Bimodal masked language modeling for bulk RNA-seq and DNA methylation representation learning

Maxence Gélard [1] [2] Hakim Benkirane [2] Thomas Pierrot [1]
Guillaume Richard [1] Paul-Henry Cournède [2]

Abstract

Oncologists are increasingly relying on multiple modalities to model the complexity of diseases. Within this landscape, transcriptomic and epigenetic data have proven to be particularly instrumental and play an increasingly vital role in clinical applications. However, their integration into multimodal models remains a challenge, especially considering their high dimensionality. In this work, we present a novel bimodal model that jointly learns representations of bulk RNA-seq and DNA methylation leveraging self-supervision from Masked Language Modeling. We leverage an architecture that reduces the memory footprint usually attributed to purely transformer-based models when dealing with long sequences. We demonstrate that the obtained bimodal embeddings can be used to fine-tune cancer-type classification and survival models that achieve state-of-the-art performance compared to unimodal models. Furthermore, we introduce a robust learning framework that maintains downstream task performance despite missing modalities, enhancing the model's applicability in real-world clinical settings. Code available at `https://github.com/instadeepai/multiomics-open-research`.

1. Introduction

The growing availability of high-throughput technologies has revolutionized molecular research, generating extensive genomic, transcriptomic, and epigenomic data that holds immense potential for personalized medicine (Ho et al., 2021; Stark et al., 2019; Dai & Shen, 2022). Cancer diagnosis and prognosis thus increasingly rely on heterogeneous patient data, and the integration of these diverse data sources remains a significant challenge, even more so when some modalities may be missing in real clinical applications.

The high dimensionality of each modality often makes classic machine learning and deep learning methods ineffective for diagnostic purposes. As a result, there is a growing tendency to first learn representations of the data, particularly using self-supervised approaches. In this context, foundation models have steadily emerged as powerful tools to learn effective and generalizable embeddings that can be applied to biological and clinical tasks. Trained with an unsupervised language modeling objective, they have already been applied to a wide range of omics data, including genomics (Dalla-Torre et al., 2025; Brixi et al., 2025), single-cell transcriptomics (Cui et al., 2024) or bulk RNA-seq (Gélard et al., 2025). These models extensively leverage the transformer architecture (Vaswani et al., 2017) which limits the maximum input length of the model due to the quadratic memory scaling of the attention mechanism. Recent models have developed new architectures to cope with these long-range sequences, either by integrating convolutional blocks (Avsec et al., 2021; Linder et al., 2025; Joshi et al., 2025) or state-space models (Popov et al., 2025).

Among multiple studies, the Cancer Genome Atlas (TCGA, `https://portal.gdc.cancer.gov/`) is a publicly available dataset that gathers multi-omics data, in particular bulk RNA-seq and DNA methylation which are the focus of this work. Including clinical information such as survival time and divided into 33 cohorts (or cancer types), this dataset is a popular benchmark for evaluating survival analysis and cancer-type classification methods. Survival analysis, or time-to-event prediction, aims at predicting the time from diagnosis to the patient's death from the disease using censored data. Though classically tackled with Cox regression (Cox, 1972), the Cox partial likelihood has more recently been reformulated as a loss used to train deep learning architectures (Ching et al., 2018; Katzman et al., 2018).

In this paper, we introduce *MOJO*, standing for **M**ulti-**O**mics **JO**int representation learning, that we here tailor

[1] InstaDeep [2] Université Paris-Saclay, CentraleSupélec, Lab of Mathematics and Computer Science (MICS). Correspondence to: Maxence Gélard <m.gelard@instadeep.com>.

Proceedings of the Workshop on Generative AI for Biology at the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

for learning joint embeddings of bulk RNA-seq and DNA methylation through bimodal masked language modeling from the TCGA dataset. We leverage a multimodal architecture that employs a mix of convolutional and transformer layers. We show that the embeddings learned by *MOJO* lead to state-of-the-art performance in various tasks from pan-cancer cancer-type classification, survival analysis, and cancer subtype clustering. Finally, we further present a framework that allows for a downstream task model to preserve its performance in the absence of a modality by introducing an auxiliary loss based on mutual information.

2. Related works

Omics representation learning

Omics representations were usually derived from statistical methods such as Principal Component Analysis (Jolliffe, 2002) or Non-negative Matrix Factorization (NMF) (Lee & Seung, 2000), the latter being particularly suited for RNA-seq and DNA methylation due to their positivity. Deep learning architectures such as Masked Auto-Encoders (Gross et al., 2024) or Mixture-of-Experts (Meng et al., 2023) have then been applied to learn omics representations used either for survival analysis or cancer-type predictions. In line with foundation models for single-cell transcriptomics (Cui et al., 2024; Yang et al., 2022), Gélard et al. 2025 developed a transformer-based model for bulk RNA-seq representation learning. Multi-modal integration is often performed using late integration, *i.e.*, each source is encoded separately before being aggregated, often using Kronecker product (Chen et al., 2020), element-wise operations (Vale-Silva & Rohr, 2021) or cross-attention (Garau-Luis et al., 2024). Variational auto-encoders (Kingma & Welling, 2013) have also been widely used for multi-omics integration, either for single-cell omics (Cao & Gao, 2022; Ashuach et al., 2023; Tu et al., 2022) or bulk omics (Benkirane et al., 2023).

Missing modalities

Improving the robustness of multimodal models under modality absence is crucial given the sensitivity of recent deep learning architectures to missing modalities (Ma et al., 2022). Part of the literature focuses on techniques that operate at the data level (Damrosch, 1995), namely leveraging modality imputation (Chen et al., 2024; Zhang et al., 2021a; Ma et al., 2021). Zhi et al. 2024 proposes a retrieval-augmented in-context learning framework to address the missing modalities issue in a low-data regime. Another path towards handling missing modalities lies in adjusting the model itself with, for example, model fusion (Wagner et al., 2011) or knowledge distillation (Saha et al., 2024). Training methods are also adapted to make multimodal models robust to missing modalities by employing modality dropout (Krishna et al., 2024; Nezakati et al., 2024) during training to simulate scenarios where a subset of modalities might be missing. In our work, we adapt a technique from Ramazanova et al. 2025, which addresses the problem of missing modalities as a test-time-adaptation problem, by incorporating an auxiliary loss during the fine-tuning of our model.

3. Multi-omics joint representation learning

3.1. Bulk RNA-seq and DNA methylation processing for language modeling

Modalities alignment Bulk RNA-seq provides an estimate of the mean expression over all cells in a sample for a large number of genes denoted N_{genes} (typically $N_{genes} \sim 10^4$). Thus, each sample of RNA-seq is composed of real values (in units of transcript per million, or TPM) per gene, $X_{rna} \in \mathbf{R}^{N_{genes}}$, to which we apply an $x \mapsto log_{10}(1+x)$ transformation. DNA methylation is the enzymatic attachment of methyl groups to DNA's nucleotide bases (usually Cytosine followed by Guanine or *CpG* site). The methylation level of a given site, obtained through the Illumina Infinium HumanMethylation450 (450K) BeadChip array (Bibikova et al., 2011), is expressed by a beta value $\beta \in [0,1]$, with the number of measured sites, N_{sites} being around $450{,}000$, resulting in a methylation sample $X_{sites\_meth} \in [0,1]^{N_{sites}}$. The first step in our modeling is to align RNA-seq and methylation data by defining a methylation value per gene $X_{meth} \in \mathbf{R}^{N_{genes}}$ as follows:

- For each gene $g \in [\![1 \ ; \ N_{genes}]\!]$, we define $sites(g)$ as all methylation sites close to the gene, as defined in the Infinium Human Methylation 450k BeadChip annotations (typically within ±1.5 kb of the transcript start site or within the gene body).
- The methylation level for a gene is then defined as:

$$X_{meth}[g] = \frac{1}{|sites(g)|} \sum_{s \in sites(g)} X_{sites\_meth}[s]$$

Therefore, a bimodal sample characterized by RNA-seq and DNA methylation is a vector $X = (X_{rna}, X_{meth}) \in \mathbf{R}^{2N_{genes}}$.

