PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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

Hakim Benkirane[1,2], Yoann Pradat[1], Stefan Michiels[2,3], Paul-Henry Cournède[1]*

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

BIBS seminar

2023. 12. 28.

Presenter: Jun Sik Kim

Table of Contents

Introduction

Introduction

- Cancer is a complex disease involving multiple genetic and environmental factors, where various factors affect biological systems on many levels.
- To better characterize a patient's molecular profile, multi-omics approaches rely on multiple dimensions simultaneously.
- However, analyzing the mix of data sources and leveraging their information to improve our understanding of the disease's underlying biological phenomena remain challenging with 2 main points.
    1. **High dimensionality of the data**
       Omics data generally suffers from the 'curse of dimensionality'
    2. **Data heterogeneity**
       Omics data is very diverse, with different scaling applied for each dataset

Introduction

- In the past few years, multi-omics integration has been a very active research subject in health science and precision medicine, giving insight into biological processes involved with cancer.
- Multiple statistical learning methods have been proposed to investigate various complex molecular systems behind cancer.
- Previous methods on multi-omics data integration can be classified into 3 main categories:
    1. Statistical learning methods
    2. Clustering methods
    3. Deep learning-based methods

Introduction

- Statistical learning methods
  - **Principal Component Analysis (PCA):** linear dimensionality reduction technique
  - **Consensus PCA:** variation of PCA for several related datasets or blocks of variables, finding a shared lower-dimensional space to represent all datasets
  - **Multi-block PCA:** applied for multiple datasets, preserving block-specific structures as well as common trends
  - **Multiple Factor Analysis:** Extension of PCA for settings where multiple sets of variables on the same set of observations exist
  - **Non-negative Matrix Factorization(NMF):** group of algorithms where a matrix is factorized into two (or more) matrices with no negative elements(only additive combinations allowed).

Introduction

- Clustering methods
  - **AutoSOME(Newman, 2010)**: clustering algorithm that combines Self-Organizing Maps (SOMs) with ensemble averaging method to create stable and robust clustering results
    - Present each node consisting of a feature vector with reduced dimensionality of the gene expression data, and cluster feature vectors with similar weights
  - **iCluster(Shen, 2009)**: integrate diverse data types to identify subtypes of diseases based on integrated molecular profiles
    - Joint latent variable model without non-negative constraint

Introduction

- Deep-learning methods
  - Autoencoders framework
    - Tong et al. : multi-view learning in survival analysis for Breast Cancer
  - Variational Autoencoder framework
    - OmiVAE – representational learning
      - Hira et al. : Ovarian Cancer study
    - OmiEmbed – multitask learning

- No benchmarking study has explored and compared different deep learning approaches and strategies for multi-omics data integration in multitask learning so far.

Introduction

- In this work, the authors first discuss strategies for integrating high-dimensional multi-source data to learn low-dimensional latent representation from multi-omics datasets.
- Then a new customizable architecture for multi-omics integration is represented, called **CustOmics**.
- CustOmics combines the advantages of the different strategies and alleviates some limitations of each methods by applying a mixed integration of the VAE structure.
- The impact of this new method is evaluated on different test cases on both classification and survival tasks by applying it to a pan-cancer dataset, and then assessing it on smaller datasets of specific subtypes like breast cancer.

Materials & 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.
- In the field of multiple omics data integration, it can help synthesize the heterogeneous distributions into a shared space, revealing the underlying interactions between different sources.
- Each patient is characterized by K omics vectors

$$(x_k)_{1 \leq k \leq K} \in \mathbb{R}^{M_k}$$

(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 $z \in \mathbb{R}^m$ with $m \ll \sum_{k=1}^{K} M_k$, the latent representation.
- The latent features are the components of the vector z.

Representation learning for multi-omics integration

- 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.

