, Liu Q, Zhao P. Intelligent Health Care: Applications of Deep Learning in Computational Medicine. *Frontiers in Genetics*. 2021; 12:607471. <https://doi.org/10.3389/fgene.2021.607471> PMID: 33912213
22. Rumelhart DE, Hinton GE, Williams RJ. In: *Learning Internal Representations by Error Propagation*. MIT Press; 1986.
23. Chaudhary K, Poirion OB, Lu L, Garmire L. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research*. 2017; 24(6):1248–59. <https://doi.org/10.1158/1078-0432.CCR-17-0853> PMID: 28982688
24. Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2019. p. 765–69.

25. Zhang X, Xing Y, Sun K, Guo Y. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. *Cancers*. 2021; 13(12):3047. <https://doi.org/10.3390/cancers13123047> PMID: 34207255
26. Simidjievski N, Bodnar C, Tariq I, Scherer P, Andres Terre H, Shams Z, et al. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *Frontiers in Genetics*. 2019; 10:1205. <https://doi.org/10.3389/fgene.2019.01205> PMID: 31921281
27. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Frontiers in Genetics*. 2019; 10:166. <https://doi.org/10.3389/fgene.2019.00166> PMID: 30906311
28. Hira M, Razzaque MA, Angione C, Scrivens J, Sawan S, Sarkar M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Scientific Reports*. 2021; 11(1):6265. <https://doi.org/10.1038/s41598-021-85285-4> PMID: 33737557
29. Tong L, Mitchel J, Chatlin K, Wang M. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC medical informatics and decision making*. 2020; 20(1):225. <https://doi.org/10.1186/s12911-020-01225-8> PMID: 32933515
30. Xie G, Dong C, Kong Y, Zhong JF, Li M, Wang K. Group Lasso Regularized Deep Learning for Cancer Prognosis from Multi-Omics and Clinical Features. *Genes*. 2019; 10. <https://doi.org/10.3390/genes10030240> PMID: 30901858
31. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*. 2013; 35(8):1798–828. <https://doi.org/10.1109/TPAMI.2013.50> PMID: 23787338
32. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*. 2020; 11.
33. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In: *International Conference on Learning Representations*; 2017.
34. Viaud G, Mayilvahanan P, Cournede P. Representation Learning for the Clustering of Multi-Omics Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2021; 19:135–45. <https://doi.org/10.1109/TCBB.2021.3060340>
35. Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*. 2021; 19. <https://doi.org/10.1016/j.csbj.2021.06.030> PMID: 34285775
36. Zhao S, Song J, Ermon S. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:170602262*. 2017;.
37. Katzman J, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*. 2018; 18(1):24. <https://doi.org/10.1186/s12874-018-0482-1> PMID: 29482517
38. Withnell E, Zhang X, Sun K, Guo Y. XOmivAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*. 2021; 22:315. <https://doi.org/10.1093/bib/bbab315> PMID: 34402865
39. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.; 2017. p. 4765–4774.
40. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*. 2016; 375(12):1109–12. <https://doi.org/10.1056/NEJMp1607591> PMID: 27653561
41. Yang Z, Xu Q, Bao S, Cao X. Learning with Multiclass AUC: Theory and Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021; PP:1–1. <https://doi.org/10.1109/TPAMI.2021.3081597>
42. McInnes L, Healy J, Saul N, Grossberger L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. 2018; 3:861. <https://doi.org/10.21105/joss.00861>
43. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019. p. 8024–8035. Available from: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
44. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*. PMLR; 2015. p. 448–456.
45. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. 2014; 15(1):1929–1958.

46. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. In: Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 2; 2012.
47. Yi J, Ren L, Li D, Wu J, Li W, Du G, et al. Trefoil factor 1 (TFF1) is a potential prognostic biomarker with functional significance in breast cancers. *Biomedicine & Pharmacotherapy*. 2020; 124:109827. <https://doi.org/10.1016/j.biopha.2020.109827> PMID: 31986408
48. Harrell J, Frank E, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA*. 1982; 247:2543–2546. <https://doi.org/10.1001/jama.1982.03320430047030>
49. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*. 1999; 18 17-18:2529–45. [https://doi.org/10.1002/\(SICI\)1097-0258\(19990915/30\)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5](https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5) PMID: 10474158