Tokenization Language models learn to estimate the likelihood of token sequences. Thus, after aligning the two modalities and obtaining a feature vector $X = (X_{rna}, X_{meth})$, each of its components is tokenized by binning their values on linear scales. The token id associated with a given RNA-seq or methylation value is its corresponding bin id, so after tokenization one sample is a vector $\widetilde{X} = (\widetilde{X}_{rna}, \widetilde{X}_{meth})$ with $\widetilde{X}_{rna} \in [\![0 \ ; \ B_{rna}-1]\!]^{N_{genes}}$ and $\widetilde{X}_{meth} \in [\![0 \ ; \ B_{meth}-1]\!]^{N_{genes}}$, B_{rna} and B_{meth} being the number of bins respectively for gene expression and methylation.

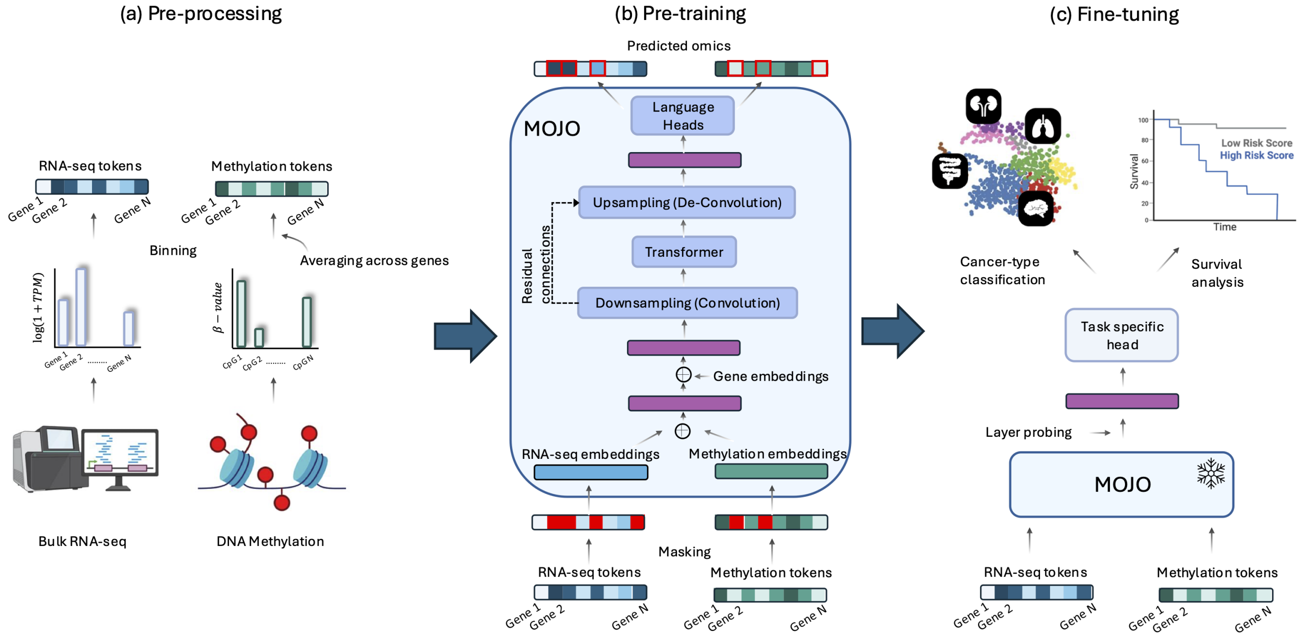

Figure 1: *MOJO* pipeline. (a) Each modality is first tokenized using linear binning. (b) *MOJO*, whose core architecture is composed of a mix of convolution and attention operations, is firstly pre-trained through bimodal masked language modeling. (c) Embeddings are probed from *MOJO* to fine-tune a task-specific head tailored for cancer-type classification or survival analysis.

3.2. A long-range model architecture for bimodal representation learning

In order to learn representations of bulk RNA-seq and DNA methylation, we propose a model that combines both convolution and transformer blocks. Inspired by Avsec et al. 2021; Linder et al. 2025; Joshi et al. 2025 to handle long-range genomic dependencies, this architecture allows us to cope with the high dimensionality of the two omics that we consider, each corresponding to a sequence of length N_{genes}. More precisely, a first bimodal embedding is obtained by passing each omic token through classic embedding layers and summing them up along with a gene embedding vector. As gene expression and DNA methylation are permutation invariant, this gene embedding acts as a positional encoding and is initialized with the *Gene2Vec* method (Zou et al., 2019) as done in Gélard et al. 2025; Yang et al. 2022. Before being fed to a transformer model made up of multi-head attention layers (Vaswani et al., 2017), the embedding is downsampled by a convolutional tower. This downscaling allows the transformer block to act on a compressed embedding vector to significantly reduce the computational cost and time. While the convolutional architecture may be counterintuitive for unordered data, it acts as an efficient mechanism for dimensionality reduction. The original sequence length is restored using a deconvolutional tower with residual connections flowing from the downsampling blocks. Separate language modeling heads predict binned gene expressions and methylation. This architecture is summarized in Figure 1.

3.3. Pre-training: bimodal masked language modeling

Self-supervision loss Our model is pre-trained through self-supervision using multimodal masked language modeling. Although this framework may be applied to more than 2 modalities that can be processed as a sequence of tokens, we present it in the bimodal case where the set of modalities $\mathbb{M} = \{rna, meth\}$. We adopt standard parameters for masked language modeling: for each sequence, 15% of the tokens are corrupted to train the model. Among these corrupted tokens, 80% are replaced with a special `<MASK>` token, 10% are substituted with random tokens, and the remaining 10% are left unchanged, but still contribute to the loss. The final heads of our model provide a set of probability distributions $p_m \in [0,1]^{N_{genes} \times B_m}$ for $m \in \mathbb{M}$. The following multimodal negative log-likelihood is then optimized:

$$\mathcal{L}_{multimodal_{MLM}} = -\sum_{m \in \mathbb{M}} \frac{1}{|\mathcal{M}_m|} \sum_{i \in \mathcal{M}_m} \log((p_m)_{i,(\tilde{X}_m)_i})$$

with for $m \in \mathbb{M}$, $\mathcal{M}_m$ corresponding to the set of masked token indices for that modality.

Experiment Our model is pre-trained using the TCGA dataset from 33 cohorts, resulting in 9,252 pairs of bulk RNA-seq and methylation, 5% being kept for testing. We selected $N_{genes} = 17{,}116$ genes by using the same set of genes as Gélard et al. 2025, from which genes with no methylation data were removed. The model is trained on a TPU v4-8 for total of 192 billion tokens using the Adam (Kingma & Ba, 2014) optimizer with gradient accumulation to reach an effective batch of 3×10^6 tokens. The complete set of hyperparameters can be found in A.1, as well as pre-training learning curves in A.2.

4. Evaluation downstream tasks

The representations learned by *MOJO* are evaluated using a panel of downstream tasks ranging from supervised classification, survival analysis, and zero-shot classification, to clustering. We compare our method to unimodal (only RNA-seq or DNA methylation) and bimodal models:

BulkRNABert (Gélard et al., 2025): A transformer-based model, pretrained on bulk RNA-seq using masked language modeling. Embeddings are extracted from the last self-attention layer, and the mean embedding over the sequence is used as input for downstream tasks. The tokenization of the RNA-seq data is the same between *MOJO* and *BulkRNABert*, *i.e.*, the same B_{rna} has been used.

MethFormer: We develop the equivalent of *BulkRNABert* for DNA methylation (averaged per gene) and we will refer to it in the results as $MethFormer$. Similarly, the same value of B_{meth} is maintained to allow for fair comparison. This model differs from *MethylBERT* (Jeong et al., 2025) which uses read-level methylome and not the 450k microarrays we are interested in.