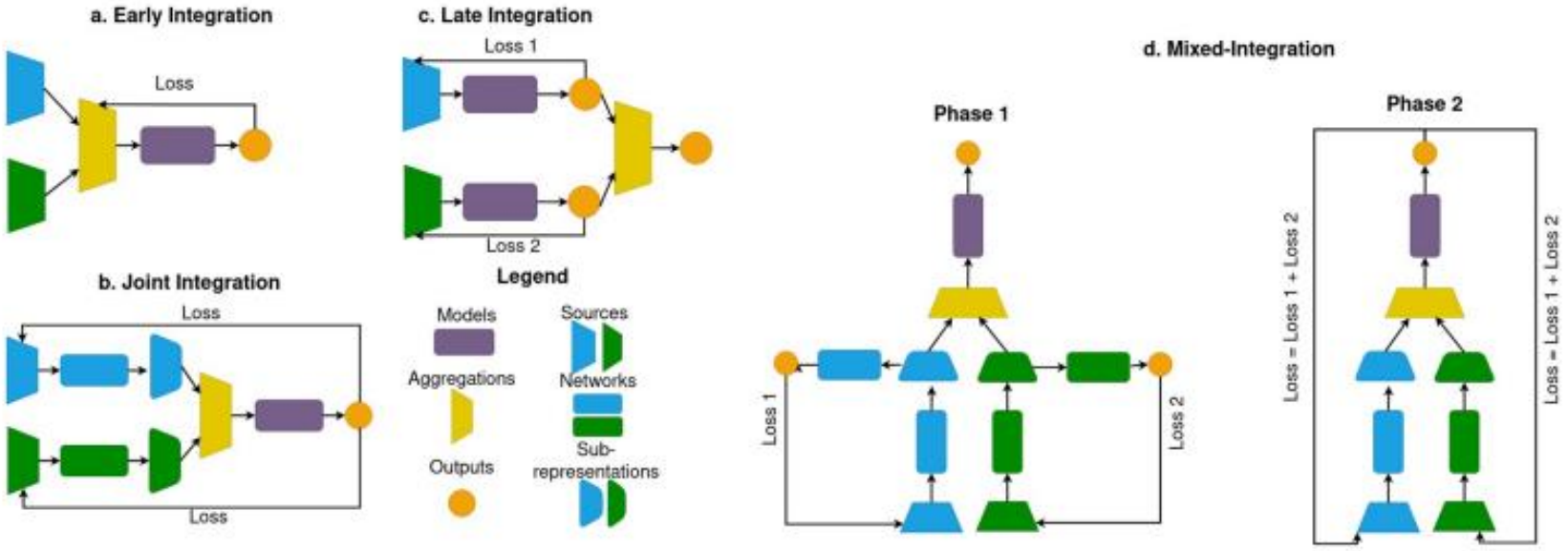


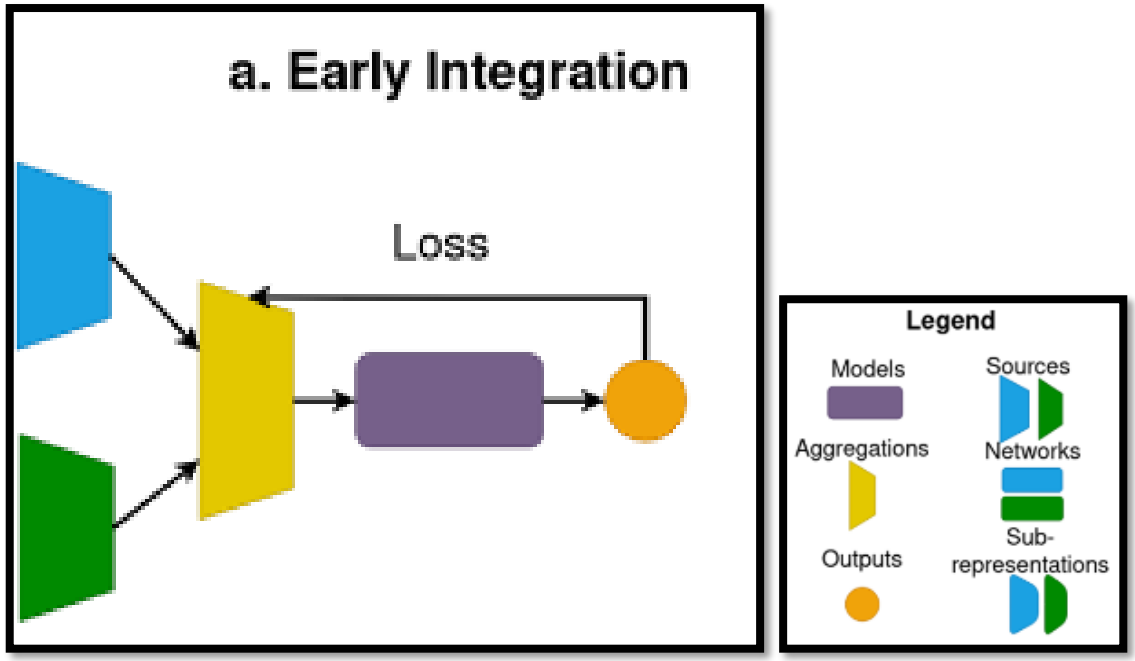
Fig 1. a. Early Integration: Sources are concatenated before being fed to a single model. **b. Joint Integration:** Sub-representations of each source are learned jointly before inputting into the model. **c. Late Integration:** Each source outputs its prediction using its independent model, and the 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

1. Early Integration(EI)

- Set of methods that aim at merging the different sources before the dimensionality reduction.
- We first concatenate the vectors of all the sources, instead of giving as model inputs a set of separate vectors for each source.

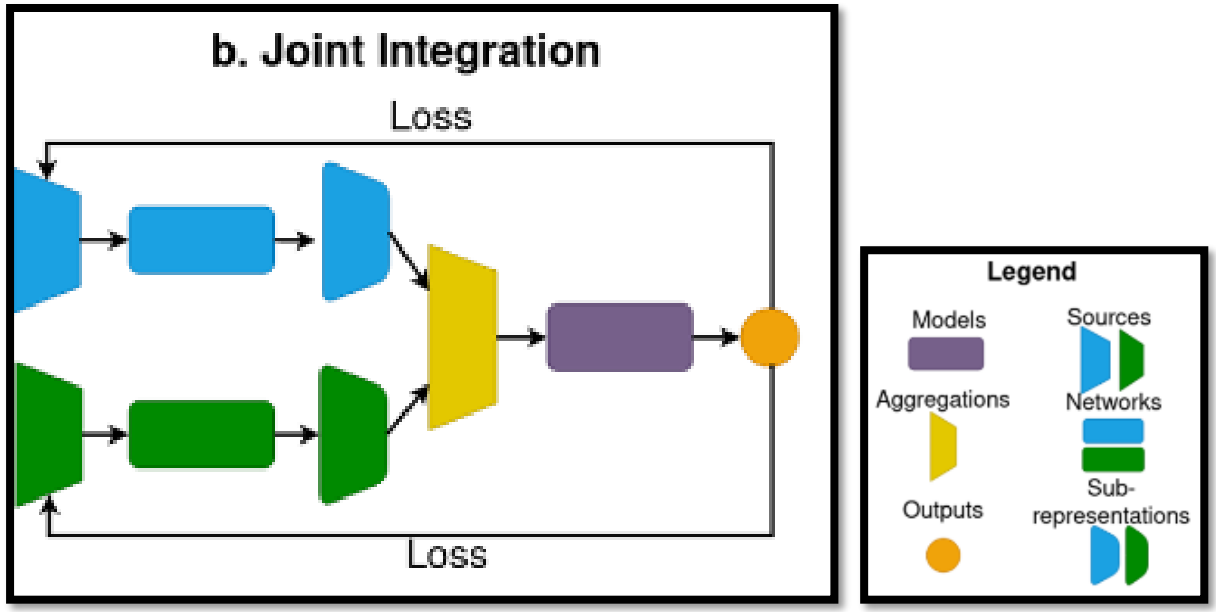
- Advantage: simplicity
- Limitations: If some sources bear more significant signals than others, it might be hard to learn interactions between sources.



2. Joint Integration(JI)

- Sub-representations are created inside the same model for each source before learning the output.

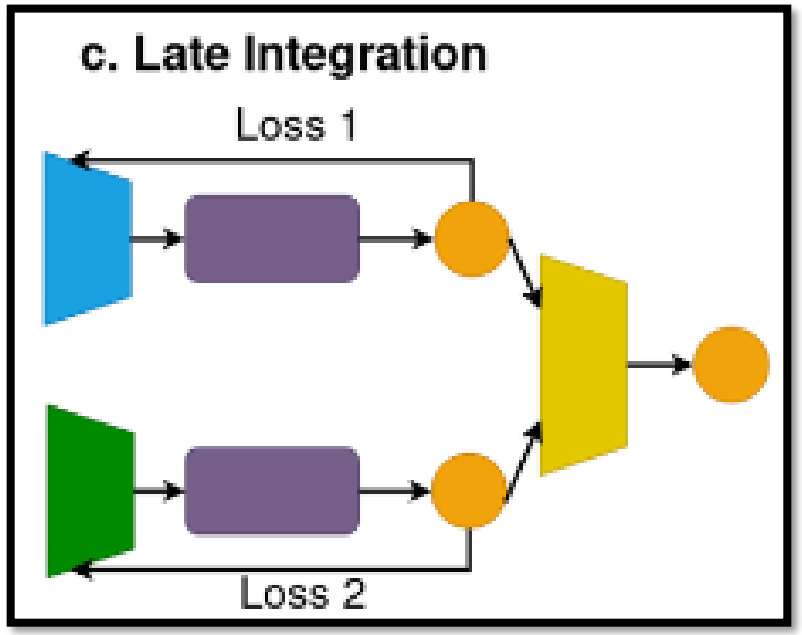
- Advantage: most theoretically promising, most widely used
- Limitations: challenged by the sources' heterogeneity, as they do not necessarily follow the same learning dynamics and may need different approaches with specific losses