Late integration: Bimodal integration resulting from the fusion of embeddings extracted from unimodal models. More precisely, we will refer to *Late Integration (concatenation)* as the concatenation of the embeddings from *BulkRNABert* (for RNA-seq) and *MethFormer* (for methylation) which have been pre-trained beforehand. *Late integration (cross-attention)* corresponds to an integration of the two embeddings with a two-step cross-attention followed by a concatenation, allowing for interaction between the two modalities. The different cross-attention modules are only trained when fitting the downstream tasks. An illustration of the late integration is provided in Figure 2.

CustOmics (Benkirane et al., 2023): A multi-omics model based on Variational Auto-Encoders and tailored for cancer-type classification and survival analysis. Although it can handle up to three modalities (bulk RNA-seq, DNA methylation, and Copy Number Variation), we are here interested in its version that can perform the downstream tasks in the bimodal setting. Two models are considered: *CustOmics (end-to-end)* that trains the VAEs and the task heads jointly, and *CustOmics (probing)* that first learns the unsupervised representation with VAEs and then uses the encoded features as input to task heads.

Multi-Omics Factor Analysis (*MOFA*) (Argelaguet et al., 2018): An unsupervised machine learning method designed to integrate multi-omics data by inferring a set of low-dimensional hidden factors, it can be seen as a multi-omics extension to PCA. *MOFA* factors are then used as input either to a Support Vector Machine (Cortes & Vapnik, 1995) for cancer-type classification or a Cox proportional model (Cox, 1972) for survival analysis.

We also integrate a more exhaustive benchmark of other feature extraction and multi-omics model integration in Appendix B.

4.1. Cancer-type classification

Methodology *MOJO* is first evaluated on the supervised task of cancer-type classification. The pan-cancer TCGA dataset is divided into 33 cohorts that make up the labels for the classification task. The last attention layer of the transformer component of *MOJO* is probed to obtain the embedding used for classification. After being downsampled by the convolutional layers, the embedding lies in $\mathbf{R}^{n_{\text{downsample}} \times \text{emb}_{\text{dim}}}$, with $\text{emb}_{\text{dim}} = 512$ and $n_{\text{downsample}} = 67$ (resulting from 8 successive downsampling operations by a factor of 2 from an initial sequence length of $N_{\text{genes}} = 17{,}116$, padded to the next power of 2, 17,152). Then a mean embedding in $\mathbf{R}^{\text{emb}_{\text{dim}}}$ is obtained by averaging over the sequence dimension. This embedding serves as the input to a Multi-Layer Perceptron (MLP) of two hidden layers (respectively of size 256 and 128) that outputs a single logit for cancer-type prediction. In addition, we add dropout (Srivastava et al., 2014) and layer normalization (Ba et al., 2016). *BulkRNABert*, *MethFormer*, and *MOJO* are further fine-tuned in addition to training the MLPs using the parameter-efficient method IA^3 (Liu et al., 2022), which introduces low-dimensional learnable parameters into the self-attention mechanisms and feed-forward networks. For *MOJO*, we also adapt this principle to the convolutional layers by adding a point-wise multiplication of the output of each convolutional operation with a learnable vector of the same dimension.

Results Cancer-type classification results on the Pan-cancer TCGA dataset are presented in Table 1. For this task, the dataset has been split into 80% for training and 20% for testing, repeating the split for 5 different seeds. We report the average and standard deviation over these 5 seeds for the different metrics. We will be using both macro F_1 and weighted F_1 scores to avoid any bias due to class imbalance (label distribution is provided in Appendix B.1). *MOJO* provides state-of-the-art results when considering the two

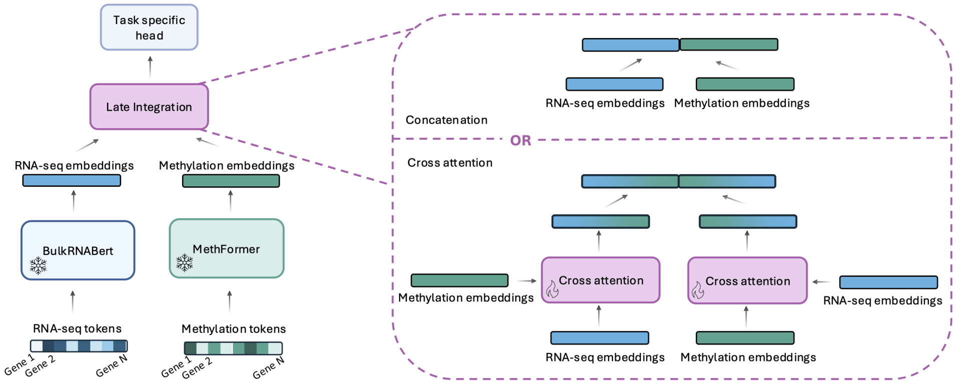

Figure 2: Late integration architecture. RNA-seq and Methylation embeddings are obtained from pre-trained transformer-based encoders (respectively *BulkRNABert* and *MethFormer*) and are fused either by concatenation or by a two-step cross-attention mechanism.

Table 1: Cancer type classification

Model	Modality	test macro-F1	test weighted-F1
BulkRNABert	RNA-seq	0.918 ± 0.008	0.943 ± 0.004
MethFormer	Methylation	0.917 ± 0.008	0.931 ± 0.006
MOFA	Bimodal	0.789 ± 0.012	0.852 ± 0.007
Late integration (concatenation)	Bimodal	0.928 ± 0.008	0.945 ± 0.007
Late integration (cross-attention)	Bimodal	0.929 ± 0.005	0.945 ± 0.002
CustOmics (probing)	Bimodal	0.887 ± 0.065	0.911 ± 0.088
MOJO (probing)	Bimodal	0.928 ± 0.009	0.945 ± 0.006
CustOmics (end-to-end)	Bimodal	0.922 ± 0.006	0.946 ± 0.006
MOJO (no pre-training)	Bimodal	0.835 ± 0.015	0.891 ± 0.006
MOJO	Bimodal	**0.935 ± 0.007**	**0.952 ± 0.006**

modalities, with better performance than *CustOmics* and late integration. For the latter, the *cross-attention* version performs slightly better on average than its *concatenation* counterpart, but not significantly. Our joint modeling of RNA-seq and methylation with *MOJO* outperforms the corresponding unimodal transformer-based models (*BulkRNABert* and *MethFormer*). Moreover, when only probing the last attention layer and fitting an SVM (*MOJO (probing)* in the table), our model shows a clear performance increase in comparison with *CustOmics (probing)*, highlighting that representations from masked language modeling exhibit stronger predictive capacity. Finally, we show that our bimodal masked language modeling pre-training produces a notable performance gain compared to a model trained from scratch (*MOJO (no pre-training)* in the table).

Training times We additionally report in Table 2 the time required by different models (*BulkRNABert*, *Late integration (cross-attention)*, *Late integration (concatenation)*, and *MOJO*) to perform a full update step (forward and backward pass) when training a pan-cancer classification model. While supporting substantially larger batch sizes compared to purely transformer-based models or late integration mechanisms, MOJO achieves approximately a 100× speedup over other benchmarked models. This highlights the computational efficiency of our hybrid architecture that combines convolutional and transformer layers, offering a more scalable alternative to fully transformer-based approaches.

4.2. Survival Analysis

Methodology We then evaluate omics embeddings on a pan-cancer survival task, also known as time-to-event prediction. This task involves predicting the survival time T_i^* for individuals who have cancer, specifically the time from diagnosis until death. A key challenge in survival analysis is right censoring, where the observed time C_i^* might be shorter than the actual survival time T_i^* due to factors like the end of a study or loss of patient contact. Consequently, the true target time used by the model, T_i, is defined as the minimum of the actual survival time and the censoring time ($T_i = min(T_i^*, C_i^*)$). One defines as well $\delta_i = \mathbb{1}_{T_i^* \leq C_i^*}$

Table 2: Average time per update step (forward + backward pass) during training of classification models on a TPU v4-8. All models are evaluated with an effective batch size of 64, achieved via gradient accumulation when necessary. For each model, we additionally report the maximum batch size supported by the model. As in classification benchmarks, parameter efficient fine-tuning is applied to *MOJO* and *BulkRNABert*.

Model	Update time (seconds)	Maximum batch size
Late integration (cross-attention)	5.819 ± 0.006	4
BulkRNABert	4.462 ± 0.006	8
Late integration (concatenation)	2.205 ± 0.004	16
MOJO	**0.059 ± 0.009**	**1,024**

(so $\delta_i = 1$ if the event occurred (death), otherwise $\delta_i = 0$), thus constituting a dataset $\mathcal{D} = \{T_i, x_i, \delta_i\}_{i=1}^{N}$, with x_i the covariates (RNA-seq and/or methylation embeddings in our study). A widely used method to tackle such time-to-event problems is the Cox proportional hazards model (Cox, 1972). This semi-parametric model focuses on modeling the hazard function $\lambda(t|x)$, which represents the instantaneous rate of an event at time t given covariates x. A Cox model expresses λ as $\lambda(t|x) = \lambda_0(t)e^{\hat{h}_\beta(x)}$, with β a vector of parameters (so in our case $x, \beta \in \mathbb{R}^{emb_{dim}}$), $\lambda_0(t)$ the hazard baseline, and in the Cox model, $\hat{h}_\beta(x) = \beta^T x$. More recent works (Katzman et al., 2018; Ching et al., 2018) allow loosening the linear combination of features by replacing $\hat{h}_\beta(x)$ by the output of a neural network and using the negative partial Cox-log-likelihood as loss:

$$\mathcal{L}_{survival} = -\sum_{i|\delta_i=1}\left(\hat{h}_\beta(x_i) - log\sum_{j\in\mathcal{R}_i} e^{\hat{h}_\beta(x_j)}\right)$$

For the survival task, we extract the embeddings from *BulkRNABert*, *MethFormer*, and *MOJO* in the same way as for the cancer-type classification task and train a similar MLP architecture on top. We do not consider the cross-attention version of the late integration here as the Cox loss requires working with a big enough batch size so that the computation of cross-attentions (given the sequence length of $N_{genes} = 17,116$) is not computationally efficient. Similarly, we do not apply IA^3 and only consider the probing experiment for the survival task.

Table 3: Pan-cancer survival analysis

Model	C-index	Weighted C-index
BulkRNABert	0.750 ± 0.004	0.657 ± 0.011
MethFormer	0.735 ± 0.006	0.618 ± 0.017
MOFA	0.648 ± 0.037	0.601 ± 0.022
CustOmics	0.686 ± 0.018	0.639 ± 0.099
Late integration	0.756 ± 0.004	0.653 ± 0.011
MOJO	**0.771 ± 0.006**	**0.670 ± 0.009**

Results Survival results on the pan-cancer TCGA dataset are presented in Table 3. The same split strategy as for the classification task is used. Two different evaluation metrics based on Harrell's C-index (Harrell et al., 1982) are reported. First, a C-index is computed on the whole test set (all cohorts) referred to as "C-index". However, in order to make sure that a pan-cancer model is able to predict survival within cohorts correctly, and not just to differentiate survival chances between cancer types, a "Weighted C-index" is also reported. This corresponds to a weighted sum of the C-indexes computed per cohort on the pan-cancer test set, with weights corresponding to the number of samples of each cohort in the test set. As for the classification task, *MOJO* exhibits higher performance than the unimodal transformer-based models and the late integration, with a significant gain over *CustOmics*. In an additional experiment, *MOJO* performance has been matched by an end-to-end version of *CustOmics* (weighted C-index of 0.669 ± 0.004). This need for end-to-end training compared to simple probing highlights the strength of *MOJO*'s learned representations. Kaplan-Meier curves are also provided in Appendix B.4, showing better patient stratification with *MOJO*.

4.3. Zero-shot pan-cancer and breast cancer sub-typing and clustering

Methodology To further assess the value of the embeddings learned by *MOJO*, we evaluate them in a fully unsupervised manner by considering their zero-shot classification and clustering capabilities on the PAM50 classification, which corresponds to breast cancer sub-typing (Luminal A, Luminal B, Basal, and HER2) (Parker et al., 2009). Inspired by Joshi et al. 2025 that evaluates single-cell RNA-seq embedding models, a zero-shot classification is performed using a k-nearest neighbors model (with $k = 5$), which is evaluated using accuracy. Compared to the supervised classification task we presented, this procedure allows for evaluating the quality of the embedding without having to fine-tune a model on top of the embeddings. Secondly, we perform a Leiden clustering (Traag et al., 2019) in the embedding space and report standard Normalized Mutual Information (NMI), and Adjusted Random Index (ARI) as clustering evaluation metrics. For this experiment, we want to understand the effectiveness of our joint modeling compared to the late integration method for bimodal data. We thus compare *MOJO* and the late integration (concatenation) method embeddings in this section. In addition to the PAM50 sub-typing problem, we also perform the same analysis with the same pan-cancer dataset (referred to as "Pan-cancer" in the results) as in 4.1.

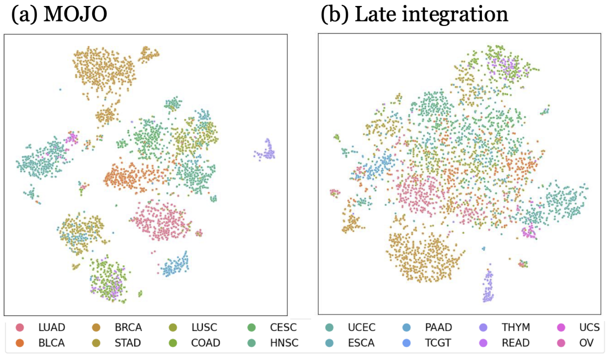

Figure 3: t-SNE representation of *MOJO* and *Late integration* embeddings, colored by cancer-type on a subset of cohorts. *MOJO*'s embeddings visually exhibit better cohort separation capacity compared to *Late Integration* ones, corroborating quantitative results from Table 4. Pan-cancer t-SNE plot is provided in Appendix B.3.

Results The zero-shot classification and clustering results are shown in Table 4, showing better performance when using *MOJO* embedding than late integration. A comparison with *CustOmics* is also added in Appendix B, with lower performance than *MOJO*. We present in Figure 3 a t-SNE (Van der Maaten & Hinton, 2008) visualization of both embeddings in the pan-cancer setting, reflecting that *MOJO* embeddings more effectively separate the cohorts.

Table 4: Zero-shot classification and clustering results on pan-cancer and PAM50 tasks. (Acc. = Accuracy, NMI = Normalized Mutual Infomation, ARI = Ajusted Rank Index).

Task	Metric	MOJO	Late integration
PAM50	Acc.	**0.777**	0.763
	NMI	**0.345**	0.291
	ARI	**0.213**	0.154
Pan-cancer	Acc.	**0.928**	0.870
	NMI	**0.862**	0.771
	ARI	**0.756**	0.620

5. Missing modalities

Integrating two modalities, from late integration to joint modeling using our *MOJO* architecture, has proven to provide performance increases on downstream tasks. However, in clinical applications, some modalities might be missing, and in the worst-case scenario, one modality may never be used by a given medical center. To this end, we aim to provide downstream task models trained on a bimodal setting but which support missing modalities and whose performance when modalities are missing remains comparable to the performance of models trained specifically on the remaining modalities. To this end, we focus on the pan-cancer cancer-type classification task and develop the following experiment framework. First, a bimodal downstream model is fine-tuned as done previously in 4.1, *i.e.* using pairs of (X_{rna}, X_{meth}) without any missing modalities from a pre-trained *MOJO* architecture. Our contribution, detailed thereafter, lies in a slight modification of the fine-tuning procedure to cope with missing modalities. Then, at test time, we compute the evaluation metrics in three ways: first without any modification of the test set, then simulating the absence of either RNA-seq or methylation by dropping that modality in $x\%$ of the (X_{rna}, X_{meth}) pairs of the test set (with $x = 100$ meaning that the modality is removed from all samples) and compute the metrics without further fine-tuning the model. The next sections detail how the missing modalities issue is addressed within this framework.