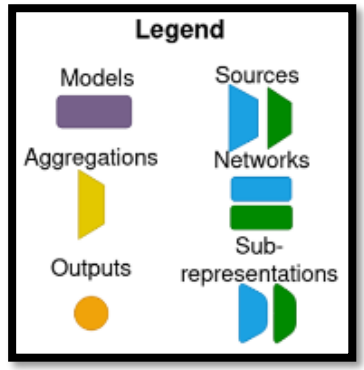


3. Late Integration(LI)

- Consists in learning the output for each source separately, with its own model.

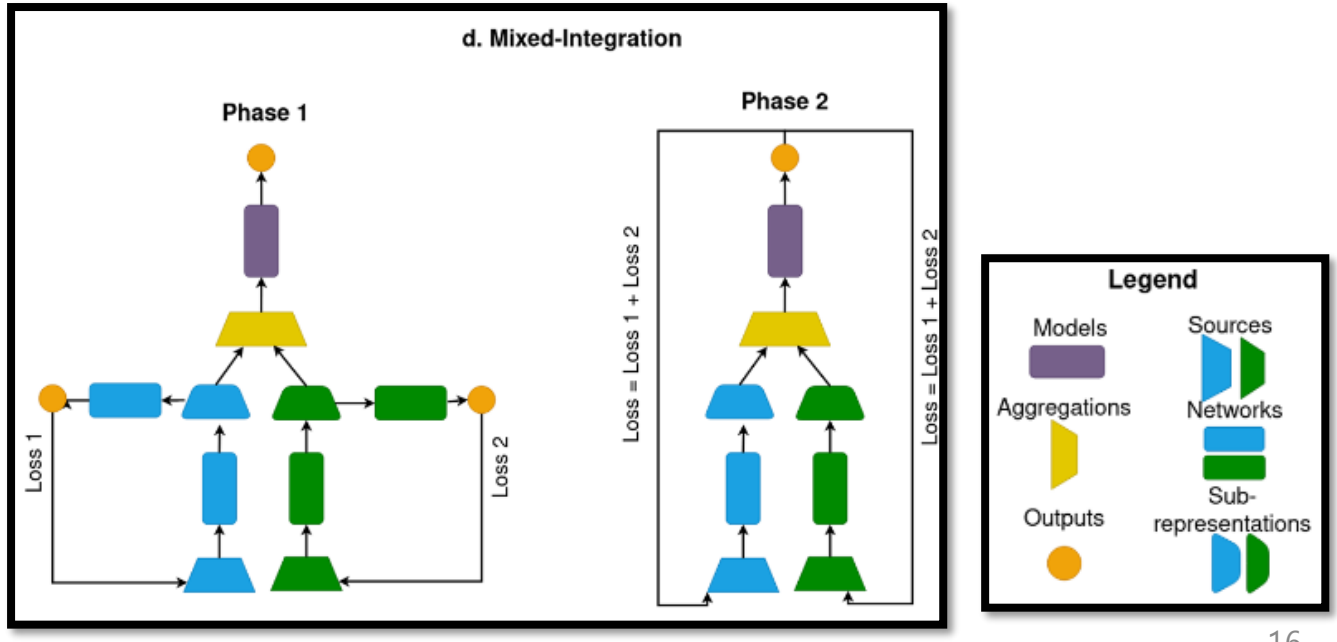
- Advantage: adopt well to the specificities of each source
- Limitations: does not retrieve any cross-modality interaction





Mixed-integration

- To alleviate the issues brought by different integration methods, the paper proposes a hybrid strategy, named **mixed-integration**, that comes halfway between joint and late integration.
- It consists of 2 learning phases that switch during a predefined epoch considered a hyperparameter: the 1st 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 2nd phase to build a global output representation.



Variational Autoencoders

- A Variational Autoencoder(VAE) is a deep generative model which can learn meaningful data representations from high-dimensional input data.
- VAE can encode a particular distribution. After the encoding phase, there is a sampling phase in which we sample points from the distribution $q_\phi(z|x)$.
  - Encoding function q: $q_\phi(z|x)$ – variational distribution(encoding distribution)
  - Decoding function p: $p_\theta(x|z)$ – posterior
- Traditionally, the distributions in the VAE architecture are supposed to be Gaussian: the encoder function will learn the two-parameter vectors μ, σ that are used to generate samples in $q_\phi(z|x)$ using the reparameterization trick:

$$z = \mu + \sigma \odot \epsilon\,, \qquad \epsilon \sim \mathcal{N}(0, I)$$

Variational Autoencoders

- The loss function for this architecture can be written as the sum of 2 distinct losses:
  1. Reconstruction loss

$$\mathcal{L}_{recon} = \mathbb{E}_{q_\phi(z|x)}[p_\theta(x|z)]$$

→ Interpretation: conditional entropy of x given z, quantifies the uncertainty one has over the joint distribution (x, z), knowing z

  2. Regularization loss: Kullback–Leibler(KL) divergence

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

→ Interpretation: measurement of difference between the variational distribution and the prior distribution

- Total loss is defined with a hyperparameter β as:

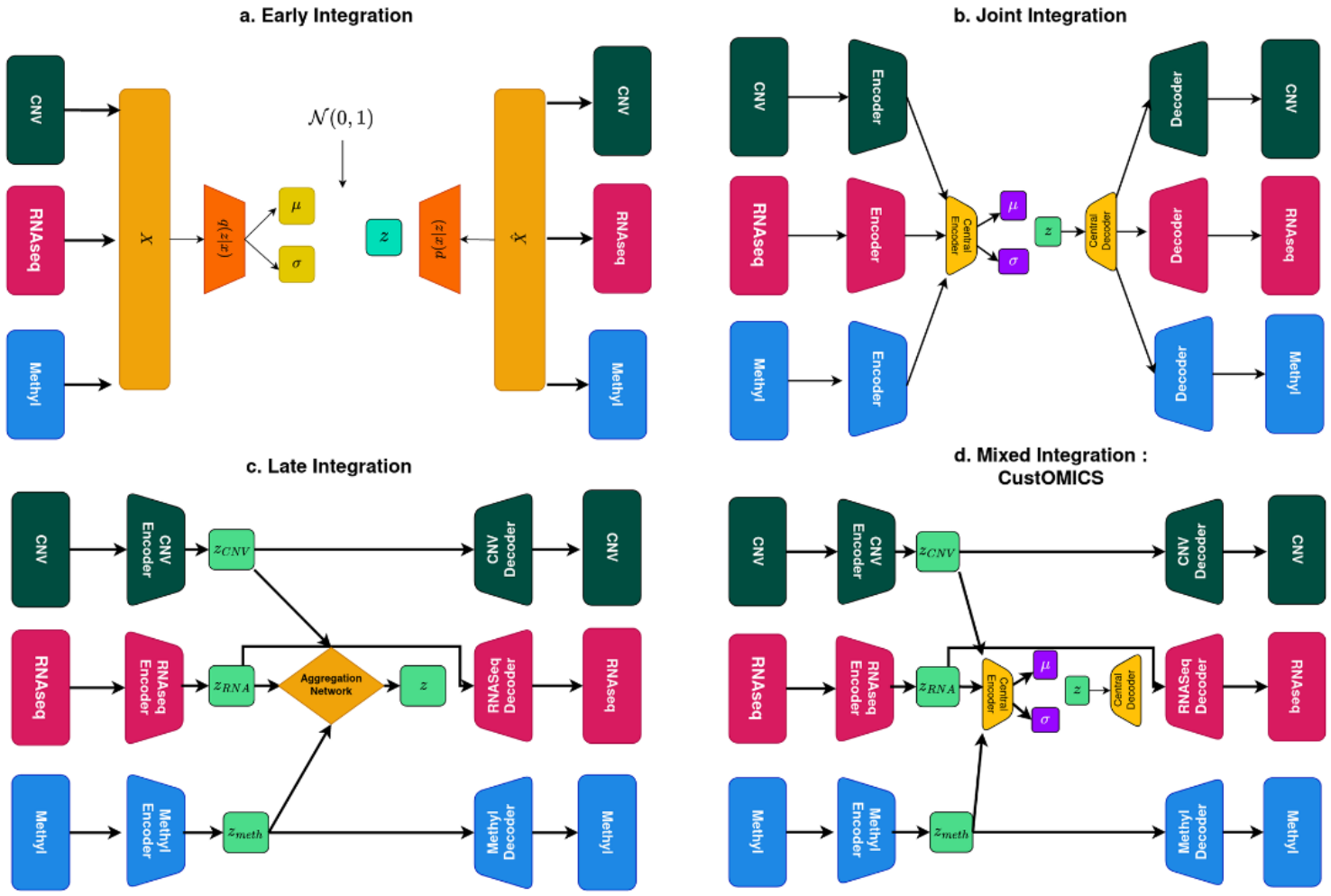
$$\mathcal{L} = \mathcal{L}_{recon} + \beta\mathcal{L}_{reg}$$

Variational Autoencoders

- The early integration, joint integration, late integration and mixed integration strategies were applied in the framework of VAE.
- Joint integration is the most commonly used setting of multi-source integration using VAE.
- This paper seeks to compare CustOmics with other common settings of encoding methods.

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.



Variational Autoencoders – Mixed-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 strategy benefits from 2 training phases:

  Phase 1: normalization process – each source train separately for a more compact representation
  Phase 2: joint integration between the learned sub-representation
- CustOmics used the Maximum Mean Discrepancy(MMD) to assess the distance between the distributions, by comparing how similar samples are within each distribution & between distributions.

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

- $\kappa(x,x')$: Gaussian kernel function where $\kappa(x,x') = \exp\left(\frac{-||x-x'||^2}{2\sigma^2}\right)$
- x, x': two sample points

Variational Autoencoders – Mixed-integration / CustOmics

- The paper used the deep learning framework 'DeepSurv' for the survival task, based on the negative partial log-likelihood formula. The loss function in this framework is:

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

- E_i: event for patient i
  - $i: E_i = 1$: summing over all individuals i for whom the event has occurred
- $\hat{\mu}(x;\theta)$: risk function associated with the risk score estimated by the output layer
- $\mathcal{R}(t)$: set of patients still at risk of failure after time t

Test cases and datasets

- In this study, the authors use datasets extracted from the Genomic Data Commons (GDC) pan-caner multi-omics study.
  - High-dimensional omics data + phenotype data from The Cancer Genome Atlas(TCGA)
- Omics data in use: Copy Number Variations (CNV), RNA-Seq gene expressions, DNA methylation
  - CNV: 19,729 genes
  - RNA-Seq expression profile: 60,484 identifiers referring to corresponding exon, measuring log2 transformed Fragments Per Kilobase of transcript per Million mapped reads (FPKM)
  - DNA methylation dataset: 485,578 probes with methylation ratio of corresponding CpG sites
- Evaluation: 5 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)
  - Uterine Corpus Endometrial Carcinoma (UCEC, n=538)

Test cases and datasets

- 1st task: classifying for different tumor types in the pan-cancer study

- 2nd task: validation & test robustness
  - Aims to perform tumor subtype classification based on the PAM50 classification

- 3rd task: survival study of the Pancancer dataset

- 4th task: evaluate survival performances of the 5 datasets for validation
  - Aims to find how well the model discriminates between risk groups

Data preprocessing

- RNA gene expression profiles: 594 exons on Y chromosome & 1,904 zero-expression & 248 missing values removed
- DNA methylation data: filter out Y chromosome & zero-expression & missing values & probes that do not map to the human reference genome → 438,831 CpG sites

- Afterwards, each combination of omics data were intersected to retrieve the maximum number of samples for each test case.
- Then the features with missing/consistently zero/NA values were identified and removed.
- The research applied normalization to non-normalized datasets such as CNVs and RN-Seq data to ensure that each omic source was scaled identically, thus would have the same importance during integration.