5.1. MOJO accepts missing modalities

Being trained by masked language modeling, *MOJO* naturally accepts missing RNA-seq or methylation information for a given subset of genes by attributing a special `<MASK>` token to these genes. One can naturally extend this procedure by attributing a sequence full of `<MASK>` tokens to a missing modality, making *MOJO* inherently capable of handling missing modalities. However, as during pre-training only a fraction of each modality is masked to account for the masked language modeling loss, the absence of a whole modality is never encountered by the model. To this end, we decided to conduct another pre-training of *MOJO* by incorporating samples from the TCGA dataset that are missing one of the two considered modalities, thus extending the initial pre-training dataset composed of $9{,}252$ pairs (X_{rna}, X_{meth}) with $2{,}022$ pairs $(X_{rna}, None)$ and 560 pairs $(None, X_{meth})$ with $None$ indicating a missing modality. We will refer to this model as *MOJO-MMO* (*MMO* = **M**issing **MO**dalities).

5.2. Mutual information auxiliary loss

As described in our experimental framework, we only fine-tune and evaluate the downstream model with samples that have both modalities, thus allowing us to get a fair comparison between all the models. Therefore, as the dataset is fixed, one needs to change the fine-tuning procedure to cope with the drop of a full modality at test time to hope for performance maintenance. To this end, we add mutual information as an auxiliary loss paired with classic cross-entropy for the pan-cancer classification task. This quantity is used in Ramazanova et al. 2025 as a test-time adaptation technique of an audio/vision model to handle missing modalities. We adapt it to be directly incorporated during the fine-tuning phase of the model to avoid any modification of the model at test time, thus saving computa-

Algorithm 1 Mutual information auxiliary (MI) loss

```
Input: Omics tokens $X = \{rnaseq : x_{rnaseq}, meth : x_{meth}\}$, true class label $y$, sequence length $N$, mask token <MASK>, mutual information coefficient $\lambda$, classification model $f_\theta$
Output: single example loss
if $noMissingModality(X)$ then
    $modalities = [rna + meth, rnaseq, meth]$
    $output = [f_\theta(X)]$
    for $m \in [rnaseq, meth]$ do
        $X' \leftarrow copy(X)$
        $X'[m] \leftarrow [\texttt{<MASK>}] * N$
        $output.append(f_\theta(X'))$
    end for
    $MILoss = MI(output, modalities)$
else
    $MILoss = 0.0$
end if
Loss = $CrossEntropy(f_\theta(X), y) + \lambda * MILoss$
```

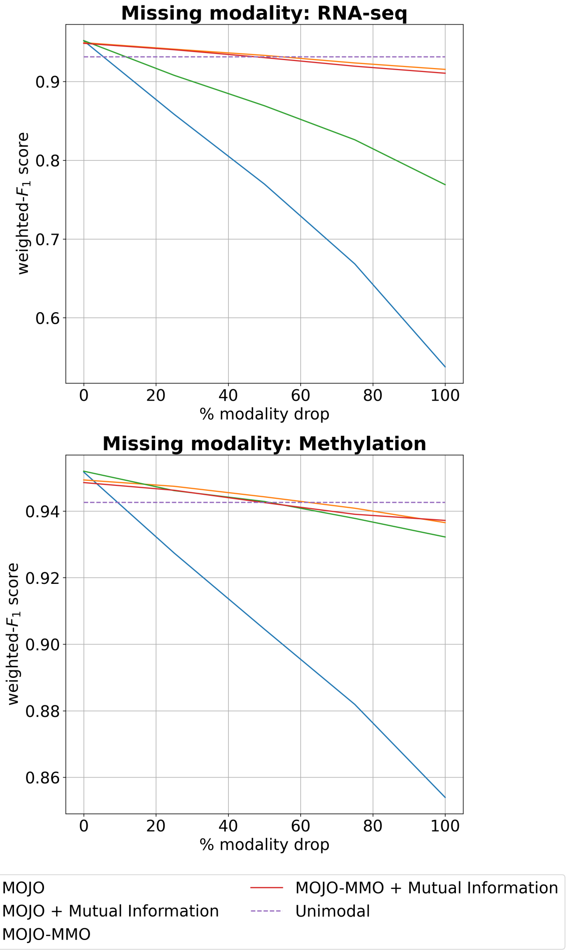

Figure 4: Missing modalities experimental results. Test weighted-F_1 score for the pan-cancer classification is reported for different methods to handle the absence of a modality in x% of the samples (left: RNA-seq, right: Methylation). Unimodal models are respectively *MethFormer* and *BulkRNABert* when RNA-seq or Methylation is missing.

tional time. Following the notation from Ramazanova et al. 2025, we denote $f_\theta(x; m)$ the output of the classification model for a given input x when modalities m are present, with $m \in \mathcal{D}_{modality} = \{rna + meth, rna, meth\}$. One would require f_θ to provide the same prediction whatever modality is given, thus satisfying the following equality: $f_\theta(x; rna+meth) = f_\theta(x; rna) = f_\theta(x; meth)$. One can satisfy such a constraint by minimizing the following loss:

$$\mathcal{L}_{aux} = \mathbb{E}_{m \in \mathcal{D}_{modality}} \left[MI(f_\theta(x, m), m) \right]$$

with $MI(X, Y) = D_{KL}[P_{(X,Y)} || P_X \otimes P_Y]$ corresponding to the mutual information (Shannon, 1948) between two random variables X and Y. The mutual information is equal to 0 when the two random variables are independent; thus, minimizing this quantity as an auxiliary loss should guide the model towards providing the same output independently of the given modality. Therefore, we end up optimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda * \mathcal{L}_{aux}$$

with $\mathcal{L}_{task}$ corresponding here to the cross-entropy for the cancer-type classification task. We detail in Algorithm 1 the procedure used to compute this loss for a single example.

5.3. Missing modalities experimental results

We present the results of applying our missing modalities framework in Figure 4 (with the exact performance figures being additionally reported in Appendix C). When considering our initial model trained on both modalities from *MOJO*, with a test weighted-F_1 of 0.952, it is no surprise that without any further intervention, when dropping either all the RNA-seq or all the methylation from the test set, one gets a significant performance decrease (respectively 0.538 and 0.854). When adding the mutual information loss as an auxiliary loss during model fine-tuning (we used $\lambda = 10$ as the ratio of the cross-entropy and the mutual information computed after model initialization), model performance without any modality drop at test time remains stable (0.949), meaning that the addition of the mutual information as an auxiliary loss does not inhibit the classification signal from the cross-entropy. The previously observed performance decrease when one modality is dropped is corrected: from 0.538 to 0.916 when RNA-seq is missing, and from 0.854 to 0.937 when methylation is missing. This recovered performance is even closer to models that have been specifically trained on one modality, for instance *MethFormer* for methylation (0.931) and *BulkRNABert* for RNA-seq (0.943). Finally, when dropping only 50% of one modality, the bimodal model fine-tuned with the mutual information as auxiliary loss performs similarly to its unimodal counterpart.

One also observes the effectiveness of extending the pre-training dataset with pairs that are missing one modality for *MOJO-MMO*, as the performance gap also narrows even without the addition of the mutual information loss. We finally conducted an experiment by combining *MOJO-MMO* and the mutual information as auxiliary loss, ending with performance similar to the auxiliary loss alone. One notices the discrepancy between the drop of methylation and the drop of RNA-seq: the performance deterioration is greater when RNA-seq is missing, suggesting this modality is more reliable than the other for the model to predict cancer type, which is corroborated by the higher performance of *BulkRNABert* compared to *MethFormer*.

6. Conclusion

We proposed a novel methodology for learning joint representations of multi-omics data and introduced the *MOJO* architecture. Our focus was on pre-training this model to learn representations from bi-omics data, specifically bulk RNA-seq and DNA methylation. By aligning these two modalities, we formulated the representation learning task as a self-supervised problem using bimodal masked language modeling. The architecture combines convolutional and attention-based components, enabling it to efficiently handle long-range sequences that arise when modeling a large number of genes, outperforming purely transformer-based approaches in this context. Compared to the late integration mechanism that requires separate pre-training of unimodal models, our joint approach allows one to learn representations of multi-omics data with a single model.