Results

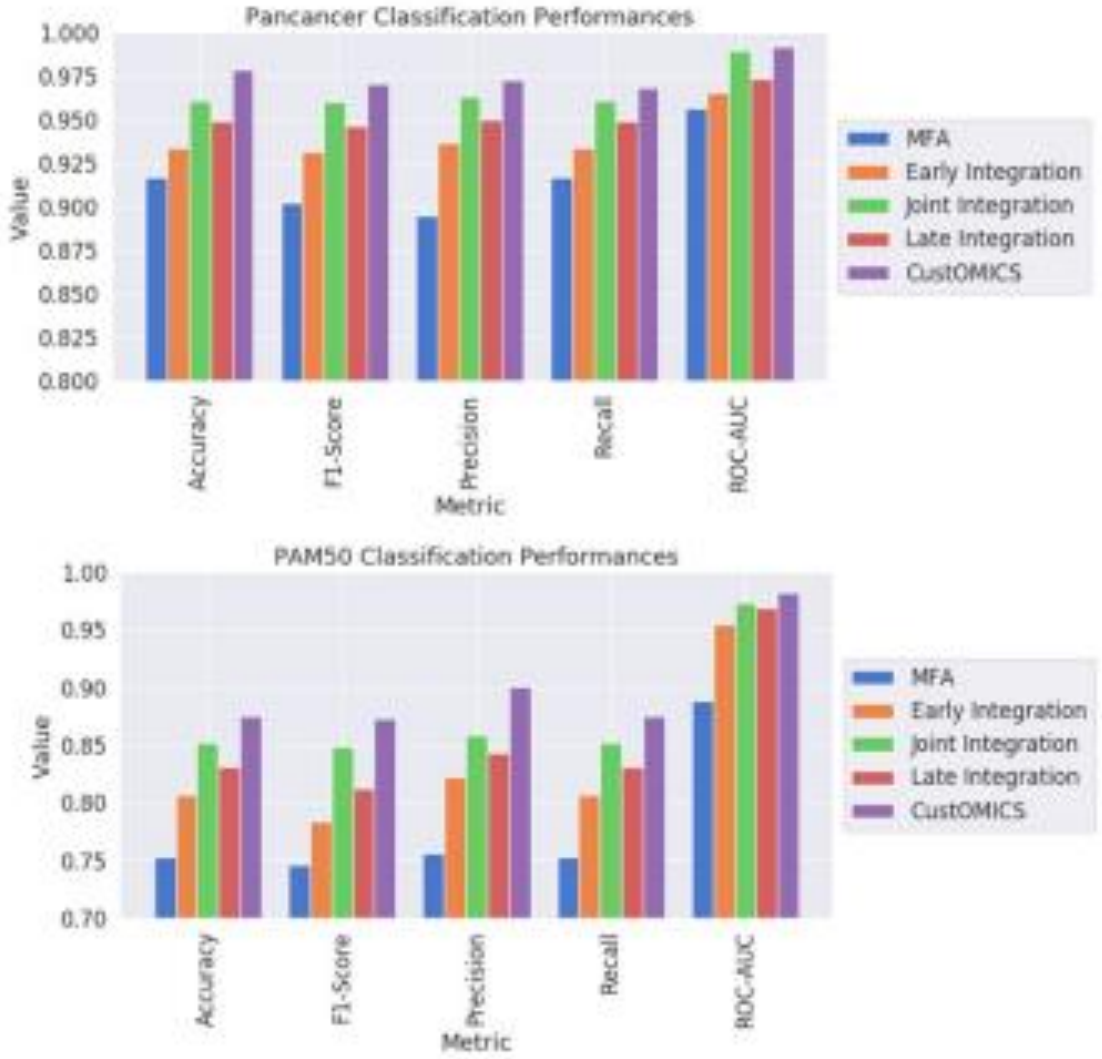
(1) Classification results

- The authors first perform the classification task on the pan-cancer dataset, where each architecture is coupled with an artificial neural network classifier composed of
  - 2 hidden layers with 256 and 128 neurons
  - ReLU activation function on the hidden layer
  - Softmax activation function on output layer
- The total loss of the model with the classification task is:

$$\mathcal{L}_{total} = \mathcal{L}_{AE} + \alpha\mathcal{L}_{task}$$

- Factorial methods does not perform as well as mot deep-learning methods, because MFA cannot uncover nonlinear relationships between different sources, unlike deep-learning architectures.

a.



(1) Classification results: Pancancer dataset

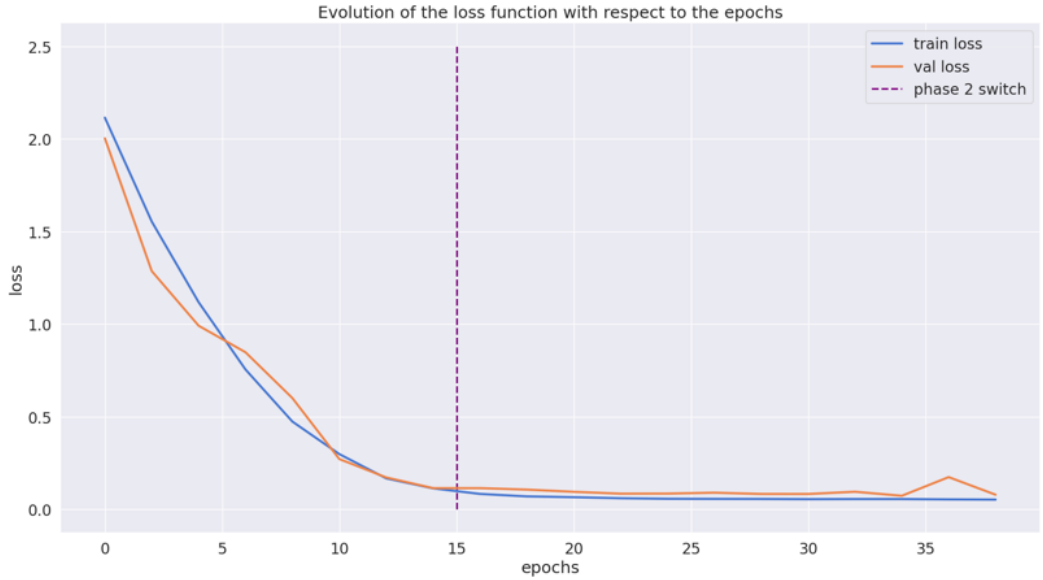
- Early integration is behind the other DL methods in terms of performance, as early integration makes the model overlook interactions between sources.
- Late integration is not optimal since interactions are not properly learned from separated architecture. (Table S3)
- Joint integration performs well in most cases, but best results are achieved by combination of only 2 sources. (Table S4)

Table 1. The classification performance for the pan-cancer dataset is evaluated with 5 standard metrics for UMAP, NMF, MFA, Unsupervised Customics 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
CustOmics	**0.9788 ± 0.0025**	**0.9705 ± 0.0033**	**0.9728 ± 0.0041**	**0.9685 ± 0.0034**	**0.9918 ± 0.0001**

(1) Classification results: Pancancer dataset

- From the results we can imply that:
  1. CustOmics gives the best performances for all the test cases without apparent overfitting.
  2. CustOmics takes advantages of the complementarity and interactions between sources, with all sources bringing additional information.
     - The model does not solely depend on one specific source of data.
     - Transcriptomics gives promising performances in most cases, as most information about tumor types and molecular subtypes is directly related and expressed in RNA data.
     - This could be done by the unique architecture of considering 2 phases in a single run.



(2) Classification results: validation datasets (BRCA; N=1,022)

- Fig.3 gives visualization of the different sources used in the model.
  - The CustOmics framework is capable of extracting meaningful relationships from the integrated omics data, leading to a clearer understanding of the underlying biological processes.

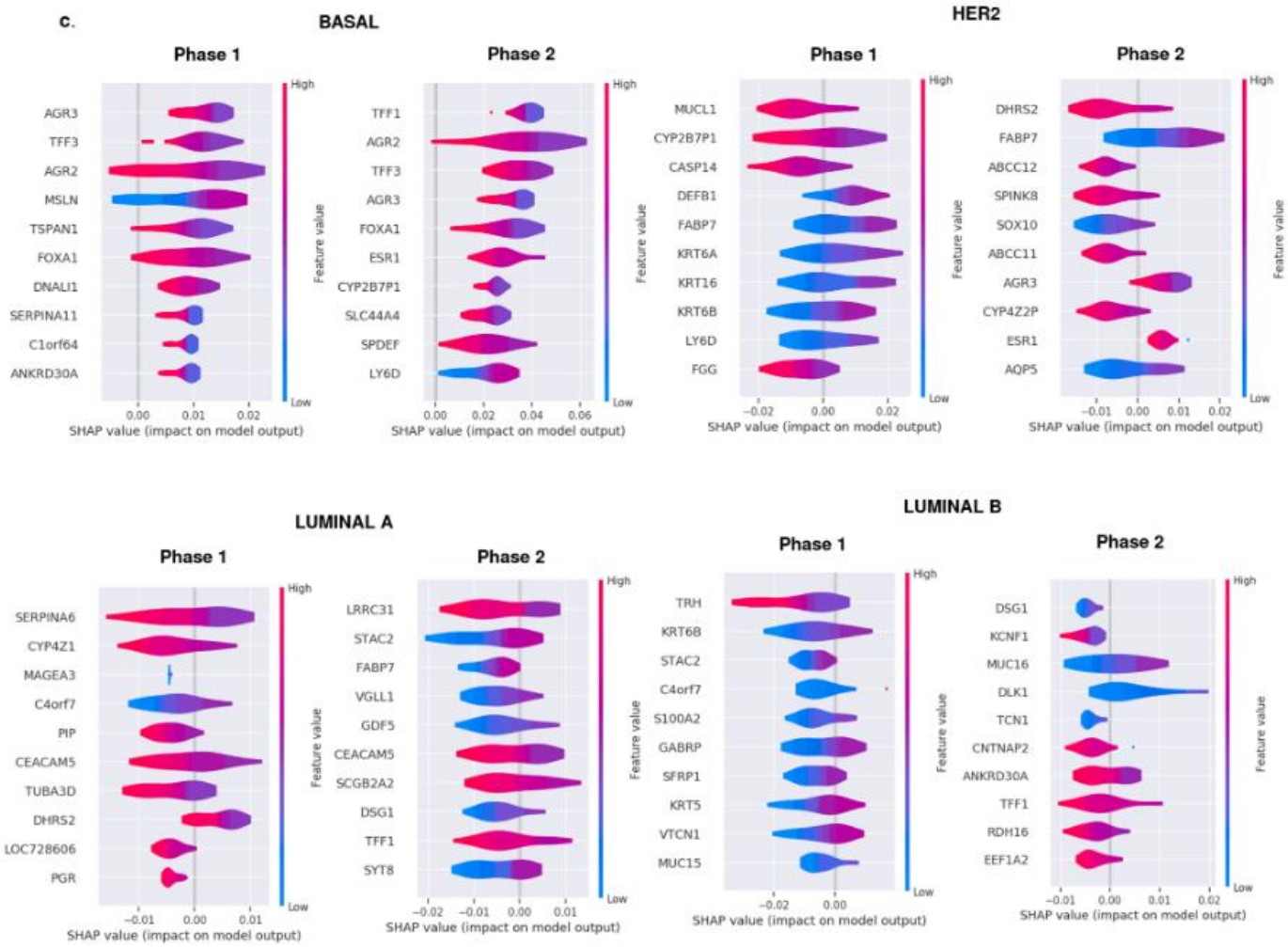
- Fig.3(b): T−SNE visualization for:
  - Each omic source separately
  - Latent representation by CustOmics

- Constructed layer representation succeeds at separating the data into 4 clusters that we could not distinguish with each omic source alone.

(2) Classification results: validation datasets (BRCA; N=1,022)

- Fig.3 gives visualization of the different sources used in the model.

- Fig.3(c): PAM50 gene importance
  - SHAP values computed on RNA-Seq data of the most relevant genes responsible for the discrimination between subtypes
  - SHAP values were adopted in the CustOmics architecture to provide interpretable results.
  - This was done for both phases:
    Phase 1: single omic source
    Phase 2: interactive source



(3) Survival analysis

- Goal: To predict the risk score associated with each patient from the corresponding high-dimensional omics data.
- Performance metrics
  - C-index: measure of probability that the predicted event times of 2 randomly selected individuals have the same relative order as their true event times

$$\text{c-index} = \frac{\sum_{i \in U} \left\{ \sum_{T_j > T_i} 1_{f_j > f_i} \right\}}{\sum_{i \in U} \left\{ \sum_{T_j > T_i} 1 \right\}}$$

- U: a set of uncensored data
- T_i: an observed survival time of sample i
- f_i: a predicted survival time of sample i
- $1_{a>b}$: indicator function(1 if $a > b$, and 0 otherwise)