After a pre-training phase, the embeddings learned by the model are used as input for clinical downstream tasks on the TCGA dataset: from supervised classification (cancer-typing) to time-to-event prediction, *MOJO* provides state-of-the-art performance compared to unimodal models. We further point up the interest in joint modeling compared to a late integration mechanism through zero-shot classification and clustering (breast cancer sub-typing). In particular, the predictive capacity of *MOJO*'s representations has been emphasized by observing a significant performance gain compared to other models in the layer probing setup.

Finally, we raise the issue of the possibility for a given modality to be missing at test time, and thus the need for a methodological solution to prevent any performance drop of the downstream model. To this end, we presented how *MOJO* can inherently cope with missing modalities, and we reformulated a test-time adaptation technique based on mutual information by incorporating it as an auxiliary loss during the fine-tuning process. Through this method, we were able to narrow the performance drop in the absence of a modality, getting results that are comparable to unimodal models.

Further work would extend this architecture to a larger class of data, especially by relaxing the need for initial modality alignment. Although already providing promising results for missing modalities, the mutual information approach may also be extended to more than two modalities, and some improvements have to be made so that performance exactly matches unimodal models.

Impact Statement

This paper presents work towards an integration of multi-omics data while handling missing modalities. This approach improves the clinical utility of deep learning-based cancer prognosis by enabling more reliable patient stratification.

Acknowledgements

Research supported by Cloud TPUs from Google's TPU Research Cloud (TRC).

References

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

Ashuach, T., Gabitto, M. I., Koodli, R. V., Saldi, G.-A., Jordan, M. I., and Yosef, N. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Benkirane, H., Pradat, Y., Michiels, S., and Cournède, P.-H. Customics: A versatile deep-learning based strategy for multi-omics integration. *PLOS Computational Biology*, 19(3):e1010921, 2023.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, pp. 2025–02, 2025.

Cao, Z.-J. and Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

Chen, Q., Zhang, J., Meng, R., Zhou, L., Li, Z., Feng, Q., and Shen, D. Modality-specific information disentanglement from multi-parametric mri for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 43(5):1958–1971, 2024.

Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F., Rodig, S. J., Lindeman, N. I., and Mahmood, F. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020.

Ching, T., Zhu, X., and Garmire, L. X. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

Dai, X. and Shen, L. Advances and trends in omics technology development. *Frontiers in Medicine*, 9:911861, 2022.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Damrosch, D. *We scholars: Changing the culture of the university*. Harvard University Press, 1995.

Garau-Luis, J. J., Bordes, P., Gonzalez, L., Roller, M., de Almeida, B., Blum, C., Hexemer, L., Laurent, S., Lang, M., Pierrot, T., et al. Multi-modal transfer learning between biological foundation models. *Advances in Neural Information Processing Systems*, 37:78431–78450, 2024.

Gélard, M., Richard, G., Pierrot, T., and Cournède, P.-H. Bulkrnabert: Cancer prognosis from bulk rna-seq based language models. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 384–400. PMLR, 15–16 Dec 2025.

Gross, B., Dauvin, A., Cabeli, V., Kmetzsch, V., El Khoury, J., Dissez, G., Ouardini, K., Grouard, S., Davi, A., Loeb, R., et al. Robust evaluation of deep learning-based representation methods for survival and gene essentiality prediction on bulk rna-seq data. *Scientific Reports*, 14(1): 17064, 2024.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Ho, W. J., Erbe, R., Danilova, L., Phyo, Z., Bigelow, E., Stein-O'Brien, G., Thomas, D. L., Charmsaz, S., Gross, N., Woolman, S., et al. Multi-omic profiling of lung and liver tumor microenvironments of metastatic pancreatic cancer reveals site-specific immune regulatory pathways. *Genome biology*, 22(1):154, 2021.

Jeong, Y., Gerhäuser, C., Sauter, G., Schlomm, T., Rohr, K., and Lutsik, P. Methylbert enables read-level dna methylation pattern identification and tumour deconvolution using a transformer-based model. *Nature Communications*, 16(1):788, 2025.

Jolliffe, I. T. *Principal component analysis for special types of data*. Springer, 2002.

Joshi, A., Boige, R., Zamparo, L., Tanielian, U., Garau-Luis, J. J., Chatzianastasis, M., Pandey, P., Sielemann, J., Seifert, A., Brand, M., et al. A long range foundation model for zero-shot predictions in single-cell and spatial transcriptomics data. 2025.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18: 1–12, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krishna, G., Dharur, S., Rudovic, O., Dighe, P., Adya, S., Abdelaziz, A. H., and Tewfik, A. H. Modality drop-out for multimodal device directed speech detection using verbal and non-verbal features. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8240–8244. IEEE, 2024.

Lee, D. and Seung, H. S. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., and Kelley, D. R. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, pp. 1–13, 2025.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. A. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35: 1950–1965, 2022.

Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., and Peng, X. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Ma, M., Ren, J., Zhao, L., Testuggine, D., and Peng, X. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18177–18186, 2022.

Ma, S., Zeng, A. G., Haibe-Kains, B., Goldenberg, A., Dick, J. E., and Wang, B. Integrate any omics: Towards genome-wide data integration for patient stratification. *arXiv preprint arXiv:2401.07937*, 2024.

Meng, X., Li, X., Yang, Q., Dai, H., Qiao, L., Ding, H., Hao, L., and Wang, X. Gene-moe: A sparsely gated prognosis and classification framework exploiting pan-cancer genomic information. *arXiv preprint arXiv:2311.17401*, 2023.

Nezakati, N., Reza, M. K., Patil, A., Solh, M., and Asif, M. S. Mmp: Towards robust multi-modal learning with masked modality projection. *arXiv preprint arXiv:2410.03010*, 2024.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8): 1160–1167, 2009.

Popov, M., Kallala, A., Ramesh, A., Hennouni, N., Khaitan, S., Gentry, R., and Cohen, A.-S. Leveraging state space models in long range genomics. *arXiv preprint arXiv:2504.06304*, 2025.

Ramazanova, M., Pardo, A., Ghanem, B., and Alfarra, M. Test-time adaptation for combating missing modalities in egocentric videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

Saha, P., Mishra, D., Wagner, F., Kamnitsas, K., and Noble, J. A. Examining modality incongruity in multimodal federated learning for medical vision and language-based disease detection. *arXiv preprint arXiv:2402.05294*, 2024.

Sánchez, A., Fernández-Real, J., Vegas, E., Carmona, F., Amar, J., Burcelin, R., Serino, M., Tinahones, F., de Villa, M. C. R., Minãrro, A., et al. Multivariate methods for the integration and visualization of omics data. In *Bioinformatics for Personalized Medicine: 10th Spanish Symposium, JBI 2010, Torremolinos, Spain, October 27-29, 2010. Revised Selected Papers*, pp. 29–41. Springer, 2012.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Stark, R., Grzelak, M., and Hadfield, J. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

Traag, V. A., Waltman, L., and Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Tu, X., Cao, Z.-J., Xia, C.-R., Mostafavi, S., and Gao, G. Cross-linked unified embedding for cross-modality representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Vale-Silva, L. A. and Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wagner, J., Andre, E., Lingenfelser, F., and Kim, J. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218, 2011.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Zhang, W., Xu, D., Zhang, J., and Ouyang, W. Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing*, 30:3293–3306, 2021a.

Zhang, X., Xing, Y., Sun, K., and Guo, Y. Omiembed: a unified multi-task deep learning framework for multi-omics data. *Cancers*, 13(12):3047, 2021b.

Zhi, Z., Liu, Z., Elbadawi, M., Daneshmend, A., Orlu, M., Basit, A., Demosthenous, A., and Rodrigues, M. Borrowing treasures from neighbors: In-context learning for multimodal learning with missing modalities and data scarcity. *arXiv preprint arXiv:2403.09428*, 2024.