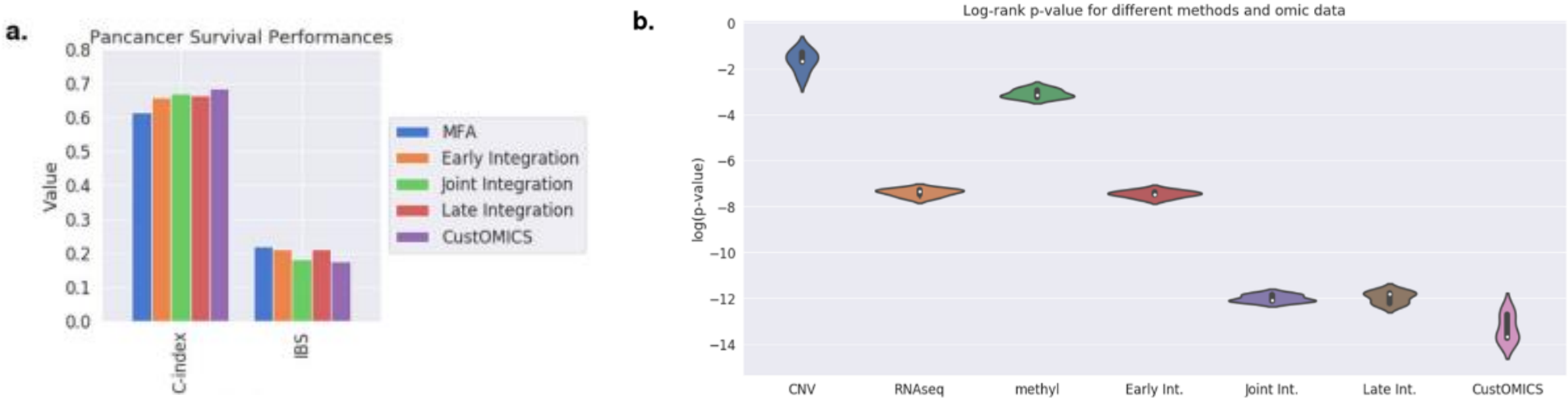
  - Integrated Brier Score(IBS): measures accuracy of probabilistic predictions for an event, using Brier score(MSE between observed outcomes and predictions) over the time period of interest

$$IBS = \frac{1}{T - t_0} \int_0^T \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2 \, dt$$

- t_0, T: start / end of study period
- n: total number of individuals in the study
- p_i: predicted probability of survival at time t
- o_i: observes status of survival at time t

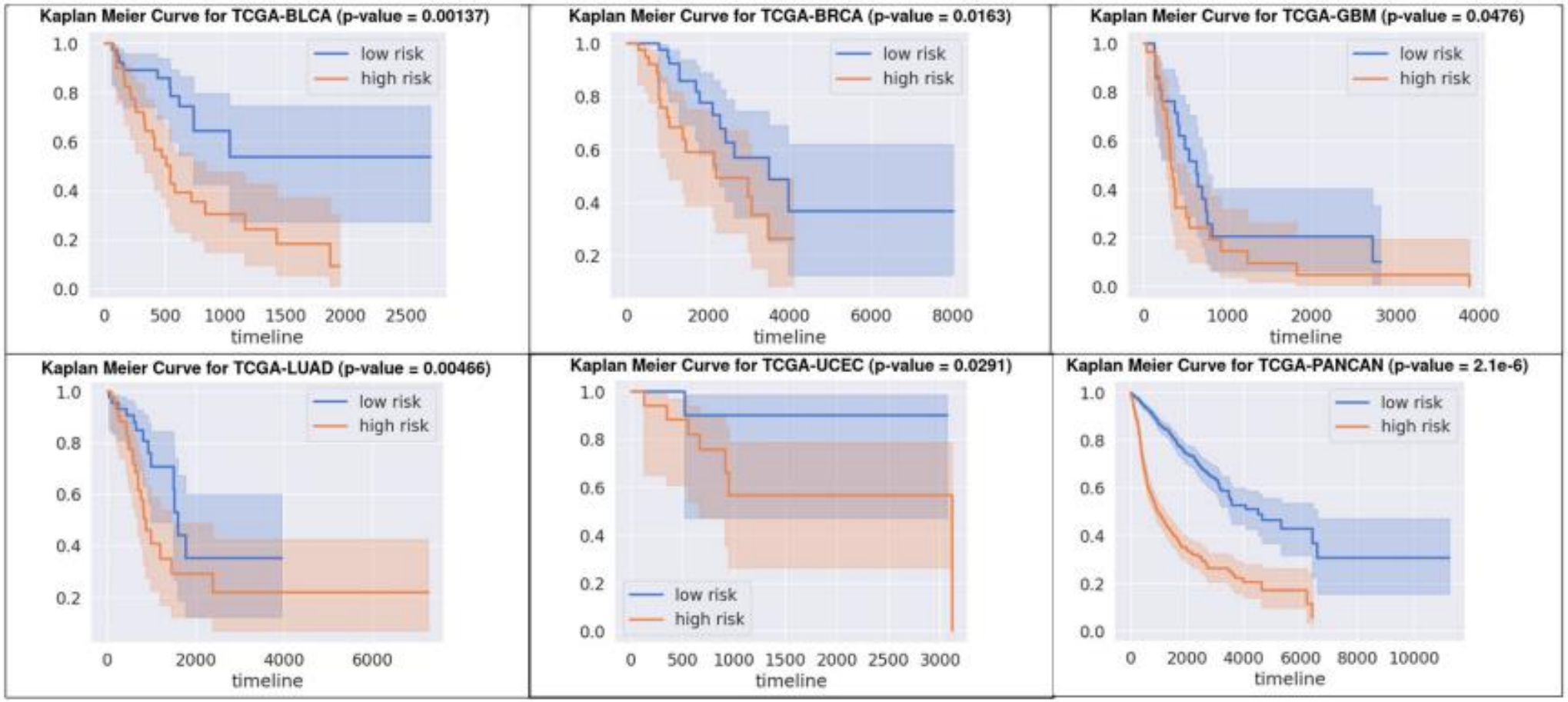
(3) Survival analysis

- Fig. 4(a): performance of survival model for pancancer dataset with C-index and IBS
- Fig. 4(b): [Log-rank test results] p-value between high and low-risk groups for every integration strategy on a validation set for the pan-cancer survival test case compared to mono-omic survival predictions



(4) Survival analysis: validation datasets

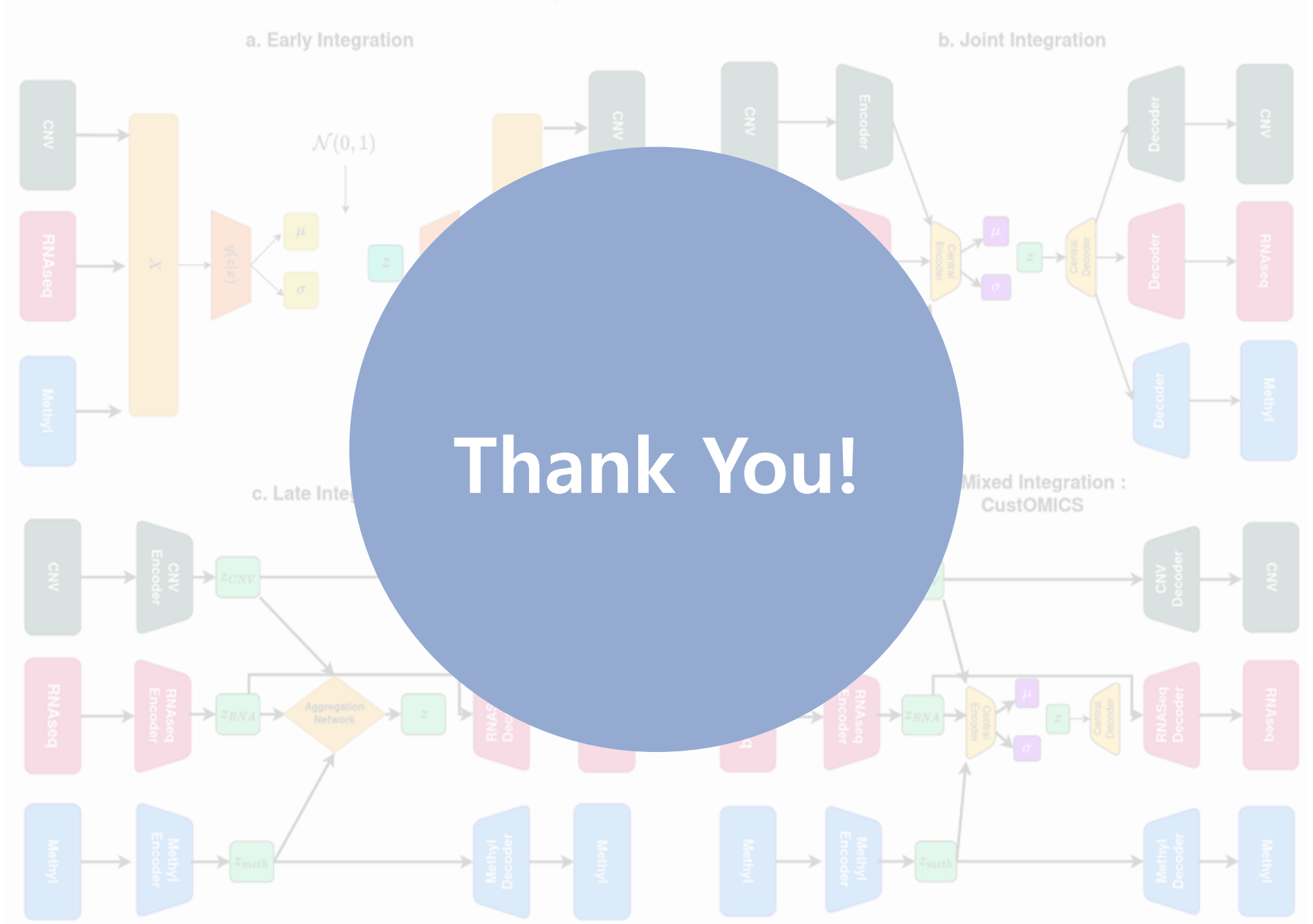
- Fig. 4(c): [Kaplan Meier Curves] KM–curves for each cancer datasets from the CustOmics model, stratifying population into high & low risk



Discussion

Discussion

- In this work, the authors 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, they 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, CustOmics 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.
- In conclusion, CustOmics 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.


a. Early Integration
b. Joint Integration
CNV
RNAseq
Methyl
Encoder
Decoder
Central Encoder
Central Decoder
Mixed Integration :
CustOMICS
CNV Encoder
RNAseq Encoder
Methyl Encoder
CNV Decoder
RNASeq Decoder
Methyl Decoder
Aggregation Network

Thank You!

(1) Classification results: Pancancer dataset

- Early integration is behind the other DL methods in terms of performance.
  - RNA-Seq data hold more signals when determining tumor types or subtypes.
  - Concatenating the sources before feeding them to the VAE overshadows other sources, thus making the model rely solely on RNA-Seq data.

Omics	Accuracy	F1-score	Precision	Recall	ROC-AUC
CNV	0.47 +- 0.03	0.47 +- 0.03	0.47 +- 0.03	0.48 +- 0.01	0.75 +- 0.02
RNAseq	0.92 +- 0.01	0.93 +- 0.01	0.93 +- 0.01	0.92 +- 0.01	0.96 +- 0.00
methyl	0.68 +- 0.02	0.68 +- 0.02	0.68 +- 0.02	0.68 +- 0.02	0.82 +- 0.01
CNV + RNAseq	0.93 +- 0.01	0.92 +- 0.01	0.92 +- 0.01	0.92 +- 0.01	0.96 +- 0.00
CNV + methyl	0.71 +- 0.02	0.69 +- 0.02	0.70 +- 0.02	0.69 +- 0.02	0.85 +- 0.01
RNAseq + methyl	0.94 +- 0.01	0.94 +- 0.01	0.94 +- 0.01	0.94 +- 0.0132	0.97 +- 0.00
CNV + RNAseq + methyl	0.98 +- 0.01	0.97 +- 0.01	0.97 +- 0.01	0.97 +- 0.01	0.99 +- 0.00

(1) Classification results: Pancancer dataset

- Performances of joint model by omics input
  - it seems that CNV data only adds noise to the latent representation, meaning that its information is not handled well with this strategy.

Omics	Accuracy	F1-score	Precision	Recall	ROC-AUC
CNV	0.47 +- 0.03	0.47 +- 0.03	0.47 +- 0.03	0.48 +- 0.01	0.75 +- 0.02
RNAseq	0.92 +- 0.01	0.93 +- 0.01	0.93 +- 0.01	0.92 +- 0.01	0.96 +- 0.00
methyl	0.68 +- 0.02	0.68 +- 0.02	0.68 +- 0.02	0.68 +- 0.02	0.82 +- 0.01
CNV + RNAseq	0.90 +- 0.02	0.89 +- 0.02	0.90 +- 0.02	0.88 +- 0.03	0.93 +- 0.00
CNV + methyl	0.70 +- 0.02	0.69 +- 0.02	0.70 +- 0.02	0.70 +- 0.02	0.85 +- 0.01
RNAseq + methyl	0.96 +- 0.01	0.96 +- 0.01	0.96 +- 0.01	0.96 +- 0.0132	0.99 +- 0.00
CNV + RNAseq + methyl	0.94 +- 0.01	0.94 +- 0.01	0.95 +- 0.01	0.94 +- 0.01	0.97 +- 0.00