Zou, Q., Xing, P., Wei, L., and Liu, B. Gene2vec: gene subsequence embedding for prediction of mammalian n6-methyladenosine sites from mrna. *Rna*, 25(2):205–218, 2019.

A. MOJO pre-training

A.1. Hyperparameters

Table 5: MOJO model and pre-training hyperparameters

Model Hyperparameters	
Number of downsamples	8
Kernel size	5
Embedding dimension	512
Number of transformer layers	8
Feed forward dimension	1,024
Number of attention heads	16
Training Hyperparameters	
Batch size	128
Gradient accumulation	4
Learning rate	5×10^{-5}
Masking ratio	15%

A.2. Pre-training learning curves

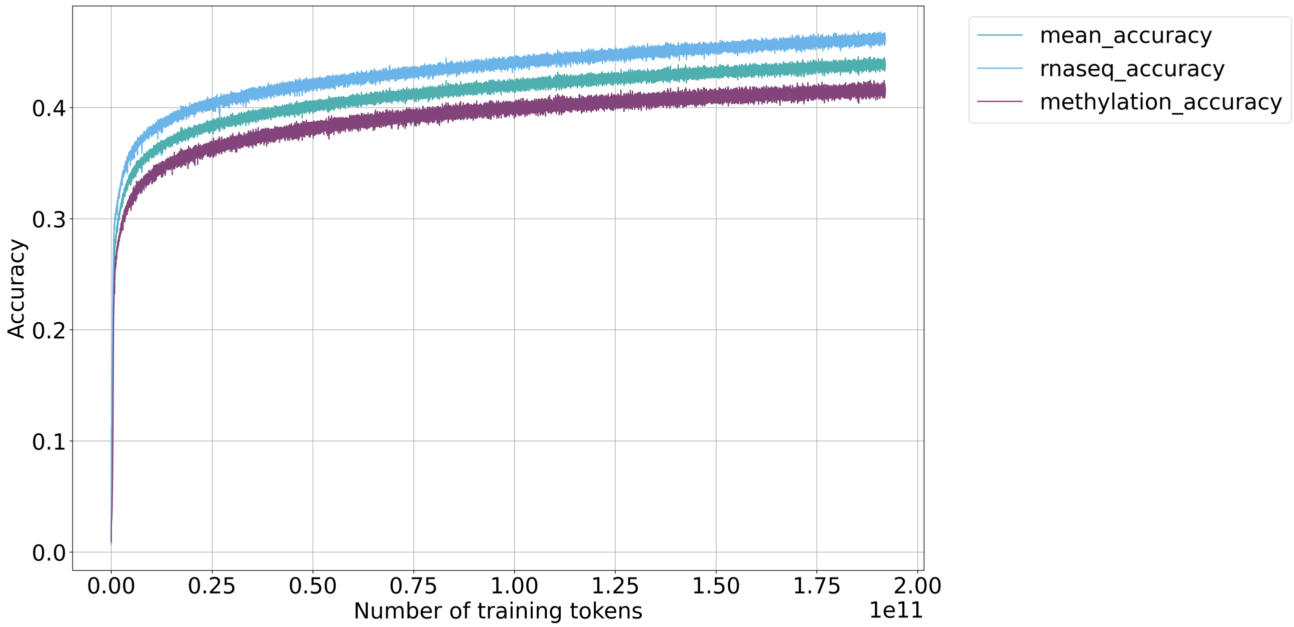

Figure 5: Bimodal masked language modeling pre-training curves of the *MOJO* architecture. The training reconstruction accuracy is represented of each omic separately as well as the average reconstruction accuracy among the different omics.

B. Downstream tasks dataset and benchmarks

B.1. Pan-cancer classification dataset

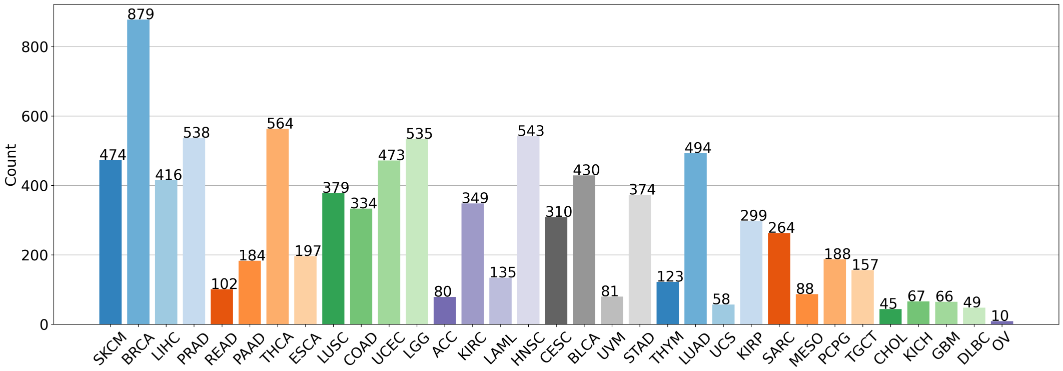

Figure 6: Pan-cancer classification label distribution.

B.2. Downstream tasks benchmarks

In addition to Table 1 (cancer-type classification) and Table 3, a more exhaustive benchmark including other representation models for RNA-seq and DNA methylation has been performed:

- Multiple Factor Analysis (MFA) (Sánchez et al., 2012), using a latent space of dimension 256.
- Non-negative Matrix Factorization (NMF) (Lee & Seung, 2000), with the same latent space dimension as for MFA.
- *OmiEmbed* (Zhang et al., 2021b): a unified multi-task deep learning framework for multi-omics data based on Variational Auto-Encoders (Kingma & Welling, 2013) from early integrated omics.
- *IntegrAO* (Ma et al., 2024): an unsupervised framework based on Graph Neural Networks (Scarselli et al., 2008) for integrating incomplete multi-omics data, tailored for classification and survival task.

Multiple Factor Analysis and Non-negative Matrix Factorization features are then fed to a Support Vector Machine (SVM) for the cancer-type classification task and to a Cox proportional model for the survival analysis task. The results are presented in Table 6 and Table 7.

Table 6: Full benchmark on cancer-type classification

Model	Modality	test macro-F1	test weighted-F1
BulkRNABert	RNA-seq	0.918 ± 0.008	0.943 ± 0.004
MethFormer	Methylation	0.917 ± 0.008	0.931 ± 0.006
MFA	Bimodal	0.753 ± 0.013	0.848 ± 0.008
NMF	Bimodal	0.725 ± 0.011	0.827 ± 0.006
MOFA	Bimodal	0.789 ± 0.012	0.852 ± 0.007
Late integration (concatenation)	Bimodal	0.928 ± 0.008	0.945 ± 0.007
Late integration (cross-attention)	Bimodal	0.929 ± 0.005	0.945 ± 0.002
CustOmics (probing)	Bimodal	0.887 ± 0.065	0.911 ± 0.088
MOJO (probing)	Bimodal	0.928 ± 0.009	0.945 ± 0.006
IntegrAO	Bimodal	0.912 ± 0.005	0.911 ± 0.015
OmiEmbed	Bimodal	0.919 ± 0.004	0.922 ± 0.016
CustOmics (end-to-end)	Bimodal	0.922 ± 0.006	0.946 ± 0.006
MOJO (no pre-training)	Bimodal	0.835 ± 0.015	0.891 ± 0.006
MOJO	Bimodal	**0.935 ± 0.007**	**0.952 ± 0.006**

Table 7: Full benchmark on pan-cancer survival analysis

Model	Modality	C-index	Weighted C-index
BulkRNABert	RNA-seq	0.750 ± 0.004	0.657 ± 0.011
MethFormer	Methylation	0.735 ± 0.006	0.618 ± 0.017
MFA	Bimodal	0.616 ± 0.033	0.593 ± 0.016
NMF	Bimodal	0.616 ± 0.040	0.591 ± 0.025
MOFA	Bimodal	0.648 ± 0.037	0.601 ± 0.022
IntegrAO	Bimodal	0.710 ± 0.008	0.624 ± 0.006
OmiEmbed	Bimodal	0.736 ± 0.006	0.631 ± 0.007
CustOmics	Bimodal	0.686 ± 0.018	0.639 ± 0.099
Late integration	Bimodal	0.756 ± 0.004	0.653 ± 0.011
MOJO	Bimodal	**0.771 ± 0.006**	**0.670 ± 0.009**

Table 8: Full benchmark on zero-shot classification and clustering results on pan-cancer and PAM50 tasks. (Acc. = Accuracy, NMI = Normalized Mutual Infomation, ARI = Ajusted Rank Index).

Task	**Metric**	**MOJO**	**Late integration**	**CustOmics**
PAM50	Acc.	**0.777**	0.763	0.765
	NMI	**0.345**	0.291	0.311
	ARI	**0.213**	0.154	0.176
Pan-cancer	Acc.	**0.928**	0.870	0.905
	NMI	**0.862**	0.771	0.830
	ARI	**0.756**	0.620	0.699

B.3. Bimodal embeddings t-SNE visualisations

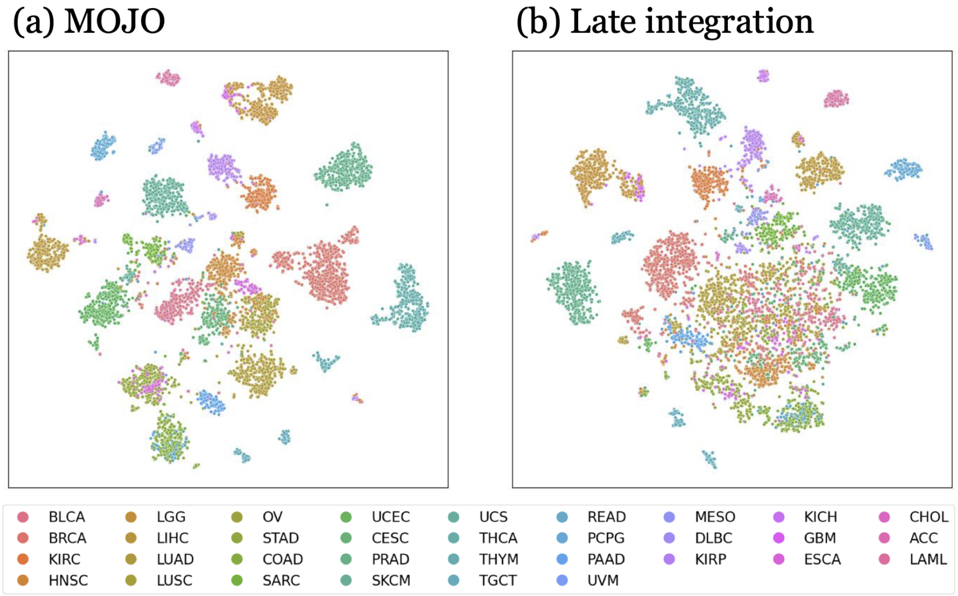

Figure 7: Pan-cancer version of the t-SNE representation of *MOJO* and *Late integration* embeddings, colored by cancer-type.

B.4. Kaplan-Meier curves

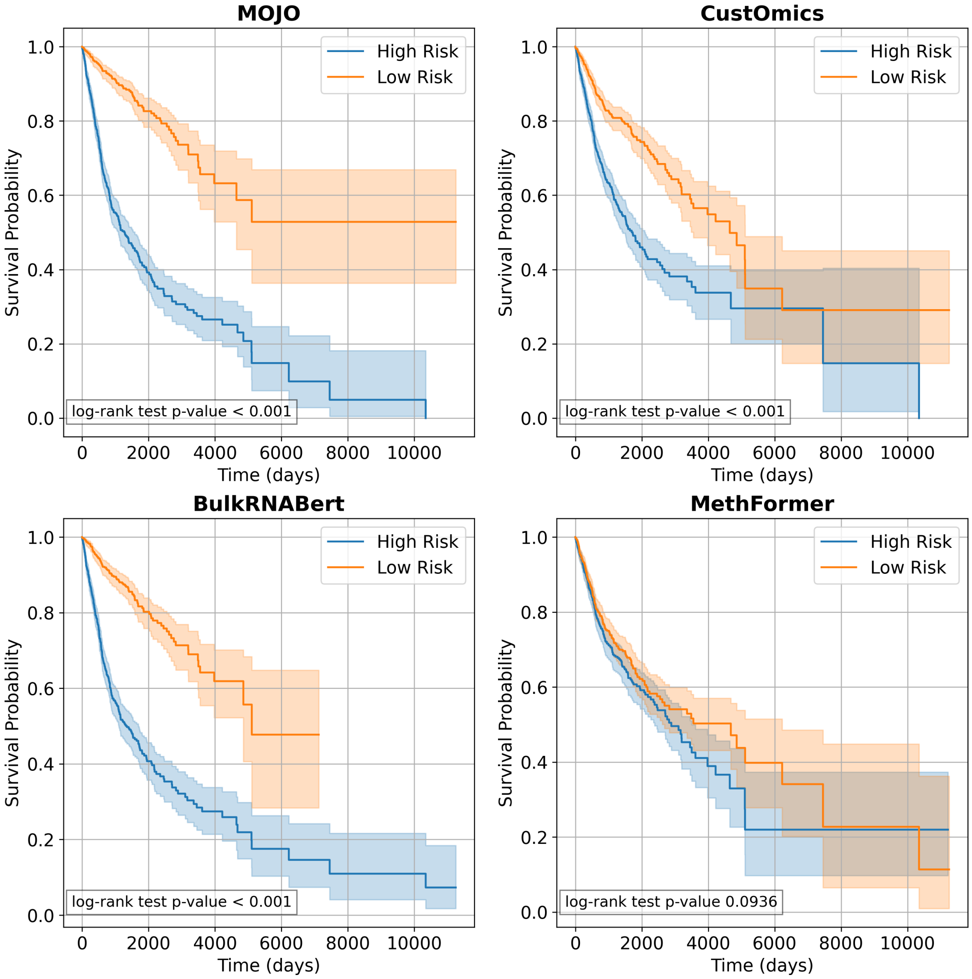

Figure 8: Kaplan-Meier curve for pan-cancer survival models for four models: *MOJO*, *CustOmics*, *BulkRNABert*, *MethFormer*.

C. Missing modalities experiments

Table 9: Missing modalities experiment: cancer type classification

Model	Add mutual information	Drop modality (test time)	macro-F1	test weighted-F1
BulkRNABert	✗	-	0.918 ± 0.008	0.943 ± 0.004
MethFormer	✗	-	0.917 ± 0.008	0.931 ± 0.006
MOJO	✗	-	0.935 ± 0.007	0.952 ± 0.006
MOJO	✗	Drop 100% of RNASeq	0.422 ± 0.022	0.538 ± 0.025
MOJO	✗	Drop 100% of Methylation	0.764 ± 0.024	0.854 ± 0.011
MOJO	✓	-	0.930 ± 0.007	0.949 ± 0.004
MOJO	✓	Drop 100% of RNASeq	0.895 ± 0.008	0.916 ± 0.007
MOJO	✓	Drop 100% of Methylation	0.911 ± 0.012	0.937 ± 0.008
MOJO-MMO	✗	-	0.933 ± 0.006	0.952 ± 0.003
MOJO-MMO	✗	Drop 100% of RNASeq	0.653 ± 0.013	0.769 ± 0.004
MOJO-MMO	✗	Drop 100% of Methylation	0.903 ± 0.010	0.932 ± 0.005
MOJO-MMO	✓	-	0.929 ± 0.006	0.949 ± 0.005
MOJO-MMO	✓	Drop 100% of RNASeq	0.883 ± 0.005	0.911 ± 0.004
MOJO-MMO	✓	Drop 100% of Methylation	0.911 ± 0.010	0.937 ± 0.006